Universidade de Aveiro Departamento de Electrónica e Telecomunicações, 2000

António Joaquim da Silva Teixeira

Síntese articulatória das vogais nasais do Português Europeu

António Joaquim da Silva Teixeira

Síntese articulatória das vogais nasais do Português Europeu

Articulatory Synthesis of European Portuguese Nasal Vowels

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Engenharia Electrotécnica, realizada sob a orientação científica do Doutor Francisco Vaz, Professor Catedrático do Departamento de Electrónica e Telecomunicações da Universidade de Aveiro

o júri
presidente

Prof. Doutor Jorge Ribeiro Frade
professor catedrático de Universidade de Aveiro
por delegação do Reitor da Universidade de Aveiro

Prof. Doutor Francisco António Cardoso Vaz
professor catedrático de Universidade de Aveiro (orientador)

Prof. Doutor José Carlos Santos Carvalho Príncipe
professor da Universidade da Florida (BellSouth Professor)

Prof. Doutora Lurdes de Castro Moutinho
professora associada da Universidade de Aveiro

Prof. Doutor Luís Miguel Veiga Vaz Caldas de Oliveira
professor auxiliar do Instituto Superior Técnico
da Universidade Técnica de Lisboa

Prof. Doutora Ana Maria Perfeito Tomé
professora auxiliar da Universidade de Aveiro

agradecimentos

Agradeço ao Professor Francisco Vaz a orientação, incentivo, e a amizade. Agradeço ter aceite ser orientador do meu mestrado e depois me ter incentivado a continuar o trabalho na área do processamento de voz.

Ao Professor José Carlos Príncipe, agradeço as suas observações e questões em várias ocasiões. Agradeço também o seu empenho para a minha visita ao *Mind Machine Interaction Research Center* (MMIRC) da *University of Florida*. Não posso deixar de agradecer a toda a sua família o acolhimento durante a minha estadia na Florida.

Agradeço ao Professor Donald Childers me ter recebido no MMIRC. Este agradecimento estende-se aos estudantes de Doutoramento presentes em 1995 no MMIRC, em especial ao M. K. Lee.

Não posso deixar de agradecer ao Tomás Oliveira e Silva, entre muitas outras coisas, a troca de impressões constante, que teve um pico na sua estadia em Gainesville, e a disponibilização de código LaTeX para as formatação das primeiras páginas da tese.

Agradeço a colaboração prestada pelo grupo de Processamento de Fala do INESC Lisboa, em especial a disponibilização dos estímulos gerados pelo sistema DIXI, sem os quais o teste de qualidade efectuado não teria sido possível.

Na estadia na Florida, a colaboração do Vitor e da Cristina não pode deixar de ter o meu mais sincero agradecimento.

Gostaria de deixar aqui o meu apreço pela pessoa responsável pelo arranque do Mestrado em Engenharia Electrónica e Telecomunicações, Professor Pedro Guedes de Oliveira. Sem a existência do Mestrado este trabalho, muito provavelmente, não teria existido.

Na fase, particularmente difícil, que mediou entre o final do meu Mestrado e o inicio do trabalho de Doutoramento o incentivo de várias pessoas foi essencial. Gostaria de reconhecer aqui o incentivo do Professor Gregor Schöner.

Agradeço a todos os que amavelmente participaram nos testes perceptuais.

O meu obrigado aos meus colegas do Departamento, em especial aqueles que mais de perto comigo conviveram durante os últimos anos. Este agradecimento não pode deixar de incluir o pessoal não docente, em especial o do ex-INESC Aveiro e actual IEETA.

Este trabalho não teria sido possível sem o financiamento do Programa PRAXIS, inicialmente sob a forma da bolsa de Doutoramento PRAXIS BD/3495/94, e na parte final pelo financiamento do projecto "Síntese Articulatória do Português" P/PLP/11222/98. A divulgação do trabalho nas várias conferências internacionais contou com a colaboração financeira da Fundação Calouste Gulbenkian e Fundação Luso-Americana para o Desenvolvimento, às quais agradeço.

Agradeço à Universidade de Aveiro, ao Departamento de Electrónica e Telecomunicações, ao IEETA, e ao seu antecessor INESC Aveiro, os meios postos à minha disposição.

Um agradecimento especial para minha esposa pelo seu apoio incondicional.

resumo

A principal motivação para este trabalho foi a obtenção de síntese com qualidade próxima da natural. Um dos caminhos mais promissores para atingir esse objectivo é a utilização de síntese articulatória, técnica baseada na modelação directa do processo de produção humano. A síntese articulatória permite realizar simulações e, funcionando como um falante versátil, obter estímulos para a realização de testes perceptuais, sendo útil em estudos indispensáveis na obtenção de novos conhecimentos sobre os mecanismos de produção e percepção.

O objectivo principal deste trabalho foi o de estudar e aplicar as técnicas da síntese articulatória ao estudo e síntese do Português Europeu. Em vez de dispersar a nossa atenção por todos os tipos de sons da língua portuguesa, optamos por nos dedicar exclusivamente ao estudo das vogais nasais do Português, sons com especial relevância na língua de Camões, e em que outras técnicas de síntese têm limitações.

Foi desenvolvido um sintetizador articulatório especialmente vocacionado para a síntese de sons nasais, utilizando técnicas já conhecidas e introduzindo um novo modelo acústico para o tracto nasal e uma fonte glotal interactiva.

O sintetizador articulatório foi utilizado em estudos, tendo por objectivo melhorar o conhecimento acerca dos sons nasais, abordando dois assuntos: a interacção entre a fonte de excitação glotal e as cavidades acima da laringe, e o efeito da dinâmica dos articuladores.

Os resultados das experiências efectuadas mostram que a variação do velo, e mesmo de outros articuladores, influencia a percepção de nasalidade. Incluindo a forma como os articuladores variam, nos vários contextos em que as vogais nasais do Português aparecem, melhora a qualidade. Com base nos resultados obtidos, consideramos que as vogais nasais, do Português Europeu, devem ser vistas como ditongos, iniciando-se com radiação pelos lábios dominante e acabando em radiação nasal dominante, sendo a transição entre as duas configurações gradual.

As alterações nas características da onda de excitação glotal, devidas ao acoplamento do tracto nasal, são muito mais visíveis no domínio do tempo do que no domínio da frequência e este efeito não tem a mesma relevância para todas as vogais. Os resultados de testes perceptuais mostram que as alterações não são de fácil identificação, existindo, no entanto, uma tendência para o efeito ser mais facilmente detectável para a vogal elevada [ĩ], produzida com uma cavidade bucal de área bastante reduzida.

No seu estado actual, o sintetizador e os conhecimentos que obtivemos permitem produzir vogais nasais e palavras contendo sons nasais com qualidade entre o razoável e bom (níveis 3 a 4 na escala MOS).

abstract

Main motivation for this work was the attainment of synthesis with natural quality. One of the promising paths to reach this objective is the use of articulatory synthesis, technique based on direct modellation of the human speech production process. Articulatory synthesis can be used in simulations, and as a versatile informant to produce stimuli for perceptual tests, indispensable in speech production and perception studies.

The main objective of this work was the application of articulatory synthesis techniques to the study and synthesis of European Portuguese. Instead of dispersing our attention with all the sound types of the Portuguese language, we decided to concentrate on the study of the Portuguese nasal vowels, sounds with special relevance in the language of Camões, and where other synthesis techniques have shown limitations.

An articulatory synthesizer, especially tailored for the synthesis of nasal sounds, was developed, using available techniques and introducing a new comprehensive nasal tract acoustic model and an interactive glottal source model.

The synthesizer was used in studies having as objective the improvement of the knowledge concerning Portuguese nasal vowels. Studies addressed two different subjects: source tract acoustic interaction, and the effect of the articulators' dynamics.

Experiments results showed that the variation of the velum and other articulators position, during nasal vowel production, influences perception of nasality. Including, in the synthesis process, information about velum and other articulators dynamics improves the quality. On the basis of the results obtained, for several contexts, we propose that European Portuguese nasal vowels can be seen as diphthongs, initiating themselves with dominant oral radiation and finishing with dominant nasal radiation, being the transition gradual.

Alterations in the glottal flux due to the extra coupling of the nasal tract are much more visible in time than in the frequency domain and the effect does not have the same relevance for all the vowels. Perceptual tests showed that the alterations are not easily perceived, existing a tendency of the effects being more detectable for the high vowels, like [ĩ], having a small area of the vocal tract regions between velum and the lips.

In its present state, the synthesizer, using the results from our studies, is capable of producing nasal vowels, and words with nasal sounds, with quality between Fair and Good (levels 3 to 4 in a MOS quality test).

À Zé,

ao Pedro,

à minha mãe, ao meu irmão Carlos

e à memória da minha tia Emília

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

> Thus in order to continue my experiments it was necessary, above all, that I should have a perfect knowledge of what I wanted to imitate. I had to make a formal study of speech and continually consult nature as I conducted my experiments. In this way my talking machine and my theory concerning speech made equal progress, the one serving as guide to the other.
>
> von Kempelen, 1791

A Voz [1] é talvez a capacidade que mais distingue o Homem das outras espécies. A Voz é para a maioria de nós, os que não estamos impossibilitados de a usar, o meio de comunicação por excelência. A comunicação pela voz é a forma de comunicação mais eficaz desenvolvida pelo Homem até hoje. É mais rápida do que qualquer outra e tem sobretudo a grande vantagem de deixar os olhos e as mãos livres, para o desempenho de outras tarefas.

Desde sempre que a espécie humana pretendeu criar máquinas que produzissem e entendessem a voz humana (Linggard, 1985; Schroeder, 1999). A linguagem, tão estritamente ligada ao pensamento, é também uma janela para o funcionamento do cérebro, como o comprovam diversos estudos das neuro-ciências baseados em perturbações da linguagem. É natural que os cientistas tenham interesse em analisar, reconhecer e produzir voz. A utilização de voz na interacção com computadores, e outros sistemas, terá vantagens a vários níveis [2]. Permite, por exemplo, a utilização de sistemas informáticos por deficientes, e o acesso via telefone a serviços de informação.

Existem dois conjuntos de motivações essenciais para fazer investigação no vasto domínio

[1] Utilizarei neste trabalho o termo "voz" para incluir, não só a transmissão de mensagens, mas também outra informação como a emoção, estado de saúde e identidade do orador. Só uma elevada qualidade de síntese deverá permitir transmitir a mensagem e os outros tipos de informação. O termo "fala" é mais adequado para trabalhos interessados "apenas" em assuntos relacionados com a linguagem. Para a distinção entre voz e fala (em Inglês *voice* e *speech*) veja-se (Childers, 1985, 2000)

[2] Num inquérito realizado em Abril de 1997 pela revista *Newsweek*, 71 % dos inquiridos, com idades entre os doze e dezassete, revelaram preferência em falar para o computador em relação à utilização do teclado.

da voz/fala (Tubach, 1996). O primeiro visa a **compreensão** profunda dos seus diversos aspectos e funções; o segundo é a **concepção e desenvolvimento de sistemas artificiais** permitindo a síntese, o reconhecimento, a compreensão da voz e a sua utilização no diálogo homem-máquina.

Um dos instrumentos mais importantes para estes estudos é a síntese de voz. A síntese articulatória, ao modelar de uma forma mais directa o processo de produção de voz humana, apresenta-se como a técnica de síntese do futuro, considerando-se que desempenhará um papel importante na obtenção de voz com qualidade natural. Parte desse papel realizar-se-á utilizando a síntese articulatória como um informador versátil (Cooper, 1961) em estudos de produção e percepção, indispensáveis na obtenção de conhecimentos.

A síntese articulatória tem ainda um longo caminho a percorrer até poder ser uma alternativa aos sistemas actuais de síntese.

No seu estado actual, o conhecimento acerca da produção e percepção de voz é ainda deficitário. A qualidade (ou falta dela) de voz sintética dos sistemas actuais é uma boa prova da necessidade de melhorar esse conhecimento. A utilização desse conhecimento em sistemas de processamento de voz é ainda mais reduzida.

Geralmente utiliza-se um modelo de produção em que a fonte e o tracto são consideradas independentes. No entanto, esta simplificação deverá ter que ser abandonada em sistemas interessados na obtenção de qualidade natural (Schroeder, 1999, pág. 34 e 88).

A produção de voz é um processo altamente dinâmico. Os órgãos móveis que contribuem para produção dos diferentes sons, designados por articuladores, não permanecem fixos durante a produção de voz. É natural que a variação das características do sinal no tempo, causadas por essa variação contínua na produção, seja utilizada na percepção. Diversos estudos comprovam esse facto para diversas classes de sons (Cooper *et al.*, 1952). Avanços recentes em métodos fisiológicos e perceptuais, associados a inovações em abordagens computacionais e modelos, levaram ao surgimento da hipótese de que a dinâmica pode conter a informação relevante. Existe, no entanto, um conhecimento parcial acerca das características dinâmicas do sinal de voz. Investigação continuada nesta área deverá permitir novos conhecimentos. Trabalhos recentes propõem a utilização de unidades inerentemente dinâmicas como base de uma teoria fonológica (Browman e Goldstein, 1990, 1992, 1989, 1995). O estudo sistemático dos efeitos de coarticulação é de especial importância para o desenvolvimento da Fonética experimental e das ciências relacionadas com o processamento de voz (Kühnert e Nolan, 1997).

O trabalho apresentado insere-se num projecto a longo prazo de desenvolvimento de um sistema de síntese a partir de texto baseado em síntese articulatória.

Este trabalho tem um carácter multidisciplinar, envolvendo conhecimento em áreas tão diversas como a anatomia e fisiologia, física de fluidos, processamento de sinal, programação e fonética.

1.1 Objectivos

A formação, em Engenharia, do autor levou a que no início do trabalho a motivação se enquadrasse "apenas" na síntese de elevada qualidade do Português. A opção pela língua portuguesa, na sua variante do Português Europeu, deve-se ao facto de haver todo o interesse em disponibilizar, para os utilizadores da língua portuguesa, serviços de elevada qualidade baseados no processamento automático da voz.

Como a síntese articulatória se afigurava como a técnica mais promissora, passamos ao estudo da sua utilização num sistema de síntese. As limitações actuais, do método e do conhecimento acerca da voz, levou-nos a enveredar por um novo caminho. Este novo caminho tem por objectivo obter mais conhecimentos e integrá-los no sistema de síntese de forma a melhorar a qualidade.

O objectivo principal deste trabalho foi o de estudar e aplicar as técnicas da síntese articulatória de voz a estudos de produção e percepção do Português Europeu.

Como não dispomos de processos para obter directamente a área do tracto vocal, como Imagens de Ressonância Magnética ou raios X, utilizamos o método, designado por inversão, de obtenção da posição dos articuladores a partir do sinal sonoro, mas estamos conscientes que a validação destes métodos deve ser feita usando dados anatómicos obtidos por medição directa.

Não sendo possível abarcar todas as classes de sons do Português, optamos por nos dedicar ao estudo das vogais nasais do Português [3]. Pode justificar-se esta escolha por diversas razões: ser o passo seguinte relativamente ao estudo das vogais orais; colocar problemas com interesse no domínio da análise devido à existência de pólos e zeros; não estar resolvido o problema de inversão para esta classe de sons; ser bastante variável o tracto nasal entre pessoas; existirem problemas de saúde associados ao tracto nasal que seria interessante estudar; o facto de a língua inglesa não usar como marca fonológica distintiva a nasalidade levou a que não se tenham feito muitos estudos; o facto de ser comum na literatura, da área da Fonética, a referência às especificidades das nasais do Português (Strevens, 1954); a riqueza em sons nasais na língua portuguesa e a certeza de que só com um sintetizador que encare este problema seriamente se obterá alta qualidade para o Português.

Pretendemos, portanto, contribuir para os conhecimentos sobre os sons nasais vocálicos do Português Europeu.

Os objectivos concretos para este trabalho foram:

1. Implementar um sintetizador articulatório adequado à síntese de sons nasais;
2. Obter sons nasais sintéticos de qualidade próxima da natural;
3. Contribuir para o aumento de conhecimento acerca dos sons nasais, em especial dos sons nasais vocálicos do Português Europeu.

[3] Até porque muito existe ainda por fazer, mesmo para as vogais orais!

1.2 Estrutura da tese

Na escrita houve uma preocupação de disponibilizar o máximo de informação de forma a permitir que trabalhos futuros necessitem de menos tempo para obter conhecimentos de base. A tese tem também um objectivo didáctico.

Como grande parte dos trabalhos científicos sobre os assuntos focados é em língua estrangeira, em especial Inglês, torna-se por vezes difícil apresentar termos em Português. O nosso objectivo foi o de utilizar, sempre que possível, um termo Português, evitando a introdução no texto de um número alargado de estrangeirismos. No entanto, nem sempre se encontra uma tradução adequada. Neste caso utilizaremos o termo estrangeiro em itálico.

No caso de acrónimos de expressões estrangeiras, optamos por utilizá-los referindo na sua primeira utilização a sua origem. São todos reunidos numa lista de acrónimos, no final.

No caso de termos científicos em Latim, como os nomes de músculos, não é feita qualquer tradução. O tipo de letra utilizado será o itálico.

1.2.1 Organização

A tese consiste em quatro partes distintas: a primeira, em que se apresentam os conhecimentos que julgamos importantes para a compreensão do trabalho por nós efectuado; a segunda, em que é descrito o sintetizador articulatório implementado; a terceira em que são apresentadas as experiências efectuadas; a quarta, e última, onde se discutem os resultados obtidos e se indicam, com base nos conhecimentos adquiridos, possíveis caminhos a explorar de futuro.

Em termos de capítulos a tese tem a seguinte organização:

- Este capítulo apresenta as motivações para o trabalho, objectivos e principais contribuições;

- No capítulo 2 apresentam-se conhecimentos acerca das vogais nasais, que julgamos necessários para o não especialista, relacionados com: o processo de produção de voz, Fonética dos sons do Português, evolução das vogais nasais em Português, características acústicas, e percepção. São também referidos alguns dos estudos mais relevantes, tendo por objecto as vogais nasais;

- No capítulo 3 dá-se uma panorâmica das técnicas utilizadas em síntese articulatória. São apresentados os principais componentes de um sintetizador articulatório, e os vários tipos de modelos existentes para cada um deles. Este capítulo tem por objectivo fornecer conhecimentos de base necessários à compreensão do capítulo seguinte;

- O sintetizador desenvolvido é apresentado, com algum detalhe em termos de implementação, no capítulo 4;

- Os estudos de interacção acústica entre a fonte glotal e as cavidades supra-laríngeas são apresentados no capítulo 5;

- O capítulo 6 apresenta resultados de estudos de percepção destinados a aferir a importância da variação no tempo dos articuladores para a produção de vogais nasais de qualidade natural;

- Sendo o objectivo inicial a obtenção de sinal de voz sintético de qualidade elevada, não se poderia deixar de avaliar a qualidade obtida. No capítulo 7 é avaliada a qualidade obtida, fazendo uso dos conhecimentos adquiridos neste trabalho;

- Encerra-se a tese, no capítulo 8, com um resumo do trabalho efectuado, discussão dos principais resultados obtidos e apresentação de algumas propostas para continuação do trabalho.

1.3 Principais contribuições

As principais contribuições originais deste trabalho são:

1. O desenvolvimento de um sintetizador articulatório para síntese e estudo dos sons nasais do Português. A aplicação de síntese articulatória no estudo e na síntese dos sons nasais portugueses constitui uma novidade e, apesar de baseado em técnicas existentes, o sintetizador inclui alguns pontos inovadores:

    - Modelo muito completo do tracto nasal, permitindo: modelos das cavidades nasais, incluindo as duas passagens paralelas e radiação por cada uma das narinas; obter radiação apenas nas narinas ou nos lábios; o cálculo da impedância do tracto, incluindo ou não a impedância nasal; a inclusão de seios paranasais; a obstrução das cavidades nasais em qualquer ponto e, de uma forma fácil, alterar a informação acerca da configuração das cavidades nasais utilizada pelo sintetizador;
    - A definição independente dos parâmetros que controlam o sintetizador. Para um determinado instante temporal apenas é necessário definir os valores dos parâmetros que contribuem para o fim em vista. Por exemplo, para a realização de uma oclusiva bilabial, apenas se torna necessário definir o parâmetro articulatório que controla a abertura dos lábios e, possivelmente, a abertura do maxilar;
    - Modelo de fonte glotal interactivo incluindo: o modelamento de irregularidades como a variação de período para período da frequência fundamental, a possibilidade de variação dos vários parâmetros de controlo ao longo do tempo e a possibilidade de incluir ou excluir, na interacção entre a fonte e o tracto, o efeito de carga das cavidades nasais e paranasais.

O sintetizador constitui uma ferramenta muito útil para a realização de estudos de produção e percepção dos sons nasais do Português, como o demonstram os diversos estudos efectuados para a elaboração deste trabalho.

2. A obtenção dos parâmetros articulatórios, utilizando apenas informação do sinal de voz natural, problema designado por inversão, das vogais orais do Português foi também realizado pela primeira vez, com resultados bastante satisfatórios. As configurações obtidas foram utilizadas na obtenção de estímulos para os testes perceptuais efectuados e permitiram também a produção de pequenas palavras, constituídas por sons nasais (consoantes e vogais), facilmente inteligíveis e de boa qualidade.

3. O estudo de interacção acústica entre as cavidades acima da glote e a onda de excitação produzida pelas cordas vocais, para sons nasais. Estudaram-se as alterações, na onda de excitação, causadas pela inclusão adicional das cavidades nasais. Este estudo, baseado em simulações, foi efectuado não só para configurações estáticas de vogais orais e nasais, mas também para vogais nasais entre oclusivas. Foi, também, investigado, através de testes perceptuais, se é possível que ouvintes portugueses detectem essas alterações. O tipo de estudo efectuado, os resultados mostrando a dependência da altura da vogal e a proposta de explicação são novos.

4. O estudo do efeito da variação no tempo dos articuladores, em especial do velo, na qualidade de vogais nasais produzidas pelo sintetizador articulatório. Os nossos estudos contemplaram o caso de vogais nasais situadas a seguir a consoantes nasais, situação geralmente não abordada em outros estudos. Com base nos resultados obtidos é proposta uma teoria para as vogais nasais portuguesas, em que estas são consideradas como compostas por duas fases, assumindo a configuração de um ditongo.

5. A obtenção, pela primeira vez para o Português Europeu, de vogais nasais e palavras contendo vogais nasais e consoantes nasais, de boa qualidade, usando síntese articulatória. O teste de qualidade efectuado reuniu, para uma primeira comparação, para o Português, exemplos de síntese articulatória, síntese de formantes e síntese por concatenação.

1.3.1 Resultados publicados

Foram já publicados:

- A descrição do sintetizador (Teixeira *et al.*, 1997b) e o modelo acústico para o tracto nasal (Teixeira *et al.*, 1998a);
- Os resultados de experiências sobre a influência da dinâmica dos articuladores (Teixeira *et al.*, 1998b, 1999b) e (Teixeira *et al.*, 2000);
- A parte inicial do estudo da interacção fonte-tracto (Teixeira *et al.*, 1999a).

Encontra-se submetido um trabalho acerca da avaliação da qualidade dos sons nasais produzidos usando síntese articulatória e outras técnicas de síntese (formantes e concatenação) (Teixeira e Vaz, 2000b).

Além destes, foram publicados trabalhos acerca de síntese articulatória, geralmente com objectivo de divulgação da área (Teixeira e Vaz, 2000a; Teixeira *et al.*, 1997a,c)

Na área deste trabalho tivemos também a oportunidade de colaborar num outro de pós-graduação (Branco, 1997) do qual resultaram as publicações (Branco *et al.*, 1997c,b,a).

1.4 Meios utilizados

O sintetizador articulatório, bem como programas auxiliares, foram desenvolvidos num sistema Linux com XWindows (XFree86), num cumputador pessoal (PC) com placa de áudio. O desenvolvimento foi feito usando C/C++ (compilador da GNU) e Tcl/Tk (Ousterhout, 1998, 1994; Welch, 1995; Harrison e McLennan, 1997). Foram utilizadas duas extensões do Tcl/Tk, BLT para gráficos, representando o sinal de excitação e de voz, e Snack para visualização de espectrogramas. Foi ainda usada a implementação de *Fast Fourier Transform* (FFT) denominada de *Fast Fourier Transform in the West* (FFTW), desenvolvida no *Massachussets Institute of Technology* (MIT) (Frigo e Johnson, 1998; Frigo, 1997), funcionalidades do Edimburgh Speech Tools (Taylor *et al.*, 1999), assim como o código do *Simulated Annealing* desenvolvido por Goffe *et al.* (1994).

A tese foi escrita em LaTeX2e. As páginas iniciais, normalizadas por regulamentação da Universidade de Aveiro, foram obtidas utilizando um conjunto de macros desenvolvidas por Tomás Oliveira e Silva.

Capítulo 2

Preliminares

> In our perspective, development on synthesis *as well as* recognition devices can profit from looking at human speech performance. However, this position must be understood correctly. We do not propose that machines should be constructed exactly like humans.
>
> Eric Keller e Jean Caelen
> (Keller, 1994, pág. 175)

> Phonetics, the study of speech, is the bedrock of scientific study of language
>
> Joseph Olive *et al.* (1993)

Como os modelos articulatórios modelam directamente o processo de produção de voz pelos seres humanos, nada mais natural do que estudar este processo antes de nos dedicarmos às aproximações criadas pelo homem.

Como estes conhecimentos se dispersam por diversas áreas do conhecimento, julgamos ser útil, em especial por este trabalho ser efectuado numa área, Engenharia Electrotécnica, com conhecimentos de base e linguagens bastante distintas.

Na secção inicial apresenta-se a descrição dos órgãos intervenientes no processo de produção de voz humano. Devido ao assunto deste trabalho descreve-se de seguida a produção de sons nasais, na secção 2.2. Segue-se a apresentação de alguns conceitos da área da Fonética e a inventariação dos sons do Português, na secção 2.5. Alguns factos acerca do processo de criação e desenvolvimento das vogais nasais são descritos na secção 2.6. Segue-se um resumo das características acústicas, secção 2.7, e percepção, secção 2.8, das vogais nasais. No final do capítulo, secção 2.10, referem-se alguns estudos sobre as vogais nasais, dando especial atenção aos estudos no âmbito da síntese articulatória.

2.1 Anatomia e Fisiologia do aparelho produtor de voz

É conveniente, e funcionalmente adequado, considerar a produção de voz em termos de três componentes: (1) a laringe que é usada como referência nesta divisão; (2) o sistema sub-laríngeo, ou sistema subglotal, incluindo os pulmões e estruturas associadas, situado abaixo da laringe; e (3) o sistema supra-laríngeo, ou supraglotal, compreendendo as cavidades faríngea, oral e nasal, situado na parte superior (Lieberman e Blumstein, 1988, pág. 3).

2.1.1 Sistema sub-laríngeo

As principais estruturas deste sistema são: a traqueia, os brônquios e os pulmões. Estas estruturas encontram-se na cavidade torácica.

Na sua parte inferior, a cavidade torácica está separada da cavidade abdominal por uma estrutura em forma de abóbada, o diafragma. Este é constituído por fibras musculares e, na sua região central, por tecido tendinoso.

Quando as costelas se elevam ou se comprimem, o volume da caixa torácica aumenta (inspiração) ou diminui (expiração), respectivamente. A variação da capacidade da caixa torácica é controlada pelos músculos respiratórios (Draper *et al.*, 1959).

A traqueia é um tubo aberto nas duas extremidades que desce da laringe, na parte anterior do pescoço, e ramifica-se nos dois brônquios, ao nível da quinta vértebra torácica. Cada brônquio alimenta um pulmão, subdividindo-se sucessivamente em ramos duplos, e por vezes triplos os bronquíolos, que terminam em pequenos "sacos", os alvéolos. Estes últimos constituem a maior parte dos pulmões.

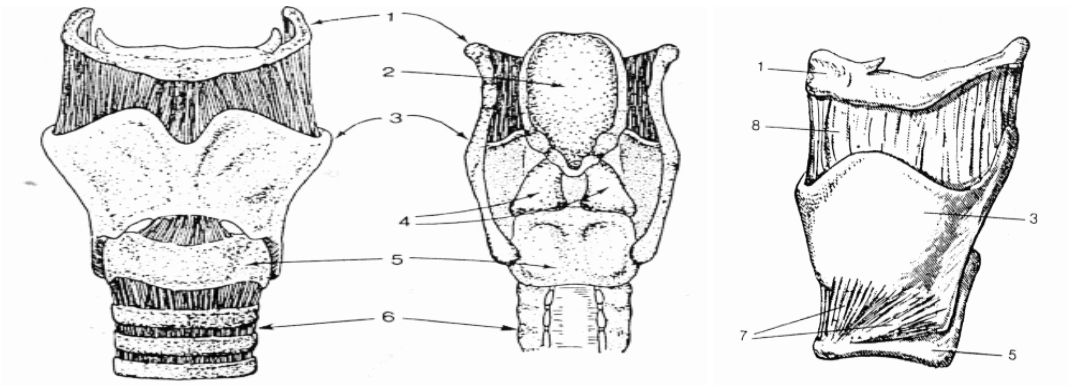
O mecanismo da respiração fornece a energia básica necessária para a produção da fala. A alternância da expansão e da redução do volume pulmonar resulta da interacção de forças elásticas, dos pulmões e caixa torácica, de forças gravitacionais e das forças dos músculos respiratórios (Ploysongsang e London, 1991) (Lieberman e Blumstein, 1988, pág. 5).

Na fala, a inspiração é mais rápida e mais profunda. A fase de expiração na fala é bastante variável, contrariamente ao que se verifica na respiração em silêncio. A grande maioria dos sons da linguagem é produzida durante a expiração. A pressão subglotal mantém-se relativamente constante entre pausas inspiratórias na fala (Lieberman e Blumstein, 1988, pág. 91). Com a saída do ar pulmonar, quando a pressão das forças deixa de ser mais elevada do que a pressão pulmonar, entram em actividade os músculos expiratórios. A fisiologia do aparelho respiratório permite, deste modo, assegurar não só o já referido alongamento da fase expiratória mas ainda a manutenção da pressão subglotal relativamente constante durante a produção de voz.

2.1.2 Laringe

A laringe é uma estrutura constituída por cartilagens, ligamentos, músculos e tecido membranoso, que estabelece a comunicação entre as vias respiratórias superiores e inferiores (Mateus *et al.*, 1990, pág. 82).

A laringe encontra-se como que suspensa, estando ligada a estruturas ósseas superiores (o osso hióide, o crânio, a mandíbula) e inferiores (o esterno e a omoplata) pelos ligamentos extrínsecos. Os chamados músculos extrínsecos controlam os movimentos de elevação (elevadores) e depressão (depressores) da laringe relativamente à sua posição normal.



1 - osso hióide	2 - epiglote	3 - cartilagem tiróide	4 - cartilag. aritenóides
5 - cartilagem cricóide	6 - traqueia	7 - músculo crico-tiroideu	8 - músculo tito-hioideu

Figura 2.1: Estrutura cartilaginosa da laringe (adaptado de Mateus *et al.*, 1990). Da esquerda para a direita, vista de frente, secção coronal pondo a descoberto as cartilagens aritenóides, e laringe vista de lado.

O esqueleto cartilaginoso, na Figura 2.1, é basicamente constituído pela cartilagem tiróide, aberta atrás e com uma forma que se assemelha à de um escudo, a cartilagem cricóide que faz lembrar um sinete, e as duas pequenas cartilagens de forma piramidal, as cartilagens aritenóides. Estas últimas estão assentes sobre a cricóide e encontram-se do lado de trás da laringe. A tiróide é comummente conhecida como "maçã de Adão", encontrando-se do lado da frente da laringe.

Na laringe encontram-se duas pregas musculares designadas por cordas vocais [1]. As cordas vocais revestem lateralmente as paredes laríngeas, uma de cada lado, e são complexamente controladas pelos músculos intrínsecos. Basicamente as cordas vocais são constituídas pelo par de músculos laríngeos tiro-aritenoideus e pelos ligamentos vocais (Figura 2.2). São estes últimos que entram em vibração na produção de certos tipos de sons. Recobre as cordas vocais uma fina membrana mucosa. O espaço entre as cordas vocais chama-se glote.

Os músculos intrínsecos (Titze, 1994, pág. 12) têm o papel fundamental de controlar a posi-

[1] O nome advém de um erro de um anatomista, pois de facto não são cordas !

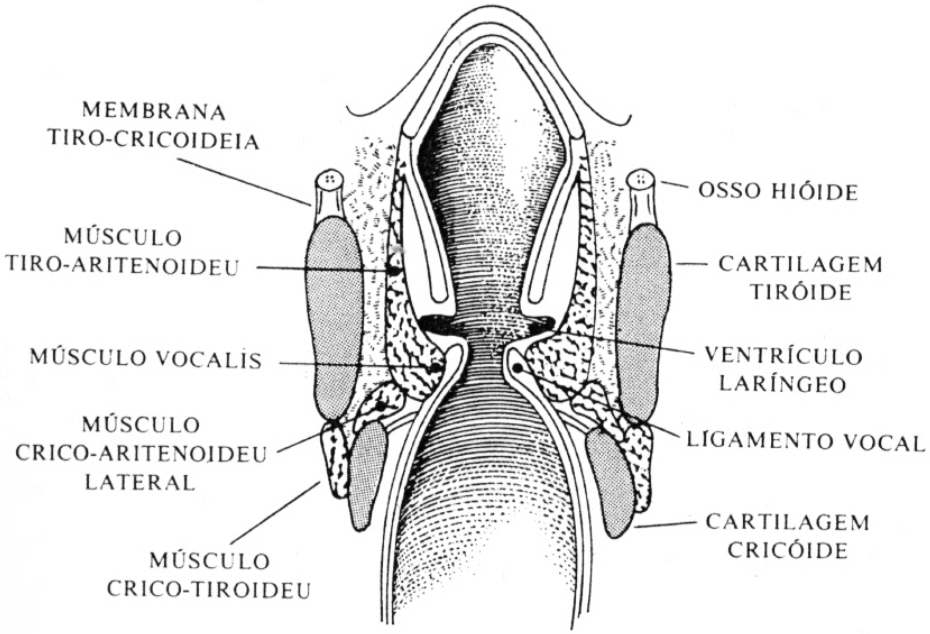


Figura 2.2: Secção coronal da laringe (adaptado de Mateus *et al.*, 1990).

ção relativa das cordas vocais e a tensão das mesmas. O músculo tiro-aritenoideu estende-se da tiróide até às cartilagens aritenóides, sendo o principal responsável pela massa das cordas vocais. Quando contrai faz as aritenóides tenderem a aproximar-se da tiróide e condiciona o grau de tensão dos ligamentos vocais. Conforme o estado dos restantes músculos, assim a contracção deste músculo determina o aumento ou redução da tensão dos ligamentos. Os músculos crico-aritenoideus laterais, ligados lateralmente às aritenóides e à cricóide, ao contraírem fazem as aritenóides rodarem para dentro resultando na adução (aproximação) das cordas vocais. A contracção dos inter-aritenoideus, que se estendem entre as duas aritenóides, também resulta na adução das cordas quando estas estão tensas. O músculo crico-aritenoideu posterior é o músculo abdutor, fazendo, quando activado, as aritenóides rodarem para fora, abduzindo (afastando) as cordas vocais. O crico-tiroideu, que se estende entre a cricóide e a tiróide, ao contrair faz as duas cartilagens aproximarem-se uma da outra. A tiróide inclina-se em direcção à cricóide e esta sobe. Deste movimento resulta um aumento da tensão longitudinal das cordas.

Teoria aerodinâmica-mioelástica da fonação

Estando o grau de tensão necessário assegurado, e sendo adequadamente aduzidas, as cordas vocais entram em actividade vibratória por acção de forças aerodinâmicas e forças elásticas dos próprios tecidos (Mateus *et al.*, 1990, pág. 87).

A iniciação da vibração necessita que as cordas vocais estejam adequadamente aproximadas e

não excessivamente tensas, e que a pressão abaixo da laringe (subglotal) seja suficientemente superior à pressão acima destas (supraglotal), para vencer a resistência das cordas, fazendo-as afastar-se uma da outra. Com o afastamento, o ar escapa-se a alta velocidade através da glote, provocando uma diminuição da pressão entre os dois extremos da glote (transglotal), devido ao efeito de Bernoulli (van den Berg *et al.*, 1957), criando condições para que as cordas se voltem a aproximar. Fechada a glote, a força subglotal volta a ser significativamente grande e o ciclo repete-se: as cordas vocais entram em vibração (Faria *et al.*, 1996, pág. 132).

Esta teoria acerca da vibração das cordas vocais, geralmente aceite actualmente, proposta por van den Berg (1958), deve o seu nome ao facto de incluir o efeito de forças aerodinâmicas e forças musculares (mio-elásticas). Mais detalhes podem ser encontrados em (Stevens, 1998; Titze, 1994; Clark e Yallop, 1990).

2.1.3 Cavidades supra-laríngeas

As cavidades supra-laríngeas incluem a cavidade oro-faríngea, ou tracto vocal, e as cavidades nasais, também designadas por tracto nasal.

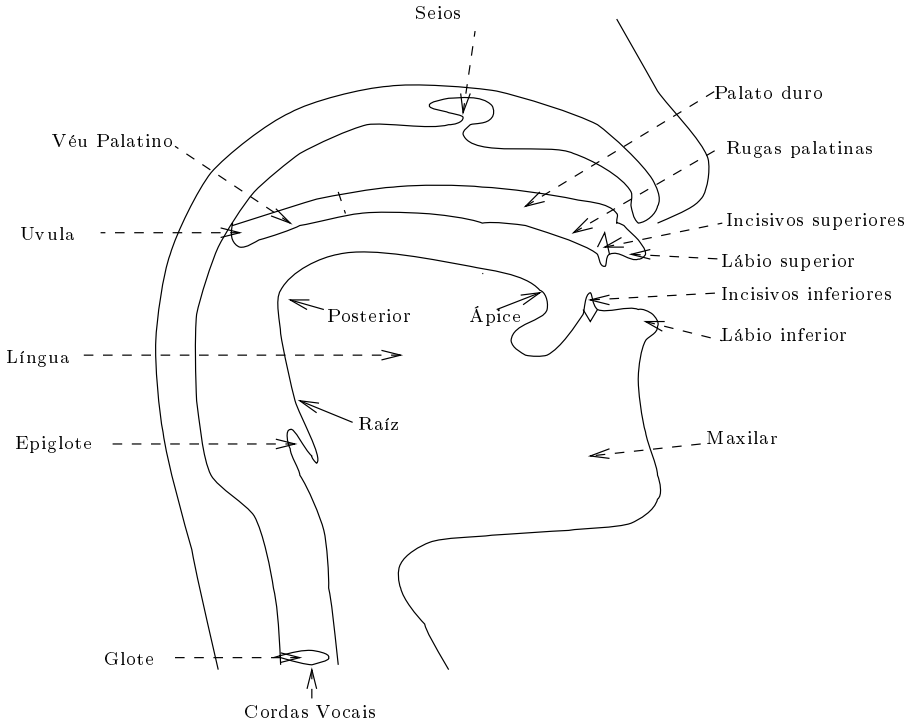


Figura 2.3: Corte sagital do sistema vocal humano. Adaptado de Jassem e Nolan (1984)

2.1.3.1 Tracto vocal

Geralmente divide-se o tracto vocal em duas zonas. A região entre a laringe e o véu palatino é denominada de faringe. A zona do véu palatino até aos lábios é designada habitualmente de cavidade bucal.

Na região superior da cavidade bucal, encontram-se: o lábio superior, os dentes incisivos superiores, os alvéolos, o palato duro e o palato mole, véu palatino ou velo, com a úvula na sua extremidade.

Na região inferior da cavidade bucal, temos: o lábio inferior, os incisivos inferiores, e a língua.

Na língua podemos distinguir três regiões principais: a coroa, o corpo e a raiz (Faria *et al.*, 1996, pág. 137). A coroa divide-se na ponta ou ápice e na lâmina. Esta é a parte imediatamente atrás do ápice que se estende, em geral, até 1 ou 2 *cm* a contar da ponta quando esta se encontra em posição de repouso. No corpo da língua temos três regiões: a frente ou região pré-dorsal, o centro ou região dorsal e a parte posterior ou região pós-dorsal. A raiz da língua encontra-se do lado oposto à parede da faringe, em frente da epiglote.

As estruturas móveis da cavidade bucal utilizadas na produção dos sons da linguagem são normalmente designadas por articuladores. São considerados articuladores os lábios, a língua, a faringe, e até certo ponto, o maxilar inferior. A língua é, de longe, o articulador de maior mobilidade e mais flexível. Funciona, na prática, como vários articuladores diferentes relativamente independentes uns dos outros. Os articuladores com maior mobilidade são o ápice e a lâmina da língua. O velo, que funciona sobretudo como uma válvula entre as regiões oral e nasal, é também um articulador.

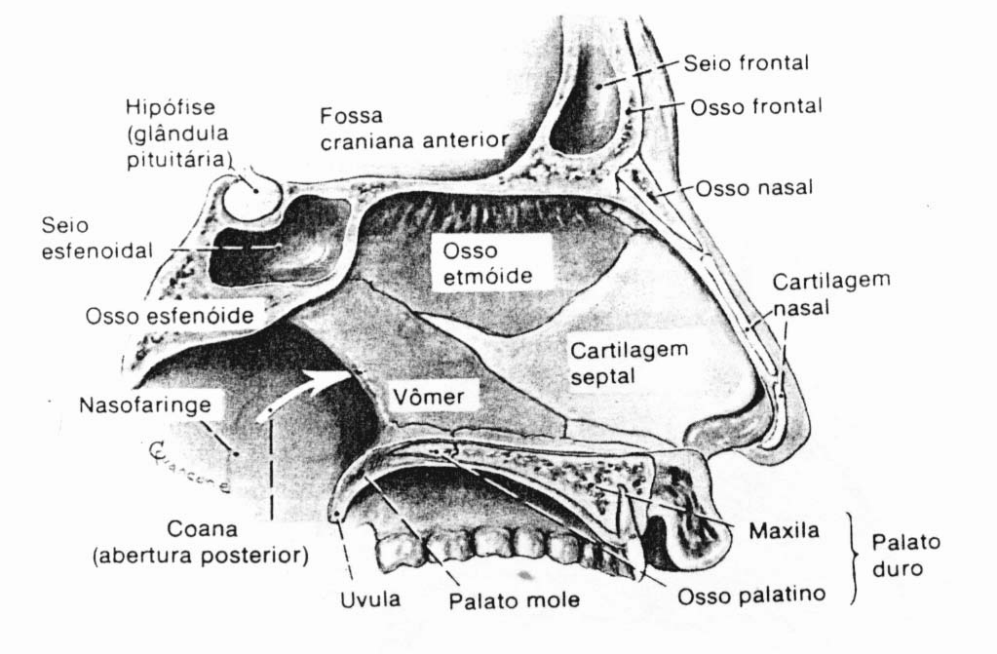
Uma apresentação mais detalhada pode ser encontrada em (Mateus *et al.*, 1990, pág. 45), (Faria *et al.*, 1996, pág. 136), e (Clark e Yallop, 1990, pág. 47).

2.1.3.2 Tracto nasal

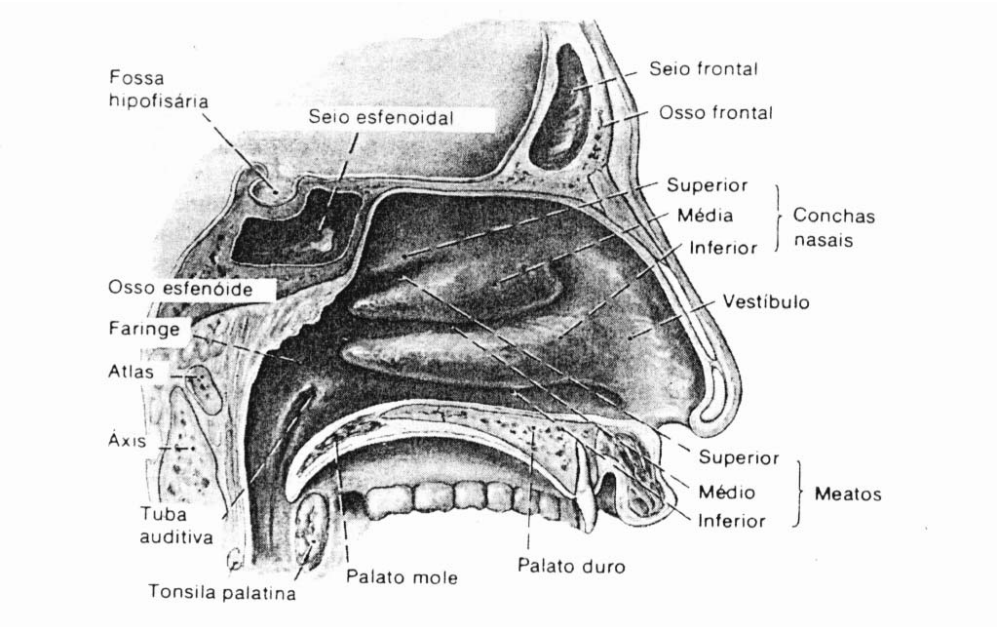
As cavidades nasais (tracto nasal), na Figura 2.4, situam-se entre o véu palatino e as narinas, por um lado, e entre o palato e o cérebro, por outro.

O tracto nasal é um labirinto complicado de passagens de ar e cavidades. Na zona do véu palatino, o tracto nasal é designado por nasofaringe, sendo a continuação da faringe. A nasofaringe termina na coana, onde ocorre uma bifurcação iniciando-se duas passagens, laterais, separadas pelo septo nasal. Estas duas passagens, que terminam nas narinas, não são simétricas, geralmente devido ao deslocamento para um dos lados do septo (Zemlin, 1988).

As cavidades nasais não são simples tubos vazios, estando cobertas por mucosa. Para aumentar a superfície da mucosa, as cavidades encontram-se parcialmente preenchidas por estruturas ósseas, as conchas, que são cobertas pela membrana mucosa. Existem três passagens principais, designadas por meatos entre essas estruturas ósseas. A mucosa, e a forma complexa das passagens, contribuem para um amortecimento acrescido do sinal acústico a todas as



(a) Mostrando o septo nasal



(b) O septo nasal foi retirado para mostrar as conchas nasais

Figura 2.4: Corte sagital da cavidade nasal (adaptado de Jacob *et al.*, 1990)

frequências (Entenman, 1976, pág. 3).

Tendo um volume e forma fixos em cada pessoa, as cavidades nasais apenas acrescentam o timbre próprio da sua ressonância aos sons, mas não permitem diferentes sons nasais. A variedade dos sons nasais é produzida por variações na cavidade bucal.

De todas as cavidades supra-laríngeas, o tracto nasal é o mais inacessível em termos de medição (Krakow e Huffman, 1993). O seu comprimento, medido usando imagens de raios-X da úvula às narinas, é de cerca de 12.5 *cm* (Fant, 1960). Dang e Honda (1994) obtiveram um valor médio de 11.6 *cm* com desvio padrão de 0.13 *cm*. Story (1995) obteve, usando ressonância magnética, um comprimento de 2.1 *cm* entre a zona de acoplamento da cavidade nasal ao tracto oral e a bifurcação. O comprimento da bifurcação às narinas é de cerca de 8 *cm* (Fant, 1960; Story, 1995). Para um adulto do sexo masculino a cavidade nasal encontra-se acoplada ao tracto vocal num ponto aproximadamente 8 *cm* acima da glote.

Diversos investigadores dedicaram-se ao estudo da configuração destas cavidades. Os primeiros estudos basearam-se em estudo de cadáveres (Bjuggren e Fant, 1964), mais tarde usaram-se raios-X, e mais recentemente ressonância magnética (Dang e Honda, 1994; Story, 1995). Bjuggren e Fant (1964) obtiveram para a superfície das cavidades nasais um valor de cerca de 3.5 vezes a de um tubo cilíndrico com a mesma área de secção. Dang e Honda (1994) descobriram assimetria entre as passagens nasais, área variável entre os indivíduos estudados, tendo, no entanto, todos eles áreas maiores na região posterior ao septo.

Seios paranasais

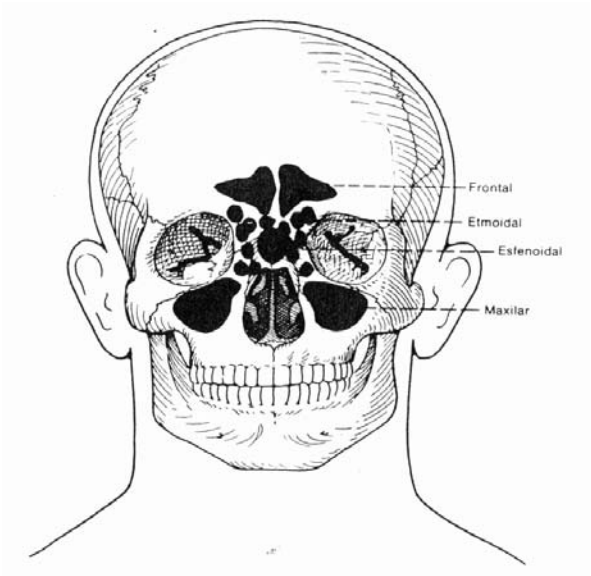


Figura 2.5: Seios paranasais (adaptado de Jacob *et al.*, 1990, pág. 104).

Os seios paranasais, nas Figuras 2.4 e 2.5, são espaços que contêm ar e que se comunicam

com as cavidades nasais (Jacob *et al.*, 1990, pág. 378). Existem quatro grupos, que ocorrem geralmente aos pares, não sendo no entanto simétricos. A principal função dos seios paranasais é o de tornar os ossos do crânio mais leves. Têm também funções secundárias, como a de fornecimento de muco para a cavidade nasal e a de actuarem como ressonâncias na produção do som (Meyerhoff e Schaefer, 1991). O seu papel na produção de sons nasais é ainda alvo de controvérsia. Para alguns, como Maeda (1982b), desempenham um papel importante no aparecimento das ressonâncias de frequência mais baixa em sons nasais; para outros, como Feng e Castelli (1996), não são suficientes para explicar essas propriedades dos sons nasais.

Os seios maxilares, situados abaixo dos olhos e lateralmente à cavidade nasal, são os de maior volume. Maeda (1982b) refere um volume total dos dois seios de 20.8 cm^3; as medições de Story (1995) apontam para valores de 11.8 cm^3 para o direito e 13.5 cm^3 para o esquerdo. Nas medições de Dang e Honda (1994), os valores médios foram de 17.3 e 15.7 cm^3 para o maxilar direito e esquerdo, respectivamente. Cada um está localizado na maxila e tendo ligação com o meato médio. As ligações são de tamanho e área reduzida, tornando muito difícil a sua medição. A título de exemplo, Story (1995) obteve para a ligação do seio maxilar esquerdo, assumindo a configuração da ligação como um cilindro de secção constante, um comprimento de 0.25 cm e área de 0.12 cm^2, não lhe tendo sido possível detectar a ligação ao seio direito.

O seio frontal, localizado no osso frontal, acima dos olhos, esvazia-se no meato médio. Pode apresentar-se divido ou como uma única cavidade. Dang e Honda (1994) obtiveram para o seio frontal direito uma média de 2.6 cm^3 e para o esquerdo 3.6 cm^3.

As células etmoidais são numerosas e irregulares, tendo ligação com os meatos superior e médio. Apresentam-se com um padrão semelhante a um favo de abelhas. Existem na parte anterior e posterior. O volume destes seios é muito difícil de medir usando as técnicas actuais, devido às suas reduzidas dimensões e por serem constituídos por várias partes.

O seio esfenoidal está situado no osso esfenóide, na parte posterior do olho, por trás da porção superior da cavidade nasal, quase na posição central do crânio. Pode ser simples ou apresentar-se dividido. A drenagem deste seio é feita para o meato superior. O volume obtido por Story (1995) foi de 6.5 cm^3. Nas medições de Dang e Honda (1994) os seios esfenoidais esquerdo e direito apresentaram volumes médios de 8.9 cm^3 e 9.0 cm^3, respectivamente.

Véu palatino

O véu palatino, ou velo, pode ser considerado como a continuação flexível do palato duro. Toma a forma de uma fina camada de fibras musculares, tecido, vasos sanguíneos, nervos e glândulas, com função principal de separar as cavidades nasais das cavidades orais. Quando abaixado, permite a passagem do ar pelas cavidades nasais. Quando totalmente subido, o velo fecha a entrada da cavidade nasal.

O velo encontra-se ligado anteriormente ao palato duro, superiormente ao crânio e inferiormente à língua e faringe.

Na ponta do velo situa-se a úvula. O seu papel não é muito importante para a produção de voz excepto como articulador na produção de vibrantes uvulares e como articulador passivo durante a produção de consoantes pós-velares (Hardcastle, 1976, pág. 121).

O grau de abertura do véu palatino varia de acordo com o contexto fonético. Para as nasais encontra-se aberto, para as vogais baixas numa posição intermédia, para as vogais altas quase fechado, para as oclusivas completamente fechado (Childers e Ding, 1991).

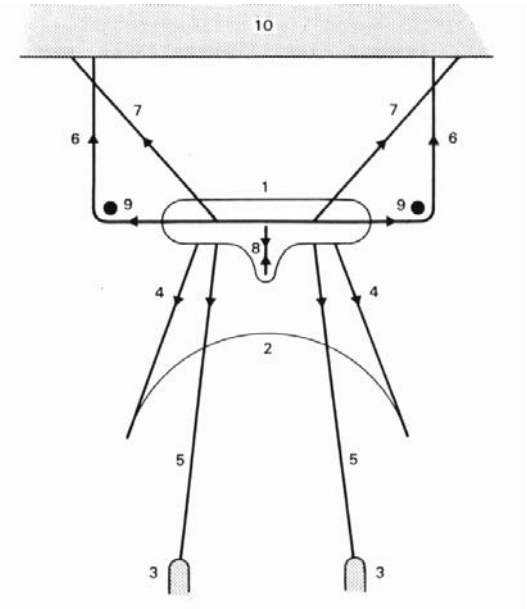
Os estudos tomográficos e cineradiográficos de Björk (1961) concluíram por uma dependência linear entre a área de abertura e o eixo sagital menor, com uma constante de proporcionalidade de 10 mm, sendo a abertura aproximadamente rectangular. Diversos investigadores (Björk, 1961; Warren, 1967) são da opinião que as dimensões úteis linguisticamente da abertura do véu palatino se encontram no intervalo entre zero e mais de 1 cm^2. Uma regra é a de que, quando o véu palatino se aproxima a menos de 2 mm da faringe (área de cerca de 20 mm^2), não existe nasalidade aparente, enquanto uma maior abertura (área de 5 a 50 mm^2) produz ressonâncias nasais, sendo o som resultante perceptualmente nasal.

Björk (1961) nos seus estudos também estudou a velocidade do véu palatino. Com base nesses estudos, em Bjork *et al.* (1961), refere-se que:

1. Em fala normal a duração média do movimento do velo entre o estado de fechado e de aberto é da ordem de 130 ms e que o movimento oposto, o fecho, demora 160 ms em média.

2. A velocidade de movimento do velo não é alterada proporcionalmente à velocidade de pronunciação em geral. Para discurso lento, normal e rápido, correspondente a durações de $100 - 200 - 300$, numa escala relativa, os movimentos do velo variaram na proporção de $100 - 130 - 160$.

Os movimentos do véu palatino são controlados por diversos músculos, apresentados na Figura 2.6, que trabalham em cooperação. A subida do velo é conseguida principalmente pela acção de dois músculos, o *levator veli palatini* (Hardcastle, 1976, pág. 122) e o *superior pharyngeal constrictor*. O primeiro liga a superfície frontal do velo com a base do crânio, sobe e retrai o velo cerca de 2 cm, em média, e para uma adulto. O segundo, que tem por função principal empurrar a comida para baixo em direcção ao esófago, pode também contrair para subir o velo. Outro músculo, o *tensor palatini* (Hardcastle, 1976, pág. 123), também actua para alongar e tornar tenso o velo ao subir. A descida é afectada por dois músculos, assistidos em parte pelo relaxamento dos músculos intervenientes na subida e pela força da gravidade. Por um lado, existe o *palatoglossus* (Hardcastle, 1976, pág. 124) que se estende para baixo, desde a parte inferior da superfície do velo, dividindo-se em dois e depois ligando-se à língua. Tanto pode baixar o velo, como subir a parte posterior da língua. Por outro lado, existe o *palatopharyngeous* (Hardcastle, 1976, pág. 124), um músculo longo e fino que liga o velo e a parte posterior da cartilagem tiróide e da parede lateral da faringe. Apesar de todos os indivíduos possuírem o mesmo conjunto de músculos, não os operam necessariamente de



1 véu palatino
2 língua
3 cartilagem tiróide
4 *palatoglossus*
5 *palatopharyngeus*
6 *tensor palatini*
7 *levator palatini*
8 *musculus uvulae*
9 processo pterigóideo (*hamular process of pterygoid bone*)
10 crânio

Figura 2.6: Diagrama da acção e localização dos músculos intervenientes nos movimento do velo (visto por trás) (adaptado Laver, 1980, pág. 71).

forma igual para atingir um certo gesto durante a produção de voz. O *palatoglossus* pode não ser activado por certos indivíduos para causar o abaixamento do velo. Também existem indicações de que o fechamento do velo provoca, em geral, um deslocamento para a parte frontal da parede posterior da faringe (Hardcastle, 1976; Lieberman e Blumstein, 1988).

2.2 Simplificações usuais

A nasalidade é tratada geralmente de uma forma simplificada, sem se explicitar quais as simplificações. Laver (1980) descreveu em detalhe as simplificações habituais, de que a seguir faremos um breve resumo.

A primeira aproximação consiste em considerar que quando o véu palatino se encontra fechado o som encontra-se livre de nasalidade, e inversamente, que quando o som é não nasal, o véu palatino se encontra fechado [2] . Experiências demonstraram que o véu palatino não fecha completamente a passagem da faringe para a cavidade nasal em situações normais. Warren em 1964 demonstrou mesmo que uma abertura até 10 mm^2 é adequada para a produção de oclusivas. Pode concluir-se que a nasalidade é causada tendo uma abertura grande para a nasofaringe em comparação com a abertura para a cavidade oral.

A segunda simplificação é a de que fluxo nasal resulta sempre em nasalidade, e inversamente

[2]Chomsky e Halle (Chomsky e Halle, 1968, pág. 316), por exemplo, escreveram que "os sons nasais são produzidos com o véu palatino abaixado o que permite a saída do ar pelo nariz; sons não nasais são produzidos com o véu palatino elevado para que o ar vindo dos pulmões pode apenas sair pela boca" (tradução do autor).

que a nasalidade requer sempre fluxo nasal. É certamente verdade que o fluxo nasal pode resultar em nasalidade, mas apenas em certas condições. Como anteriormente referido é possível produzir oclusivas orais com uma abertura para a cavidade nasal, que produz fluxo nasal. É também possível obter vogais sem nasalidade mesmo com passagem de ar pelo nariz, como no caso das vogais baixas inglesas. Claramente, o fluxo de ar pela cavidade nasal não é uma condição necessária ou suficiente para a produção de nasalidade audível. A nasalidade é essencialmente uma condição de ressonância. A cavidade nasal pode ser uma ressonância sem passagem de ar, basta reparar na capacidade que possuímos de produzir vogais nasais fechando completamente as narinas. Um factor vital para existir ressonância da cavidade nasal é a proporção entre as aberturas oral e nasal. Em sons não nasais a proporção entre a abertura nasal e oral é de $1:11$, e em voz nasal de $8,8:3,1$ (Kaltenborn, 1948 citado em Laver (1980)).

A terceira, e última, simplificação usual é a de que a ressonância da cavidade nasal é a única ressonância responsável pela produção de nasalidade, e inversamente que a nasalidade requer sempre ressonância nasal. O efeito de nasalidade pode ser produzido no tracto vocal sem a intervenção do tracto nasal. Foi demonstrado que características espectrais semelhantes às obtidas pelo acoplamento do tracto nasal podem ser conseguidas utilizando outras cavidades formadas pelo sistema muscular na parte inferior e superior da faringe (Laver, 1980, pág 84). Vogais adjacentes a sons com elevado fluxo de ar podem também ser percebidas como nasais (Sampson, 1999, pág. 11).

2.3 Transcrição Fonética

Antes de estudar os sons da linguagem, precisamos de tratar o problema da sua representação gráfica, por meio dos chamados alfabetos fonéticos.

Na ortografia, a mesma letra representa muitas vezes sons diferentes e o mesmo som é representado por letras diferentes. Por outro lado, usam-se por vezes duas letras consecutivas para representar um som. Para evitar as confusões e incertezas que daí resultam, usam-se os alfabetos fonéticos, nos quais um som corresponde sempre a um símbolo de uma forma biunívoca; marca-se sempre a sílaba acentuada com um sinal de acento agudo, colocado antes do símbolo que representa o primeiro som da sílaba acentuada; para indicar que se trata de uma transcrição fonética coloca-se tudo entre parêntesis rectos (Barbosa, 1994).

Vários alfabetos foram, e são, utilizados [3]. O mais divulgado, e o mais completo, é o Alfabeto Fonético Internacional, tendo sido o adoptado neste trabalho. Na descrição dos fonemas do Português serão apresentados os símbolos usados neste trabalho.

[3] O facto de na literatura portuguesa haver quase um alfabeto por cada autor dificulta bastante o estudo de Fonética para pessoas de outras áreas, como é o nosso caso. Esta observação não pretende ser uma crítica mas uma chamada de atenção para a necessidade de uniformização.

2.4 Classificação articulatória dos sons

Os sons dividem-se em vogais, sons em cuja articulação o ar proveniente do pulmões não encontra qualquer obstáculo que produza oclusão ou fricção, e consoantes, sons em que existe um obstáculo na cavidade laríngea ou bucal [4].

2.4.1 Consoantes

Tradicionalmente as consoantes são classificadas pelo modo de passagem do ar pelo tracto vocal, o modo de articulação, e a região do tracto vocal onde se situa a maior constrição, o ponto de articulação. O véu palatino também influi, dividindo as consoantes em nasais e não nasais. A vibração das cordas vocais produz os sons sonoros por oposição aos sons surdos [5], produzidos sem vibração destas.

2.4.1.1 Modo de articulação

A classificação do modo de articulação das consoantes é função do grau de aproximação relativa dos articuladores, da duração dessa aproximação ou, ainda, da modificação da configuração do tracto causada pela aproximação dos articuladores superiores e inferiores (Mateus *et al.*, 1990, pág. 48). Com interesse para a língua portuguesa, existem consoantes oclusivas, nasais, fricativas, africadas (sons combinados oclusivo-fricativos), laterais e vibrantes.

Oclusivas

A articulação destas consoantes implica o fecho completo da passagem do ar pelo canal bucal, encontrando-se também bloqueada a entrada do ar nas cavidades nasais. As oclusivas podem produzir-se com vibração das cordas vocais, designando-se por sonoras ou vozeadas, ou sem vibração, sendo designadas por surdas ou não-vozeadas.

Nasais

Se existir obstrução na cavidade bucal mas o véu palatino estiver descido, o ar pode passar pelas cavidades nasais, produzindo-se uma consoante nasal. Alguns autores (Mateus *et al.*, 1990, por exemplo) consideram as consoantes nasais como casos particulares das oclusivas. Considero que o seu processo de produção e propriedades acústicas são suficientemente diferentes para uma classificação separada. Ao contrário das oclusivas, as consoantes nasais são

[4] As denominações de "vogal" e "consoante" não correspondem a estas definições, mas pretendem assinalar apenas a função destes sons dentro da sílaba: em grego e latim só uma "vogal" podia ser ápice de sílaba, isto é, "sonora", enquanto que as "consoantes" só podem ser vale silábico crescente ou decrescente do ápice silábico, portanto um acompanhamento "consoante" do ápice silábico (Lausberg, 1981, pág. 59).

[5] Alguns autores utilizam as designações vozeado e não vozeado.

sons contínuos, perceptíveis durante todo o tempo das articulação, podendo ser prolongados. As consoantes nasais são sempre sonoras.

Fricativas

Na produção destas consoantes os articuladores provocam uma constrição numa zona do tracto, provocando um fluxo de ar turbulento nessa zona, que tem por consequência a produção de ruído que é depois propagado ao longo do tracto e radiado. Podem existir fricativas surdas e sonoras. No caso destas últimas além do ruído produzido na zona de constrição existe também a fonte de excitação glotal, devida à vibração das cordas vocais.

Laterais

As consoantes laterais são pronunciadas com uma obstrução parcial do fluxo de ar provocada pela língua com o palato ou os alvéolos, deixando aberturas laterais para a passagem do ar.

Vibrantes

Na produção destes sons não existe apenas uma oclusão mas sim várias, intercaladas por aberturas. O articulador móvel, por exemplo a língua, começa por formar uma oclusão. O aumento de pressão provocado pela oclusão acaba por voltar a abrir a passagem, restabelecendo-se o fluxo de ar. Devido à força exercida pelo fluxo de ar (efeito de Bernoulli), é novamente formada uma oclusão. Este processo pode repetir-se várias vezes. A designação destas consoantes provém do facto de o órgão articulador móvel tocar repetidamente no outro articulador, num movimento vibratório.

Africadas

São consoantes em que há uma obstrução completa do tracto vocal seguida de constrição do tipo fricativo.

2.4.1.2 Ponto de articulação

Para além do modo de articulação, a identificação da localização de constrição no tracto vocal, dada por um articulador passivo, e/ou dos articuladores activos utilizados é importante para a descrição das diferentes classes de sons. Esta dimensão é designada por ponto de articulação (Faria *et al.*, 1996, pág. 139). A designação do ponto de articulação obtém-se pelo nome do órgão que se desloca (activo) seguido do órgão em direcção ao qual se desloca (passivo), ou em casos em que facilmente se subentende o articulador activo apenas pela indicação do articulador passivo. Com base no articulador passivo, podem identificar-se os

seguintes pontos de articulação (Faria *et al.*, 1996, pág. 140): labial, dental, alveolar, pós-alveolar, palatal, velar, uvular, faríngeo, e glotal. Exemplos de articuladores activos são: a língua, o véu palatino e os lábios.

2.4.2 Vogais

As vogais também estão associadas a determinadas configurações do tracto, podendo ser descritas em função da posição dos articuladores que intervêm na sua produção, os lábios, a língua, e o maxilar inferior. O dorso da língua desempenha o papel principal.

Classificam-se as vogais em função do grau de abertura do tracto durante a sua produção, dependente simultaneamente da altura do dorso da língua e da abertura do maxilar inferior. São geralmente distinguidos quatro graus de abertura: fechado, semi-fechado, semia-berto e aberto. As vogais semi-fechadas e semi-abertas são por vezes designadas conjuntamente por médias. É também comum usar a posição do dorso para designar as vogais. Em lugar de vogais fechadas fala-se de vogais elevadas e em vez de vogais abertas têm-se vogais baixas. A posição intermédia pode designar-se por central.

Para um mesmo grau de abertura, o dorso da língua pode mover-se no plano horizontal, avançando ou recuando. Distinguem-se em geral três graus: frontal (ou anterior), central e posterior (Faria *et al.*, 1996, pág. 143).

Outra classificação deve-se à posição dos lábios, tendo-se num extremo vogais produzidas com os lábios arredondados, e no outro os lábios abertos e não arredondados, havendo posições intermédias entre estas duas.

As vogais podem pronunciar-se com o véu palatino fechado ou aberto. Quando o véu está aberto, isto é, pendente em posição relaxada, o ar sai tanto pela boca como pelo nariz, produzindo-se vogais nasais. Em princípio, todas as vogais podem ser nasais, no entanto é usual serem menos as vogais nasais que as orais.

Existem línguas que exploram a diferença de duração possuindo vogais longas e vogais breves. Outras características podem também ser usadas para descrição de vogais como por exemplo a posição da raiz da língua e o facto de serem ou não tensas (Ladefoged e Maddieson, 1995).

2.4.3 Semivogais

As semivogais possuem características articulatórias semelhantes às das vogais, mas têm uma duração menor, não podendo constituir núcleo de sílaba e ocorrendo sempre junto de uma vogal com a qual formam os ditongos. Podem também receber a denominação de semiconsoantes (Mateus *et al.*, 1990, pág. 52).

2.5 Sons do Português Europeu

A minha Pátria é a Língua Portuguesa

Fernando Pessoa

Cada língua utiliza apenas um subconjunto dos sons que o aparelho fonador humano tem capacidade para produzir. De seguida apresentam-se os sons utilizados no Português padrão falado em Portugal, designado, por vezes, por Português Europeu.

2.5.1 Consoantes

As várias consoantes existentes no Português são descritas de seguida de forma muito sumária. Na Tabela 2.1(a) apresenta-se de uma forma compacta a classificação apresentada para as consoantes.

Oclusivas

Em Português existem as oclusivas: [b] de (**b**ote), [d] (**d**ote), [g] (**g**ato), [p] (**p**ato), [t] (**t**olo), e [k] (**c**aro). As três primeiras são sonoras, as três últimas surdas.

Distinguem-se no Português oclusivas labiais, apicais e dorsais. As oclusivas labiais são o [p] surdo e o [b] sonoro. Nas oclusivas apicais, o [t] e [d], o lugar de articulação é o bordo posterior dos dentes, daí se designarem também pós-dentais ou só dentais. O pós-dorso com o palato posterior, véu palatino, forma as oclusivas velares [k] e [g].

Nasais

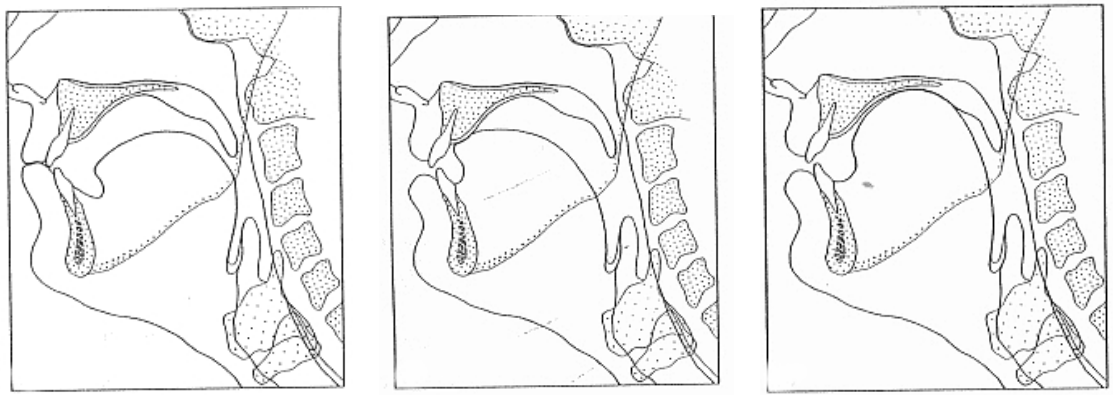
Em Português existem três consoantes nasais: [m] (**m**ar), [n] (**n**ata) e [ɲ] (apa**nh**a). São todas sonoras. Configurações aproximadas da sua realização encontram-se na Figura 2.7.

O [m] é bilabial, sendo produzido pelo movimento de fecho dos lábios. O [n] é produzido com a ponta da língua junto aos incisivos ou na zona alveolar. Alguns autores classificam-no como alveolar outros como dental (Mateus *et al.*, 1990, pág. 50). O [ɲ] é produzido pelo contacto da língua na região do palato, sendo portanto palatal.

Fricativas

Temos as fricativas sonoras [v] (**v**aca), [z] (a**z**ul), [ʒ] (a**g**ir) e as surdas [f] (**f**aca), [s] (**s**aca) e [ʃ] (**ch**ave).

Em Português o lábio superior aproxima-se dos incisivos inferiores para produzir o [v] e [f], sendo estes designados de lábio-dentais. A coroa da língua aproxima-se da região dento-alveolar para produzir as fricativas apico-dentais, ou apico-alveolares [z] e [s]. O [ʃ] e [ʒ] são produzidos pela língua na região palato-alveolar (Mateus *et al.*, 1990, pág. 49).



(a) [m], bilabial (b) [n], alveolar (c) [ɲ], palatal

Figura 2.7: Configurações para as consoantes nasais portuguesas (adaptado de Laver, 1994, pág. 210).

As fricativas [s] e [z] são, por vezes, chamadas de sibilantes e as [ʃ] e [ʒ] de chiantes.

Laterais

Na língua portuguesa existem as laterais: [l] (**l**ado), [ɫ] (ma**l**), e [ʎ] (fo**lh**a). As duas primeiras são produzidas com o ápice da língua, designando-se por apicais, a terceira com o dorso da língua, designando-se por dorsal.

A lateral [l] é apico-alveolar, a ponta da língua forma uma oclusão pós-dental, na zona alveolar, deixando espaço aberto em cada lado. A posição da língua na articulação do [l] varia muito em cada individuo. Alguns falantes em vez de colocarem a língua no centro colocam-na num dos lados de forma a que a corrente de ar saia pelo lado oposto.

A lateral [ʎ] é dorso-palatal, forma-se pela oclusão central entre o pré-dorso e o palato anterior, deixando-se um espaço em ambos os lados.

Existe ainda uma lateral em Português, representada por [ɫ], em cuja produção, além de existir a obstrução formada pela ponta da língua junto dos alvéolos, ocorre uma articulação secundária criada pela elevação do dorso da língua em direcção ao véu palatino. Por essa razão designa-se por velar ou velarizado. Ocorre em Português Europeu em final de sílaba ou posição final absoluta (exemplos saltar, animal) (Mateus *et al.*, 1990, pág. 49), sendo, por isso, denominada de variante contextual.

Vibrantes

Em Português existem três destes sons: [ɾ] (pa**r**a), [r] (pa**rr**a), e [R] (pa**rr**a). Os dois primeiros são produzidos na parte frontal e o terceiro na parte posterior. Pode existir só um batimento,

como no [ɾ], ou batimentos múltiplos (dois ou mais) como acontece para o [r] e [R]. O primeiro caso designa-se por vibrante simples, ou *tap*, o segundo por vibrante múltipla ou *trill* (Laver, 1994).

A vibrante múltipla não é produzida de forma igual por todos os portugueses (Barbosa, 1994). Alguns usam uma articulação apical, com vários batimentos do ápice da língua elevada em direcção aos alvéolos ([r]). Outros usam uma articulação constritiva do dorso na região velar-uvular ([R]). Outros utilizam uma articulação vibrante uvular múltipla. São variantes do mesmo fonema, dependendo o seu uso de hábitos individuais.

Modo	Sonoridade	bilabial	labio-dental	apico-dental	apico-alveolar	palato-alveolar	dorso-palatal	dorso-velar
Oclusiva	sonora	[b]		[d]				[g]
	surda	[p]		[t]				[k]
Nasal	sonora	[m]		[n]			[ɲ]	
Lateral	sonora				[l]		[ʎ]	
Vibrante	sonora (simples)				[ɾ]			
	sonora (múltipla)				[r]			[R]
Fricativa	sonora		[v]	[z]		[ʒ]		
	surda		[f]	[s]		[ʃ]		

(a) Consoantes

	Não arredondadas			Arredondadas		
	Anterior	Central	Posterior	Anterior	Central	Posterior
Fechada	[i]					[u]
Meia-fechada	[e]					[o]
Média		[ɨ]				
Meia-aberta	[ɛ]	[ɐ]				[ɔ]
Aberta		[a]				

(b) Vogais Orais

Tabela 2.1: Classificação dos sons do Português Europeu.

2.5.2 Vogais e afins

Vogais orais

As nove vogais orais do Português são: [i] (p**i**co), [e] (p**ê**co), [ɛ] (p**é**ga), [a] (p**á**po), [ɐ] (c**a**da), [ɔ] (p**ó**te), [o] (t**ô**po), [u] (c**u**bo), [ɨ] (p**e**di).

Se abrirmos a boca e fizermos quase assentar a língua na sua base, muito próxima da posição de repouso, daremos à cavidade bucal a sua forma e volume maiores, e este volume será idêntico ao da cavidade faríngea; o resultado será a vogal [a]. Elevando um pouco a língua na direcção do palato duro, isto é, para a frente, produz-se o [ɛ]. Elevando-a um pouco mais na mesma direcção, obtém-se a vogal [e]. Mais um pouco de elevação, sempre na mesma direcção, resulta na vogal [i]. Passou-se assim gradualmente de [a] a [i] diminuindo a cavidade bucal

e aumentando a cavidade faríngea. Com a excepção do [a], que é central, todas as outras são designadas por anteriores, por a língua ocupar nelas a parte anterior da cavidade bucal (Barbosa, 1994, pág. 52).

Se partirmos do [a] e formos fechando a cavidade bucal elevando a língua na direcção central, produziremos as vogais centrais. Um pouco acima do [a] produziremos o [ɐ], subindo um pouco mais o [ɨ] (Barbosa, 1994, pág. 53).

Se, mais uma vez, partindo de [a], elevarmos o dorso da língua em direcção à parte posterior da cavidade vocal, ou seja em direcção ao palato mole ou véu palatino, diminuir-se-á a cavidade faríngea e aumentar-se-á o volume da cavidade bucal, obtendo-se assim o [ɔ]. Mais acentuado, esse movimento dá lugar ao som mais fechado [o]. Indo mais longe no movimento obter-se-á a vogal [u]. Para se ser mais preciso, deve notar-se que, na realidade, para o [u], a parte posterior do dorso da língua se dirige para a zona velar, enquanto que para o [o] e [ɔ] se dirige mais em direcção à faringe. De qualquer modo, estas vogais são chamadas de posteriores (Barbosa, 1994, pág. 53).

As vogais [ɔ],[o] e [u] são pronunciadas com arredondamento dos lábios, pelo que se designam arredondadas (Mateus *et al.*, 1990, pág. 52).

O resumo da classificação é apresentado na Tabela 2.1(b).

Ditongos e Tritongos orais

Existem em Português ditongos crescentes e decrescentes [6]. Exemplos de ditongos são [aj] (p**ai**), [aw] (p**au**), [ej] (d**ei**), [ew] (d**eu**), [oj] (b**oi**), [ɔj] (d**ói**), [ow] (d**ou**), [iw] (v**iu**), [ɛw] (c**éu**) e [uj] (f**ui**).

Existem, segundo alguns autores (Laver, 1994), encontros vocálicos constituídos por semivogal mais vogal mais semivogal, os denominados tritongos. Como exemplo [7] temos [wow] (averig**uou**), [wej] (averig**uei**).

Vogais nasais

Em Português é geralmente referida a existência de cinco vogais nasais (Sampson, 1999, por exemplo): o [ɐ̃] (c**an**to), [ẽ] (t**en**ta), [ĩ] (p**in**ta), [õ] (p**on**te), e [ũ] (f**un**do).

Almeida (1976) refere a existência de outras três vogais, o [ɛ̃], o [ɔ̃] e o [ã]. A existência do [ã] resultante da crase de [ɐ] com [ɐ̃] como em *a andar* é referida desde Viana (1892). Também Barbosa (1994), página 56, considera a existência nalgumas zonas do Norte de Portugal de [ɔ̃] e [ɛ̃], por exemplo nas palavras *vende* e *ontem*, e que, o primeiro destes, se tem vindo a

[6] Alguns linguistas apenas consideram ditongos os decrescentes, visto que, em muitos casos, os potencialmente crescentes podem ser pronunciados com uma verdadeira vogal no início (Mateus *et al.*, 1990, pág. 53).

[7] Desde que proferidos suficientemente rápido.

generalizar cada vez mais, pelo menos neste termo, no Português dito padrão.

Não existe muita informação sobre a sua produção. Devido ao menor número de vogais nasais, relativamente às orais, é possível que a vogal [ɐ̃] seja produzida por configurações correspondentes às vogais orais [a] e [ɐ], a vogal [ẽ] por configurações entre o [e] e [ɛ], e o [õ] por [o] ou [ɔ].

Segundo diversos estudos de fonética, as vogais nasais do Português Europeu diferem das vogais nasais de outras línguas, como o Francês, devido a serem apenas fortemente nasalizadas junto ao fim (Stevens *et al.*, 1987; Barbosa, 1965; Mateus, 1975).

Ditongos e Tritongos nasais

Existem também ditongos nasais crescentes [wɐ̃] (q**uan**do), [wẽ] (ag**uen**tar), [jɐ̃] (fi**an**do), [wĩ] (r**uim**). Sendo os dois últimos instáveis, passando a hiato se pronunciados lentamente (Silva, 1995). Ditongos nasais decrescentes são [ɐ̃j̃] (q**uem**), [õj̃] (p**õe**), [ũj̃] (m**ui**to), [ɐ̃w̃] (p**ão**). Alguns autores (Sampson, 1999, por exemplo) apenas referem a existência destes últimos, os decrescentes. Como exemplo de um tritongo nasal temos o [wɐ̃w] (enxag**uam**) (Cunha 1976, citado em Silva, 1995).

2.5.3 Semivogais

Existem no Português duas semivogais, o [u] e o [i] que se representam por [w] (pa**u**) e [j] (pa**i**).

2.6 Desenvolvimento das vogais nasais em Português

Referem-se de seguida alguns factos relacionados com o aparecimento e desenvolvimento das vogais nasais na língua portuguesa. O exposto baseia-se em Cunha (1982) e Sampson (1999).

Em geral, o processo de nasalização envolve dois estados sucessivos. Primeiro, a vogal é nasalizada por efeito de uma consoante, resultando numa vogal alofonicamente nasal. Depois, a consoante pode desaparecer, ficando uma vogal nasal. Se a consoante nasal se encontra depois da vogal o processo de nasalização é obtido por assimilação regressiva, ao encontrar-se antes a assimilação diz-se progressiva. Em Português o aparecimento de vogais nasais deu-se geralmente como resultado de assimilação regressiva. Apenas num número reduzido de casos aconteceu o aparecimento de vogais nasais devido a uma assimilação progressiva.

O processo evolutivo é complexo e dependente de variados factores. Para facilitar a descrição, são geralmente considerados três contextos distintos para cobrir as situações principais em que ocorreu o desenvolvimento de vogais nasais nas línguas românicas, pelo processo regressivo (Sampson, 1999, pág. 33):

- /VNC/, onde uma vogal precede uma consoante nasal que é seguida de outra consoante;

- /VN#/, onde a vogal precede uma consoante nasal final;
- /VNV/, onde a vogal precede uma consoante nasal intervocálica.

2.6.1 Vogais nasais resultado de assimilação regressiva

Todos os três contextos causadores de nasalidade por assimilação regressiva existiam no Galaico-Português no início da idade média. As formas dos contextos /VNC/ e /VNV/ foram herdadas do Latim, como em VĔNTUM e LŪNA. O contexto /VN#/ também se tornou bem representado, devido às formas latinas como NŌN e QUĔM, a que foram adicionadas as formas verbais da terceira pessoa do plural resultantes da simplificação de -NT, como em SŬNT > [son], e também como resultado de apócope de sequências (N)NEM, como em FĪNEM e IOHĂNNEM (Sampson, 1999, pág. 180).

Existem indicações que o processo de nasalização afectou certas vogais no período pré-literário. Nas formas /VNV/, o [n] intervocálico, resultante do N latino simples, foi progressivamente enfraquecendo no período pré-literário, ocorrendo ao mesmo tempo uma nasalização forte da vogal que o precede. Este enfraquecimento resultou no desaparecimento do [n] pela altura dos primeiros textos em Português, que datam do início do século XIII. Por exemplo, na palavra *corona* houve primeiro a nasalização da vogal que precede o *n*, donde *corõna*; em seguida, o *n* caiu e ficou-se com *corõa*, tendo-se hoje *coroa*. Assim todos os *n* intervocálicos desapareceram depois de terem nasalizado a vogal precedente; ex.: *vinu* > *vĩo*. Em todas estas palavras a vogal nasal e a que veio a segui-la pertenciam a duas sílabas diferentes. De referir que este fenómeno de queda do *n* é particular ao Galaico-Português, não tendo ocorrido nas outras línguas peninsulares. Em numerosas palavras de origem árabe permanecem os *n* intervocálicos etimológicos; ex.: *azeitona, alfenin, atafona*, etc.

No início do período literário (século XIII), todas as variantes do Galaico-Português tinham já desenvolvido, ou encontravam-se em vias de desenvolver, um conjunto de cinco vogais nasais /ĩ ẽ ã õ ũ/ resultantes do contexto /VNV/, de que são exemplo ['lũa] < LŪNA e ['mão] < MĂNUM. No contexto /VNC/, pode assumir-se que as vogais são tipicamente nasalizadas mas de uma forma alofónica, existindo ainda a consoante nasal, tendo-se ['vẽndo] < VĒNDO. No contexto /VN#/, a presença da consoante nasal é menos clara. No início do século XIII, pode ter já perdido a sua oclusão, ou ser realizada, pelo menos por alguns falantes, como uma nasal com fraca oclusão oral (Sampson, 1999, pág. 186).

No que se refere aos *m* e *n* latinos que passaram a ocorrer em Português em final de palavra ou antes de consoante, no século XIII a representação gráfica preponderante é *n*, embora apareçam também *m* e o til [8], este último só muito raramente omitido. (Nobiling, 1974, pág. 99).

[8] O til – um acento gráfico – começou a ser utilizado no século XII para notar vogais nasais, por essa altura ainda em concorrência com a consoante *n*: é exemplo disso a palavra *razõ* (hoje em dia, razão), que de início co-ocorria com a forma *razon* (Walter, 1996, pág. 207).

No contexto /VN#/ a consoante nasal foi progressivamente enfraquecendo até desaparecer criando-se uma consoante nasal. Depois do desaparecimento, as vogais altas evoluíram de forma diferente das não altas. As vogais altas [ĩ] e [ũ] devem ter desenvolvido sons de transição que foram posteriormente absorvidos. Um exemplo é FĪNEM > fĩn > fĩ> *fim*. As vogais não altas deram origem a ditongos, de que são exemplos BĔNE > bẽn > bẽ > bẽj̃ > *bem* e CĂNEM > kãn > kã > kãw̃ > *cão* (Sampson, 1999, pág. 187).

Durante o período medieval, a vogal nasal baixa [ã] subiu tomando uma posição central média-baixa [ɐ̃]. Esta subida não ocorreu nos dialectos do Minho e noroeste de Portugal e Galiza.

Resultantes da queda do [n] intervocálico apareceram muitas sequências de vogal nasal seguida de vogal, pertencendo as duas a sílabas diferentes. Estes grupos de vogais, em hiato, são, por natureza, muito instáveis, e a maior parte deles será eliminado ulteriormente. Já nos textos medievais ocorrem certas evoluções: por exemplo, *pinho* por *pĩ-o* (desenvolvimento do *ĩ* em hiato numa consoante nasal) ou *alheo* por *alhẽo* (desnasalização da vogal).

Das várias evoluções da fonética do Português desde o século XIV interessa-nos a eliminação de encontros vocálicos (pág. 189-194 Sampson, 1999, para uma descrição detalhada). Esta eliminação foi feita de três formas, quando existiam vogais nasais envolvidas. No caso de sequências com *-ĩ-o* e *-ĩ-a*, desenvolve-se uma consoante nasal entre as duas vogais, obtendo-se *-inho* e *-inha*. Exemplos de palavras onde isto ocorreu são *vinho* < vĩ-o < VĪNU , *galinha* < galĩ-a < GALLĪNA. A consoante nasal [nh] [9], surgida de [ĩ] em hiato, separa as duas vogais, eliminando a instabilidade. Outra forma consiste na contracção das duas vogais numa vogal única. Quando uma das duas vogais é nasal, o resultado é uma vogal nasal; ex.: *lã-a* > *lã*. As vogais resultantes já existiam na língua não sendo afectado o sistema fonológico. A contracção de uma vogal nasal e de uma vogal oral, em que a altura destas é diferente, resultou em ditongos nasais, como em *mã-o* > *mão, cã-es* > *cães*.

O destino do *n* latino intervocálico foi o seguinte: após nasalizar a vogal anterior, caiu, e as duas vogais que entraram em contacto passaram a actuar uma sobre a outra em diferentes processos de assimilação [10]. Se as articulações bucais eram semelhantes, tornaram-se iguais, a vogal nasal na maioria dos casos nasalizou a oral, e fundiram-se; outras vezes, sobretudo quando átona, a vogal nasal passou a oral. Se havia acentuada diferença entre as duas articulações ocorreu a nasalização da vogal oral, e ambas formaram um ditongo. Quando a segunda vogal era mais intensa que a primeira e diferentes as articulações bucais, mantiveram-se em sílabas separadas e desenvolveu-se um som de transição (Nobiling, 1974).

Pelo final do século XVI, o sistema de vogais nasais tinha já tomado uma forma semelhante à actual. Ocorreram, no entanto, algumas alterações significativas nos últimos quatro séculos. Assistiu-se ao reaparecimento do contraste pré-nasal entre [ã] e [ɐ̃], ao aparecimento de um

[9] As grafias lh e nh surgiram já depois de 1250 para notar por escrito as articulações palatais que ocorrem em palavras como velho e Catalunha. São de origem Provençal (Walter, 1996, pág. 207).

[10] Em (Barbosa, 1994, Figura 29) é descrito o processo de passagem de *lana* a *lã*

novo ditongo, o [ũj̃], e à fusão de [ẽj̃] com [ɐ̃j̃] (Sampson, 1999, pág 197).

Nos contextos /VNC/ a consoante nasal resistiu muito mais ao desaparecimento. No entanto, mais recentemente também foi enfraquecendo gradualmente, em especial antes de consoantes contínuas (fricativas) resultando num som de transição, de característica definidas essencialmente pelo segmento seguinte, ou desaparecendo totalmente. Antes de oclusivas, pode apresentar uma breve oclusão homorgânica com o ponto de articulação da consoante seguinte, que devido ao enfraquecimento pode actualmente consistir num breve som de transição. Fagan (1988) propôs que este som de transição pode ter resultado de epêntese em tempos recentes. Segundo ele a evolução teria sido da forma $\tilde{\text{V}}\text{NC} > \tilde{\text{V}}\text{C} > \tilde{\text{V}}\tilde{\text{G}}\text{C}$, onde $\tilde{\text{G}}$ indica uma transição nasal. O *m* e o *n* latinos são ainda hoje pronunciados diante de oclusiva, não só, quando já no latim, a precediam directamente, como em *chumbo, vende*, mas também, quando somente chegaram a esta posição no decorrer da evolução da palavra, como em *senda ['sẽnda]* < *semitam*; *manga* < *manicam* (Nobiling, 1974, pág. 95).

2.6.2 Vogais nasais resultantes de assimilação progressiva

Um conjunto limitado de formas com vogais nasais apareceram, em Português, resultado de assimilação progressiva (Sampson, 1999, pág. 185). O /m/ e, menos comum, o /n/ nasalizaram a vogal seguinte, particularmente quando esta era alta e acentuada. Exemplos são MĬHĪ > [mĩ] > *mim*, MĔA > ['mĩa] > *minha*, e NĪDUM > ['nĩo] > *ninho*.

O aparecimento e difusão de tais formas e a sua integração na língua padrão tem sido lenta, estendendo-se por vários séculos. Uma realização completamente nasal da vogal acentuada, nas evoluções de MĔA e NĪDUM, por exemplo, era aceite em falares educados no início do período literário, tendo essas vogais sofrido evolução semelhante ao [ĩ] resultante da queda de [n] intervocálico. À semelhança de *vinho* < ['viɲu] < ['vĩo] < VĪNUM, também ['nĩo] resultou em ['niɲu] *ninho*.

No entanto, demorou bastante para outras palavras semelhantes ganharem aceitação. Mãe < MĀTREM, ainda no século XVI aparece como *may* e *nai*, apenas se tornando habitual a variante com a vogal nasal mais recentemente. Outro exemplo, Camões ainda rima *muito* com *fruito*.

Para outras formas lexicais, a co-existência de variantes orais e nasais não resultou na entrada para a língua padrão da variante nasal. A variante nasal ['mẽzɐ] de *mesa*, existente no século XVI, persiste apenas em dialectos do sul. Gramáticos do século XVII e XVIII referem outras formas, nessa época relativamente usuais, como *menxa, menxer, mexiricar* que no Português actual correspondem a *ameixa, meixer*, e *mexericar*

2.7 Propriedades acústicas das vogais nasais

Em termos acústicos, as vogais resultam da ressonância nas cavidades orais, faríngea e bucal. A fonte de energia para a ressonância é a excitação glotal , causada pela vibração periódica das cordas vocais. Associadas a cada vogal existem um conjunto de formantes, ou ressonâncias, localizadas em frequências específicas. As diferenças para os valores das formantes de vogal para vogal devem-se à diferente configuração das cavidades durante a sua produção.

A adição da cavidade nasal complica a situação ao adicionar novas ressonâncias (pólos) e, em especial, ao fazer aparecer anti-ressonâncias (zeros). O resultado é complexo e não se pode facilmente relacionar as ressonâncias e anti-ressonâncias com uma das cavidades.

Os seios paranasais também contribuem para a modificação das ressonâncias (Lindqvist-Gauffin e Sundberg, 1976; Feng, 1987).

Foi feita uma investigação exaustiva das características espectrais das vogais nasais ao longo das últimas quatro décadas (Delattre, 1954; House e Stevens, 1956; Hattori *et al.*, 1958; Fant, 1960; Fujimura e Ludqvist, 1971; Bell-Berti e Baer, 1983; Hawkins e Stevens, 1985; Bognar e Fujisaki, 1986; Maeda, 1993). Os resultados foram diversificados e por vezes mesmo contraditórios entre si. Existem, no entanto, alguns pontos de acordo. As marcas principais de nasalidade de uma vogal parecem ser: a modificação do espectro nas baixas frequências, particularmente na vizinhança da primeira formante oral; a existência de uma formante nasal cerca dos 250 Hz (característica também das consoantes nasais); a existência de um zero que interage com a primeira formante oral, reduzindo a sua amplitude e aumentando a sua largura de banda; modificação nas frequências mais elevadas, resultando numa distribuição mais difusa da energia (Sampson, 1999, pág. 7).

2.8 Percepção de vogais nasais

> ... nasality is above all else an auditory concept, ...
>
> John Laver (Laver, 1980, pág. 77)

Questões centrais acerca da percepção de vogais nasais são (Sampson, 1999, pág. 9):

1. Qual a propriedade, ou conjunto de propriedades, que permitem a um ouvinte perceber uma vogal como sendo nasal?

2. A percepção da nasalidade é independente da vogal e da língua?

3. Que outros aspectos da qualidade da vogal são influenciados pela nasalização ?

2.8.1 Propriedades influenciadoras da percepção de nasalidade

Uma vogal é rapidamente percebida como nasal se a proeminência na área da primeira formante é reduzida e a largura de banda aumenta, alterações que, como já foi referido, constituem as características acústicas principais da nasalização de uma vogal. Já na década de 1950 foi mostrado por Delattre (1954), com a ajuda de síntese, que a nasalidade pode ser assinalada pela simples redução da intensidade da primeira formante de uma vogal oral. Mais tarde, o mesmo investigador, notou que, para a vogal [ɛ], a redução de F_1 de 12 dB resultava na sua percepção, por ouvintes franceses, como [ɛ̃] (Delattre, 1968). O estudo realizado por Hawkins e Stevens (1985), usando voz sintética apresentada a falantes de Gujerati, Hindi, Bengali, e Inglês, aponta para que a redução da proeminência e aumento da largura de banda na zona da primeira formante sejam independentes da língua considerada.

Outras propriedades, espectrais, temporais e contextuais, parecem também influenciar a percepção de nasalidade.

Uma é o aumento progressivo do nível de nasalidade durante a articulação de uma vogal nasal. Este facto foi brevemente referido por Hattori *et al.* (1958). Reenen, num trabalho de 1982 (citado em Sampson, 1999, pág. 10) atribui muita importância a esta característica dinâmica, vendo-a como crucial para a percepção da nasalidade numa vogal, independentemente da língua. Uma experiência realizada por Linthorst, em 1983 (referido em Sampson, 1999, pág. 10), fornece também suporte para a importância desta propriedade na percepção. Foram gravadas três palavras Francesas *même, dais, baie*, extraídas as vogais e, depois, adicionadas em várias combinações, às consoantes iniciais das palavras usadas. Representando a vogal parcialmente nasal da primeira palavra por [E] e a vogal oral das outras duas por [ɛ], os nove ouvintes franceses perceberam quase sempre a sequência [ɛE] como vogal nasal, e as outras combinações, incluindo o [E] e [EE], como vogal oral. Não foram realizados estudos similares para outras línguas.

Evidências fonológicas da relação entre a duração das vogais e nasalização motivaram experiências em que se manipulou a duração para determinar os efeitos na nasalização das vogais. Delattre e Monnot em 1968, usando o *Pattern Playback* sintetizaram nove versões de uma sequência CVC diferindo apenas na duração da vogal. Ouvintes franceses e americanos identificaram as vogais mais breves com sendo orais e as vogais mais longas como nasais. Neste estudo o grau de nasalização foi mantido constante e igual a um valor intermédio entre o de uma vogal oral e a vogal nasal. Mais recentemente, Whalen e Beddor, 1989 (citados em Beddor (1993)), variaram, quer a duração, quer o grau de nasalização. Usando síntese articulatória, geraram as vogais /a/, /i/ e /u/ com cinco durações e várias aberturas do véu palatino. Ouvintes americanos julgaram os estímulos com maior abertura da passagem para a cavidade nasal como mais nasais, mas também julgaram os estímulos de maior duração como mais nasais. Esta relação parece ser independente da vogal. O efeito da duração das vogais também parece ser independente da língua materna do ouvinte. Para ouvintes franceses, as vogais mais longas são perceptualmente nasais o que é consistente com o facto de as vogais

nasais serem mais longas que as suas correspondentes orais. A duração das vogais nasais portuguesas também é superior às vogais orais (Silva, 1995). Para ouvintes americanos este efeito perceptual opõe-se ao facto conhecido das vogais do Inglês, em contexto nasal, terem uma duração inferior à das vogais orais.

Dois tipos de resultados experimentais sugerem uma relação entre a altura da vogal e a percepção de nasalidade. Primeiro, estudos usando voz natural descobriram que vogais baixas têm uma maior tendência a serem percebidas como nasais (Brito, 1975, por exemplo). Segundo, estudos com voz sintetizada mostram que vogais baixas requerem um maior acoplamento nasal para serem percebidas como nasais (House e Stevens, 1956; Maeda, 1982b). Esta aparente contradição deve ter origem na utilização de diferentes aberturas do velo. Em voz natural, as vogais baixas tendem a ser produzidas com uma posição do velo mais baixa mesmo em contextos orais. Hawkins e Stevens (1985) descobriram que, em condições de nasalidade ambígua, a percepção do contínuo [o-õ] era largamente dependente das diferenças da altura percebida da vogal.

A percepção da nasalidade de uma vogal é também influenciada pelo contexto fonético em que a vogal ocorre. Por exemplo, ouvintes americanos julgam vogais nasais como mais nasais quando em contextos não nasais (Beddor, 1993, pág. 179). Kawasaki (1986) descobriu que a percepção de nasalidade de uma vogal era aumentada ao serem atenuadas consoantes nasais adjacentes (ou seja, [N$\tilde{\text{V}}$N] ->[$\tilde{\text{V}}$]). De uma forma similar, Krakow e Beddor, em 1991, (referido em Beddor, 1993, pág. 179) descobriram que vogais nasais eram mais vezes julgadas como nasais, quando extraídas de contextos nasais e apresentadas de forma isolada ([$\tilde{\text{V}}$]) ou quando ocorriam num contexto oral ([C$\tilde{\text{V}}$C]), do que quando no seu contexto nasal original ([N$\tilde{\text{V}}$N]). Uma das explicações possíveis (Beddor, 1993, pág. 179) para estes resultados é a de que o conhecimento pelos ouvintes dos efeitos de coarticulação leva-os a atribuir a nasalidade às consoantes nasais adjacentes, ouvindo vogais nasais, num contexto nasal, como não nasais.

2.8.2 Independência da vogal

Apesar das características acústicas variáveis das vogais nasais de alturas diferentes, parece existir uma propriedade correlacionada com a nasalidade independente da vogal, a proeminência relativa ou forma plana (*flatness*) da região de baixas frequências (Beddor, 1993, pág. 172). Estudos recentes dedicaram-se a um estudo quantitativo desta medida de nasalidade (Maeda, 1993). Este investigador propôs a distância entre os dois picos do espectro nas baixas frequências como medida de dispersão espectral e descobriu que, em geral, a percepção de nasalidade aumenta com o aumento dessa distância.

2.8.3 (In)dependência da língua

Estudos como o de Hawkins e Stevens (1985) revelam que ouvintes com línguas nativas em que não existe oposição fonológica entre vogais orais e nasais conseguem discriminar estas duas

classes. Mas será que falantes de diferentes línguas usam as mesmas características espectrais para efectuarem a distinção oral-nasal ?

Estudos com o Inglês e três línguas faladas na Índia (Hawkins e Stevens, 1985), e com Inglês, Francês e Português (Stevens *et al.*, 1987) sugerem que os ouvintes respondem às mesmas propriedades acústicas para classificar uma vogal como oral ou nasal, não sendo relevante o facto de a língua materna usar a nasalidade como oposição fonológica.

Estudos da capacidade dos ouvintes de "discriminar" diferenças entre vogais orais e nasais sugerem que esta capacidade está mais relacionada com a experiência linguística. Em geral, os ouvintes para os quais a distinção é fonémica exibem uma função de discriminação categorial, com boa discriminação de diferenças que cruzam a separação oral-nasal e uma discriminação pobre para diferenças dentro das categorias. Em contraste, os ouvintes para os quais a distinção é alofónica mostram uma boa capacidade de discriminação nos dois casos (Beddor e Strange, 1982; Hawkins e Stevens, 1985).

Os julgamentos por parte dos ouvintes da naturalidade da nasalização das vogais é também dependente da língua. Em (Stevens *et al.*, 1987) é referido que ouvintes portugueses, franceses e ingleses preferiram quantidades diferentes e padrões diferentes de nasalização. Estas diferenças são consistentes com as características acústicas da nasalização nestas línguas.

Resumindo, a identificação de vogais orais e nasais é muito semelhante para as várias línguas, mas a discriminação e julgamento de naturalidade destas vogais mostram diferenças dependentes da língua (Beddor, 1993, pág. 177).

2.8.4 Efeitos secundários da nasalidade

Tem sido dada atenção em anos recentes a efeitos perceptuais secundários provocados pela adição de nasalidade a uma vogal (Sampson, 1999, pág. 12).

Um desses efeitos é a tendência da nasalidade "obscurecer" as qualidades da vogal oral. A adição de acoplamento nasal altera a região da primeira formante, crucial na identificação de uma vogal, tendo por resultado que diferenças entre vogais próximas mas distintas podem ser diluídas causando a associação de vogais originalmente diferentes.

Outro efeito relaciona-se com a altura percebida de uma vogal (Beddor, 1993, pág. 180). Quando é adicionada nasalidade a uma vogal, o ouvinte pode perceber também uma modificação da altura da vogal, relacionada em termos articulatórios com a altura da língua. Quando se adiciona nasalidade a vogais altas aparece energia adicional abaixo da primeira formante da vogal oral, provocando um abaixamento perceptual da altura da vogal. De forma inversa, para vogais baixas, a adição de nasalidade desloca o centro de gravidade na zona da primeira formante, causando a subida da vogal em termos perceptuais.

Também a percepção da posição no eixo anterior-posterior de uma vogal é afectada pela nasalidade (Beddor, 1993, pág. 182). Num estudo realizado por Wright em 1986, as vogais não baixas [ĩ],[ẽ] e [ɛ̃] eram percebidas como mais recuadas que as suas versões orais. As vogais

baixas não exibiram um desvio uniforme: enquanto o [õ] foi percebido como mais frontal do que o [o], o [ũ] foi considerado mais recuado.

Em resumo, estudos perceptuais mostraram que a nasalização de uma vogal influencia outros aspectos da vogal para além da nasalidade. As alterações do espectro, na zona das baixas frequências, alteram a percepção da altura da vogal, e também, de forma muito menos notória, a percepção de frontalidade da vogal.

2.9 Utilização da nasalidade em vogais

As vogais nasais ocorrem como alofones, por ajustamentos contextuais, ou contrastando fonologicamente com as vogais orais. O primeiro caso ocorre virtualmente em todas as línguas do mundo. A utilização como fonemas contrastantes a nível fonológico é mais raro, apesar de não muito raro, ocorrendo em línguas como o Hindi, Polaco, Francês, e, claro, o Português (Laver, 1994, pág. 291). Estudos indicam que um pouco menos de um quarto das línguas do mundo possuem vogais nasais. Destas línguas nenhuma possui maior número de vogais nasais que vogais orais. Para metade destas línguas o número de vogais orais e nasais é igual (Hombert 1986, citado em (Laver, 1994, pág. 293)).

2.10 Estudos de vogais nasais

A nasalidade em vogais foi estudada por muitos investigadores usando as mais variadas técnicas. Referem-se nesta secção alguns estudos representativos das técnicas utilizadas, sem se pretender efectuar um levantamento exaustivo. Não se inclui, em geral, informação acerca dos resultados dos trabalhos citados devido a já terem sido referidos nas secções anteriores. Estudos das vogais nasais do Português são tratados em separado.

2.10.1 Técnicas utilizadas

Os primeiros estudos das vogais nasais foram realizados utilizando técnicas de análise do sinal de voz. Por exemplo, Hattori *et al.* (1958) efectuou estudos comparativos de vogais nasais e de consoantes nasais com a ajuda de espectrogramas.

O estudo do espectro de vogais nasais, nomeadamente o comportamento dos pólos e zeros da função de transferência para vogais nasais, foi efectuado por Fujimura (1960), utilizando um modelo simplificado das cavidades oral e nasal. Utilizou um método gráfico para obtenção das ressonâncias e anti-ressonâncias. Para tornar o problema tratável não considerou as perdas no seu modelo.

A medição directa da função de transferência para vogais nasais, em seres humanos, foi efectuada por Båvegård *et al.* (1993) usando uma técnica designada de *sweeptone* (Fujimura e Ludqvist, 1971). Colocaram uma fonte de vibração no pescoço, muito próximo da posição da

laringe, e mediram a radiação dos lábios e narinas em conjunto. Variando a frequência de oscilação da fonte de vibração, mediram a função de transferência entre os 100 e 5000 Hz.

Diversos estudos utilizaram sinal natural editado (Benguerel e Lafarge, 1981; Kawasaki, 1986). Num exemplo deste tipo de estudos, Kawasaki (1986) para estudar a hipótese de que a nasalização numa vogal nasal deve ser mais evidente perceptualmente, ao atenuarem-se as consoantes nasais adjacentes, gravou três sílabas constituídas pela nasal [m] e uma vogal, atenuando de seguida as consoantes nasais. Os estímulos assim obtidos foram utilizados num teste perceptual, obtendo-se resultados que suportam a hipótese.

O movimento do velo durante a produção de vogais nasais foi investigado através de técnicas como: o *velotrace* (Horiguchi e Bell-Berti, 1987), endoscopia (Benguerel *et al.*, 1977), *Electro-Magnetic Midsagittal Articulography* (EMMA) (Schönle *et al.*, 1987). Uma apresentação das várias técnicas instrumentais utilizadas em estudos de sons nasais pode ser encontrada em (Krakow e Huffman, 1993).

Uma parte, substancial, dos estudos acerca de vogais nasais consistiu na realização de testes perceptuais utilizando estímulos obtidos usando técnicas de síntese. Podem dividir-se estes estudos em dois tipos: um, manipulando directamente as características espectrais do som, o outro, utilizando síntese articulatória.

Os primeiros estudos baseados na manipulação directa do espectro foram realizados por Delattre (1954), usando o *Pattern Playback* desenvolvido nos Laboratórios Haskins. Conseguiu obter vogais nasais adicionando energia na vizinhança da frequência fundamental. Takeuchi *et al.* (1975) adicionou pares de pólos e zeros ao espectro de voz natural a várias frequências, descobrindo que os sons eram considerados mais vezes como nasais, quando a adição era feita na vizinhança da primeira formante. Diversos estudos utilizaram o sintetizador de formantes (Klatt, 1980; Klatt e Klatt, 1990) para o estudo das vogais nasais (Chen, 1995; Stevens *et al.*, 1987, 1985; Hawkins e Stevens, 1985).

Estudos usando síntese articulatória

Diversos estudos utilizaram síntese articulatória para, através de simulação, obter a função de transferência, respectivas formantes e anti-formantes, e/ou para obter os estímulos necessários para a realização de testes perceptuais.

O primeiro estudo deste tipo foi o realizado por House e Stevens (1956), usando o sintetizador desenvolvido por Stevens *et al.* (1953). Efectuaram medições de impedâncias, de entrada do tracto nasal e do conjunto do tracto nasal e oral, para diferentes áreas de acoplamento. Estudaram o efeito nas primeiras formantes de diferentes aberturas do velo. Realizaram, ainda, testes perceptuais usando estímulos produzidos pelo sintetizador. Com base nos estudos realizados concluíram, entre outras coisas, que o acoplamento da cavidade nasal durante a produção da vogal resulta numa redução da amplitude da primeira formante, num aumento da largura de banda da mesma formante, e numa redução geral do nível da vogal.

Também Fant (1960) estudou vogais nasais usando o seu sintetizador analógico LEA. Estudou, para várias vogais, o efeito no espectro de diversos factores, como: a área de acoplamento, a área de radiação, a existência de obstruções, e a alteração da área da passagem oral pelo abaixamento do velo.

Maeda (1982b) estudou o efeito da adição ao modelo do tracto nasal de seios paranasais, obtendo resultados que apontam para a possibilidade de os seios desempenharem um papel importante na definição do espectro de vogais nasais.

Krakow *et al.* (1988) estudaram a influência do contexto na percepção de vogais nasais, comparando a percepção na presença e ausência de uma consoante nasal adjacente. Os estímulos para os testes perceptuais foram obtidos utilizando o sintetizador desenvolvido nos Laboratórios Haskins (Rubin *et al.*, 1981).

Childers e Ding (1991) e Ding (1990) efectuaram experiências com vogais e consoantes nasais, usando síntese articulatória. Investigaram o efeito dos seios na produção de consoantes e vogais nasais, obtendo resultados contrários aos de Maeda (1982b). Os seus resultados indicam que o seio maxilar afecta pouco a qualidade da vogal. Estudaram, também, a correlação entre a área de abertura do véu palatino e a nasalidade, concluindo pela existência de relação entre a percepção de nasalidade e a abertura do velo.

A procura de uma medida acústica, independente da vogal, para a nasalidade percebida, foi efectuada por Maeda (1993), com sucesso parcial, utilizando síntese articulatória. Realizou simulações, para obtenção de funções de transferência para várias vogais nasais, e testes perceptuais em que ouvintes classificaram o grau de nasalidade numa escala de cinco pontos.

O estudo dos efeitos da assimetria, das duas passagens laterais das cavidades nasais, foi estudado por Lin (1994), através da inclusão de um modelo nasal constituído por três tubos no sintetizador desenvolvido anteriormente pelo mesmo investigador (Lin, 1990).

2.10.2 Estudos das vogais nasais portuguesas

Diversos estudos, usando metodologias diversas, e com objectivos também bastante diferentes, foram já realizados acerca das vogais nasais do Português. Resumem-se, de seguida, algumas contribuições importantes para o conhecimento dos sons nasais vocálicos do Português, em especial, na sua variante europeia.

Os primeiros estudos foram realizados usando apenas a capacidade auditiva dos investigadores. Viana (1883) refere a diferença das vogais nasais do Português em relação às do Francês. Segundo ele, a nasalização em Português não é acompanhada de guturalização; o timbre das vogais nasais é o mesmo das vogais orais e não há em Português nenhuma vogal nasal equivalente, em timbre, a qualquer vogal nasal francesa. Considera que a nasalidade portuguesa é mais fraca do que a francesa (Viana, 1892). Noutro estudo (citado em Lacerda e Head, 1966, pág. 12), o mesmo autor refere que apenas existem vogais e ditongos nasais puros antes de repouso, de uma vogal ou de uma consoante contínua. Quando a uma vogal se segue explo-

siva, além dessa vogal nasal ouve-se, atenuada, uma consoante nasal, homorgânica com essa explosiva (Viana, 1892).

Os estudos de Louro (1954-1955), usando valores de formantes e radiografias, levam-no a considerar as vogais nasais como sons oro-nasais, visto que a ressonância nasal se junta à vogal emitida normalmente pela boca. Para ele apenas o [ə̃] pode ser oro-nasal ou exclusivamente nasal, neste caso apenas emitido pelas fossas nasais. As outras vogais, quando mediais ou no interior das frases, são geralmente ligadas (ou mesmo substituídas na sua parte final) por um ə̃ (formando com elas uma espécie de ditongo decrescente). Para Louro são as vibrações deste ə̃ que fazem pensar na existência de verdadeiras consoantes nasais, em fim de sílaba interna, em Português. Relativamente a essas consoantes nasais finais, antes de oclusiva, apesar de serem representadas na actual ortografia, não existe a consciência de se pronunciarem, não lhes correspondendo quaisquer movimentos articulatórios activos, próprios. Especialmente no caso da nasal ser seguida de consoante oclusiva labial, pode surgir o equivalente fonético da consoante nasal, do *m*. Mas este *m* é também inconsciente e passivo, resultando apenas do mesmo ə̃ e da oclusão da boca para a pronúncia da labial seguinte.

Lacerda e Strevens (citados em Lacerda e Head, 1966, pág. 9), em 1956, utilizaram um extensor sonoro (*speech-stretcher*) e verificaram que as vogais nasais em Português têm um segmento inicial cujo grau de nasalidade varia grandemente. Em alguns casos a nasalidade pode ser tão diminuta que na prática se pode considerar inexistente, considerando-se o início como oral. A natureza deste segmento inicial é dependente do contexto, havendo contextos que favorecem mais a existência de um início oral do que outros.

Para Barbosa (1961), antes de uma pausa, a vogal nasal é em princípio pura, podendo no entanto existir na fase final uma oclusão da passagem oral na região velar. Para este autor, essa consoante nasal passa despercebida a ouvidos não treinados em análise fonética, apesar de ser bem notória a sua presença em espectrogramas. Antes de oclusiva, considera a existência de uma consoante nasal nítida e audível, com ponto de articulação dependente da consoante seguinte. Barbosa (1961) refere nunca ter encontrado exemplos de vogal nasal seguida de oclusiva oral sem a existência de consoante nasal intermédia.

Em 1964 Head (citado em Almeida, 1976, pág. 357) estudou as características acústicas das vogais nasais portuguesas usando espectrogramas. Refere uma maior densidade de formantes nas vogais nasais do que nas correspondentes vogais orais. Para a maioria das vogais, /ĩ, ẽ, õ, ũ/, existe, segundo Head, uma formante adicional entre as duas primeiras; para o /ɐ̃/ a formante adicional aparece abaixo da posição da formante mais baixa do /a/.

Lacerda e Head (1966) efectuaram um estudo instrumental, usando o denominado pneumocromográfico (Hammarström, 1952), o que lhes permitiu ter informação simultânea da radiação nasal e da radiação oral em traçados separados. Foi efectuado o estudo da nasalidade de vogais nasais em posição final, antes de oclusiva e antes de restritiva. Como resultado das suas análises obtiveram que as vogais nasais são sempre parcialmente orais, com nasalidade médio-final ou final, excepto no caso de ocorrer uma consoante nasal anterior. Neste caso são

inteiramente nasais, isto é, com radiação nasal desde o início até ao final.

A análise acústica efectuada por Martins em 1973 (citada em Mateus, 1975, pág. 94) distinguiu, nas vogais de palavras como *canto, campo*, uma primeira área sem ressonância nasal, seguida de uma área em que a ressonância nasal aumenta.

Stevens *et al.* (1987) efectuaram testes perceptuais usando um conjunto de estímulos gerados por um sintetizador de formantes. Os estímulos foram obtidos variando a duração da vogal, a nasalização, e a duração do murmúrio nasal que segue a vogal. Foi pedido a falantes de Português, Inglês e Francês, para (1) assinalar a presença ou ausência de nasalização e (2) classificar o estímulo em relação à sua naturalidade. Foi estudado, apenas, o caso de vogal nasal entre duas oclusivas, o [t]. Os falantes de Português preferiram estímulos em que existia murmúrio nasal e mais nasalização. Apesar de preferirem mais nasalização, não era necessário que a nasalização ocorresse durante toda a realização da vogal.

Silva (1995) efectuou a análise de vogais nasais em vários contextos [11], utilizando o método *Recursive Least Squares* (RLS) (Haykin, 1996) para obtenção dos parâmetros de um modelo auto-regressivo. Obteve informação acerca da duração e da variação ao longo da realização das vogais nasais das primeiras três formantes, respectivas amplitudes e larguras de banda. As suas análises levaram-no a concluir pela existência de dois estados estáveis, além de uma zona de transição. Silva (1995) considera que a pronunciação das vogais nasais, em Português, é efectuada da seguinte forma:

- Posicionamento dos articuladores para a pronunciação da vogal oral correspondente. Esta pronunciação terá entre 60 e 100 ms de duração;
- Abertura da passagem velo-faríngea, dando origem a uma fase de transição, com a duração de 30 a 50 ms;
- Aparecimento de uma segunda fase estável, correspondente à situação de acoplamento do tracto nasal. Esta fase possui características muito semelhantes para as várias vogais nasais.

Galvão (1998) efectuou medidas de *nasalance* [12] de sete falantes do Português, obtendo confirmação para a existência de um murmúrio nasal ou traço de uma consoante nasal, na fase final de vogais nasais, quando situadas antes de consoantes oclusivas. Confirmou, também, que vogais elevadas possuem uma maior percentagem de *nasalance* do que vogais médias.

Para além dos estudos da Fonética outros tipos de estudos poderão ser úteis para a compreensão das vogais nasais. Um tipo de estudo com interesse é o da forma como se deu o aparecimento e a evolução deste tipo de sons em Português (Sampson, 1999; Fagan, 1988; Nobiling, 1974; Cunha, 1982), nas línguas românicas (Sampson, 1999) e mesmo nas línguas

[11] Infelizmente não aparecendo representado o caso de vogal nasal depois de consoante nasal.
[12] *Nasalance* é o quociente da energia nasal pela energia total, $N/(N+O)$.

em geral (Hajek, 1997). Outro tipo de estudos considera a Fonologia (Barbosa, 1994; Parkinson, 1983; Almeida, 1976; Mateus, 1975; Barbosa, 1965). Nesta área ainda não existe um consenso, existindo várias teorias. Alguns autores consideram a existência de vogais nasais ao nível fonológico, outros consideram que as vogais nasais se podem derivar de sequências compostas por vogal e consoante nasal, Parkinson (1983) propõe que as vogais nasais devem ser vistas como ditongos. Alguns estudos recentes utilizam representações multilineares, como a Fonologia Autosegmental (Goldsmith, 1990), para tentar superar as dificuldades de análise fonológica das vogais nasais em Português (d'Andrade e Kihm, 1988). Nestes trabalhos a nasalidade é representada numa fiada (*tier*) separada. Os estudos da variante do Português falada no Brasil podem também dar informações úteis (de Moraes, 1997; de Sousa, 1994; Cagliari, 1977; Brito, 1975).

Capítulo 3

Síntese Articulatória

> The next generation of text-to-speech will probably be based on vocal tract line analogues or a parallel formant synthesis designed for automatic and complete simulation of a line analogue.
>
> Gunnar Fant
> (Fant, 1991, pág. 77)

Neste capítulo apresentam-se as técnicas utilizadas em síntese articulatória. A informação tem como destinatários não especialistas, dando-se especial atenção a assuntos com utilização directa no sintetizador por nós desenvolvido, descrito no capítulo seguinte.

O capítulo inicia-se definindo síntese articulatória e apresentando os vários componentes que geralmente constituem um sintetizador articulatório. Na segunda secção apresenta-se uma breve história de síntese de voz utilizando modelos directamente relacionados com o processo humano de produção. O modelamento das estruturas constituintes do aparelho de produção de voz é apresentado na terceira secção. Os modelos acústicos são descritos na quarta e quinta secções. As fontes de excitação são descritas na sexta secção. Os métodos para obtenção dos parâmetros articulatórios, representando os articuladores, são descritos na sétima secção. Na última secção, apresentam-se exemplos de vários tipos de aplicações dos sintetizadores articulatórios.

3.1 Síntese articulatória

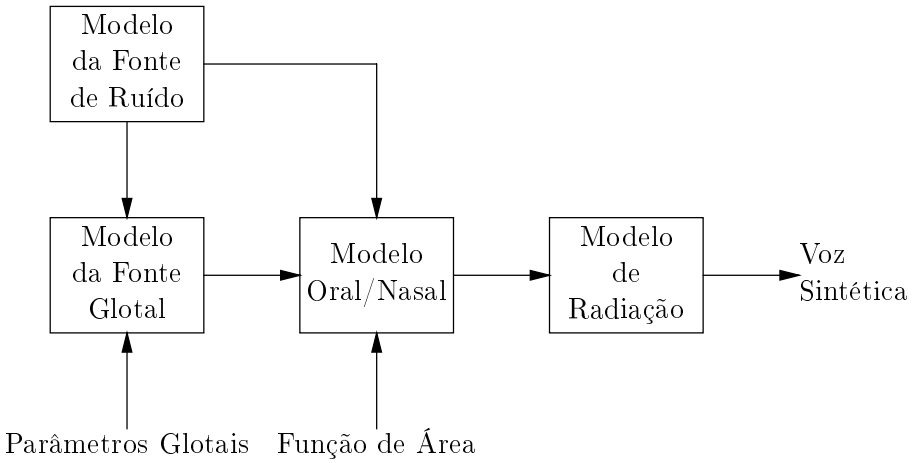


Figura 3.1: Estrutura básica da síntese articulatória. Adaptado de (Hsieh, 1994).

A síntese articulatória gera o sinal de voz através da modelação das características físicas, anatómicas e fisiológicas do aparelho produtor de voz humano. A grande diferença para outros sistemas, como a síntese de formantes (Lalwani, 1991; Klatt, 1980), é que nesta técnica se modela directamente o sistema em lugar de se modelar o sinal ou as suas características acústicas. Nas abordagens baseadas no sinal [1] o objectivo é reproduzir o sinal de voz natural o mais fielmente possível com poucas, ou nenhumas, preocupações para a forma como este é produzido. Por contraste, um modelo baseado no sistema produtor utiliza leis da física para descrever a propagação no tracto, e modela os fenómenos de mecânica e física de fluidos para descrever a oscilação das cordas vocais.

Para implementar um sintetizador articulatório num computador digital precisa-se de um modelo matemático do sistema vocal. Geralmente os sintetizadores incluem dois subsistemas: um modelo anatómico-fisiológico das estruturas envolvidas na produção de voz, e um modelo da produção e propagação do som nessas estruturas. O primeiro modelo transforma as posições dos articuladores, como o maxilar, língua, e velo, na área de secção do tracto vocal. O segundo modelo consiste num conjunto de equações que descrevem as propriedades acústicas do sistema vocal. Geralmente é constituído por vários submodelos para simular diferentes fenómenos, como: a criação de uma fonte de excitação periódica por oscilação das cordas vocais; fontes de som causadas pelo fluxo turbulento no caso de existência de zonas de área bastante reduzida ao longo do tracto; propagação do som nas cavidades acima e abaixo das

[1] Consideramos que a descrição destes métodos, mesmo sem grande profundidade, não se justificaria nesta tese, remetendo os leitores menos familiarizados com estes métodos para outros trabalhos. Descrições genéricas dos vários métodos podem ser encontradas em Benoît (1997); Carlson e Granström (1996); Oliveira (1996); Carlson (1994); Klatt (1987). Diversas obras apresentam descrições detalhadas dos vários métodos existentes (Childers, 2000; Sproat, 1998; Dutoit, 1997; van Santen *et al.*, 1996; Flanagan, 1972).

cordas vocais; radiação nos lábios e/ou narinas.

Os parâmetros para os modelos podem ter várias origens. Podem ser obtidos directamente de sinal de voz por um processo de inversão por optimização, serem definidos manualmente pelo investigador, ou serem a saída da parte de processamento linguístico de um sistema de conversão de texto para fala (*Text to Speech* (TTS)).

Estes sintetizadores ainda não atingiram o desenvolvimento necessário para serem uma alternativa aos métodos actualmente utilizados em sistemas de conversão de texto para fala. Isto deve-se a diversos factores: a dificuldade de obter informação acerca do tracto vocal e das cordas vocais durante a produção de voz em seres humanos [2]; as técnicas de medição directa geralmente apenas nos darem valores para configurações estáticas, não sendo fácil obter informação acerca da dinâmica dos articuladores; não existe, ainda, um processo de análise para obtenção dos parâmetros articulatórios a partir de voz natural; os cálculos necessários são complexos e demorados, havendo problemas de estabilidade na resolução numérica.

Apesar das desvantagens, a síntese articulatória apresenta algumas vantagens importantes: os parâmetros do sintetizador estão directamente relacionados com os mecanismos articulatórios humanos, sendo portanto muito úteis em estudos de produção e percepção de voz (Rubin *et al.*, 1981); como os parâmetros variam lentamente no tempo são bons candidatos ao uso em processos de codificação eficientes; este método pode produzir consoantes nasais e vogais nasais com elevada qualidade (Maeda, 1982b); os parâmetros são mais fáceis de interpolar que os parâmetros LPC e os dos sintetizadores de formantes (Sondhi e Schroeter, 1987), pequenos erros nos sinais de controlo não provocam geralmente sons de baixa naturalidade, pelo facto dos valores interpolados serem fisicamente realizáveis; a interacção entre fonte e o tracto, que é essencial para um som natural, pode ser convenientemente modelada (Rothemberg, 1981; Koizumi *et al.*, 1985), por se simular o movimento das cordas vocais e do tracto como um sistema único.

3.2 Breve história da síntese de voz baseada no modelamento articulatório

> To understand a science it is necessary to know its history.
>
> Auguste Comte (1798–1857)

De uma forma resumida, apresentam-se de seguida alguns passos na síntese de voz, em especial os relacionados com a síntese articulatória.

Há muitos anos que o Homem demonstra curiosidade acerca da produção de voz. Essa curiosidade levou-o a investigar se seria ou não capaz de produzir voz artificial.

[2] O desenvolvimento de outras técnicas de síntese é feita à custa de análise do sinal de voz muito mais fácil de obter.

Os princípios da teoria acústica da produção de fala já eram conhecidos no século XVIII. Já nessa época a laringe era considerada como a principal fonte sonora utilizada na fala.

O Professor Kratzenstein, na Rússia, construiu em 1769 cinco tubos acústicos que excitados por palhetas produziam as vogais /a,e,i,o,u/ (Linggard, 1985, pág. 8).

O primeiro sintetizador de voz deve-se ao barão Wolfgang von Kempelen, um nobre Austríaco. Em 1791 demonstrou, em Viena, a sua máquina mecânica falante que imitava vogais e algumas consoantes, incluindo nasais. O seu sintetizador, capaz de produzir cerca de 20 sons diferentes, era composto por um fole, uma caixa de ar comprimido, um ressoador de couro e apitos accionados por alavancas. Embora a qualidade deixasse certamente muito a desejar, estes eram suficientemente próximos dos sons da fala para poderem ser identificados como vogais e consoantes. As vogais eram produzidas alterando manualmente o volume do ressoador de couro. A produção de consoantes exigia um maior virtuosismo por parte do operador que tinha de accionar as alavancas para criar orifícios por onde passava o ar, ao mesmo tempo que, com os dedos, controlava o grau de fechamento e a forma desses orifícios. Apesar de rudimentar, esta máquina abriu caminho para futuras explorações. Mais detalhes podem ser encontrados em (Mateus *et al.*, 1990, pág. 147)) , (Linggard, 1985, pág. 4) e (Dudley e Tarnoczy, 1950).

Steward (1922) foi o primeiro a produzir vogais utilizando um dispositivo eléctrico.

Um dos primeiros sintetizadores eléctricos foi demonstrado em 1936 por Homer Dudley. O seu *Voder* (ou *Voice Operation Demonstrator*) conseguiu, pela primeira vez, sintetizar voz contínua usando circuitos eléctricos. Este dispositivo foi demonstrado na Feira Mundial de Nova Iorque, em 1939, onde operadores especialmente treinados produziram frases a pedido dos visitantes.

O *Pattern Playback* (Cooper *et al.*, 1951) que apareceu em 1950 nos Laboratórios Haskins é o primeiro exemplo de um sintetizador moderno, não articulatório. A evolução das formantes era desenhada numa placa de vidro, depois varrida (*scanned*) para produzir voz. Este dispositivo, conhecido como sintetizador opto-electrónico, produzia o som descrito pelo espectrograma. O uso extensivo desta ferramenta promoveu muito o estudo da produção e percepção de voz.

Chiba e Kajiyama (1958) publicaram estudos da resposta do tracto utilizando integração numérica da equação de Webster.

Num trabalho percursor, Dunn (1950) recorreu à teoria das linhas de transmissão eléctricas para desenvolver uma descrição quantitativa da acústica do tracto vocal. Construiu um modelo análogo eléctrico. É considerada a primeira simulação do tracto vocal. Este modelo consistia de 25 secções em T de 0.5 cm de comprimento e área igual a 6 cm^2. Uma indutância variável podia ser inserida entre duas secções para simular a língua. Outra indutância variável representava a constrição nos lábios. A radiação era simulada medindo a tensão na saída aos terminais de uma pequena indutância. Para sons vozeados, o sintetizador era excitado com uma onda triangular de que se podia controlar a frequência fundamental. O espectro da fonte

era ajustado de forma a ter um decréscimo de $-12\ dB/oitava$. Para simular os sons surdos e murmurados, uma fonte de ruído era aplicada num ponto apropriado da linha.

Foi efectuado um modelo eléctrico melhorado por Stevens *et al.* (1953). Mais tarde Rosen (1958) construiu um modelo mais detalhado incluindo o tracto nasal. Para o estudo do sistema subglotal van den Berg (1960) construiu outro modelo eléctrico. A variação contínua dos elementos da linha de transmissão por meios electrónicos permitiu a estes dispositivos sintetizar sons contínuos (Rosen, 1958). Outro exemplo de modelo eléctrico foi o sintetizador FLEA, desenvolvido por Fant (1960).

Todos os sintetizadores iniciais usando linhas de transmissão utilizaram redes analógicas na sua implementação. No entanto as técnicas digitais, tornadas possíveis com o desenvolvimento do computador, oferecem vantagens em termos de estabilidade e precisão. Um dos primeiros sintetizadores digitais utilizou os coeficientes de reflexão nas junções dos elementos cilíndricos (Kelly Jr. e Lochbaum, 1962).

Outra implementação em computador simulou as propriedades das linhas de transmissão usando equações diferença equivalentes. Com esta formulação foi possível estudar a interacção acústica entre o tracto vocal e as cordas vocais. Esta técnica foi usada num sintetizador completo para sons surdos e sonoros por Flanagan e Landgraf (1968); Flanagan e Cherry (1969).

Também na década de sessenta tiveram lugar as primeiras tentativas de obtenção da configuração do tracto com base no sinal acústico. As primeiras abordagens basearam-se na relação entre as áreas dos diversos tubos que podem ser usados para aproximar o tracto e os coeficientes de reflexão. Estes coeficientes são facilmente derivados dos coeficientes de predição linear (*Linear Predictive Coding* (LPC)). Outra técnica utilizada baseou-se na medição da resposta impulsional nos lábios.

Os primeiros modelos representando a cavidade oral no plano sagital são apresentados no final da década de sessenta (Coker, 1967; Henke, 1966, são dois exemplos). Um dos modelos mais utilizados, ainda hoje, foi proposto por Mermelstein em 1973.

Na década de oitenta os modelos foram sendo melhorados (Liljencrants, 1985; Prado, 1991) e é proposto o modelo híbrido por Sondhi e Schroeter (1987).

Com a melhoria das técnicas computacionais e de obtenção de dados acerca do processo de produção tem-se assistido, nos últimos anos, ao desenvolvimento de modelos tridimensionais do tracto e a utilização de novos métodos de simulação dos fenómenos acústicos. Em relação à inversão, o poder de cálculo permitiu: a utilização de métodos baseados em optimização (Prado *et al.*, 1992), utilizando, por exemplo, algoritmos genéticos; a utilização de redes neuronais (Rahim e Goodyear, 1990); e melhorar os processos baseados na procura em tabelas.

3.3 Modelamento das cavidades

O primeiro aspecto do processo de produção de voz que é necessário modelar é a geometria dos tractos oral e nasal. O tracto nasal é essencialmente constante. O tracto oral, no entanto, varia continuamente a sua forma. Devido às suas características específicas, um e outro são modelados de forma diferente.

3.3.1 Modelos para o tracto vocal

A geometria do tracto vocal pode ser convenientemente descrita em termos da posição dos articuladores: a língua, lábios, glote, maxilar, etc. Modelos baseados neste tipo de descrição são designados por modelos articulatórios (Schroeter e Sondhi, 1992, pág. 233).

Um grande número de modelos articulatórios pode ser encontrado na literatura. Podem ser classificados em dois tipos principais: modelos paramétricos da área e modelos sagitais.

3.3.1.1 Modelos paramétricos da área

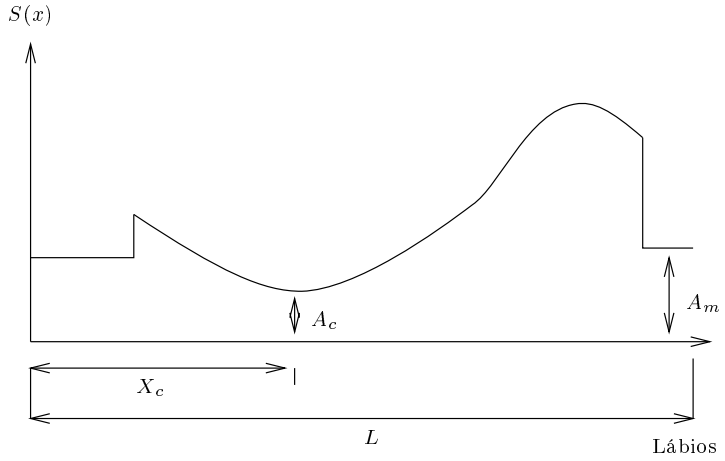


Figura 3.2: Modelo paramétrico da área de Atal et al., 1978. X_c é a distância entre a glote e o ponto de constrição máxima, A_c a área na zona de constrição (em cm^2), A_m a área de abertura da boca (em cm^2), L o comprimento total do tracto vocal.

Os modelos paramétricos da área não representam as posições dos articuladores directamente, concentram-se no modelamento da área ao longo do tracto vocal. Um grande número de modelos deste tipo foi utilizado (Stevens e House, 1955; Fant, 1960; Atal *et al.*, 1978; Flanagan *et al.*, 1980; Lin, 1990; Yu, 1993; Båvegård, 1996). A sua característica comum é especificarem a área, A_c, e a posição, X_c, de máxima constrição. A área é geralmente representada por funções contínuas como hipérboles, parábolas ou sinusóides (Lin, 1990). A título de exemplo na Figura 3.2 apresenta-se o modelo de Atal *et al.* (1978).

Parte destes modelos é baseada nas características acústicas, como o modelo *Distinctive Region Model* (DRM) proposto por Mrayati em 1988 (Hardcastle e Marchal, 1990, pág. 224). Neste modelo o tracto é dividido em várias regiões sendo a delimitação das regiões baseada na teoria acústica. Quando se altera a área numa dada região provocam-se alterações nas formantes. Estas alterações do valor das formantes não é independente da região onde se altera a área, antes pelo contrário, existindo zonas em que pequenas alterações de área causam grandes alterações nas formantes e outras regiões em que as formantes são relativamente estáveis ao alterar-se a área. Nalgumas zonas, aumentos da área provocam aumentos das frequências das formantes, noutras, diminuição. Representando as alterações das formantes ao longo do tracto, para uma pequena perturbação, os limites das regiões são obtidos com base nos zeros desta função.

Este tipo de modelos, modelando directamente a área, contemplou inicialmente apenas os sons vocálicos, só mais recentemente foi feita a sua extensão para configurações consonânticas (Båvegård, 1995b, por exemplo).

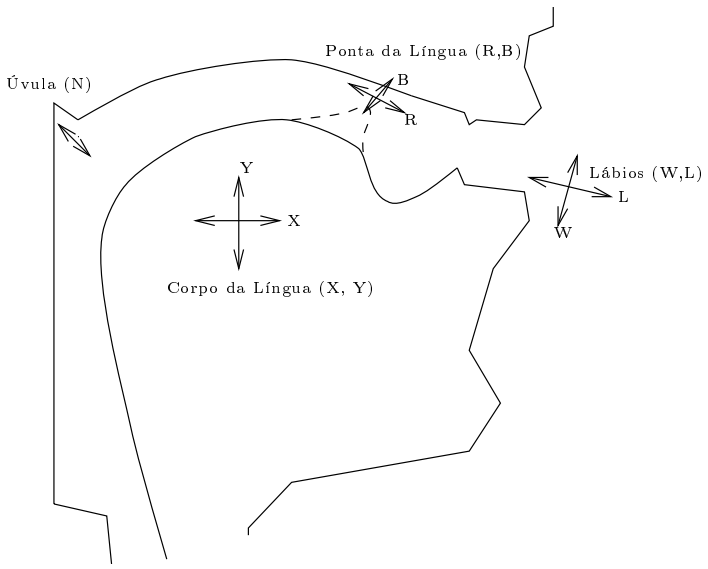


Figura 3.3: Modelo sagital de Sete Parâmetros, Flanagan et al. 1970. L e W controlam a protrusão e abertura dos lábios, X e Y a posição horizontal e vertical do corpo da língua, N a posição da úvula em relação à parede posterior da faringe, e R e B a posição do ápice da língua.

3.3.1.2 Modelos sagitais

Os modelos sagitais são baseados numa representação no plano sagital como o de uma imagem de raios X. Descrevem o movimento dos órgãos empregues na produção de voz num plano sagital. Todos os modelos deste tipo incluem as limitações do tracto vocal. Por exemplo, a língua não pode passar através do palato. A visualização e a interpretação do estado dos articuladores são as principais vantagens destes modelos. A Figura 3.3 representa um

destes modelos. Estes modelos podem dividir-se em estáticos ou dinâmicos, descritivos ou funcionais (Båvegård, 1996). Outra classificação, utilizada em Bouabana (1995), divide-os em: geométricos, estáticos, estatísticos e fisiológicos.

Um exemplo de um modelo dinâmico funcional é o de Henke (1966). É controlado por gestos (*gesture*) ou alvos (*targets*) articulatórios que são controlados por equações do movimento dos articuladores. Outros exemplos são os modelos de Perkell (1974) e o desenvolvido nos Laboratórios Haskins (Saltzman e Munhall, 1989, por exemplo).

Modelos articulatórios estatísticos, baseados na extracção de componentes principais de imagens de raios X e medições da abertura dos lábios, foram propostos por Kiritani e Maeda (1982a). O modelo de Maeda é descrito em detalhe em Bouabana (1995) e Maeda (1990).

Os modelos de mais fácil compreensão são os modelos descritivos estáticos como os desenvolvidos por Mermelstein (1973), Coker (1967, 1976) e Lindblom e Sundberg (1971).

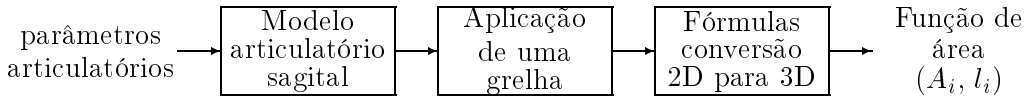


Figura 3.4: Processo de obtenção da função de área (comprimento e área das várias secções em que é dividido o tracto oral) em modelos articulatórios sagitais

Os modelos deste tipo apenas representam a configuração do tracto no plano sagital médio. Para os modelos acústicos é necessária informação tridimensional. A forma como é feita esta conversão encontra-se representada na Figura 3.4.

Antes da passagem de duas a três dimensões, o plano sagital é decomposto em várias secções para as quais se determina o comprimento e a distância entre os contornos superior e inferior no plano sagital. Utiliza-se, na decomposição, uma grelha onde cada secção corresponde à zona do tracto compreendida entre dois segmentos de recta que definem a secção. Utilizam-se diversos tipos de grelhas, sendo no entanto as mais utilizadas baseadas no sistema de coordenadas proposto por Heinz e Stevens (1964). Este sistema de coordenadas divide o tracto em três zonas: a primeira entre a glote e a parte superior da faringe, consistindo de linhas paralelas horizontais; a segunda, entre a faringe e a parte média da cavidade bucal, usando linhas radiais convergindo no ponto de origem das coordenadas; e a última representando as zonas restantes do tracto até aos lábios, usando linhas paralelas verticais. Mais detalhes sobre este processo são apresentados na secção 4.1.1.2, na página 84.

Diversos autores estudaram a obtenção da função de área (área e comprimento das várias secções ao longo do tracto) com base nas distâncias sagitais (Sundberg *et al.*, 1987; Baer *et al.*, 1991; Beautemps *et al.*, 1995, por exemplo). Geralmente a conversão entre a distância sagital e a área de secção é efectuada usando uma formula do tipo

$$\text{Área} = a \times (\text{largura no plano sagital})^b \ , \tag{3.1}$$

em que os coeficientes a e b são determinados empiricamente de medições do tracto, usando métodos directos como raios-X ou *Magnetic Ressonance Imaging* (MRI). A relação não é no entanto simples, pois os coeficientes são bastante variáveis ao longo da laringe (Ladefoged *et al.*, 1971) e os coeficientes variam de estudo para estudo.

3.3.1.3 Modelos tridimensionais

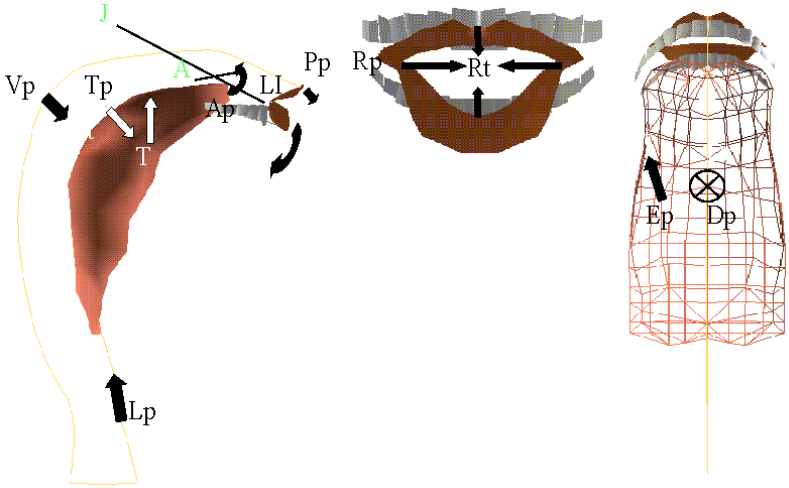


Figura 3.5: Exemplo de um modelo tridimensional (Engwall, 1999). Apenas se representa o modelo referente ao tracto oral.

Os modelos descritos, até agora, são bidimensionais. O tipo de dados disponíveis na altura em que foram desenvolvidos não permitia a inclusão da terceira dimensão. Mais recentemente, técnicas melhoradas usando MRI contribuíram para conhecimento da geometria tridimensional (Baer *et al.*, 1991; Dang e Honda, 1994; Suzuki *et al.*, 1995), tendo aparecido modelos tridimensionais como o de Engwall (1999), apresentado na Figura 3.5. O modelo consiste de uma malha tridimensional de polígonos repartidos por cinco áreas, representando as paredes do tracto oral e nasal, lábios, dentes e língua. A malha tem 750 vértices e aproximadamente 1000 polígonos. Para reduzir a complexidade foi assumido que existe simetria em relação ao plano sagital médio. Os parâmetros articulatórios utilizados neste modelo seguem, em larga medida, os do modelo de Mermelstein (1973), modificados para o caso tridimensional. Os parâmetros permitem controlar a altura da laringe, abertura do maxilar, protrusão dos lábios, arredondamento dos lábios, posição do velo, e os movimentos da língua. O modelo considera a língua como um todo. Os movimentos do ápice e dorso são sobrepostos ao modelo base. Apesar de não atingir a sofisticação de modelos da língua como os propostos por Wilhelms-Tricarico (1995) é um modelo bastante detalhado.

3.3.2 Modelos das cavidades subglotais

Não são geralmente modeladas directamente as dimensões destas cavidades, optando-se por modelar usando equivalentes acústicos, descritos mais adiante na secção 3.4.5. Uma excepção é o modelamento efectuado por Boersma (1998). Este investigador utilizou uma sequência de 29 tubos com comprimento fixo e área dependente da região subglotal a modelar (Boersma, 1998, pág. 46, para mais detalhes).

3.3.3 Modelos do tracto nasal

Ao longo dos anos, vários modelos do tracto nasal foram sendo usados em síntese articulatória. Os primeiros usaram dados provenientes de cadáveres (House e Stevens, 1956) e de moldes do tracto nasal (Fant, 1960). Estes primeiros modelos apenas modelavam as cavidades nasais não incluindo os seios paranasais e juntavam as duas passagens laterais, não considerando as assimetrias. Foi sugerido por Fujimura e Ludqvist (1971) que as cavidades paranasais seriam necessárias para explicar o espectro de vogais naturais. Um dos primeiros a incluir no seu modelo o efeito dos seios foi Maeda (1982b) que obteve as suas dimensões por um processo de análise-síntese. Maeda considerou apenas uma cavidade. Outros investigadores estudaram estas cavidades, como Masuda, em 1992, dissecando mais de 20 crânios e estudando as consequências acústicas de obstrução da passagem (ostia) (citado em (Dang e Honda, 1994)). Recentemente foi efectuado um estudo detalhado usando MRI (Dang e Honda, 1994). Este estudo obteve informação tridimensional da área do tracto nasal e dimensões dos seios. Os valores da área diferem consideravelmente dos anteriormente publicados (House e Stevens, 1956; Fant, 1960), em especial na zona média.

Apresentam-se, de seguida, resumidamente, alguns destes modelos.

3.3.3.1 Modelo de House e Stevens (1956)

As dimensões deste modelo foram baseadas largamente em atlas anatómicos, crânios, e imagens de raios-X laterais. O modelo analógico nasal era acoplado ao modelo do tracto vocal 8 *cm* acima da glote. Não era feito qualquer ajuste à área oral ao fazer variar a área de acoplamento nasal.

3.3.3.2 Modelo DANA

Este modelo, representado na Figura 3.6, foi desenvolvido por Hecker (1961, 1962) para ser integrado no sintetizador eléctrico desenvolvido no MIT por Rosen (1958). O nome de DANA adveio-lhe da denominação inglesa *Dynamic Analog of the NAsal cavities.*

Consistia em 9 secções com um comprimento total, fixo, de 12.5 *cm.* As secções 1 e 2, representando a nasofaringe, operam em conjunto e constituem um secção de 3 *cm* com área variável electronicamente (de aproximadamente 0.05 a 5.0 cm^2). A área da secção 3 era

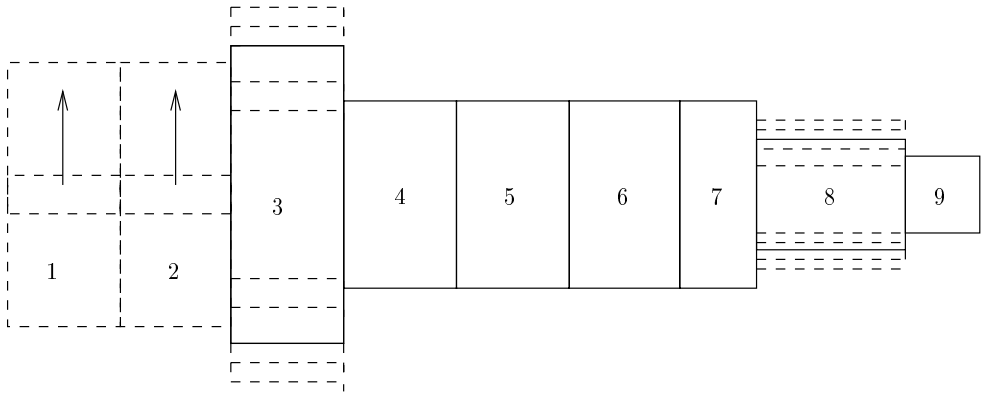


Figura 3.6: Representação do modelo DANA. A área das secções 1 e 2, representando a nasofaringe, é variável entre aproximadamente 0.05 cm^2 e 5.0 cm^2. As áreas das secções 3 e 8 são ajustáveis manualmente. As secções 4, 5, 6 e 7 têm área fixa igual a 2.6 cm^2, enquanto a secção 9, representando as narinas, tem uma área de 0.42 cm^2 (Hecker, 1961, Figura XVIII-11, pág. 190).

manualmente variável (2.0, 4.0, 6.0, 8.0 e 10.0 cm^2). As secções 4 a 7 representavam uma região de área aproximadamente constante (2.6 cm^2), e a secção 8 oferecia controlo manual (0.4, 0.8, 1.2, 1.6 e 2.0 cm^2). A secção 9 tinha área igual a 0.42 cm^2. Para um adulto do sexo masculino, as cavidades nasais eram acopladas aproximadamente 8 cm acima da glote.

3.3.3.3 Modelo de Fant (1960) e seus derivados

Baseado nos dados anatómicos de Fant (1960), este modelo continuou em uso no *Kungl Tekniska Högskolan* (KTH) (Lin, 1990; Båvegård *et al.*, 1993).

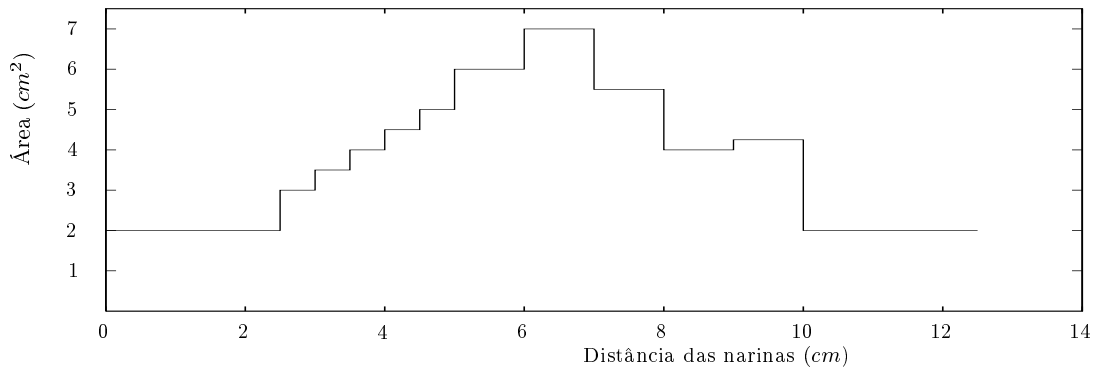


Figura 3.7: Configuração nasal, baseada nos dados anatómicos de (Fant, 1960), usada por Lin (Lin, 1990, Figura 2.12, pág. 37).

Lin (1990) utilizou uma aproximação aos dados anatómicos de Fant, como se representa na figura 3.7. Dois seios foram incorporados no modelo. Estes representando, presumivelmente,

os seios maxilares e frontais foram inseridos a 6 cm e 8 cm das narinas, respectivamente. Cada um é representado por um circuito RLC. As frequências de ressonância variam com os indivíduos, e segundo dados de Lindqvist-Gauffin e Sundberg (1976) para os seios maxilares encontram-se entre 200 e 600 Hz e para os frontais entre 500 e 2000 Hz. Lin escolheu os valores médios destes intervalos, respectivamente 500 Hz e 1400 Hz. Os valores dos componentes dos circuitos RLC, volume e ressonâncias encontram-se na Tabela 3.1.

	Volume cm^3	R_{seio} $dine \cdot s/cm^5$	L_{seio} g/cm^4	C_{seio} $cm^4 \cdot s^2/g$	F Hz	B Hz
Maxilares	43	1.1	$3.42 \cdot 10^{-3}$	$29.7 \cdot 10^{-6}$	500	50
Frontais	1.6	7.2	$11.4 \cdot 10^{-3}$	$1.13 \cdot 10^{-6}$	1400	100

Tabela 3.1: Dados para os seios nasais, segundo (Lin, 1990, Tabela 2.6, pág. 38)

3.3.3.4 Modelo de Maeda, 1982

Neste modelo o tracto nasal tem um comprimento de 11 cm e é representado por 11 secções de 1 cm de comprimento com as áreas apresentadas na Figura 3.8. As primeiras 3 secções têm área variável, sendo a área da primeira secção a área de acoplamento nasal e a área das secções 2 e 3 obtida por interpolação linear entre a área da primeira e da quarta secção.

Os seios foram representados por uma única cavidade, os seios maxilares, com um volume de 20.8 cm^3 acoplado ao tracto nasal por um tubo de 0.5 cm de comprimento e 0.1 cm^2 de secção, a uma distância de 7 cm do véu palatino. Esta cavidade foi modelada por uma concatenação de várias secções (Maeda, 1982b). O efeito do seio na resposta do tracto nasal, considerando a abertura na zona de acoplamento nula, pode ver-se na Figura 3.9.

Diversos investigadores usaram os dados de Maeda para a área do tracto nasal (Childers e Ding, 1991; Ding, 1990, são dois exemplos). Este modelo, com adaptações, foi utilizado por Sondhi e Schroeter (1987) no seu modelo híbrido. Modelaram os seios paranasais usando um circuito ressonante RLC com impedância:

$$Z_{seio} = R_{seio} + j\omega L_{seio} + \frac{1}{j\omega C_{seio}} \ , \tag{3.2}$$

representanto o ressoador de Helmholtz constituído pela cavidade dos seios e a ligação, de área reduzida, destes com as cavidades nasais (Borden *et al.*, 1994, pág. 15).

3.3.3.5 Modelos com área de radiação reduzida

Diversos autores propuseram, e utilizaram, modelos das cavidades nasais em que a área de radiação, isto é, a área das narinas é mais reduzida do que a medida em seres humanos. Como base para esta escolha encontra-se a necessidade de ter modelos com características acústicas

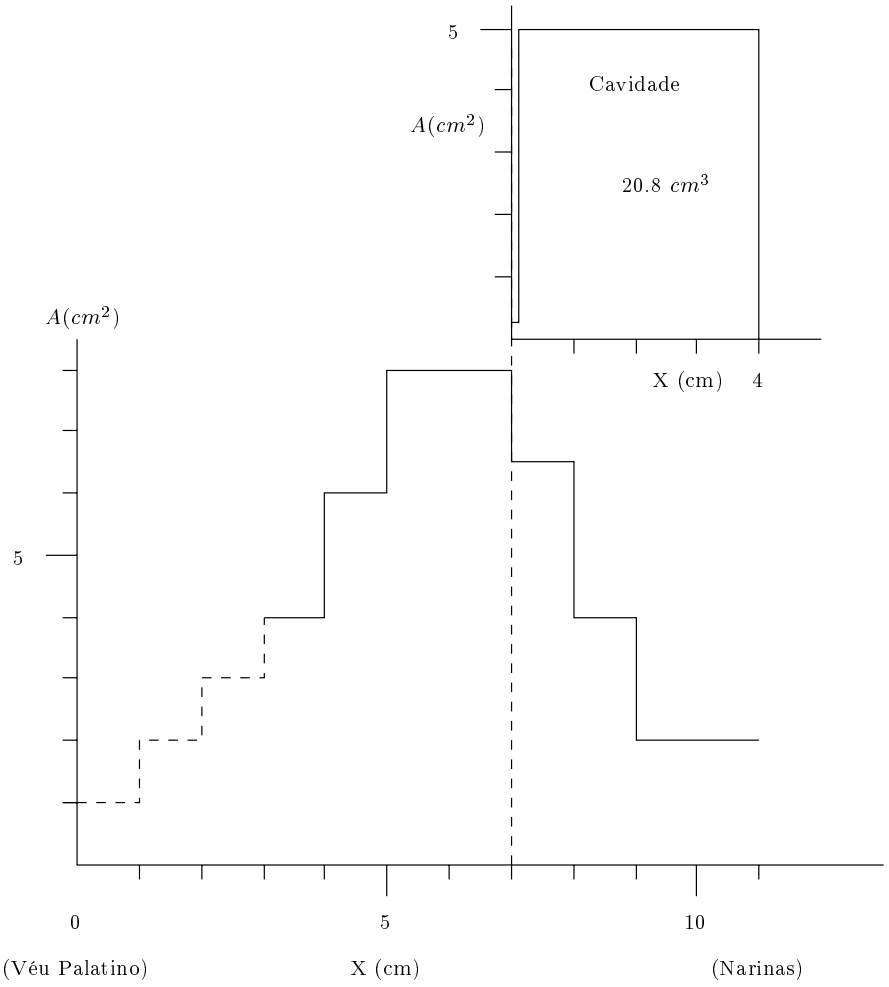


Figura 3.8: A função de área do tracto nasal segundo Maeda, 1982. A tracejado indicam-se as secções com área variável. Adaptado de (Childers e Ding, 1991, Figura 3)

equivalentes às do tracto nasal humano.

Os primeiros a utilizar este tipo de modelos foram House e Stevens (1956), que utilizaram uma área de 0.23 cm^2.

Feng (1987) fez um estudo exaustivo, concluindo pela necessidade de utilização deste tipo de modelos. Apresenta, também, uma possível explicação anatómica para este tipo de modelos. Segundo este autor a utilização de uma área reduzida das narinas justifica-se pela existência de uma zona de passagem relativamente estreita, um pouco antes das narinas, designada por *limen nasi*. Nos seus trabalhos de simulação utilizaram um valor de 0.6 cm^2 (Feng, 1987; Feng e Castelli, 1996).

Chen (1997), baseando-se em dados de Stevens (1998) e Dang e Honda (1994), também utilizou uma área de radiação reduzida, 0.5 cm^2.

Båvegård *et al.* (1993) utilizaram os dados anatómicos de Fant (1960) reduzindo para metade

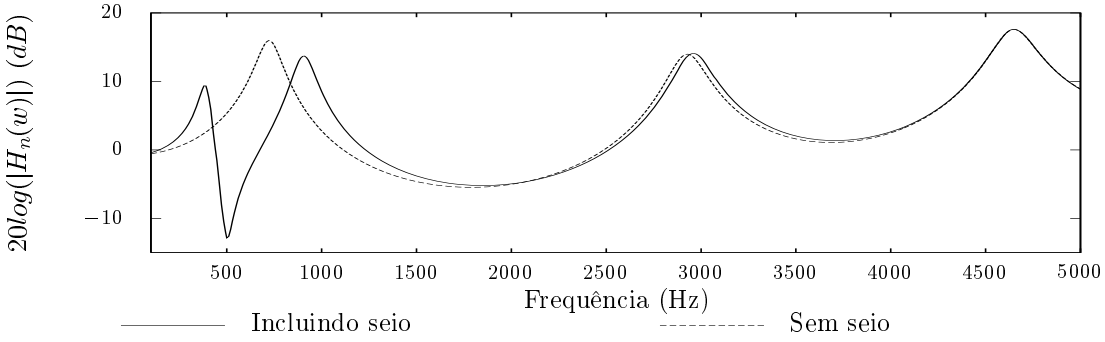


Figura 3.9: Resposta do tracto nasal usando os dados de Maeda, 1982. Foi calculada a resposta incluindo ou não os seios maxilares. O factor S para as perdas usado foi de 2, modelo de radiação de Flanagan, 1972 (Maeda, 1982b).

a área das secções, cobrindo os primeiros 4 cm a contar das narinas. Chamaram a este modelo "nariz estreito" (do Inglês *narrow nose*).

3.3.3.6 Modelo assimétricos

A utilização de ressonância magnética permitiu a obtenção de dados anatómicos mais detalhados das cavidades nasais. Tornou possível a medição em condições mais próximas das reais (sem aplicação de soluções destinadas a diminuir a cobertura mucosa), a obtenção de dados acerca das duas passagens laterais e, ainda, dados acerca das dimensões das cavidades paranasais e suas ligações às cavidades nasais.

Um dos estudos mais relevantes foi o efectuado por Dang e Honda (1994). Foram medidas áreas de secção, e perímetros das passagens nasais para 4 indivíduos. Foram também obtidos dados relativamente aos seios paranasais maxilares e esfenoidais.

As principais conclusões dos estudos efectuados por Dang e Honda (1994) foram:

1. as diferenças entre indivíduos são maiores para o volume do que para o comprimento da cavidade nasal;
2. os valores de área diferem grandemente dos anteriormente publicados (House e Stevens, 1956; Fant, 1960), em particular na parte média do tracto nasal;
3. é necessário implementar as cavidades paranasais num modelo do tracto nasal para descrever adequadamente as suas propriedades acústicas;
4. o tracto nasal tende a ser assimétrico, sendo necessário estudar os efeitos dessa assimetria em sons nasais.

Experiências com um modelo assimétrico, baseado nos dados de Dang e Honda (1994), foram efectuadas por (Lin, 1994).

Os valores obtidos por Story (1995), usando também MRI, confirmam a assimetria das passagens nasais esquerda e direita.

3.3.3.7 Problemas no modelamento das cavidades nasais e paranasais

O modelamento das cavidades nasais e paranasais coloca diversos problemas:

- É muito difícil a obtenção de dados anatómicos das cavidades nasais e cavidades paranasais porque estas cavidades são de mais difícil acesso do que as cavidades orais, têm uma forma mais complexa e encontram-se cobertas por uma mucosa;
- O facto de a cavidade ser revestida por uma mucosa leva a grandes alterações de volume por perturbações de saúde. Torna também incorrectos dados anatómicos obtidos em cadáveres;
- As dimensões, especialmente reduzidas, do canal de ligação dos seios paranasais torna muito difícil, ou mesmo impossível, a sua medição mesmo com MRI;
- O formato dos dois canais é bastante complexo o que torna difícil o cálculo das perdas que dependem da superfície;
- Existe, também, grande variabilidade das cavidades nasais entre indivíduos, sendo especialmente variável o volume.

3.4 Modelos acústicos

O modelo acústico do sistema vocal humano engloba vários submodelos. Os modelos do tracto oral e nasal simulam a propagação nesses tractos. A fonte de excitação glotal representa e gera a onda de excitação glotal. O fluxo turbulento de ar numa constrição nas fricativas e oclusivas é produzido pelo modelo da fonte de ruído. O modelo de radiação simula a radiação da energia acústica dos lábios e narinas.

O tracto vocal é um tubo acústico tridimensional curvo com forma lentamente variável ao longo do tempo. As paredes do tubo são flexíveis, existem perdas provocadas por atrito e condução de calor. As condições de fronteira, nos pontos de radiação e glote, são também variáveis. Existe a possibilidade de acoplamento de um tubo adicional representando o tracto nasal. O acoplamento é feito na parte superior da faringe. O tracto nasal tem dimensões fixas, mas o acoplamento é variável.

Investigações preliminares demonstraram que a equação de Navier-Stokes para fluxo de fluidos pode caracterizar as não linearidades envolvidas na produção de som pelas cordas vocais; a produção de fricativas surdas por fluxo turbulento em constrições; e efeitos de radiação condicionados por propagação num tubo não-uniforme, com perdas e de paredes flexíveis (Thomas, 1986; Hegerl e Höge, 1991). No entanto, os resultados têm sido limitados pelas

exigências computacionais necessárias à resolução da equação de Navier-Stokes numa grelha tempo-espaço realística. Estas limitações levaram os investigadores à procura de modelos simplificados.

3.4.1 Simplificações

> Truth is much too complicated to allow anything but approximations.
>
> John von Neuman

A primeira simplificação que é geralmente efectuada no modelamento acústico do tracto consiste em "esticar" o tracto vocal. Segundo estudos de Sondhi (1986) a variação nas formantes provocada por esta simplificação situa-se no intervalo 2 a 4 %, para frequências inferiores a 4 kHz. O tracto vocal pode ser representado por um tubo direito sem grande perda de precisão.

A segunda aproximação é considerar a propagação das ondas como sendo planar ao longo do tubo. Existem duas razões que justificam esta aproximação: o tecido ao longo do tracto contraria a propagação radial; e as dimensões laterais médias na ordem dos 2.0 cm levam a que outros modos de propagação só ocorram para frequências próximas ou acima do limite superior das frequências [3] com informação do sinal de voz. Em Rossing e Fletcher (1994), pág. 175, deduzem-se as expressões para as frequências a partir das quais existem outros modos de propagação, para o caso de um tubo cilíndrico infinito. Os primeiros três modos de ordem superior têm frequências angulares de corte iguais a $1.84c/a$, $3.05c/a$ e $3.80c/a$, sendo a o raio do tubo e c a velocidade do som. Para uma área, relativamente elevada [4], de 15 cm^2 , a frequência de corte, do primeiro modo, é de cerca de 4700 Hz. Por esta razão os algoritmos, baseados na aproximação planar da propagação, são considerados válidos até $4000 - 5000$ Hz (Story, 1995, pág. 30). Felizmente, a maior parte da informação do sinal de voz encontra-se abaixo dos 4000 Hz.

Mesmo desprezando as perdas por fricção, condução e as resultantes das paredes flexíveis, as equações daí resultantes, em geral, apenas podem ser resolvidas numericamente. Precisamos, pois, de mais uma aproximação. Uma abordagem habitual é dividir o tracto num conjunto de secções cilíndricas contíguas. Faz-se, portantoi, uma discretização espacial do tubo. Se o número de secções for elevado, estes elementos de comprimento reduzido constituem uma boa aproximação da função de área contínua. As frequências de ressonância do conjunto de tubos são muito próximas das obtidas no caso contínuo. O tubo cilíndrico uniforme torna-se de muito mais fácil análise. Mesmo assim, as primeiras análises não incluíram as perdas.

[3] Geralmente 4 a 5 kHz para sons não fricativos e 8 kHz para fricativas.

[4] O valor máximo da área para as vogais americanas, segundo os dados de Story (1995), não atinge os 8 cm^2.

3.4.2 Equação de onda

O tracto vocal constitui um tubo acústico com forma variável. Considerando, numa primeira aproximação, as paredes rígidas, a teoria acústica linear (Rossing e Fletcher, 1994; Morse, 1991; Morse e Ingard, 1968) descreve a propagação do som através das equações de continuidade e conservação do momento (Sinder, 1999, pág. 7)

$$\frac{\partial p}{\partial t} + \rho c^2 \frac{\partial v_i}{\partial x_i} = 0 \tag{3.3}$$

$$\rho \frac{\partial p}{\partial t} + \frac{\partial p}{\partial x_i} = 0 \tag{3.4}$$

Nestas equações ρ é a densidade do meio, c a velocidade de propagação do som, v_i a velocidade da partícula na direcção x_i, e p representa a pressão.

Assumindo propagação planar, apenas é necessário considerar formas das equações utilizando uma dimensão. Reescrevendo as equações, com substituição das velocidades pelo fluxo, obtém-se

$$\frac{\partial p}{\partial t} + \rho c^2 \frac{\partial}{\partial x} \frac{u}{A(x)} = 0 \tag{3.5}$$

$$\rho \frac{\partial}{\partial t} \frac{u}{A(x)} + \frac{\partial p}{\partial x} = 0 \tag{3.6}$$

em que, u é o fluxo, e $A(x)$ é a área de secção que é função de x. Estas duas equações combinadas resultam na conhecida equação de Webster (1919)

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{1}{A(x)} \frac{\partial}{\partial x} \left[A(x) \frac{\partial p}{\partial x} \right] . \tag{3.7}$$

Esta equação não inclui perdas. A falta de uma solução analítica desta equação para geometrias arbitrárias levou, nos primeiros modelos de sintetizadores articulatórios desenvolvidos, à utilização de analogias com linhas de transmissão, assunto da próxima secção.

3.4.3 Modelo para um tubo usando a analogia com uma linha de transmissão

Dunn (1950) propôs um modelo em que o tracto é aproximado por uma série de tubos com área constante. Cada tubo foi modelado recorrendo à analogia com linhas de transmissão, relacionando a resistência acústica, inertância e complacência com a resistência, indutância e capacidade eléctricas.

Para um tubo de área constante A, não incluindo perdas, as equações anteriores simplificam-

se, obtendo-se,

$$\frac{\partial p}{\partial t} + \frac{\rho c^2}{A}\frac{\partial u}{\partial x} = 0 \quad (3.8)$$

$$\frac{\rho}{A}\frac{\partial u}{\partial t} + \frac{\partial p}{\partial x} = 0 \quad (3.9)$$

Os leitores familiarizados com a teoria das linhas de transmissão recordarão que, para uma linha de transmissão uniforme sem perdas, a tensão v e a corrente i na linha satisfazem as equações

$$\frac{\partial v}{\partial x} + L\frac{\partial i}{\partial t} = 0 \quad (3.10)$$

$$\frac{\partial i}{\partial x} + C\frac{\partial v}{\partial t} = 0\ , \quad (3.11)$$

onde L e C são a indutância e capacitância por unidade de comprimento, respectivamente. A teoria de linhas de transmissão aplica-se ao estudo da transmissão num tubo acústico, se usarmos as analogias apresentadas na Tabela 3.2.

Grandeza acústica		Grandeza eléctrica análoga	
p	pressão	v	tensão
u	fluxo (*volume velocity*)	i	corrente
ρ/A	inertância (ou indutância acústica)	L	indutância
$A/(\rho c^2)$	complacência (ou capacitância acústica)	C	capacitância

Tabela 3.2: Analogias entre grandezas acústicas e eléctricas.

3.4.3.1 Inclusão de perdas no modelo

As perdas, desprezadas por Dunn (1950), foram adicionadas ao modelo por Stevens *et al.* (1953). É possível representar as perdas provocadas pelo fluxo laminar por uma resistência em série e outra em paralelo (Linggard, 1985, pág. 43). A resistência em série representa as perdas devidas à viscosidade, proporcionais ao quadrado do fluxo; a condutância em paralelo representa as perdas devidas à transmissão de calor, proporcionais ao quadrado da pressão.

O mecanismo que provoca as perdas por viscosidade é o atrito. Se uma camada de ar junto à parede do tubo se pode considerar estacionária e o ar no centro do tubo se move com uma velocidade v, então existe um gradiente radial de velocidade. A fricção pode considerar-se como ocorrendo entre anéis concêntricos de ar, cada um movendo-se a uma velocidade ligeiramente diferente dos seus vizinhos. A "resistência acústica" por unidade de comprimento

é dada por (Flanagan, 1972),

$$R = \frac{S}{A^2}\sqrt{\frac{\omega\rho\mu}{2}} \; . \tag{3.12}$$

Atente-se que R depende da frequência angular não sendo portanto uma resistência no sentido usado em Electrotecnia. Também depende de S.

A condutância em paralelo introduz perdas proporcionais ao quadrado da tensão, representando as perdas no tubo acústico por condução de calor nas paredes. Este processo é difícil de visualizar porque se considera o tubo a uma temperatura uniforme. Isto é apenas verdade ao nível macroscópico. As variações rápidas e adiabáticas da pressão causam variações de temperatura. Flanagan (1972) mostrou que a "condutância acústica" é

$$G = \frac{S(\eta - 1)}{\rho c^2}\sqrt{\frac{\lambda\omega}{2\xi\rho}} \tag{3.13}$$

sendo λ a condutividade térmica e ξ o calor específico do ar. Como R, também G depende de ω e S.

Um problema surge no que respeita à escolha do perímetro para calcular estas resistências. Geralmente o tubo acústico é considerado circular, o que implica $S = 2\sqrt{\pi A}$ para a circunferência. Fant (1960) duplicou esse valor, o que corresponde a uma forma elíptica ou, no caso de uma forma circular, a um aumento do atrito. Este valor duplo foi adoptado por exemplo por Wakita e Fant (1978). Num modelo mais detalhado, uma conversão entre a área e o perímetro dependente da localização no tracto poderia ser usada. No entanto, são necessários mais dados anatómicos e acústicos para se poder fazer esse refinamento do modelo.

3.4.3.2 Paredes flexíveis

Até este momento consideraram-se as paredes do tubo rígidas. No tracto vocal esta aproximação não é válida. Não só as paredes vibram devido à onda de pressão, como também o volume do tubo é alterado com a variação da pressão. O efeito das paredes é primordial no caso das oclusivas sonoras. Flanagan *et al.* (1975) adicionaram novos elementos à analogia das linhas de transmissão para modelar as paredes flexíveis, assim como o som radiado, devido às vibrações das paredes do tracto.

As variações de pressão no interior do tracto submetem as paredes a uma força variável. Como as paredes são elásticas, a área do tubo irá variar. Assumindo que a reacção das paredes é local, o movimento normal à superfície, de um segmento das paredes, depende apenas da pressão acústica nesse segmento e é independente de qualquer outro segmento; a vibração da parede é simulada por um modelo mecânico com massa, viscosidade e complacência. O circuito equivalente é um circuito RLC série, com $L_p = \frac{m_p}{S^2}$, $C_p = \frac{S^2}{k_p}$ e $R_p = \frac{b_p}{S^2}$ (Rabiner e Schafer, 1978; Maeda, 1982a), onde S representa, mais uma vez, o perímetro do tubo.

Na tabela 3.3 encontram-se valores para os parâmetros propostos por vários investigadores.

Fonte	Região	b_p g/s	m_p g	k_p $dine/cm$	Observações
Ishizaka *et al.* (1975)	Face tensa	800	2,1	33300	Medição directa
Ishizaka *et al.* (1975)	Face relaxada	1060	1,5	33300	idem
Ishizaka *et al.* (1975)	Pescoço	2320	2,4	49100	ibidem
Flanagan *et al.* (1975)		1600	1,5	não considerado	
Maeda (1982a)		1400	1,5	30000	Baseado em (Ishizaka *et al.*, 1975)
Lin (1990)		1600	1,4	não considerado	

Tabela 3.3: Parâmetros para as paredes flexíveis, por unidade de área.

Habitualmente, na literatura da especialidade estas constantes foram especificadas em termos de área unitária. Em tais caso, a massa total das paredes varia com a configuração. Se forem especificados por unidade de comprimento, a massa total sofre pouca variação, visto que as variações do comprimento são pequenas, especialmente quando comparadas com a superfície. Considerando tal facto Maeda (1982a) propôs que estes parâmetros sejam definidos por unidade de comprimento. Estimou os valores das constantes, usando os dados de Ishizaka *et al.* (1975), assumindo uma área seccional de 4 cm^2.

A impedância das paredes pode ser incluída em cada secção como um elemento distribuído (Flanagan, 1972; Flanagan e Ishizaka, 1976; Flanagan *et al.*, 1975, 1980; Ishizaka *et al.*, 1975; Maeda, 1982a; Hsieh, 1994) ou inserida como duas impedâncias discretas, uma na faringe e outra ao nível do queixo (Wakita e Fant, 1978; Badin e Fant, 1984; Lin, 1990). Pode ainda usar-se um factor de correcção (Lin, 1990). O modelo discreto, que é independente da configuração do tracto, pode não dar resultados satisfatórios. O condensador foi eliminado em alguns estudos devido a ter um efeito muito reduzido (Wakita e Fant, 1978). Mais informação pode ser encontrada em (Ding, 1990; Maeda, 1982a; Lin, 1990; Linggard, 1985; Hsieh, 1994).

3.4.3.3 Modelo equivalente

Incluindo os vários componentes descritos anteriormente obtém-se o circuito equivalente para uma secção de tubo elementar, com área de secção constante, representado na Figura 3.10.

Fazendo $z = R + j\omega L$, $y = G + j\omega C + 1/Z_p$ e $Z_p = R_p + j\omega L_p + \frac{1}{j\omega C_p}$, a constante de propagação γ é dada por

$$\gamma = \sqrt{zy} \tag{3.14}$$

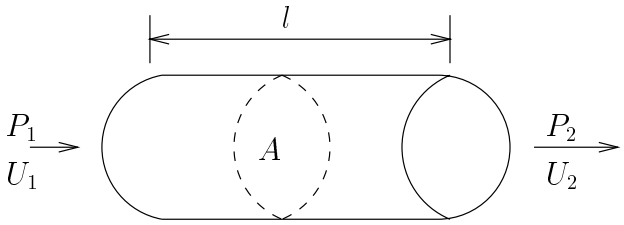
e a impedância característica Z por

$$Z = \sqrt{\frac{z}{y}}\,. \tag{3.15}$$

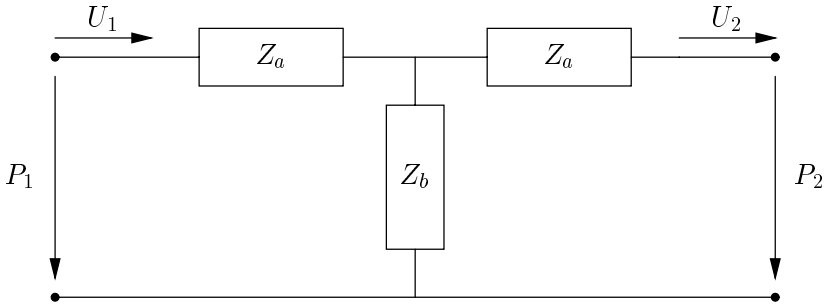
Os elementos Z_a e Z_b do circuito equivalente em T são,

$$Z_a = Z \tanh\left(\frac{\gamma l}{2}\right) \qquad Z_b = \frac{Z}{\sinh(\gamma l)} \tag{3.16}$$

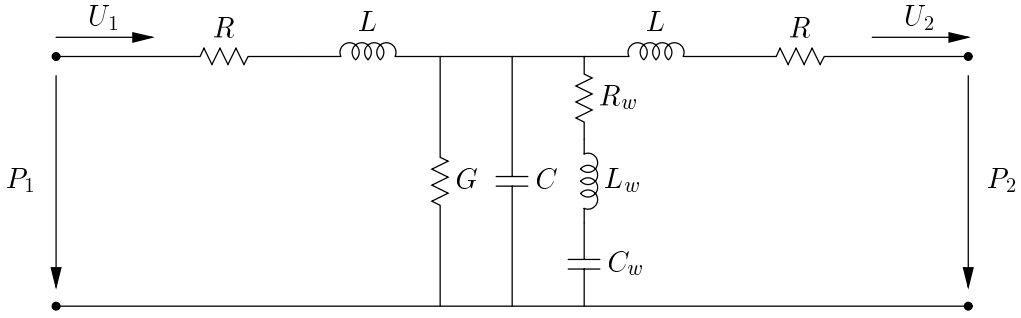
Utilizando estes valores, as relações entre as correntes e tensões para o circuito em T podem ser facilmente derivadas, quer no domínio do tempo (Ishizaka e Flanagan, 1972), quer no domínio da frequência (Sondhi e Schroeter, 1987).



(a) Tubo uniforme, de comprimento l e área de secção A, com indicação de pressões e fluxos de entrada e saída.



(b) Circuito equivalente em T.



(c) Circuito equivalente, incluindo os vários componentes.

Figura 3.10: Circuito equivalente de um tubo com perdas. Adaptado de (Flanagan, 1972)

3.4.4 Modelo alternativo proposto por Sondhi

Sondhi e Schroeter (1987), com base em Sondhi (1974), derivaram um outro circuito equivalente. Algumas correcções foram propostas depois em Schroeter e Sondhi (1992). O resultado final é semelhante ao obtido por Flanagan (1972). No entanto, segundo os autores, a derivação é mais rigorosa. A dedução entra em linha de conta com as perdas devidas a viscosidade, condutividade e paredes flexíveis, sem ter de aproximar o tracto por uma cascata de secções uniformes. Os parâmetros do modelo podem ser determinados a partir de medições acústicas.

Remetem-se os leitores interessados em mais detalhes para as referências citadas, apenas se referindo aqui os resultados. A relação entre as pressões e fluxos nos dois extremos de uma secção do tracto homogénea é dada por,

$$\begin{pmatrix} P_s \\ U_s \end{pmatrix} = \begin{pmatrix} A & C \\ B & D \end{pmatrix} \times \begin{pmatrix} P_e \\ U_e \end{pmatrix} = K \times \begin{pmatrix} P_e \\ U_e \end{pmatrix} \ , \tag{3.17}$$

onde a entrada se encontra do lado glotal e a saída do lado dos lábios ou narinas. P_s representa a pressão na saída e U_s o fluxo na saída. P_e e U_e representam as mesmas grandezas, mas agora na entrada do tubo. Para um tubo de comprimento l e área A os elementos da matriz de transmissão K são dados pelas expressões (Sondhi e Schroeter, 1987, pág. 959),

$$A = \cosh(\sigma l/c) \tag{3.18}$$

$$B = -\frac{\rho c}{A}\lambda \sinh(\sigma l/c) \tag{3.19}$$

$$C = -\frac{A}{\rho c}\frac{\sinh(\sigma l/c)}{\lambda} \tag{3.20}$$

$$D = \cosh(\sigma l/c). \tag{3.21}$$

As variáveis complexas σ e λ definem-se como

$$\lambda = \sqrt{\frac{a + j\omega}{\beta + j\omega}} \tag{3.22}$$

$$\sigma = \lambda(\beta + j\omega) \ , \tag{3.23}$$

com

$$\alpha = \sqrt{j\omega c_1} \tag{3.24}$$

$$\beta = \frac{j\omega\omega_0^2}{(j\omega + a)j\omega + b} + \alpha \ . \tag{3.25}$$

Foram propostos por Sondhi e Schroeter (1987) os seguintes valores para os parâmetros: $a = 130\pi$, $b = (30\pi)^2$, $\omega_0^2 = (406\pi)^2$, $c_1 = 4$ para o tracto oral, $c_1 = 72$ para o tracto nasal. Em Schroeter e Sondhi (1992) os valores para os parâmetros referentes ao tracto nasal foram alterados. Passou a ter-se c_1 apenas metade e ω_0 nasal o dobro.

Comparando as expressões dos elementos da matriz de transmissão apresentados com o caso geral, em função de γ e Z, de um modelo de tubo com inclusão de perdas (Scavone, 1997, por exemplo), em que $A = \cosh(\gamma l)$ e $B = -Z\sinh(\gamma l)$, facilmente se obtém a constante de propagação e impedância característica como sendo,

$$\gamma = \frac{\sigma}{c} \qquad Z = \frac{\rho c}{A}\lambda \; . \tag{3.26}$$

Caso se esteja interessado no circuito em T, Z_a e Z_b obtêm-se da mesma forma que no modelo anterior (equações 3.16).

3.4.5 Modelos das cavidades subglotais

O sistema subglotal, que inclui a traqueia e os pulmões, é geralmente omitido nas simulações, pois o seu efeito nas características espectrais é considerado pequeno, excepto para sons surdos, em que a abertura da glote é grande (Ishizaka *et al.*, 1976).

Foram efectuadas medições da impedância de entrada do sistema subglotal (Ishizaka *et al.*, 1976). Ananthapadmanabha e Fant (1982) usaram os dados destas medições e representaram o sistema subglotal por uma cascata de ressonâncias, representadas por circuitos paralelos RLC. Utilizaram apenas três circuitos para representar as três primeiras ressonâncias das cavidades subglotais. As formantes do sistema situam-se em 640, 1335, e 2110 Hz, com larguras de banda de 246, 155 e 140 Hz, respectivamente.

Outros valores foram propostos por Fant, Ishizaka, Lindqvist e Sundberg em 1972 (citados em (Wakita e Fant, 1978, pág. 14)). As frequências de ressonância situam-se em 600, 1350 e 2160 Hz com larguras de banda de 240, 180 e 190 Hz, respectivamente.

Os efeitos foram estudados por Ananthapadmanabha e Fant (1982), Badin e Fant (1984) e Lin (1990), concluindo que o efeito do sistema subglotal é pequeno, excepto para sons surdos, onde a abertura glotal é grande.

Além deste tipo de modelamento acústico simplificado, motivado pela escassez de dados acerca das configurações, foram também usadas técnicas semelhantes às utilizadas para as cavidades supraglotais (van den Berg, 1960; Boersma, 1998).

3.4.6 Modelos de radiação

A energia acústica abandona o tracto vocal pelos lábios. No caso dos sons nasais, parte da energia é também libertada pelas narinas. A pressões normais [5], a radiação das paredes da garganta é geralmente desprezável, excepto para oclusivas sonoras A radiação neste caso foi simulada por Flanagan *et al.* (1975), colocando uma impedância em cada secção do modelo

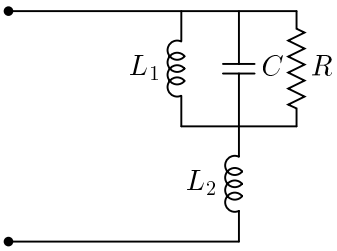
[5] Para voz de mergulhadores esta radiação é também importante devido à menor rigidez das paredes do tracto vocal.

do tracto vocal. Como neste trabalho não abordaremos as oclusivas, não consideramos este efeito.

Nos análogos que representam o tracto como uma linha de transmissão, os lábios e as narinas são representadas por impedâncias de radiação que carregam o tracto vocal e nasal. Estas impedâncias possuem uma parte resistiva e outra reactiva. A primeira, que é responsável pelo consumo de energia, aumenta por um factor superior a w^2 e é portanto um factor preponderante nas larguras de banda das formantes com frequências mais elevadas. A parte reactiva representa a massa efectiva posta em vibração em frente aos lábios e/ou narinas, tornando o comprimento efectivo do tracto superior às suas dimensões físicas.

Um tratamento matemático preciso desta impedância pode ser obtido considerando-a como um pistão vibrante (do Inglês *vibrating piston*), a abertura dos lábios ou narinas, numa esfera, a cabeça (Flanagan, 1972). Este modelo é conhecido por *Piston in Sphere* (PIS) . No entanto este modelo não é computacionalmente eficiente envolvendo o cálculo de séries. Outro modelo mais simples é o de radiação de uma abertura circular num plano infinito. Este modelo é válido, pois a abertura de radiação, nos lábios ou narinas, é muito menor que a cabeça.

a) Modelo SKF



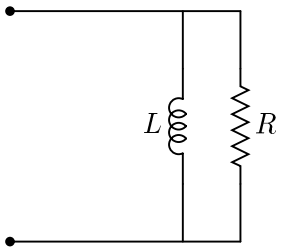
$L_1 = 4.23 \times 10^{-4} \times \sqrt{A_0}$

$R = \frac{45.9}{A_0}$

$C = 1.033 \times 10^{-7} \times A_0^{3/2}$

$L_2 = \frac{7.11 \times 10^{-5}}{\sqrt{A_0}}$

a) Modelo de Flanagan



$R = \frac{128\rho c}{9\pi^2 A_0}$

$L = \frac{8\rho}{3\pi A_0}\sqrt{\frac{A_0}{\pi}}$

Figura 3.11: Modelos de Radiação. (a) Modelo SKF (Stevens *et al.*, 1953). (b) Modelo paralelo de Flanagan (Flanagan, 1972)

Várias simplificações foram propostas para aplicações práticas. Uma delas foi proposta por Stevens, Kasowski e Fant, em 1953, utilizando uma resistência e três outros componentes dependentes da frequência. Este modelo ficou conhecido pelas iniciais dos autores, modelo SKF. Outras propostas, usando elementos dependentes da frequência, foram feitas por Fant (1960) e Wakita e Fant (1978). Flanagan (1972) apresentou uma aproximação ao modelo de radiação num plano infinito, usando uma associação em paralelo de dois componentes independentes da frequência. A Figura 3.11 apresenta o modelo SKF e o modelo de Flanagan (1972).

Como o ouvido humano é sensível às variações de pressão do ar, a pressão sonora a uma distância d dos pontos de radiação é o objectivo final dos cálculos. A pressão sonora a uma distância d, $p_r(t)$, está relacionada com o fluxo radiado, $u_r(t)$. A relação depende da forma da boca e narinas assim como da cabeça do locutor. Fant (1960) usando o domínio da frequência propôs a seguinte relação:

$$\frac{P_r(w)}{U_r(w)} = \frac{\rho\omega}{4\pi d} K_T(w). \tag{3.27}$$

O factor $K_T(w)$ é um ênfase de cerca de 1.5 dB por oitava de 312 a 5000 Hz. Representa dois efeitos: o do reflector (em Inglês *baffle*) e o aumento da resistência de radiação para além da sua proporcionalidade com a frequência. Devido à falta de verificação experimental, $K_T(w)$ é geralmente considerado unitário, sendo a relação, no tempo,

$$p_r(t) = \frac{\rho}{4\pi d} \frac{\partial u_r(t - \frac{d}{c})}{\partial t}, \tag{3.28}$$

geralmente aproximada pela derivada de $u_r(t)$ (Badin e Fant, 1984; Stevens, 1998).

O leitor com interesse em mais detalhes sobre este assunto poderá consultar (Morse, 1948; Morse e Ingard, 1968), (Flanagan, 1972, pág. 36), (Fant, 1960), (Linggard, 1985, pág. 48), (Rabiner e Schafer, 1978, pág. 71).

3.5 Métodos de resolução do modelo acústico

Tendo modelado a configuração dos tractos e os fenómenos acústicos de excitação, propagação e radiação, torna-se necessário obter a informação desejada que consiste, geralmente, no sinal radiado.

Em geral, são usadas três abordagens principais em sintetizadores articulatórios: filtros de onda digitais; resolução das equações diferenciais; método híbrido. Com a disponibilidade crescente de meios poderosos de cálculo, têm sido, em anos mais recentes, tentadas outras técnicas com menos limitações e menos simplificações.

3.5.1 Filtros de onda digitais (*Wave Digital Filters*)

A equação de Webster (equação 3.7 da página 59) no caso de a área se manter constante reduz-se a:

$$\frac{\partial^2 p}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} \tag{3.29}$$

D'Alembert publicou uma solução geral para esta equação em 1747, com a forma (Farlow, 1993),

$$p(x,t) = f(t - x/c) + g(t + x/c). \tag{3.30}$$

As funções $f(.)$ e $g(.)$ são completamente gerais e contínuas, podendo ser interpretadas como ondas de forma arbitrária, mas fixa, que se propagam em direcções opostas ao longo do eixo dos xx com velocidade c. A pressão $p(x,t)$ em qualquer ponto é dada pela soma de dois componentes, uma onda propagando-se para a frente $p^+(x,t)$ (sentido positivo do eixo), e outra para trás $p^-(x,t)$.

A equação 3.29 apenas é válida para uma secção de área constante do tracto, sendo o tracto aproximado pela concatenação de vários tubos de área constante. Na junção de duas secções, com impedâncias diferentes, devido a áreas diferentes, cada onda sofre os efeitos da descontinuidade. Parte da onda $p^+(x,t)$ continua a propagação no mesmo sentido, a parte restante $rp^+(x,t)$ é reflectida, propagando-se no sentido contrário, somando-se a $p^-(x,t)$. O factor r é chamado coeficiente de reflexão, e é definido por:

$$r = \frac{Z_f - Z_t}{Z_f + Z_t} \ , \tag{3.31}$$

onde Z_f é a impedância característica para a frente da junção, e Z_t é a impedância para trás da junção.

O mesmo tipo de considerações aplica-se à onda $p^-(x,t)$ excepto o cálculo de r, onde se tem de trocar os termos Z_f e Z_t devido à direcção de propagação.

A modelação acústica do tracto usando filtros de onda digitais baseia-se neste conceitos. Mais detalhes podem ser obtidos em (Rabiner e Schafer, 1978; Scavone, 1997; Linggard, 1985, por exemplo). Refira-se que nesta técnica é utilizada directamente a função de área, não sendo necessário construir um modelo análogo.

Este modelo foi originalmente proposto por Kelly Jr. e Lochbaum (1962), tendo sofrido, ao longo dos anos, diversos melhoramentos. Foi modificado para incluir efeitos da variação dinâmica da área por Strube (1982). Rubin *et al.* (1981) modificaram o modelo de Kelly e Lochbaum para representar uma terminação não ideal na glote, lábios e narinas. Calcularam os coeficientes de reflexão e a função de transferência no domínio z. Baseando-se na função de transferência, implementaram filtros digitais.

Uma apresentação mais elegante do modelo de Kelly e Lochbaum foi proposta por Fettweis e Meerkötter em 1975. Ficou conhecida por filtros de onda digitais (em Inglês *wave digital filters*) (Fettweis, 1986; Lawson e Mirzai, 1990).

Em geral, esta abordagem é a mais rápida, sendo também adequada a implementações paralelas. Foi mesmo realizado um sistema completo em tempo real, usando *hardware* especial por Meyer *et al.* (1989).

Os maiores problemas desta abordagem são: a dificuldade de modelar as perdas dependentes da frequência; dificuldade de inclusão da interação entre a fonte glotal e o tracto; a dificuldade em ter um comprimento do tracto arbitrário.

O problema do comprimento pode ser combatido utilizando variação da frequência de amostragem (Wright e Owens, 1993) ou *fractional delay wave digital filters* (Välimäki, 1995; Välimäki *et al.*, 1994).

No que respeita à inclusão das perdas, alguns progressos foram conseguidos nos últimos anos (Story, 1995; Meyer *et al.*, 1989; Nagai, 1990; Story, 1995), mas ainda é necessária mais investigação.

3.5.2 Métodos no tempo

Nestes métodos, começa-se por discretizar espacialmente o tracto, constrói-se de seguida um circuito equivalente (utilizando os modelos anteriormente descritos), sendo as equações diferenciais parciais que relacionam a pressão e o fluxo (ou os seus análogos eléctricos), discretizadas no tempo. O conjunto de equações diferença obtido é depois resolvido para cada instante de tempo, por forma a se obter a pressão e fluxo em cada ponto da linha de transmissão (Flanagan e Cherry, 1969; Flanagan e Landgraf, 1968; Flanagan *et al.*, 1975; Flanagan e Ishizaka, 1976; Flanagan *et al.*, 1980). Os valores da pressão e fluxo, num instante no tempo, são usados para calcular parâmetros do circuito equivalente, a utilizar nos cálculos para o instante seguinte. Esta abordagem foi designada por resolução no tempo (em Inglês *time-domain*). A Figura 3.12 representa esquematicamente esta técnica.

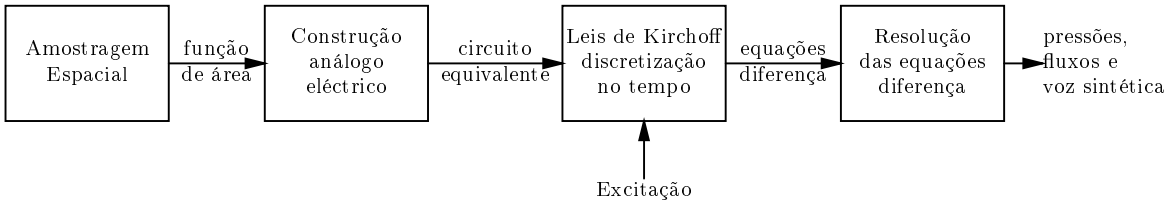


Figura 3.12: Síntese articulatória no domínio do tempo.

Nestes sintetizadores é usada uma frequência de amostragem bastante elevada, para evitar *frequency-warping* (Wakita e Fant, 1978). Os componentes dependentes da frequência são simulados a uma frequência fixa. Os efeitos desta aproximação foram estudados por Hsieh Hsieh (1994). Apesar destas aproximações, o som obtido é natural.

O modelo inicial de Flanagan foi simplificado por Maeda (1982a), substituindo o modelo mecânico vibrante das cordas vogais por um modelo representando a área de abertura da glote; não incluindo as fontes de ruído e omitindo os efeitos dos seios nasais. Estas simplificações tornaram as simulações muito mais rápidas. Outras simplificações foram introduzidas por Bocchieri e Childers (1984), reduzindo o número de fontes de ruído e usando um terminal gráfico para desenhar o contorno do tracto vocal. Baseado no trabalho de Maeda, foi desenvolvido um sintetizador por Childers e Ding (1991); Ding (1990), usando um circuito equivalente e

convertendo as equações acústicas em equações algébricas lineares. Hsieh (1994) rederivou as equações para incluir o sistema subglotal, a impedância glotal, o ruído de turbulência e os seios paranasais.

Um dos exemplos mais completos de aplicação desta técnica é o sintetizador desenvolvido por Boersma (1998).

3.5.3 Métodos híbridos

Este método, proposto por Sondhi e Schroeter (1987), e representado de forma muito resumida na Figura 3.13, difere dos dois anteriores, ao utilizar o domínio da frequência para modelar as cavidades supraglotais. Enquanto a glote é modelada no domínio do tempo, devido à sua natureza altamente não linear, o tracto vocal e o tracto nasal são modelados na frequência, aproveitando o facto para modelar, de forma mais precisa, as perdas e a radiação, fenómenos que, como já foi referido, dependem da frequência. Os dois modelos, da fonte e tracto, são interligados, na proposta inicial, através da transformada inversa de Fourier e convolução. A utilização de informação acerca da impedância de entrada do tracto permite a realização de sistemas com interacção entre a fonte glotal e o tracto. A designação de método híbrido resulta da utilização simultânea do domínio tempo e domínio frequência. Informações complementares acerca deste método podem ser encontradas no capítulo 4.

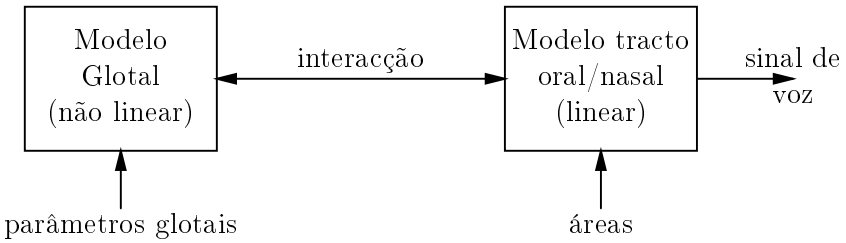


Figura 3.13: Descrição geral do método híbrido de síntese articulatória, adaptado de Sondhi e Schroeter (1987).

Como a utilização de transformada inversa e convolução requerem um tempo considerável, foi proposto por Lin (1990) a aproximação da resposta na frequência por um conjunto de filtros de segunda ordem, substituindo-se a convolução pelo processo de filtragem. Este método, apesar de mais rápido, necessita da obtenção dos filtros de segunda ordem, tarefa bastante complicada no caso de existência de antiformantes.

3.5.4 Outras técnicas

A aplicação de métodos variacionais à equação de Webster e às condições fronteira do tracto foi proposta por van Praag (1997) e Jospa e Praag (1999). Esta técnica permite tratar configurações com diversos ramos, com aplicação na simulação de sons envolvendo uma cavidade nasal incluindo a assimetria entre as duas passagens que terminam nas narinas. Permite

também a obtenção de pólos e zeros da função de transferência o que poderá ser útil para nasais.

Métodos geralmente aplicados na simulação de sistemas electromagnéticos como o *Transmission Line Model* ou *Transmission Line Matrix* (TLM) têm também sido aplicados aos modelos acústicos do tracto (El-Masri *et al.*, 1996).

O acesso, recentemente, a poderosos meios computacionais tem permitido a utilização de métodos usando elementos finitos na resolução da equação de Navier Stokes (Thomas, 1986; Richard *et al.*, 1995; Slimon *et al.*, 1996; Sinder *et al.*, 1997). Este tipo de simulações é especialmente interessante para sons em que a propagação se torna turbulenta, como as fricativas. Servem também os resultados deste método para validação de modelos mais simplificados. De facto, resultados para vogais mostram como geralmente válidas as aproximações habitualmente utilizadas (Sinder *et al.*, 1997).

Os métodos usando elementos finitos e o TLM permitem a utilização de informação tridimensional disponibilizada por técnicas como MRI.

3.6 Fontes de excitação

3.6.1 Fonte de excitação glotal

Existem três tipos de modelos (Cummings e Clements, 1995): modelos glotais paramétricos não-interactivos, em que não existe interacção entre a fonte glotal e o tracto vocal; modelos glotais mecânicos e paramétricos interactivos, que incluem, implícita ou explicitamente, a interacção entre a fonte e o tracto vocal; e modelos glotais fisiológicos, baseados em teorias de comportamento fisiológico das cordas vocais.

Modelos paramétricos

São os mais simples pois assumem que a fonte e o tracto vocal são separáveis não existindo iteração entre os dois. Baseiam-se pois na teoria fonte-filtro, proposta por Fant (1960). São muito usados em codificação e em síntese acústica de voz.

O modelo trigonométrico (Rosemberg, 1971), representado na Figura 3.14, é definido por:

$$u_g(t) = \begin{cases} \frac{\alpha}{2}\left(1 - \cos\left(\frac{t\pi}{T_P}\right)\right) & \text{para } 0 \leq t \leq T_P \\ \alpha \cos\left(\frac{\pi}{2}\frac{t-T_P}{T_N}\right) & \text{para } T_P \leq t \leq T_P + T_N \end{cases} . \tag{3.32}$$

Como o efeito da radiação pode ser modelado usando a primeira derivada, pode modelar-se a derivada do fluxo glotal. O modelo mais utilizado é o modelo LF, proposto por Liljencrants e Fant (Fant *et al.*, 1985b). A sua popularidade deve-se a diversos factores, dos quais ressalta

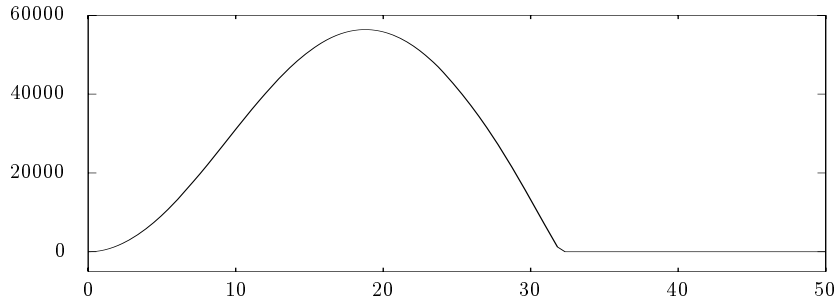


Figura 3.14: Modelo do fluxo glotal trigonométrico de Rosemberg (1971). $\alpha = 56383.798$, $T_P = 18.784$ e $T_N = 13.216$, valores de (Cummings e Clements, 1995, Fig. 5).

a facilidade em obter os seus parâmetros. A forma do modelo é

$$u_g^{'}(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g(t)) & \text{para } 0 \leq t \leq T_e \\ -\left(\frac{E_e}{\epsilon T_a}\right)\left(e^{-\epsilon(t-T_e)} - e^{-\epsilon(T_c-T_e)}\right) & \text{para } T_e \leq t \leq T_P + T_c \end{cases} \tag{3.33}$$

A Figura 3.15 mostra um período do modelo LF. O modelo LF original tem 5 parâmetros: E_0, α, ω_g, T_a e F_0.

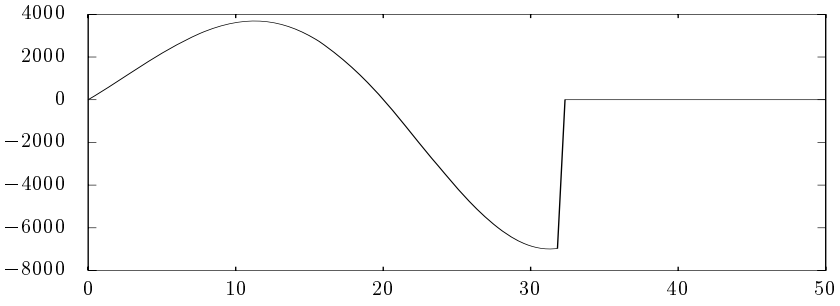


Figura 3.15: Modelo LF do fluxo glotal (Fant *et al.*, 1985b). $E_0 = 2622.799$, $\alpha = 0.032$, $f_g = 0.025$, $T_e = 32.0$, $\epsilon = 2.0$, valores de (Cummings e Clements, 1995, Fig. 6).

Modelos glotais mecânicos e paramétricos interactivos

Exemplo de um modelo paramétrico interactivo é o modelo proposto por Allen e Strong (1985). Este modelo parametriza a área utilizando uma fórmula proposta por Titze: A área glotal, $A_g(\theta)$, é calculada segundo

$$A_g(\theta) = \begin{cases} A\left(\left(\frac{\theta}{\theta_m}\right)^{-\theta_m \cot\theta_m}\left(\frac{\sin\theta}{\sin\theta_m}\right)\right)^{\beta} & \text{para } \theta \leq \pi \\ 0 & \text{para } \theta \geq \pi \end{cases}, \tag{3.34}$$

onde $\theta = \frac{\pi t}{\gamma T}$; $\theta_m = \frac{\pi \delta}{(1+\delta)}$; A é a área glotal máxima; T o período; γ o quociente de velocidade; δ a simetria da forma de onda; e β o declive.

Exemplos de outros modelos interactivos são: o modelo de uma massa (Flanagan e Landgraf, 1968); o modelo analítico desenvolvido para estudo da interacção fonte-tracto por Anantha-padmanabha e Fant (1982); o modelo utilizando parametrização da condutância de Rothemberg (1981).

O modelo de duas massas, desenvolvido por Ishizaka e Flanagan (1972) é, o mais utilizado dos modelos interactivos mecânicos. O diagrama deste modelo encontra-se na Figura 3.16. O movimento das duas massas, m_1 e m_2, é controlado pelas forças aerodinâmicas e pelas forças mioelásticas, representadas por molas e amortecedores.

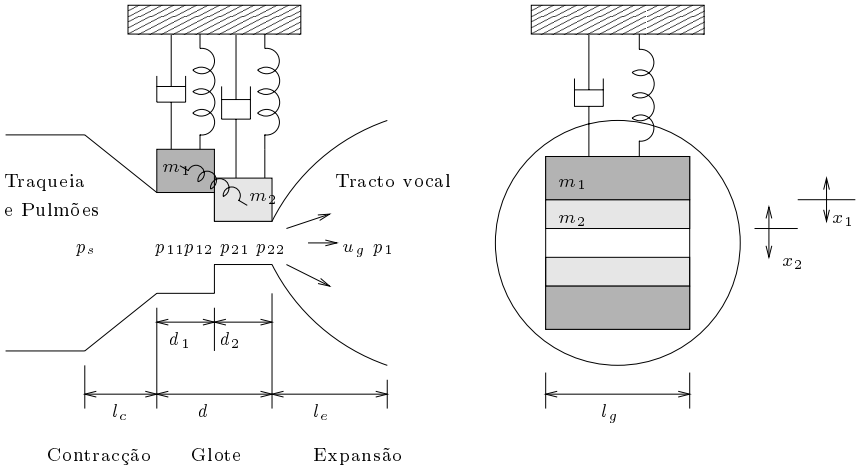


Figura 3.16: Modelo de duas massas das cordas vocais (Ishizaka e Flanagan, 1972).

As equações que controlam os movimentos são:

$$m_1 \frac{d^2 x_1}{dt^2} + r_1 \frac{dx_1}{dt} + k_1 x_1 + k_c(x_1 - x_2) + F_1 = 0 \tag{3.35}$$

$$m_2 \frac{d^2 x_2}{dt^2} + r_2 \frac{dx_2}{dt} + k_2 x_2 + k_c(x_2 - x_1) + F_2 = 0 \tag{3.36}$$

onde x_i representa o deslocamento lateral das massas, F_i representa as forças aerodinâmicas exercidas em cada massa, r_i a resistência devida à viscosidade, $i = 1$ para a massa inferior, e $i = 2$ para a massa superior. No modelo as molas possuem características não lineares. Durante a fase em que a glote se encontra fechada existe uma força de contacto. O valor da frequência fundamental neste modelo é controlado por um parâmetro, Q, representando a tensão das cordas.

O circuito acústico equivalente encontra-se na Figura 3.17. R_c representa a contracção abrupta à entrada; Rv_1 e Rv_2 representam as perdas por viscosidade no bordo inferior e superior das cordas, respectivamente; R_{12} representa a variação da energia cinética por unidade de volume

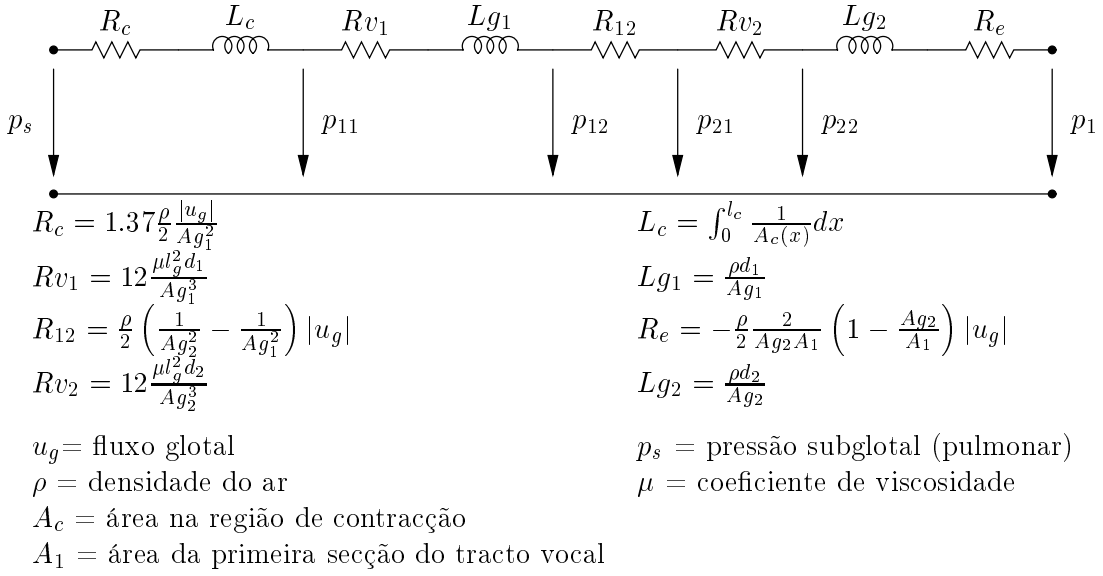


Figura 3.17: Circuito equivalente do modelo de duas massas (Flanagan *et al.*, 1975).

na junção das duas massas; R_e a expansão; L_c, Lg_1 e Lg_2 as inertâncias das massas de ar nas três zonas. Os componentes são função da área da primeira secção do tracto vocal, $A1(t)$, e das áreas de abertura glotal de cada uma das massas, $Ag_1(t)$ e $Ag_2(t)$, obtidas com base nos deslocamentos laterais, relativamente a uma posição de repouso, através de

$$Ag_1(t) = Ag_{0_1} + 2l_g x_1(t) \quad (3.37)$$

$$Ag_2(t) = Ag_{0_2} + 2l_g x_2(t)\ , \quad (3.38)$$

onde as áreas de repouso, Ag_{0_1} e Ag_{0_2}, são geralmente iguais. A impedância total pode ser representada por uma indutância, $L_g(t)$, em série com uma resistência, $R_g(t)$. A interacção entre o modelo glotal e o tracto é feita através da pressão supraglotal $p_1(t)$.

Modelos glotais fisiológicos

Os modelos desta categoria são geralmente utilizados em aplicações em que é necessário grande precisão, pois são muito exigentes computacionalmente (Cummings e Clements, 1995, pág. 31). Os modelos mais conhecidos são os desenvolvidos por Titze (1989). Outro modelo foi proposto por Hegerl e Höge (1991) baseado na resolução numérica da equação de Navier-Stokes.

3.6.2 Modelo de fonte de ruído

Uma área suficientemente reduzida provoca a mudança de fluxo laminar para um regime turbulento. Para fluxo do ar turbulento, a analogia eléctrica não existe. No entanto, Stevens,

Kasowski, e Fant, em 1953, mostraram que inserindo um gerador de ruído no ponto da constrição a analogia pode ser mantida. A potência e a resistência interna da fonte depende da área da constrição e do fluxo (Flanagan, 1972). A inserção de uma fonte de ruído filtrado, no modelo baseado na analogia com linhas de transmissão, produziu resultados muito aceitáveis, para algumas fricativas pelo menos, apesar de ser apenas uma simples aproximação de um fenómeno complexo e não linear.

Uma abordagem baseada na teoria aeroacústica permitiu a Sinder (1999) não considerar o fenómeno de excitação em sons fricativos como algo de separado da propagação e radiação.

Não sendo neste trabalho abordada a síntese de sons necessitando deste tipo de excitação, não entrarei em mais detalhes. Estes podem ser obtidos em: (Riegelsberger, 1997, pág. 42 e seguintes), (Flanagan, 1972, pág. 53), (Childers, 2000, pág. 413) e (Stevens, 1971).

3.7 Obtenção da posição dos articuladores

> There is a set of fundamental problems the solutions to which require the determination of vocal tract shape from the acoustic parameters of the speech signal.
>
> Victor N. Sorokin (1992)

Os modelos descritos permitem definir configurações do aparelho de produção e obter os sons correspondentes, designado por problema directo. Iremos agora tratar do problema inverso, obter a posição dos articuladores com base no sinal de voz.

Desde a década de 1950, quando se efectuaram as primeiras simulações em computador de produção de voz, que os investigadores desejam conhecer a forma do tracto vocal ao longo do tempo. Este conhecimento é muito importante para os Linguistas, tendo também aplicações em Engenharia e Medicina. Em Engenharia oferece possibilidades de melhorar a codificação, reconhecimento e síntese de fala. No entanto, um processo de análise para obter a área é necessário para tornar essas possibilidades realidade.

Dados sobre o tracto vocal são essenciais. A teoria acústica de produção de voz (Fant, 1960) considera o tracto como um tubo com área variável. Existem basicamente dois tipos de métodos para obter a área: medição directa ou estimação com base no sinal acústico.

3.7.1 Medição directa

Existem técnicas que permitem obter imagens de zonas completas do aparelho de produção humano, outras que apenas possibilitam seguir o comportamento no tempo de alguns pontos e, ainda, técnicas que permitem medição de comportamentos complexos como a área de contacto da língua com o palato (Hardcastle e Laver, 1996, capítulo 1).

Os métodos directos baseados em raios-X, utilizados já por investigadores como Chiba e

Kajiyama (1958) e Fant (1960), constituem ainda fonte de dados para muitos sintetizadores. Infelizmente esta técnica é laboriosa, necessita de dosagens elevadas, e apenas nos dá imagens bidimensionais. Em especial, pelo risco que a radiação representa, raramente é empregue actualmente.

Mais recentemente, começaram a usar-se técnicas como a Tomografia Axial Computorizada e MRI (Baer *et al.*, 1991). A Ressonância Magnética não sofre dos problemas dos raios-X, parece ser a melhor técnica para recolha dos dados necessários. As desvantagens advêm do facto de problemas de resolução e dificuldades com estruturas calcificadas com pouco hidrogénio móvel que não se distinguem. As imagens obtidas por esta técnica têm sido essencialmente de configurações estáticas do tracto. Foram efectuadas medições para vogais (Yang, 1999; Story *et al.*, 1995), de fricativas (Narayanan *et al.*, 1995; Story *et al.*, 1995), liquidas (Narayanan *et al.*, 1997a), laterais (Narayanan *et al.*, 1997b), do tracto nasal (Dang e Honda, 1994; Story, 1995) e da posição do velo (Demolin *et al.*, 1998). Desenvolvimentos recentes, dos equipamentos associados a técnicas de processamento, permitem obter imagens dinâmicas (Shadle *et al.*, 1999). Imagens obtidas utilizando endoscopia podem também fornecer informação útil (Esling *et al.*, 1998).

Não basta ter uma técnica de obtenção de imagens do tracto, é preciso extrair delas, de preferência de forma automática, informação acerca da configuração. Diversos trabalhos têm-se dedicado a esta difícil tarefa. Constituem alguns exemplos o processamento automático de imagens de raios X por Thimm e Luettin (1999) e a análise de imagens de ressonância magnética (MRI) por Soquet *et al.* (1998).

Técnicas como a articulografia electromagnética, designada por *ElectroMagnetic Articulography* (EMA) ou *ElectroMagnetic Midsagittal Articulography* (EMMA), aparecendo o segundo M para explicitar que as imagens são no plano sagital médio (em Inglês *Midsagittal*), permitem obter informação acerca da posição de um conjunto de pontos ao longo do tempo (Perkell *et al.*, 1992, 1993; Schönle *et al.*, 1987; Hoole e Nguyen, 1997). Têm sido utilizadas para estudar o comportamento da língua, lábios e, em alguns casos, do velo (Engelke *et al.*, 1996; Wrench, 1999). Os sistemas actualmente em funcionamento apenas permitem obtenção de dados acerca de pontos situados num mesmo plano sagital, limitação que será corrigida pelos novos sistemas em fase final de desenvolvimento (Zierdt *et al.*, 1999; Horn *et al.*, 1999).

Para a medição de certos articuladores foram desenvolvidos métodos especiais. Exemplo deste caso é o *velotrace* desenvolvido para medir directamente a posição do velo (Bell-Berti *et al.*, 1993; Horiguchi e Bell-Berti, 1987).

Uma técnica recente que promete ser da máxima utilidade é a proposta por Burnett *et al.* (1999), baseada na utilização de um radar de baixa potência. Esta técnica pode ser utilizada para obter informação, tanto acerca da área de abertura glotal, como acerca da posição da língua durante a produção de voz.

3.7.2 Métodos indirectos

O mapeamento do domínio acústico para o domínio articulatório consiste em estimar a forma do tracto vocal, usando apenas o sinal acústico. Nesta secção apresenta-se um resumo da investigação efectuada no passado e no presente.

As várias abordagens podem agrupar-se nas seguintes classes: métodos analíticos, métodos de procura em tabelas, redes neuronais e métodos de optimização usando realimentação.

Uma das maiores dificuldades do mapeamento acústico-articulatório é a presença de múltiplas soluções. Este facto foi já bem documentado teórica e empiricamente (Flanagan, 1972; Fant, 1960).

3.7.2.1 Métodos analíticos

Vários investigadores propuseram métodos analíticos para obter a função de área com base no sinal acústico. As técnicas baseiam-se nos coeficientes de Predição Linear (LPC) ou resposta impulsional do tubo. A abordagem LPC deriva as áreas dos coeficientes de reflexão obtidos por filtragem inversa (Wakita, 1973, 1979). O grande problema desta abordagem é o "efeito de ventríloquo", diferentes formas podem produzir as mesmas formantes (Schroeder, 1967; Mermelstein, 1967; Atal *et al.*, 1978; Schroeder *et al.*, 1979; Charpentier, 1984).

O método utilizando a resposta impulsional nos lábios, designado por *Lips Impulse Response* (LIR), assenta no facto de que se a função de transferência do tracto vocal é conhecida então $A(x)$ pode ser obtida (Schroeder, 1967; Gopinath e Sondhi, 1970; Sondhi, 1979; Sondhi e Resnick, 1983).

O LIR possui duas limitações principais: é necessário saber as condições de fronteira na glote; para a estimação dos valores principais (*eigenvalues*) é necessário usar um período de tempo longo (10 a 20 ms), perdendo-se a estacionaridade. O *Linear Prediction Acoustic Tube* (LPAT) tem diversos problemas: incerteza na fonte; a função de área obtida não é única; exclui nasais e sons surdos; a condição de fronteira, não considerando a carga de radiação, é irrealista; não considera perdas; obriga a estimar correctamente o comprimento do tracto vocal.

3.7.2.2 Métodos de procura em tabelas

Faz-se a amostragem dos parâmetros articulatórios e constroem-se tabelas (*codebooks*) com as formas do tracto vocal e as respectivas representações acústicas (Atal *et al.*, 1978). Schroeter *et al.* (1990) fez alguns melhoramentos para a geração dos *codebooks* e utilizou procura por programação dinâmica. Muitos autores se dedicaram, e dedicam, ao melhoramento do processo de geração das tabelas (Ouni e Laprie, 1999, por exemplo) e aos métodos de procura (Silva *et al.*, 1999, exemplo recente).

Esta técnica tem diversos problemas: carga computacional; sensibilidade à fonte de excitação; ambiguidade no mapeamento; limitações do modelo acústico (Schroeter e Sondhi, 1994). Estas

técnicas são geralmente adequadas para derivar configurações iniciais dos articuladores para processos de inversão baseados em optimização.

3.7.2.3 Redes neuronais

Uma abordagem, mais recente, consiste na aplicação de redes neuronais. A rede é treinada com um conjunto grande de parâmetros acústicos e articulatórios. Um padrão acústico é depois utilizado para obter o padrão articulatório correspondente (Xue *et al.*, 1990; Båvegård *et al.*, 1993; Rahim *et al.*, 1993). No entanto, o processo de treino continua um desafio (Xue *et al.*, 1990) e não existem ainda vantagens claras destas técnicas em relação a outras (Schroeter e Sondhi, 1994). A capacidade actual das redes neuronais é a de fornecer valores iniciais para os parâmetros articulatórios, para um pequeno conjunto de treino. Têm sido utilizados vários tipos de redes.

3.7.2.4 Métodos de optimização usando realimentação

Tentativas para uma resolução analítica do problema produziram resultados insatisfatórios (Wakita, 1979; Schroeder, 1967). As soluções podem não ser únicas, devido ao sinal ser limitado em frequência e a assunção de ondas planas não ser válida a frequências elevadas (Sondhi, 1979). Têm portanto de se utilizar métodos numéricos. Os processos convencionais de optimização são processos iterativos, tipicamente utilizando alguma forma de procura por gradiente. Técnicas não usando gradiente, como o *simulated annealing* (Hsieh, 1994) e algoritmos genéticos (McGowan, 1994) são também usadas.

A optimização pode ser feita em apenas uma *frame* temporal, ao longo de várias (Gupta e Schroeter, 1993), ao longo de uma trajectória parametrizada (Parthasarathy e Coker, 1992), ou em termos de configurações alvo (em Inglês *targets*) (McGowan, 1994).

Nos métodos utilizando apenas uma *frame*, o sinal é dividido em secções de 5 a 40 ms, onde se pode considerar o sinal como estacionário. Esta forma, conveniente, é a usada na maioria dos trabalhos publicados. A configuração do tracto vocal é estimada independentemente para cada *frame*. Não se utiliza o facto de a configuração do tracto vocal variar lentamente ao longo do tempo.

Se os parâmetros forem estimados conjuntamente para várias *frames*, a correlação entre elas pode ser explorada através de restrições ou parametrização das trajectórias dos articuladores. As trajectórias podem representar o tracto vocal ao longo de muitas *frames* eficientemente e podem aliviar o problema do mapeamento não ser unívoco (Shirai e Kobayashi, 1986; Parthasarathy e Coker, 1992). O problema é que a eficácia dos processos de optimização diminui com o aumento do número de parâmetros a optimizar, a chamada "praga da dimensionalidade".

Como alternativa, o movimento dos articuladores pode ser representado como um sistema dinâmico de que se estimam as entradas. Desta forma, o movimento pode ser representado por alvos, facilmente relacionados com fonemas, ou *gestures* (Parthasarathy e Coker, 1992; McGowan, 1994).

As representações da entrada e saída influenciam os resultados da optimização através do significado das propriedades, a dimensão das propriedades, a métrica escolhida para o erro e as restrições (em Inglês *constraints*). Devido à existência de várias soluções e para evitar mínimos locais utilizam-se uma variedade de restrições, técnicas de inicialização e regularização.

Têm sido empregues processos de optimização como: algoritmo de Hookes e Jeeves (Flanagan *et al.*, 1980; Parthasarathy e Coker, 1990, 1992; Gupta e Schroeter, 1991, 1993); algoritmo de gradiente óptimo (Levinson e Schmidt, 1983); combinações do método de Fletcher-Reeves e aproximações sucessivas (Prado, 1991; Prado *et al.*, 1992); métodos estocásticos como algoritmos genéticos (McGowan, 1994) e *simulated annealing* (Hsieh, 1994).

Muitas representações foram propostas para representar a configuração do tracto de uma forma mais eficiente do que utilizando directamente a função de área. Foi, por exemplo, utilizada a decomposição em série de Fourier da forma da língua e da função de área (Yehia e Itakura, 1994; Mermelstein, 1967). Os modelos articulatórios sagitais de Mermelstein (1973) e Coker (1976) são os mais populares (Prado, 1991; Prado *et al.*, 1992; Hsieh, 1994), mas muitos outros são utilizados. Mesmo os modelos paramétricos de área continuam a ser utilizados (Båvegård, 1996).

Muitas representações do domínio acústico e métricas foram utilizadas. A distância Euclidiana entre as primeiras 3 a 5 frequências é muito comum (Schroeder, 1967; McGowan, 1994; Mermelstein, 1967; Charpentier, 1984; Prado, 1991; Prado *et al.*, 1992), distâncias espectrais lineares e logarítmicas (Flanagan *et al.*, 1980; Levinson e Schmidt, 1983), distâncias cepstrais (Shirai e Kobayashi, 1986; Parthasarathy e Coker, 1992; Meyer *et al.*, 1991; Gupta e Schroeter, 1991). Outras representações incluem distância Euclidiana entre o logaritmo das frequências das formantes, algumas vezes complementados com a amplitude das formantes (Atal *et al.*, 1978; Sorokin, 1992; Charpentier, 1984) e distâncias LPC. Recentemente, foram utilizadas medidas de erro utilizando informação perceptual, como a distância das 3 primeiras formantes em Barks (Båvegård e Fant, 1995). Alguns investigadores propuseram critérios de erro múltiplos (Parthasarathy e Coker, 1992; Gupta e Schroeter, 1991).

Nem sempre adicionar mais informação acústica conduz a melhores resultados, como Sorokin descobriu, ao utilizar os logaritmos das primeiras 3 e 4 frequências, com piores resultados no segundo caso (Sorokin, 1992).

3.7.2.5 Mapeamento acústico-articulatório de consoantes e sons nasais

Geralmente, apenas se tem abordado a inversão de sons sonoros orais não obstruentes. Pouco trabalho foi feito para consoantes e sons nasais.

Sorokin (1994) e Shirai abordaram a inversão de fricativas surdas com resultados razoáveis. O caso das fricativas sonoras foi abordado, recentemente, por Riegelsberger (1995, 1997). A obtenção do lugar de constrição para oclusivas foi investigada por Galván-Rodríguez (1997).

Em relação aos sons nasais é também reduzido o número de trabalhos, estando o problema

longe de resolvido. Foi estudada a obtenção da posição do velo utilizando uma rede neuronal treinada com dados de articulografia electromagnética (EMMA) por Richmond (1999). Rossato efectuou experiências relacionadas com a inversão de vogais nasais, mas não utilizou voz natural, apenas propriedades acústicas derivadas do próprio modelo articulatório usado (Rossato *et al.*, 1998; Rossato e Feng, 1999).

3.8 Aplicações

Têm sido desenvolvidos ao longo dos anos vários sintetizadores articulatórios. A variedade de técnicas, aplicações a que se destinam e limitações é muito grande.

Apesar do estado de desenvolvimento deste tipo de sintetizadores não os tornar ainda utilizáveis em sistemas comerciais de conversão de texto em fala existem já primeiros protótipos deste tipo (Coker, 1997; Coker *et al.*, 1973a; Parthasarathy e Coker, 1992; Coker *et al.*, 1973b; Coker, 1967; Scully, 1987).

Muitos dos sintetizadores articulatórios têm sido utilizados em estudos de Fonética e Fonologia. Sem o acesso a esta ferramenta teria sido muito difícil, ou mesmo impossível, o aparecimento e desenvolvimento de algumas destas teorias que tanto têm contribuído para o aumento de conhecimento acerca dos processos de produção e percepção de voz. Têm sido propostas, nos últimos anos, teorias fonológicas baseadas total ou parcialmente na utilização de descrições articulatórias. Exemplos destas teorias são a Fonologia Articulatória (Browman e Goldstein, 1995, 1992, 1990, 1989), desenvolvida fazendo uso do sintetizador *Configurable Articulatory Synthesizer* (CASY) dos laboratórios Haskins Rubin *et al.* (1981) e a Fonologia Funcional proposta por Boersma (1998) que levou o autor a desenvolver um sintetizador articulatório muito completo.

Outra área de aplicação de interesse é a sua utilização em codificação de voz. O reduzido número de articuladores, aliado à lenta variação no tempo das suas posições, torna-os candidatos ideais para codificação. O grande problema reside na obtenção automática dos articuladores apenas com base no sinal acústico, problema ainda não completamente resolvido. Exemplo deste tipo de aplicação é o projecto de desenvolvimento de um *voice mimic* por J. Flanagan e colaboradores (Silva *et al.*, 1999; Chennoukh *et al.*, 1997; Zussa *et al.*, 1995; Zussa, 1995).

Tem, também, sido estudada a forma de tornar um sintetizador articulatório capaz de aprender a falar, imitando o processo de aprendizagem de uma criança (Bailly *et al.*, 1997)

Embora não relacionado com a voz de uma forma directa, as técnicas referentes ao modelamento acústico do tracto são úteis no modelamento de instrumentos musicais de sopro (Scavone, 1997, por exemplo). Neste tipo de aplicação, as técnicas baseadas em filtros de onda digitais são geralmente as utilizadas, devido às paredes rígidas dos instrumentos musicais.

Capítulo 4

Sintetizador articulatório

> The use of such a [articulatory] synthesizer has much to commend it in phonetic studies
>
> Franklin S. Cooper (1961)

Um sintetizador articulatório vocacionado para os sons nasais precisa modelar pelo menos três aspectos do processo de produção de voz. O primeiro deles é a modelação da geometria das cavidades, acima da glote, que contribuem para as características dos sons. O segundo aspecto prende-se com a necessidade de modelar a propagação das ondas sonoras nas cavidades. A necessidade de modelar as cavidades nasais coloca problemas adicionais devido à sua forma complexa. O terceiro aspecto a modelar é a fonte de excitação glotal. Estes três aspectos constituem as três primeiras secções deste capítulo. O processo completo de obtenção do sinal de voz, com base na informação dos vários modelos, é descrito na quarta secção.

Apenas podendo obter valores para os parâmetros articulatórios se torna útil o sintetizador. A obtenção da posição dos articuladores a partir do sinal de voz é abordada na quinta secção deste capítulo. Apresentam-se, nessa secção, alguns exemplos de configurações obtidas para vogais orais.

Na sexta secção, apresenta-se informação acerca da implementação prática do sintetizador.

No final do capítulo, sétima secção, resumem-se as características do sintetizador desenvolvido.

4.1 Modelamento da geometria dos tractos

Para sons não nasais apenas se tem de considerar o tubo de área variável entre a glote e os lábios, designado por tracto vocal. Para sons nasais temos de considerar também o tracto nasal. As dimensões destes dois tubos são da mesma ordem de grandeza tornando-se possível aplicar técnicas semelhantes no seu modelamento acústico. O tracto nasal é essencialmente constante, com a excepção da zona do véu palatino. O tracto vocal varia continuamente e a sua forma tem de ser especificada em intervalos não maiores que alguns milisegundos (Schroeter e Sondhi, 1992).

4.1.1 Modelamento do tracto vocal

A geometria do tracto é convenientemente descrita em termos das posições dos articuladores: a língua, lábios, maxilar, etc. Optamos, por essa razão, por utilizar um modelo articulatório para a representação do tracto vocal. Dos vários modelos existentes optamos por um modelo derivado do originalmente proposto por Mermelstein (1973). Este modelo, sagital geométrico, foi desenvolvido no *Mind Machine Interaction Research Center* (MMIRC) da *Univerity of Florida* por D. Childers e colegas (Prado, 1991; Hsieh, 1994). As modificações introduzidas, por estes investigados, no modelo originalmente proposto consistiram em: (1) melhoramento da representação da parte mais baixa da faringe; (2) e da região entre o ápice da língua e o maxilar. A parte mais baixa da faringe passou a ter parâmetros ajustáveis. O ápice da língua passou a ser definido de uma forma mais independente do corpo da língua e do maxilar. A adopção deste modelo deveu-se essencialmente a termos tido acesso a informação necessária à sua implementação.

De seguida, fazemos uma breve descrição do modelo implementado. A descrição com mais detalhe do modelo articulatório pode ser encontrada em (Branco, 1997) e nas fontes originais (Hsieh, 1994; Prado, 1991; Mermelstein, 1973).

4.1.1.1 Modelo articulatório utilizado

O modelo, apresentado na Figura 4.1, é constituído por 3 partes distintas: uma parte fixa, uma parte ajustável e a parte variável definida pela posição dos articuladores.

Manteremos na descrição a denominação original, proveniente do Inglês, dos pontos e dos parâmetros articulatórios.

Constituem a parte fixa: o ponto fixo F sobre o qual roda o maxilar; a parede posterior da faringe (pontos G, G1, G2 e W); a parte do palato duro (entre N e M) e incisivos superiores (ponto U). O contorno posterior-superior é fixo, excepto para a zona do palato mole, representada pelo arco M-V'. É também fixo o ponto mais elevado do velo, fechando a passagem para o tracto nasal (ponto V) e a inclinação da recta ao longo da qual se desloca a extremidade do velo. A distância entre o maxilar e o ponto fixo F, designada por sj, é também mantida fixa,

assim como o raio do arco de circunferência utilizada na representação do corpo da língua.

A parte inferior da faringe pode ter as suas dimensões alteradas variando-se 3 parâmetros. São eles:

- A distância, na horizontal, do ponto H, representando a intersecção da parte anterior da epiglote com a parte superior do osso hióide, á parede posterior da faringe, denominada **wh**;
- A distância, na vertical, do mesmo ponto H à posição da glote, representada por **hk1**;
- A distância, na horizontal, entre os pontos K e G1, representada por **g1k**. O ponto K representa uma estimativa da posição da extremidade anterior da laringe.

Estes parâmetros na nossa implementação comportam-se como 3 parâmetros articulatórios adicionais.

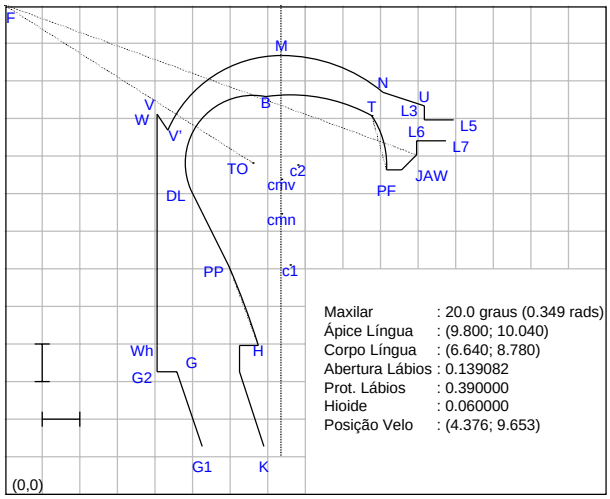


Figura 4.1: Modelo Articulatório implementado. Apresentam-se identificados os pontos necessários à sua construção. Os quadrados têm 1 *cm* de lado.

O resto do modelo depende das posições dos articuladores, definidas pelos parâmetros articulatórios. Os articuladores contribuem para a definição de pontos utilizados no modelo da seguinte forma:

- O corpo da língua é representado pelo arco de circunferência DL-B com o centro móvel e raio fixo. As coordenadas rectangulares do centro (**tbodx**, **tbody**) [1] constituem parâmetros articulatórios;
- O ápice da língua é representado pelas coordenadas rectangulares (**ttx**, **tty**) [2], do ponto T. Os arcos B-T e T-PF representam o contorno. Como o ponto B varia com a posição

[1] tbod do Inglês *tongue body*
[2] tt do Inglês *tongue tip*

do centro da língua (tbodc) e o ângulo do maxilar (jaw), a zona definida pela ponta da língua é afectada pela posição destes dois outros parâmetros articulatórios;

- O maxilar é representado, em coordenadas polares, por (sj,thetaj). Como já referido a distância sj é mantida constante. O parâmetro articulatório **jaw** é igual ao ângulo thetaj. A zona junto ao ponto que define o maxilar é aproximada por uma sequência de segmentos de recta PF-PS-JAW-L6.

- Os lábios são representados pelos pontos L5 (lábio superior) e L7 (lábio inferior). Relativamente ao ponto **jaw**, as coordenadas do lábio inferior são representadas por (**lipp**,**lipo**) [3] que representam, respectivamente, a protrusão e abertura dos lábios. A utilização destes dois parâmetros como variáveis separadas permite ter lábios fechados, lábios separados ou configuração arredondada. O lábio superior, representado por L5, tem as mesmas coordenadas mas em relação ao ponto U;

- A posição do hióide é definida pelo parâmetro **hyoid**, representando a distância entre o ponto PP e o segmento de recta H-DL. O ponto PP encontra-se na perpendicular ao segmento H-DL que passa pelo ponto médio deste. No modelo, DL-PP é representado por um segmento de recta e PP-H por um arco;

- O estado do véu palatino é representado pela posição do ponto V', representando a ponta da úvula que se move ao longo do segmento de recta V-V'. A abertura velar é proporcional à distância entre o ponto V e a posição mais elevada do véu palatino. No modelo, esta distância é especificada pelo parâmetro **velum**. O arco M-V', com centro na linha vertical que passa pelo ponto M [4], é afectado pela posição do véu palatino.

Para facilitar a utilização do modelo, tornando desnecessário entrar sempre em linha de conta com os valores possíveis para cada um dos articuladores, os parâmetros articulatórios são representados por um número entre 0 e um valor máximo, igual a 1000, excepto para o ângulo do maxilar, que tem por máximo 800, e para a abertura do velo, com máximo igual a 519.

4.1.1.2 Obtenção da área

A informação final do modelo articulatório, necessária para os modelos acústicos, é a função de área, isto é, o comprimento e área das várias secções do tracto. É necessário obter esta informação, tridimensional, com base no modelo sagital bidimensional.

A primeira fase consiste em obter, aplicando uma grelha, as distâncias sagitais g_j, definidas como o comprimento das linhas da grelha entre os contornos posterior/superior e anterior/inferior (Figura 4.2). Optamos por uma grelha variável com as posições dos articuladores, semelhante à utilizada por Hsieh (1994) e Prado (1991), para que as linhas limite de cada

[3] lipp do Inglês *lip protrusion* e lipo do Inglês *lip open*.
[4] representado na figura como cmv

secção sejam o mais possível perpendiculares à propagação do som. Desta forma obtêm-se melhores estimativas das distâncias sagitais e consequentemente das áreas de secção. A complexidade adicional não se torna proibitiva. A grelha utilizada divide o tracto em 60 secções repartidas por 6 zonas.

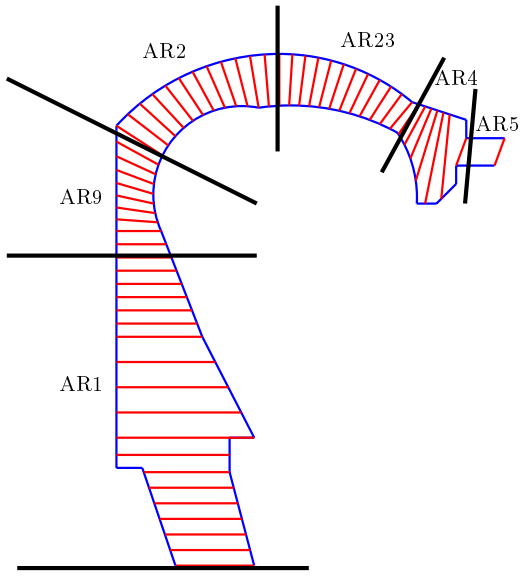


Figura 4.2: Grelha e as zonas utilizadas na obtenção da função de área com base no contorno sagital do tracto.

A distância entre os pontos médios de duas linhas sagitais consecutivas, g_j e g_{j+1}, dá-nos o comprimento da secção j, designado por l_j.

As distâncias sagitais g_j são convertidas em áreas de secção utilizando um conjunto de fórmulas empíricas baseadas em dados publicados por Mermelstein (1973). Em geral, a área de secção, A_j, é obtida como

$$A_j = F(j, g_j) \cos \alpha_j \ , \tag{4.1}$$

onde $F(j, g_j)$ é uma função empírica com fórmulas diferentes para as várias regiões do tracto, e α_j é o ângulo entre a direcção de propagação da onda e a normal à j-ésima linha da grelha (Mermelstein, 1973; Rubin *et al.*, 1981).

Estas fórmulas têm duas formas diferentes, seguindo as propostas de Heinz e Stevens (1964). Na zona da faringe e laringe, a forma da secção é aproximada por uma elipse. Para a zona palatal a área seccional é obtida a partir da distância g_j com uma fórmula do tipo

$$F(j, g_j) = \alpha_j g_j^{\beta_j} . \tag{4.2}$$

Na zona da faringe (AR1 e AR9), segundo Hsieh (1994); Childers (2000),

$$F(j, g_j) = \pi g_j(g_j + \Delta_g) \text{ com } \Delta_g \in [1.5, 3] \tag{4.3}$$

b_j aumenta desde a laringe.

As fórmulas para as zonas do palato foram obtidas de modo a aproximar os dados de Ladefoged *et al.* (1971).

Na zona do palato mole (AR2) a fórmula é:

$$F(j, g_j) = 2.0 g_j^{1.5}. \tag{4.4}$$

Para o palato duro (AR23),

$$F(j, g_j) = 1.6 g_j^{1.5}. \tag{4.5}$$

Para a outra região (AR4), zona alveolar, a fórmula é:

$$F(j, g_j) = \begin{cases} 1.5 g_j & g_j < 0.5 \\ 0.75 + 3(g_j - 0.5) & 0.5 < g_j < 2 \\ 5.25 + 5(g_j - 2) & g_j > 2 \end{cases} . \tag{4.6}$$

Para a região labial (AR5) a área é novamente assumida como elíptica (Mermelstein, 1973; Mermelstein e Maeda, 1971), tendo-se

$$F(j, g_j) = g_j[2.0 + 1.5(lipo - lipp)]. \tag{4.7}$$

4.1.2 Modelamento do tracto nasal

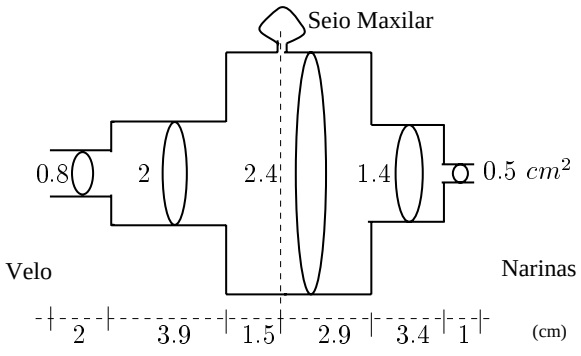


Figura 4.3: Modelo nasal utilizado. Baseado em Chen (1997).

A escolha do modelo para a configuração do tracto nasal foi motivada, essencialmente, pela possibilidade de obtenção das ressonâncias características de vogais nasais e de consoantes nasais naturais, em especial nas baixas frequências.

Diversos estudos (Feng e Castelli, 1996; Feng, 1987) mostram que para se obter as formantes características das vogais nasais e simultaneamente as formantes de consoantes nasais não é suficiente a inclusão no modelo dos seios paranasais. O problema surge em especial ao tentar obter um valor para F_1 próximo do 300 Hz para um consoante nasal velar. Os mesmos investigadores propuseram a utilização de uma área de radiação reduzida, de 0.6 cm^2, para obter os resultados pretendidos.

Para verificar a validade desta solução foram por nós calculadas (Teixeira *et al.*, 1999b) as formantes para uma configuração faringo-nasal, obtida com o tracto configurado para a vogal [ə] abaixando o velo até obstruir completamente a passagem oral. Obtiveram-se, para área de radiação de 0.6 cm^2, $F_1 = 340\ Hz$, $F_2 = 1000\ Hz$ e $F_3 = 2055\ Hz$, valores muito próximos dos referidos por Ohala e Ohala (1993), $F_1 = 300\ Hz$, $F_2 = 1000\ Hz$ e $F_3 = 1900\ Hz$, para esta configuração. Se a área de radiação for alterada para 2 cm^2 (valor utilizado no modelo proposto por Maeda (1982b)), mantendo o resto da configuração nasal, obtém-se para F_1 um valor desadequado de 428 Hz.

A simplicidade e existência de dados foram também consideradas na escolha. O modelo por nós adoptado foi o apresentado por Chen (1997) baseado nos estudos de Dang e Honda (1994) e Stevens (1998). As dimensões encontram-se representadas na Figura 4.3. Note-se a área de radiação pelas narinas igual a 0.5 cm^2 e a inclusão de um seio paranasal, o maxilar. O modelo utiliza apenas um tubo, assumindo simetria das duas passagens nasais [5].

4.2 Modelamento da propagação do ar no tracto

4.2.1 Escolha do método

Os métodos de resolução dos modelos acústicos no domínio do tempo (veja-se a secção 3.5.2 na página 69) são lentos; é fácil incluir perdas, mas estas são independentes da frequência; a radiação é também independente da frequência (circuito paralelo proposto por Flanagan); o comprimento das secções tem de ser igual e não é possível uma variação contínua do comprimento total do tracto.

Para os métodos usando linhas de transmissão digitais (secção 3.5.1 na página 67), os seus pontos fortes são a rapidez e o facto de ser fácil gerar sons contínuos; os seus problemas resultam das dificuldades de incluir as perdas e o comprimento das secções estar relacionado com a frequência de amostragem.

Alternativamente, os métodos de análise na frequência e síntese no domínio do tempo (secção 3.5.3 na página 70), permitem incluir perdas dependentes da frequência, variação contínua do comprimento do tracto e simular a interacção entre a fonte glotal e o tracto. Existem duas variantes na forma como é feita a passagem da frequência para o tempo. O primeiro método,

[5] O modelo acústico desenvolvido permite, como veremos, modelos mais complexos.

usando transformada inversa de Fourier, proposto por Sondhi e Schroeter (1987), ao usar convolução para a obtenção do sinal de voz, é lento e a interface com a fonte é complexa. O outro método, baseado na aproximação da resposta na frequência por filtros em paralelo, proposto por Lin (1995), torna a síntese mais expedita. O grande problema resulta da dificuldade em garantir a continuidade das formantes. Existem também duas formas de calcular a resposta em frequência: usando matrizes; ou um processo iterativo.

Depois de analisados os prós e os contras das várias técnicas [6], optamos por um sistema usando síntese na frequência por possibilitar uma simulação mais realista das perdas e permitir a obtenção da resposta na frequência de uma forma fácil.

Para a síntese, o processo de convolução, embora lento, evita o problema da continuidade das formantes e permite utilizar toda a informação espectral, tendo sido por isso o escolhido. O uso de matrizes é adequado para obter as funções de transferência e impedância de entrada do tracto, necessárias no processo de síntese.

O processo iterativo de cálculo da função de transferência (Lin, 1990) foi utilizado para a análise necessária na inversão. Este método é adequado para se obter os pólos e zeros, usualmente utilizados nos processos de inversão (ver secção 4.5), pois permite separar a função de transferência em duas partes, uma contendo os zeros, outra os pólos.

Apresenta-se de seguida o método de síntese utilizado, deixando-se a descrição do processo iterativo de cálculo da função de transferência para a secção sobre a inversão.

4.2.2 Análise na frequência utilizando matrizes

Na exposição não se assume qualquer modelo para o tubo elementar. O método tanto é válido para o modelo de Sondhi utilizado como para qualquer outro em que tenhamos uma matriz ABCD.

4.2.2.1 Modelo de um tubo elementar

Cada secção do modelo acústico pode ser representada, na frequência, como uma função de transferência representada na forma matricial por uma matriz ABCD,

$$\begin{bmatrix} P_s(\omega) \\ U_s(\omega) \end{bmatrix} = \begin{bmatrix} A(\omega) & B(\omega) \\ C(\omega) & D(\omega) \end{bmatrix} \begin{bmatrix} P_e(\omega) \\ U_e(\omega) \end{bmatrix} \tag{4.8}$$

$$= K(\omega) \begin{bmatrix} P_e(\omega) \\ U_e(\omega) \end{bmatrix} . \tag{4.9}$$

A matriz relaciona a pressão, $P_s(\omega)$, e a velocidade de volume, $U_s(\omega)$, à saída, com a pressão,

[6] E mesmo experimentá-las, como fiz com o sintetizador, usando as técnicas no tempo do *Mind Machine Interaction Research Center* (MMIRC), University of Florida. Um simulador, usando a técnica na frequência para vogais, foi também implementado com a minha colaboração (Branco, 1997)

$P_e(\omega)$, e velocidade de volume, $U_e(\omega)$, na entrada do tubo. Designemos esta matriz por $K(\omega)$. Os elementos $A(\omega)$, $B(\omega)$, $C(\omega)$, $D(\omega)$, variam com a frequência, incluindo o efeito de vários tipos de perdas. São função do comprimento e área seccional do tubo. Neste trabalho utilizamos o modelo proposto por Sondhi e Schroeter (1987) já descrito na secção 3.4.4. Esta escolha deveu-se essencialmente à dificuldade de obtenção de valores para os parâmetros de outros modelos. O facto do modelo de Sondhi e Schroeter (1987) estar descrito em mais detalhe na literatura, associado à sua maior generalidade e obtenção dos parâmetros com base em medidas acústicas, levou a que tenha sido utilizado por diversos investigadores (veja-se por exemplo Riegelsberger (1997)).

No resto da exposição do modelo não incluiremos a dependência dos elementos da matriz com a frequência para simplificar as expressões. Representa-se na Figura 4.4 o modelo descrito para uma secção.

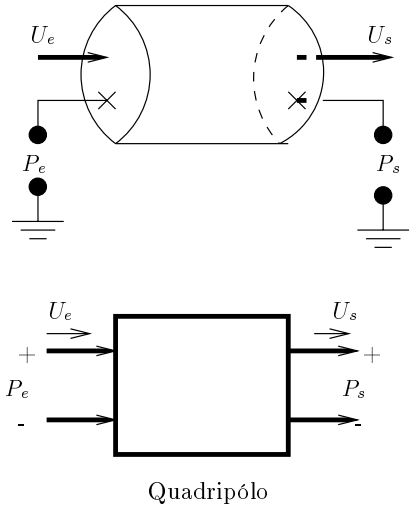


Figura 4.4: Representação de uma secção por um quadripólo.

4.2.2.2 Modelo com vários tubos

Um modelo composto por várias secções, num total de N, pode ser representado pelo produto de N matrizes, cada uma representando uma secção,

$$K_{Nsecs} = \prod_{i=1}^{N} K_i = \begin{bmatrix} A_{Nsecs} & B_{Nsecs} \\ C_{Nsecs} & D_{Nsecs} \end{bmatrix} . \tag{4.10}$$

Com base na matriz K_{Nsecs} a função de transferência do conjunto, terminado por uma impedância de carga Z_c, pode ser obtida por:

$$H_{Nsecs} = \frac{U_s}{U_e} = \frac{A_{Nsecs}D_{Nsecs} - C_{Nsecs}B_{Nsecs}}{A_{Nsecs} - C_{Nsecs}Z_c}, \tag{4.11}$$

e a impedância de entrada do conjunto é

$$Z_e = \frac{P_e}{U_e} = \frac{D_{Nsecs} Z_c - B_{Nsecs}}{A_{Nsecs} - C_{Nsecs} Z_c}. \tag{4.12}$$

Note-se a igualdade do denominador nas duas expressões anteriores. De uma forma similar podem obter-se as funções P_s/U_e, P_s/P_e e U_s/P_e.

4.2.2.3 Modelo completo do tracto

Para poder simular todas as cavidades supraglotais, incluindo as cavidades nasais, torna-se necessário decompor o tracto em várias regiões. Na Figura 4.5 estão representadas as zonas utilizadas. Note-se que apenas estamos interessados neste trabalho em sons com excitação glotal. Para outros sons, como por exemplo as fricativas, teria de considerar-se ainda a decomposição da zona oral entre os lábios e a zona de acoplamento do tracto nasal (Riegelsberger, 1997).

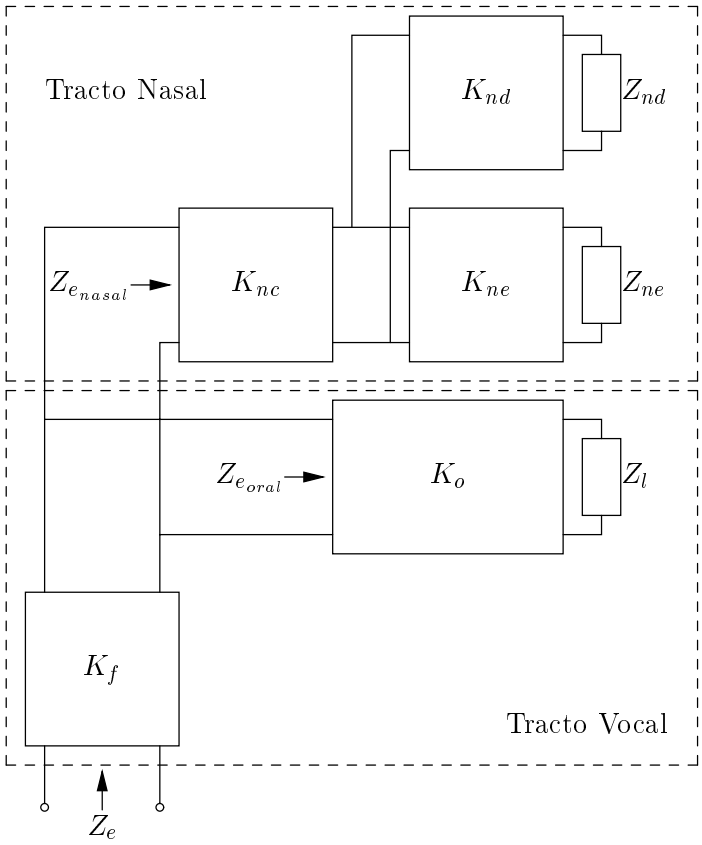


Figura 4.5: Modelo acústico completo.

O tracto vocal, terminado pela impedância de radiação dos lábios Z_l, divide-se em duas regiões:

1. A região faríngea, entre a glote e a zona de acoplamento do tracto nasal, representada pela matriz K_f. Não são permitidas oclusões nesta região em Português.

2. A região oral, entre a zona de acoplamento do tracto nasal e os lábios, representada por K_o.

O modelo geral para o tracto nasal é constituído por três regiões:

1. A região nasal comum, que constitui a continuação da faringe, até à bifurcação do tracto nasal nas suas duas passagens, representada por K_{nc}.

2. A passagem nasal esquerda, que termina na narina esquerda, representada por K_{ne}.

3. A passagem nasal direita, representada por K_{nd}.

No modelo existem duas impedâncias de radiação, uma para cada narina, representadas por Z_{ne} e Z_{nd}.

No caso de se considerar o tracto nasal simétrico deixa de ser necessário considerar as três regiões. No modelo implementado neste caso, apenas se considera a existência da zona nasal comum e uma única impedância de radiação nasal Z_n.

No modelo do tracto nasal as cavidades paranasais (seios) são representadas por circuitos ressonantes de Helmholtz, inseridos em paralelo. A impedância, $Z_{Helmholtz}$, representativa destes circuitos é incluída no cálculo das matrizes nasais utilizando a matriz

$$K_{seio} = \begin{bmatrix} 1 & 0 \\ -1/Z_{Helmholtz} & 1 \end{bmatrix}. \tag{4.13}$$

4.2.2.4 Função de transferência total e impedância de entrada

Tem-se, em geral, três pontos de radiação: as duas narinas e os lábios [7]. Desprezando os efeitos dos diferentes trajectos desde o ponto de radiação até ao ponto de medição da pressão sonora total, utilizaremos a soma das radiações nestes diferentes pontos como pressão total. Para obter o sinal radiado em cada um destes pontos, torna-se necessário obter as matrizes entre a glote e esse ponto.

A matriz entre a glote e os lábios, K_{gl}, é dada por

$$K_{gl} = K_o \times K_{an} \times K_f \ , \tag{4.14}$$

[7] Não se incluem neste trabalho as radiações pelas paredes do tracto, devido á sua reduzida relevância para o caso das vogais.

onde K_{an} é a matriz referente ao acoplamento do tracto nasal,

$$K_{an} = \begin{bmatrix} 1 & 0 \\ -1/Z_{e_{nasal}} & 1 \end{bmatrix}, \tag{4.15}$$

sendo $Z_{e_{nasal}}$ a impedância do tracto nasal vista do velo. Quando o velo se encontra posicionado de forma a fechar a passagem para o tracto nasal $Z_{e_{nasal}}$ é infinita e K_{an} torna-se a matriz identidade. Também se se pretender não incluir o efeito da carga nasal no cálculo de K_{gl} pode fazer-se esta matriz, K_{an}, igual á matriz identidade.

A matriz entre a glote e a narina esquerda, K_{gne}, é dada por

$$K_{gne} = K_{ne} \times K_{and} \times K_{nc} \times K_{ao} \times K_f \ , \tag{4.16}$$

onde K_{ao} é a matriz de acoplamento representando a impedância de entrada da região oral na zona de acoplamento do tracto nasal, sendo K_{and} a matriz de acoplamento representativa da impedância de entrada da passagem nasal direita.

De uma forma similar obtém-se a matriz entre a glote e a narina direita como sendo

$$K_{gnd} = K_{nd} \times K_{ane} \times K_{nc} \times K_{ao} \times K_f. \tag{4.17}$$

A função de transferência completa é

$$H_{tot} = \frac{U_{ne} + U_{nd} + U_l}{U_g} = \frac{U_{ne}}{U_g} + \frac{U_{nd}}{U_g} + \frac{U_l}{U_g} \ , \tag{4.18}$$

representando U_g o fluxo glotal. As respostas parciais obtêm-se de K_{gl}, K_{gnd}, e K_{gne}, utilizando as relações apresentadas na equação 4.11.

Como o ouvido humano é sensível às variações de pressão, o objectivo final dos cálculos é obter a pressão radiada. O efeito da radiação pode ser representado, de forma aproximada, pela derivada do fluxo radiado (Fant, 1960). Na prática, é usual efectuar a derivada do fluxo à entrada do tracto em vez de a efectuar no fluxo radiado, técnica que adoptamos neste trabalho. Utilizando como excitação a derivada do fluxo glotal, a função de transferência, H_{tot}, permite obter directamente a pressão.

4.2.2.5 Obstruções

A área do tracto, oral ou nasal, pode reduzir-se a zero em algumas secções. Quando esta situação ocorre a impedância dessa secção é infinita e a secção anterior tem por impedância de carga essa impedância infinita. A função de transferência da região, contendo essa secção, será nula, pois não existe fluxo na saída. No entanto, continua a ser necessário calcular a impedância de entrada da região. Aplicando a equação 4.12, no caso de Z_c infinita, à secção

antes da oclusão (secção mais próxima da glote), para calcular a impedância de entrada

$$Z_{e_{oc-1}} = \frac{D_{Nsecs}}{-C_{Nsecs}}. \tag{4.19}$$

A matriz representativa da região é calculada para as restantes secções, na direcção da glote. Para o cálculo da impedância de entrada da região é utilizada a matriz obtida e $Z_{e_{oc-1}}$ como impedância de carga.

4.2.2.6 Implementação do cálculo da função de transferência e impedância de entrada

O cálculo da função de transferência e da impedância de entrada, para o caso geral, e com base nas explicações anteriores, é efectuado pelos seguintes passos:

1. Cálculo das matrizes para cada uma das regiões, tendo em conta possíveis oclusões;
2. Cálculo das impedâncias de radiação, ou no caso de oclusões, cálculo da impedância de carga a utilizar para a região;
3. Cálculo da impedância de entrada das duas passagens nasais e da cavidade bucal. Para modelos nasais simétricos, esta tarefa reduz-se ao cálculo da impedância de entrada da região oral;
4. Cálculo da impedância de entrada do tracto nasal. No modelo geral, com três regiões, o paralelo das impedâncias de entrada das duas passagens constitui a impedância de carga da região nasal comum. No caso simétrico, a impedância de carga é a impedância de radiação nasal Z_n;
5. Utilizando as impedâncias de entrada calcular as matrizes de acoplamento;
6. Calcular as matrizes entre a glote e os vários pontos de radiação;
7. Utilizando as matrizes, obtidas no passo anterior, calcular as três funções de transferência. As regiões contendo oclusões têm, obviamente, função de transferência igual a zero. As várias funções de transferência podem ser adicionadas se se pretender o efeito total ou guardadas individualmente se se pretender obter, por exemplo, o som radiado por uma das narinas.
8. Calcular a impedância de entrada do tracto. Esta é facilmente calculada utilizando como carga, para a região faríngea, o paralelo da impedância de entrada da região oral, $Z_{e_{oral}}$ com a impedância de entrada do tracto nasal, $Z_{e_{nasal}}$. Este processo facilita a não inclusão do efeito da impedância de entrada do tracto nasal no cálculo da impedância de

entrada do tracto, caso se deseje investigar o efeito do acoplamento nasal na impedância de entrada [8].

4.2.2.7 Obtenção da resposta impulsional

Na secção anterior, descrevemos como obter a resposta das cavidades supraglotais para uma frequência. Para sintetizar um som é necessário a resposta impulsional para efectuar a convolução com a onda de excitação glotal.

O processo utilizado é o seguinte: N amostras da função de transferência são obtidas entre 0 e metade da frequência de amostragem, com intervalo constante. Na implementação actual, em que a frequência de amostragem é de 10 kHz, $N = 256$, dando uma resolução de aproximadamente 19.5 Hz e uma resposta impulsional com 512 amostras. A resposta em frequência é filtrada com o filtro utilizado por Schroeter e Sondhi (1992)

$$H_f(z) = \frac{1 + z^{-1}}{1 + 0.95z^{-1}} \ . \tag{4.20}$$

À resposta, depois de filtrada, é aplicada uma Transformada Inversa de Fourier (Brigham, 1988), utilizando-se uma implementação rápida desenvolvida por Frigo (1997).

Como passo final, é aplicada à resposta impulsional obtida uma janela de Hanning.

O mesmo procedimento é aplicado à impedância de entrada, $Z_e(w)$, para se obter $z_e(n)$ necessária para implementação de interacção entre a fonte glotal e o tracto, como veremos na secção seguinte.

4.3 Modelamento da excitação glotal

Estando apenas interessados no estudo de sons com excitação glotal, apenas foi implementado um modelo de excitação glotal.

Os requisitos base para o modelo da excitação foram: permitir o estudo da interacção entre a fonte e as cavidades supra-laríngeas; permitir o controlo directo de parâmetros como a frequência fundamental; contribuir para a obtenção de som sintético de qualidade natural; não ser demasiado pesado computacionalmente.

4.3.1 Modelamento dos vários subsistemas

Para a obtenção da excitação glotal, $u_g(t)$, torna-se necessário modelar os vários subsistemas envolvidos: pulmões, cavidades subglotais, a glote e o tracto supraglotal. O esquema da Figura 4.6 representa estes subsistemas.

[8] Esta facilidade foi muito importante para o estudo da interacção fonte-tracto e que abordaremos no capítulo 5.

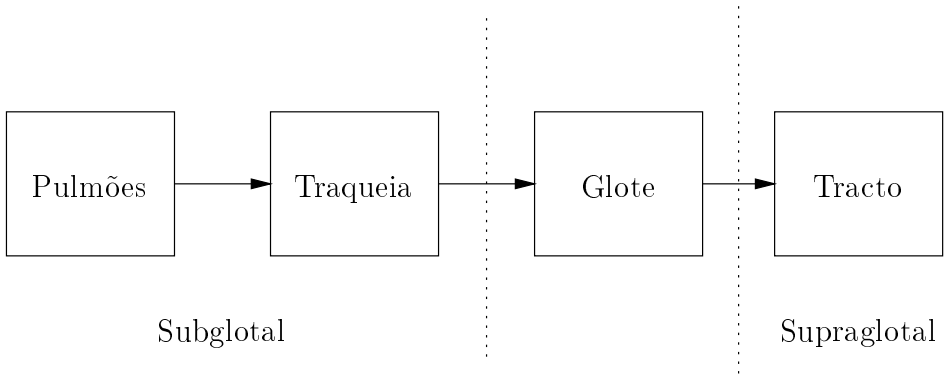


Figura 4.6: Vários subsistemas intervenientes na obtenção da onda de excitação glotal.

4.3.1.1 Modelamento dos pulmões

O papel dos pulmões é o de fonte de pressão, quase constante, sendo representados no nosso modelo por uma fonte de pressão pulmonar p_p em série com uma resistência R_p. Um valor típico para a pressão pulmonar é o de 10 $cm\ H_2O\ =\ 9806\ dine/cm^2$ o que leva alguns autores a usar 10000 $dine/cm^2$. A pressão pulmonar pode ser obtida dos parâmetros geralmente utilizados para controlo de sintetizadores de formantes (Pinto *et al.*, 1989). Allen e Strong (1985) usaram $R_p\ =\ 18\ \Omega\ cgs$, mas chegaram à conclusão de que se tratava de um valor demasiado elevado. Usamos neste estudo o valor de $R_p = 8\ \Omega\ cgs$.

Parâmetro	Valor	Unid.	Parâmetro	Valor	Unid.	Parâmetro	Valor	Unid.
R_{sg1}	36.7	Ω	R_{sg2}	53.6	Ω	R_{sg3}	53.9	Ω
L_{sg1}	3.80	mH	L_{sg2}	0.72	mH	L_{sg3}	0.27	mH
C_{sg1}	17.6	μF	C_{sg2}	19.2	μF	C_{sg3}	21.1	μF

Tabela 4.1: Valores para os circuitos RLC utilizados no modelamento das cavidades subglotais (Hsieh, 1994).

4.3.1.2 Modelamento da traqueia e brônquios

Para a representação da parte subglotal, incluindo a traqueia, utilizamos a abordagem de Ananthapadmanabha e Fant (1982), com três circuitos RLC ressonantes. A simulação da parte subglotal utilizando modelos similares aos utilizados no caso das cavidades supraglotais não foi tentada, devido à falta de informação detalhada acerca das dimensões. Os valores dos componentes dos circuitos RLC são os utilizados por Ananthapadmanabha e Fant (1982) inicialmente propostos por Ishizaka *et al.* (1976), encontrando-se na Tabela 4.1.

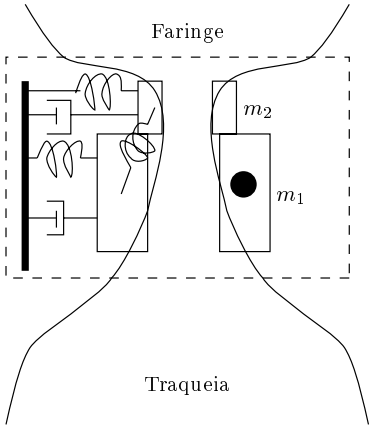


Figura 4.7: Esquema muito simplificado do modelo de duas massas proposto por Ishizaka e Flanagan (1972).

4.3.1.3 Modelamento das cordas vocais

Várias abordagens (veja-se a secção 3.6.1 na página 71) foram utilizadas para modelar as cordas vocais: modelos auto-oscilantes, modelos com área glotal parametrizada, etc. Pretendia-se um modelo que permitisse elevada qualidade e com bases fisiológicas, como o modelo de duas massas (Figura 4.7), mas que fosse também não muito exigente em termos computacionais. O modelo deveria ainda permitir o controlo directo de parâmetros como a frequência fundamental. Foi utilizado o modelo proposto por Prado (1991) em que se parametriza directamente as áreas glotais do modelo de duas massas.

4.3.1.4 Modelamento do efeito de carga das cavidades supraglotais

Os sistemas que se encontram acima da glote, o tracto, podem ser modelados por uma impedância de entrada $z_e(t)$ (ou a pressão $p_{supra}(t)$ que se obtém pela convolução dessa impedância de entrada e o fluxo glotal) ou aproximados por uma cascata de circuitos RLC.

A utilização da impedância de entrada permite modelar melhor as perdas dependentes da frequência (Allen e Strong, 1985, pág. 59). Foi por isso escolhido este método para o nosso modelo. A impedância de entrada é obtida do modelo acústico das cavidades supralaríngeas, pelo processo descrito anteriormente. Interessa aqui referir que, na implementação efectuada do cálculo da impedância, é possível calcular a impedância de entrada, para sons nasais, desprezando a impedância de entrada do tracto nasal. Esta facilidade é da máxima utilidade para estudar o efeito adicional do acoplamento do tracto nasal nas características da onda de excitação glotal. No caso de não se pretender incluir o efeito de carga supraglotal $z_e = 0$.

4.3.2 Fonte Implementada

Depois de efectuadas as escolhas para a forma de modelar cada um dos subsistemas envolvidos, obtemos o circuito apresentado na Figura 4.8.

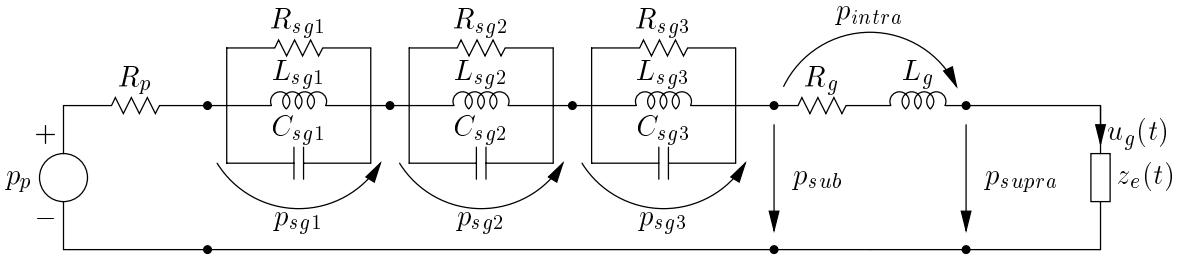


Figura 4.8: Análogo eléctrico utilizado na obtenção da onda de excitação glotal, $u_g(t)$.

4.3.2.1 Modelo paramétrico das áreas

No modelo, a resistência R_g e indutância L_g, representando as cordas vocais, dependem da área de abertura glotal. Como já foi referido, optamos por utilizar um modelo paramétrico de duas massas baseado no trabalho de Prado (1991).

As áreas glotais, $Ag_1(t)$ e $Ag_2(t)$, obtêm-se através de

$$Ag_1(t) = \begin{cases} A_1\left(0.5 - 0.5\cos\left(\frac{\pi t}{T_a}\right)\right) + Ag_0 & 0 < t < T_a \\ A_1\cos\left(\frac{\pi(t-T_a)}{2T_f}\right) + Ag_0 & T_a < t < T_a + T_f \\ Ag_0 & T_a + T_f < t < T_0 \end{cases} \quad (4.21)$$

$$Ag_2(t) = \begin{cases} Ag_0 & 0 < t < \tau \text{ ou } T_a + T_f + \tau < t < T_0 \\ A_2\left(0.5 - 0.5\cos\left(\frac{\pi(t-\tau)}{T_a}\right)\right) + Ag_0 & \tau < t < T_a + \tau \\ A_2\cos\left(\frac{\pi(t-T_a-\tau)}{2T_f}\right) + Ag_0 & T_a + \tau < t < T_a + T_f + \tau \end{cases} \quad (4.22)$$

em que T_0 é o período de excitação glotal ($T_0 = 1/F_0$); T_a a duração do movimento de abertura das cordas vocais; T_f o tempo que as cordas demoram a fechar; Ag_0 abertura mínima da glote; A_2 e A_1 aberturas máximas; e $\tau = \frac{\Phi T_0}{360}$ com Φ a diferença de fase entre Ag_1 e Ag_2.

A_1 toma o valor do parâmetro Ag_{max} e A_2 obtém-se subtraindo a Ag_{max} o valor do parâmetro $A_2 - A_1$.

O valor de T_a e T_f, representados na Figura 4.9, obtêm-se do quociente de abertura, OQ (do Inglês *Open Quotient*), e quociente de velocidade, SQ (do Inglês *Speed Quotient*), usando as seguintes expressões:

$$T_a = OQ \times T_0 \times SQ/(SQ+1) \quad (4.23)$$

$$T_f = T_a/SQ\ , \quad (4.24)$$

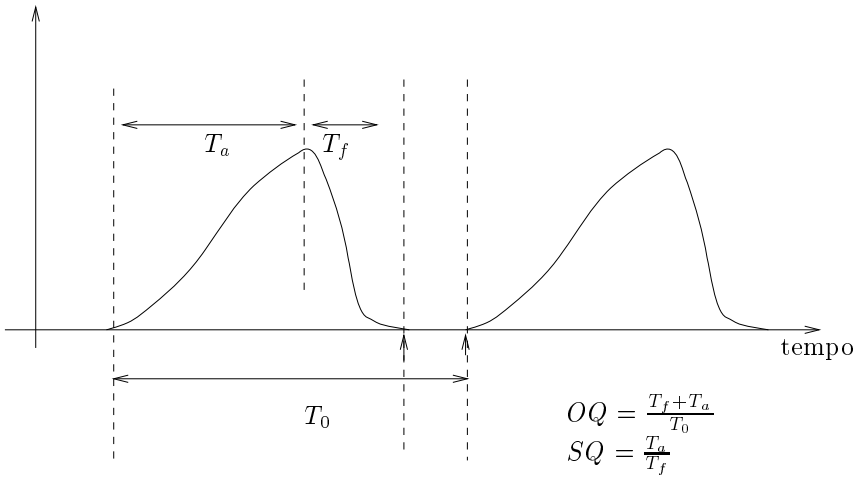


Figura 4.9: Definição dos parâmetros da onda glotal quociente de abertura (OQ) e quociente de velocidade (SQ).

4.3.2.2 Cálculo de R_g e L_g

Com base nas áreas, $Ag_1(t)$ e $Ag_2(t)$, obtêm-se os valores para R_g e L_g, usando as expressões (Schroeter e Sondhi, 1992, equações 6 e 7):

$$R_g = \frac{\rho}{2}\left[\frac{0.37}{Ag_1^2} + \frac{1 - 2\frac{Ag_2}{area_1}\left(1 - \frac{Ag_2}{area_1}\right)}{Ag_2^2}\right]|u_g| + 12\mu l_g^2\left(\frac{d_1}{Ag_1^3} + \frac{d_2}{Ag_2^3}\right) + R_{12} \tag{4.25}$$

$$L_g = \rho\left(\frac{d_1}{Ag_1} + \frac{d_2}{Ag_2}\right), \tag{4.26}$$

em que R_{12} representa as perdas causadas pela diferença de fase entre as duas massas e é dada por (Schroeter e Sondhi, 1992, equação 9a)

$$R_{12} = \frac{\rho}{2}\eta_{12}\left(\frac{1}{Ag_1} - \frac{1}{Ag_2}\right)^2|u_g|\ , \tag{4.27}$$

com

$$\eta_{12} = \begin{cases} 0.4 & \text{se } Ag_1 \geq Ag_2 \\ 1.0 & \text{se } Ag_1 < Ag_2 \end{cases}\ . \tag{4.28}$$

Nas equações 4.25 a 4.28 Ag_1, Ag_2, $area_1$ e u_g são todas variáveis no tempo, tendo-se omitido essa dependência para simplificar as expressões.

4.3.2.3 Cálculo do fluxo glotal

As variações de pressão ao longo do circuito podem ser representadas por:

$$p_p - R_p u_g(t) - p_{sg1} - p_{sg2} - p_{sg3} - \frac{d(L_g u_g(t))}{dt} - R_g u_g(t) - p_{supra}(t) = 0 \ , \tag{4.29}$$

onde p_{sgk} , $k = 1, 2, 3$ se obtêm de,

$$u_g(t) = C_{sgk}\frac{d(p_{sgk}(t))}{dt} + \frac{p_{sgk}(t)}{R_{sgk}} + \frac{1}{L_{sgk}}\int_0^t p_{sgk}(\tau)d\tau \ . \tag{4.30}$$

A obtenção de $u_g(t)$, ou melhor do seu equivalente discreto $u_g(nT_s)$, implica a resolução das equações diferenciais. A resolução numérica destas equações consegue-se usando uma aproximação por equações diferença. As derivadas e integrais são aproximados por

$$\frac{d}{dt}f(t) \cong \frac{f(t_i) - f(t_{i-1})}{T_s} \tag{4.31}$$

e

$$\int_0^t f(t)dt \cong T_s \sum_{j=0}^{i-1} f(t_j) \tag{4.32}$$

onde $T_s = t_i - t_{i-1}$ representa o intervalo de amostragem. Ishizaka e Flanagan (1972) usaram amostragens entre 10 e 30 kHz.

A equação 4.29 resulta em [9]

$$\begin{aligned} p_p(n) - R_p u_g(n) - p_{sg1}(n) - p_{sg2}(n) - p_{sg3}(n) - \frac{L_g(n)(u_g(n) - u_g(n-1)}{T_s} \\ - \frac{u_g(n)(L_g(n) - L_g(n-1))}{T_s} - R_g(n)u_g(n) - p_{supra}(n) = 0 \ , \end{aligned} \tag{4.33}$$

e a Eq. 4.30 em

$$u_g(n) = \frac{C_{sgk}(p_{sgk}(n) - p_{sgk}(n-1))}{T_s} + \frac{p_{sgk}(n)}{R_{sgk}} + \frac{T_s}{L_{sgk}}\sum_{j=0}^{n-1} p_{sgk}(j) \ , k = 1, 2, 3 \ . \tag{4.34}$$

Resolvendo a Eq.4.34 para obter $p_{sgk}(n)$

$$p_{sgk}(n) = \frac{u_g(n)}{a} + \frac{C_{sgk}p_{sgk}(n-1)}{aT_s} - \frac{T_s}{aL_{sgk}}\sum_{j=0}^{n-1} p_{sgk}(j) \ , \tag{4.35}$$

onde $a_k = \frac{C_{sgk}}{T_s} + \frac{1}{R_{sgk}}$.

[9] Para simplificar as fórmulas utiliza-se n em lugar de nT_s e $n-1$ para representar $(n-1)T_s$.

A pressão supraglotal, $p_{supra}(n)$, pode ser obtida como a convolução da impedância de entrada do tracto $z_e(t)$, com o fluxo glotal $u_g(t)$ (Allen e Strong, 1985, pg. 60):

$$p_{supra}(t) = u_g(t) * z_e(t). \tag{4.36}$$

A convolução discreta pode ser decomposta na soma de dois termos, a saber:

$$p_{supra}(n) = \sum_{j=0}^{\infty} u_g(n-j) z_e(j) \tag{4.37}$$

$$= u_g(n) z_e(0) + \sum_{j=1}^{\infty} u_g(n-j) z_e(j). \tag{4.38}$$

Pela substituição das equações 4.35 e 4.38 na Eq. 4.33, seguida de resolução em ordem a $u_g(n)$ obtém-se:

$$\begin{aligned} u_g(n) = \Bigg(& p_p(n) + \frac{L_g(n) u_g(n-1)}{T_s} - \sum_{j=1}^{N} u_g(n-j) z_e(j) \\ & - \frac{C_{sg1} p_{sg1}(n-1)}{a_1 T_s} - \frac{C_{sg2} p_{sg2}(n-1)}{a_2 T_s} - \frac{C_{sg3} p_{sg3}(n-1)}{a_3 T_s} \\ & + \frac{T_s}{a_1 L_{sg1}} \sum_{j=0}^{n-1} p_{sg1}(j) + \frac{T_s}{a_2 L_{sg2}} \sum_{j=0}^{n-1} p_{sg2}(j) + \frac{T_s}{a_3 L_{sg3}} \sum_{j=0}^{n-1} p_{sg3}(j) \Bigg) \\ & \times \left(R_p + \frac{L_g(n)}{T_s} + \frac{L_g(n) - L_g(n-1)}{T_s} + R_g(n) + z_e(0) + \frac{1}{a_1} + \frac{1}{a_2} + \frac{1}{a_3} \right)^{-1} \end{aligned} . \tag{4.39}$$

4.3.2.4 Cálculo das pressões

Depois de obtido $u_g(n)$, para o instante actual, pode calcular-se facilmente a pressão supraglotal utilizando a equação 4.38. Note-se que apenas se tem de adicionar o primeiro termo, $u_g(n) z_e(0)$, ao somatório anteriormente calculado no processo de obtenção de $u_g(n)$.

Tendo o valor de $u_g(n)$, a diferença de pressão entre os dois extremos da glote, $p_{intra}(n)$, é dada por

$$p_{intra}(n) = R_g(n) u_g(n) + L_g(n) \frac{u_g(n) - u_g(n-1)}{T_s} + u_g(n) \frac{L_g(n) - L_g(n-1)}{T_s} . \tag{4.40}$$

A pressão subglotal, $p_{sub}(n)$, é simplesmente a soma de $p_{supra}(n)$ e $p_{intra}(n)$.

4.3.3 Irregularidades

A forma de onda em períodos sucessivos de $u_g(t)$ não é igual. Na literatura sobre o assunto aparecem termos como *jitter*, variação aleatória de período para período na duração do período (Horii, 1979); *shimmer*, referente à variação da amplitude do pulso glotal entre dois períodos

consecutivos; diplofonia; etc. Vozes normais têm *jitter* entre 0.5 % e 1.0 % (Hollien citado em (Klatt e Klatt, 1990, pág. 839)).

O modelo da fonte inclui a possibilidade de modelar algumas irregularidades, nomeadamente flutuações da frequência fundamental e da abertura máxima da glote.

A nossa abordagem baseou-se em (Lalwani, 1991) [10]. Ao valor, obtido por interpolação, de F_0 é adicionado um valor dependente do parâmetro *jitter*

$$F_{0_{com\ jitter}} = F_0 + random \times 2 \times F_0 \times jitter/100.0\ . \tag{4.41}$$

sendo $random$ um valor aleatório entre -0.5 e 0.5. O valor de T_0 correspondente é depois arredondado para o instante de amostragem mais próximo. A implementação é semelhante para o *shimmer*:

$$Ag_{max_{com\ shimmer}} = Ag_{max} + random \times 2 \times Ag_{max} \times shimmer/100.0\ . \tag{4.42}$$

4.3.4 Aspiração

Foi também incluído no modelo a geração de ruído de aspiração, seguindo as propostas de Schroeter e Sondhi (1992).

Utilizando a área glotal Ag_2 e o fluxo glotal, é calculado o quadrado do número de Reynolds, Re^2, segundo a fórmula

$$Re^2(n) = \left[\frac{l_g \rho}{\mu A g_2} u_g(n)\right]^2 \tag{4.43}$$

e o fluxo resultante da aspiração, que apenas existe para $Re^2 > Re^2_{critico}$, é igual a

$$u_{asp}(n) = G \times random(n) \frac{Re^2 - Re^2_{critico}}{R_g(n)}\ , \tag{4.44}$$

com $G = 0.5 \times 10^{-7}$, $Re^2_{critico} = 2700^2$, e $random(n)$ um número aleatório entre -0.5 e 0.5.

4.3.5 Parâmetros do modelo

O modelo de fonte implementado é controlado por dois tipos de parâmetros. O primeiro tipo de parâmetros é passível de variar ao longo do tempo, comportando-se de forma análoga aos parâmetros articulatórios do tracto. No processo de síntese podem ser utilizados para adicionar entoação e maior naturalidade. Estes parâmetros incluem: a pressão pulmonar; os parâmetros relacionados com a duração de cada período de excitação; os parâmetros que definem o período de abertura das cordas vocais em cada período glotal; as áreas de aber-

[10] Outra abordageim, baseada em (Klatt e Klatt, 1990), foi utilizada em (Oliveira, 1996, pág. 165).

tura mínimas e máximas; parâmetros relacionados com fenómenos aleatórios e a aspiração. Colocando a zero o parâmetro Asp o efeito de aspiração não é incluído. Valores diferentes de zero servem para alterar o valor do parâmetro G na implementação da aspiração. Este primeiro tipo é apresentado na Tabela 4.2(a), indicando-se valores típicos para cada um dos parâmetros.

Os valores para a resistência pulmonar, as dimensões da glote e os valores de k_1 e k_2 podem ser alterados editando um ficheiro de configuração. Não podem, no entanto, ter valores variáveis no tempo. As constantes ρ e μ também podem ver os seus valores alterados no ficheiro de configuração. No entanto, estes valores só deverão ser alterados para simular situações, pouco comuns, como por exemplo produção de voz depois de inalar hélio. Estes parâmetros, resumidos na Tabela 4.2(b), constituem o segundo tipo.

Parâmetro	Descrição	Valor típico	Unidade	Ficheiro
p_p	Pressão pulmonar	10000	$dine/cm^2$	lungs
F_0	Frequência fundamental	$100-200$	Hz	f0
OQ	Quociente de abertura	60	% de T_0	openq
SQ	Quociente de velocidade	2		speed
A_{g0}	Valor mínimo da área glotal	0	cm	ag0
A_{gmax}	Valor máximo da área glotal	0.3	cm	agmax
$A_2 - A_1$		0.03	cm	slope
Jitter	variação aleatória de F_0	2	%	jitter
Shimmer	variação aleatória de A_{gmax}	5	%	shimmer
Asp	Aspiração	1		aspiration

(a) Parâmetros do modelo de excitação variáveis no tempo. São definidos pelo utilizador de forma semelhante aos articuladores que definem a configuração do tracto. Na última coluna, apresenta-se o nome do ficheiro (com extensão .xmg) utilizado na definição do parâmetro.

Parâmetro	Descrição	Valor defeito	Unidade
R_p	Resistência pulmonar	8	Ω cgs
Φ	diferença de fase entre Ag_1 e Ag_2	45	graus
d_1	espessura de m_1	0.25	cm
d_2	espessura de m_2	0.05	cm
l_g	comprimento da glote	1.4	cm
k_1		1.37	
k_2		0.3	
ρ	densidade do ar	1.14×10^{-3}	g/cm^3
μ	viscosidade do ar	1.86×10^{-4}	$dine.s/cm^2$

(b) Parâmetros do modelo de excitação que podem ser ajustados pelo utilizador, editando um ficheiro de configuração do sintetizador. São apresentados os valores utilizados por defeito.

Tabela 4.2: Parâmetros do modelo de fonte glotal implementado.

4.3.5.1 Obtenção dos parâmetros

Embora não tenhamos feito qualquer trabalho no que concerne à obtenção dos parâmetros para o modelo com base no sinal de voz, ou outros sinais como o obtido através de *Electroglotography* (EGG) (Childers, 2000; Mahshie, 1993; Childers e Larar, 1984), referem-se, muito resumidamente, algumas técnicas existente para a sua obtenção.

A obtenção da frequência fundamental tem sido abordada de muitas e variadas formas (Hermes, 1993). Métodos de obtenção de F_0 período a período com elevada precisão permitem obter informação acerca do *jitter*.

Valores para os parâmetros OQ e SQ podem ser obtidos utilizando os picos no sinal EGG (Prado, 1991, pág. 111). Quando não se dispõe de EGG, técnicas de filtragem inversa podem ser utilizadas (Lee, 1992; Childers *et al.*, 1995; Alku *et al.*, 1998; Alku, 1993, 1992).

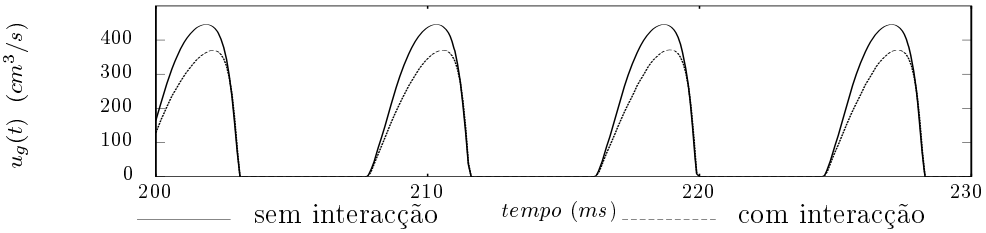


Figura 4.10: Onda de excitação glotal para uma vogal oral, o [i]. Apresentam-se, simultaneamente, os resultados das simulações com e sem interacção. Foram utilizados na simulação $Ag_{max} = 0.3\ cm^2$, $Ag_0 = 0$, $OQ = 60$ % e $SQ = 2$.

Informação acerca das áreas glotais pode ser obtida utilizando medidas directas, como estroboscopia ou métodos indirectos baseados na energia do sinal sonoro (Prado, 1991, pág. 108).

A pressão pulmonar pode ser obtida utilizando equipamentos dedicados (Titze, 1994) ou estimada com base na energia média e frequência fundamental média (Pinto *et al.*, 1989).

4.3.6 Exemplos de utilização da fonte

Apresentam-se, a título exemplificativo das facilidades do modelo, alguns exemplos. Mais exemplos e detalhes acerca da implementação e utilização do modelo da fonte podem ser encontrados em Teixeira e Vaz (1999).

Exemplo 1 - A onda glotal calculada incluindo, e não incluindo, o efeito de carga das cavidades acima glote, para uma vogal oral, o [i], é apresentada na Figura 4.10.

Exemplo 2 - Para a mesma vogal oral, o efeito do não fechamento completo das cordas vocais, $Ag_0 > 0$, é mostrado na Figura 4.11

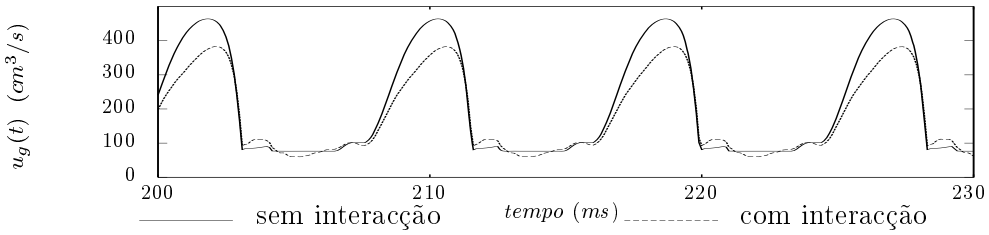


Figura 4.11: Onda de excitação glotal para uma vogal oral com $Ag_0 > 0$. Usou-se $Ag_{max} = 0.3\ cm^2$, $OQ = 60\%$ e $SQ = 2$.

Exemplo 3 - Para a vogal nasal correspondente à utilizada nos exemplos anteriores, pode ver-se, não só o efeito da interacção, como também em que medida o tracto nasal é responsável por esses efeitos. Veja-se a figura 4.12(a).

Pode também ver-se o efeito na frequência, para um período, na Figura 4.12(b). O efeito da interacção torna-se mais notório se se olhar para a derivada de $u_g(t)$, designada por $u_g^{'}(t)$, e representada na Figura 4.12(c).

4.4 Processo de síntese

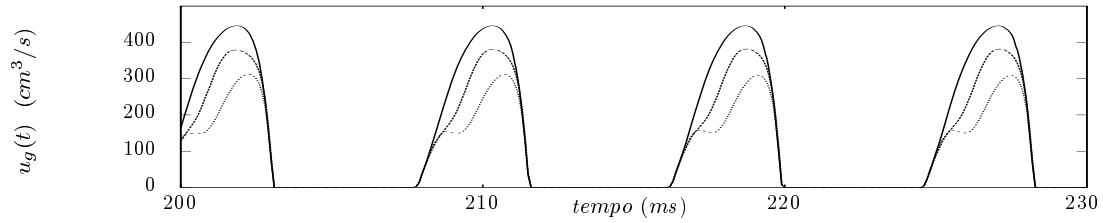
Vejamos como se integram os modelos descritos nas três secções anteriores de forma a se obter o sinal de voz com base em parâmetros articulatórios e parâmetros para a excitação, variáveis no tempo.

Como os articuladores variam no tempo a resposta impulsional do sistema, dada pelo modelo acústico, vai também variar. Idealmente deveria obter-se a resposta para cada instante de amostragem (no nosso caso uma frequência de amostragem de 10 kHz). No entanto isto seria muito demorado em termos computacionais. Aproveitando o facto de os articuladores variarem lentamente, consideramos a forma do tracto "fixa" durante uma período de oscilação das cordas vocais. Apenas se calcula a resposta no início de cada novo período. Caso os articuladores se mantenham fixos não existe a necessidade de recalcular. Desta forma, para a síntese de uma configuração com os articuladores completamente estáticos, apenas se tem de calcular uma única resposta.

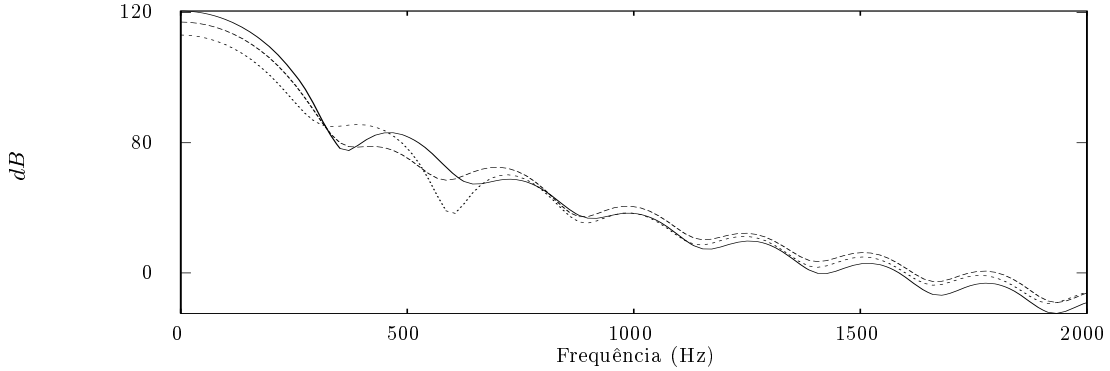
Na implementação efectuada, cada parâmetro articulatório pode ser definido de forma independente para uma maior flexibilidade [11]. O valor dos parâmetros articulatórios entre instantes no tempo em que este está definido é obtido por interpolação linear.

Considerando uma frequência de amostragem $F_s = 1/T_s$, o processo de obtenção do sinal de voz é o seguinte:

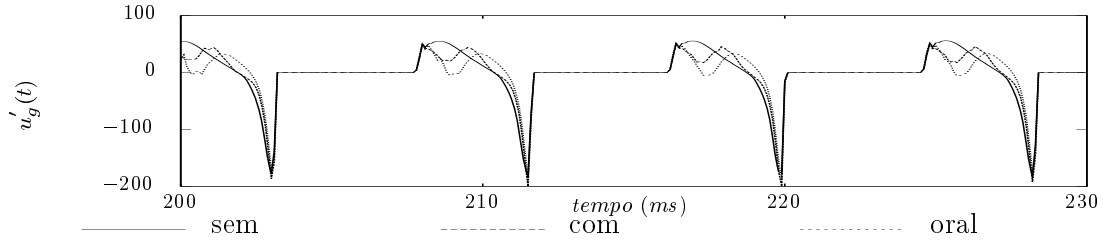
[11] Implementações como o sintetizador ASY, desenvolvido nos Laboratórios Haskins (Rubin *et al.*, 1981), obrigavam à definição de todos os parâmetros para cada instante de tempo necessário à definição do som.



(a) Fluxo glotal.



(b) $20\log(|U_g(f)|)$



(c) Derivada do fluxo glotal.

Figura 4.12: Onda de excitação glotal para uma vogal nasal. Usou-se $Ag_{max} = 0.3\ cm^2$, $Ag_0 = 0$, $OQ = 60\%$ e $SQ = 2$. Apresentam-se três simulações: sem interacção; com interacção total; e interacção considerando apenas a carga das cavidades orais.

1. Para cada novo período da onda de excitação:

   (a) Obter o valor da frequência fundamental (F_0) no instante de tempo, que designaremos de t_i, correspondente ao início de um período de oscilação das cordas vocais. O inverso [12] do valor de F_0 nesse instante será a duração do período, chamemos-lhe

[12] Por razões de implementação, não se utiliza exactamente o inverso, mas sim o valor arredondado para o instante de amostragem mais próximo, para se ter um número de amostras inteiro.

$T_0(t_i)$. Neste intervalo teremos N_i amostras.

O início do período glotal seguinte será $t_{i+1} = t_i + T_0(t_i)$;

(b) Obter, por interpolação, os valores para os parâmetros articulatórios nos instantes de tempo t_i e t_{i+1}. Obter as funções de área usando o modelo articulatório e de seguida as respostas impulsionais, $h(t_i)$ e $h(t_{i+1})$, relativas a estes dois instantes. Obter também as impedâncias de entrada $z_e(t_i)$ e $z_e(t_{i+1})$;

(c) Obter, por interpolação, os valores dos parâmetros do modelo de excitação para o instante de início t_i. Com base nestes valores, calcular as N_i amostras das áreas de abertura das duas massas, usando o modelo paramétrico. Calcula-se de uma só vez todas as amostras correspondentes a um período de excitação.

2. Para cada instante de amostragem nT_s:

(a) Obter a resposta impulsional $h(nT_s)$ por interpolação de $h(t_i)$ e $h(t_{i+1})$

$$h(nT_s) = h(t_i) + \frac{nT_s - t_i}{N_i T_s} \left(h(t_{i+1}) - h(t_i)\right) \ ;$$

(b) Obter $z_e(nT_s)$ por interpolação de $z_e(t_i)$ e $z_e(t_{i+1})$;

(c) Utilizando as áreas glotais, anteriormente calculadas, e $z_e(nT_s)$ obter a excitação $u_g^{'}(nT_s)$. Ao utilizar-se a derivada da onda glotal como excitação, evita-se derivar o sinal final para simular o efeito de radiação;

(d) Convoluir a excitação, $u_g^{'}(nT_s)$, com a resposta impulsional $h(nT_s)$. Obtém-se, assim, uma amostra do sinal de voz, $s(nT_s)$.

Repetindo o processo descrito para todos os períodos obtém-se a totalidade do sinal.

Controlando a forma como se obtém a função de transferência no modelo acústico, processo descrito na secção 4.2.2.4 na página 91, pode obter-se o sinal radiado apenas pela boca, o sinal radiado pelas narinas ou, ainda, o sinal total resultante.

No processo de síntese pode não se incluir a carga do tracto na obtenção do sinal de excitação. Pode ainda não se incluir, no cálculo da impedância de entrada, a impedância de entrada do tracto nasal.

Além do sinal de voz, $s(nT_s)$, é também guardado o sinal de excitação, $u_g(nT_s)$, para possibilitar o estudo da interacção entre a fonte de excitação glotal e as cavidades supraglotais.

4.5 Obtenção dos parâmetros articulatórios do sinal de voz

Neste capítulo, já descrevemos como, dada uma configuração, podemos obter o som respectivo, o designado problema directo.

Precisamos, no entanto, da configuração do tracto, definida pelos parâmetros articulatórios, como entrada. Torna-se portanto necessário ter um processo, ou vários, para obter essas

configurações. Como já foi apresentado no capítulo anterior, este é um problema a que se tem dado muita atenção, mas está longe de se apresentar completamente resolvido. No nosso caso, necessitamos da configuração do tracto durante a pronunciação das vogais, orais e nasais, assim como das várias consoantes nasais existentes em Português. Não tendo tido acesso à utilização de métodos directos, a única solução foi recorrer a um processo de inversão. O processo escolhido foi um método baseado em optimização. Este processo de optimização foi aplicado por Hsieh (1994) a inversão de vogais do Inglês Americano com bons resultados e daí a nossa escolha. É também um processo simples de implementar.

4.5.1 Inversão como um problema de optimização

O problema de inversão pode ser formulado da seguinte forma:

- Dado um vector acústico y, qual é o vector articulatório, x, tal que a distância entre y e $y' = f(x)$ é mínima ?

A inversão consiste num **problema de optimização não-linear com restrições**.

Para o resolver precisa-se de uma técnica de optimização e de uma forma de calcular o erro. Para processo de optimização escolhemos *simulated annealing*. O erro baseia-se em diferenças entre as formantes geradas pelo nosso modelo e valores de formantes extraídas de voz natural. O processo encontra-se representado na Figura 4.14, na página 113.

De seguida, descreve-se o algoritmo de *simulated annealing*, as medidas de erro, o processo de cálculo das formantes do modelo e, finalmente, são apresentados alguns resultados obtidos.

4.5.2 *Simulated Annealing*

O algoritmo *simulated annealing* é uma técnica de optimização que pode: processar funções de custo com graus arbitrários de não linearidade, descontinuidades e estocacidade (em Inglês *estochasticity*); processar condições de fronteira e restrições arbitrárias; ser implementada facilmente relativamente a outras técnicas de optimização; e garantir estatisticamente a obtenção da solução óptima (Ingber, 1993).

Algumas características negativas destes algoritmos são: o processo de optimização poder ser muito demorado e ser difícil fazer o ajuste dos parâmetros do algoritmo em problemas concretos. No entanto, muitos investigadores usam-no pela facilidade com que as restrições e funções complexas podem ser abordadas e codificadas (Ingber, 1993).

O algoritmo baseia-se numa analogia com a termodinâmica, mais especificamente com a forma como os líquidos e os metais arrefecem e cristalizam (em Inglês *cool and anneal*). A temperaturas elevadas, as moléculas dos líquidos movem-se livremente. Se o líquido é arrefecido lentamente, a mobilidade é perdida. Os átomos são geralmente capazes de se alinharem por forma a formarem um cristal. Este cristal é o estado de energia mínima para o sistema. O

facto extraordinário é que, para sistemas arrefecidos lentamente, a natureza é capaz de atingir o estado de energia mínima. Se o metal líquido é arrefecido com demasiada rapidez, situação designada em Inglês por *quenching*, não é atingido o estado de energia mínima quedando-se num estado poli-cristalino ou amorfo com energia mais elevada.

A essência do processo é o arrefecimento lento, dando tempo suficiente para a redistribuição dos átomos, à medida que eles perdem mobilidade (Press *et al.*, 1992, pág. 444).

O algoritmo foi usado em muitas disciplinas científicas. Os primeiros problemas abordados foram o problema do caixeiro-viajante e a optimização de interligações em circuitos integrados (Kirkpatrick *et al.*, 1983). Desde essa altura foi aplicado em Matemática, para problemas de grafos; análise de dados; processamento de imagem, nomeadamente reconstrução e filtragem; redes neuronais; Biologia; Física; Geofísica; Finanças; e aplicações militares (Ingber, 1993).

4.5.2.1 Origem do algoritmo

Foi proposto por Metropolis *et al.* (1953) um método para calcular o equilíbrio de um conjunto de partículas num banho quente, usando um método de simulação por computador. Para o sistema em equilíbrio térmico a uma dada temperatura T, assumiram que a probabilidade $\pi_T(c)$ de o sistema se encontrar numa dada configuração c depende da energia $E(c)$ da configuração e segue uma distribuição de Boltzmann

$$\pi_T(c) = \frac{e^{-\frac{E(c)}{kT}}}{\sum\limits_{s \in C} e^{-\frac{E(s)}{kT}}} \ , \tag{4.45}$$

onde k é a constante de Boltzmann e C o conjunto de todas as configurações possíveis. A configuração do sistema é dada pela posição espacial das partículas. Foi desenvolvida uma técnica de relaxação estocástica para simular o comportamento do sistema. Estando o sistema na configuração C_t no instante de tempo t, é gerada aleatoriamente uma configuração candidata C_n para o sistema no instante $t+1$. O critério para aceitar ou rejeitar C_n como nova configuração depende da diferença de energia entre C_n e C_t. Definindo p como o quociente entre a probabilidade do sistema se encontrar na configuração C_n e a probabilidade de se encontrar na configuração C_t como

$$p = \frac{\pi_T(C_n)}{\pi_T(C_t)} = e^{-\frac{E(C_n)-E(C_t)}{kT}} \ , \tag{4.46}$$

aplica-se de seguida um critério que ficou conhecido como critério (ou algoritmo) de Metropolis, para decidir acerca da aceitação de C_n.

Critério de Metropolis : Se $p > 1$, isto é, se a energia de C_n é inferior à energia de C_t, então a configuração C_n é automaticamente aceite como a nova configuração para $t + 1$. Se $p \leq 1$, a nova configuração é aceite com probabilidade p. É portanto aceite a transição para

estados de energia mais elevada, mas de uma forma limitada.

Repetindo o processo para um número suficientemente grande de movimentos, independentemente da configuração inicial, demonstra-se que a configurações geradas convergem para a distribuição de Boltzmann.

4.5.2.2 Arrefecimento

Uma questão fundamental surge em mecânica estatística quando o sistema se aproxima de baixas temperaturas. Para atingir uma configuração cristalina, com baixa energia, ir reduzindo a temperatura não é suficiente. É necessário um método de arrefecimento [13], no qual a temperatura do sistema é aumentada, e depois gradualmente reduzida, permanecendo tempo suficiente em cada temperatura para garantir que é atingido o equilíbrio termodinâmico. Se o sistema permanecer tempo insuficiente a cada temperatura, em especial nas baixas temperaturas, então a probabilidade de se atingir um estado cristalino, de baixa energia, é grandemente reduzida.

4.5.2.3 Aplicação do algoritmo em problemas de optimização

Para utilizar o algoritmo num problema concreto de optimização é necessário:

1. Identificar os análogos dos conceitos físicos. A função de energia torna-se a função custo. A configuração das partículas passa a ser a combinação das variáveis independentes. A modificação da configuração das partículas passa a ser o melhoramento iterativo da função custo pela alteração de valores das variáveis independentes. A configuração de baixa energia equivale a uma solução quase-óptima e a temperatura torna-se o parâmetro de controlo de todo o processo;

2. Um processo de gerar os novos estados candidatos. Geralmente são gerados usando um processo aleatório;

3. Um processo de seleccionar um novo estado. A utilização de um processo de selecção probabilístico permite subidas de energia ocasionais. Exemplos de processo de selecção utilizados com sucesso são a máquina de Boltzmann e o algoritmo de Metropolis.

   Máquina de Boltzmann : $p(\Delta E) = \frac{1}{1+e^{-\frac{\Delta E}{T}}}$

   Critério de Metropolis : $p(\Delta E) = \begin{cases} 1.0 & \Delta E \leq 0 \\ e^{-\frac{\Delta E}{T}} & \Delta E > 0 \end{cases}$

   onde $\Delta E = E(\text{novo estado}) - E(\text{estado actual})$ é a diferença de energia entre o novo e o estado actual. A máquina de Boltzmann aproxima melhor a metáfora física, mas é

[13] em Inglês *cooling schedule*.

mais exigente em termos computacionais. No nosso trabalho utilizamos o algoritmo de Metropolis;

4. Um processo de arrefecimento composto por:
   (a) Valor inicial do parâmetro de controlo, isto é, o valor inicial para a temperatura artificial T;
   (b) Uma função para o decréscimo do valor do parâmetro de controlo, também designada por taxa de arrefecimento;
   (c) Valor final para o parâmetro de controlo que funciona como critério de paragem;
   (d) Número de movimentos para cada valor do parâmetro de controlo, ou seja, tempo passado a cada temperatura.

4.5.2.4 Extensão do algoritmo para variáveis contínuas

Devido ao sucesso do algoritmo em problemas de optimização com variáveis discretas, foi estudado, por diversos investigadores, o seu potencial para problemas envolvendo variáveis contínuas. Uma das implementações, proposta por Corana *et al.* (1987), oferece uma boa combinação de facilidade de aplicação e de robustez, tendo sido utilizada em econometria e na obtenção de configurações articulatórias, para vogais do Inglês Americano, com base em formantes de fala natural por Hsieh (1994).

4.5.2.5 Algoritmo de Corana *et al.* (1987)

O algoritmo proposto por Corana *et al.* (1987) encontra-se representando na Figura 4.13.

Seja $\vec{x}$ um vector com M dimensões e componentes $[x_1, x_2, \ldots, x_M]$. Seja $\varepsilon(\vec{x})$ a função custo, e $li_j \leq x_j \leq ls_j$, $j = 1, \ldots, M$, as M variáveis com os correspondentes limites inferior e superior.

Primeiro, é avaliada a função de custo para o ponto inicial $\vec{x}$ e o valor de ε guardado. De seguida, é gerado um novo candidato, $\vec{x_n}$, variando o elemento i de $\vec{x}$

$$\vec{x_{n}}_i = x_i + r \times v_i \ . \tag{4.47}$$

A variável r é um número aleatório gerado por uma distribuição uniforme entre -1 e 1 e v_i é o elemento i do vector $\vec{v}$, o vector passo. A nova função ε_n é depois calculada. Se ε_n é inferior a ε, $\vec{x}$ é aceite, $\vec{x}$ toma o valor de $\vec{x_n}$, ε passa a ter o valor de ε_n e a procura prossegue. Se ε_n é maior ou igual, a probabilidade $p = e^{(\varepsilon - \varepsilon_n)/T}$ é utilizada para decidir, usando o critério de Metropolis. A probabilidade de um movimento implicando aumento de energia diminui com a diminuição da temperatura T.

Depois de N_S passagens por todos os elementos de $\vec{x}$, o vector $\vec{v}$ é ajustado por forma a que 50% de todos os movimentos sejam aceites. O objectivo á fazer o algoritmo acompanhar a

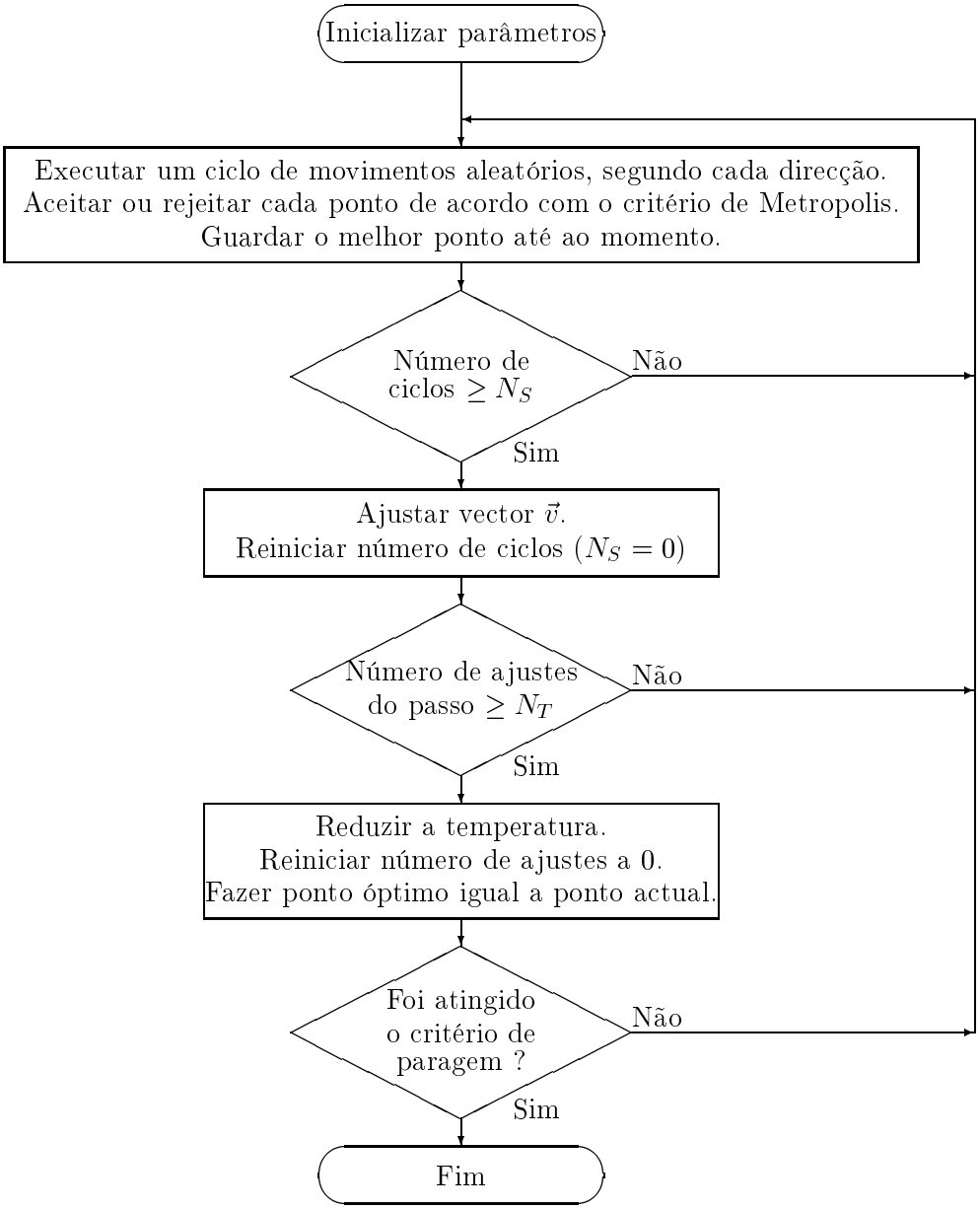


Figura 4.13: Algoritmo de *Simulated Annealing* proposto por Corana *et al.* (1987).

função de custo (Corana *et al.*, 1987). Uma maior percentagem de pontos aceite significa que os candidatos estão demasiado perto do ponto actual. Neste caso, o vector $\vec{v}$ é aumentado. No caso oposto, uma grande percentagem de pontos rejeitados, os pontos candidatos encontram-se demasiado longe. O comprimento do passo é reduzido.

Depois de N_T passagens pelos ciclos anteriores, atinge-se o equivalente ao equilíbrio térmico, e a temperatura, T, é reduzida com base na seguinte equação:

$$T_n = r_T \times T \ . \tag{4.48}$$

O coeficiente de redução r_T toma valores entre zero e um. O ponto inicial na nova temperatura é o ponto óptimo obtido na temperatura anterior. A procura começa assim no ponto mais favorável. Como a temperatura caracteriza a grau de "excitação" do sistema, a temperaturas mais baixas diminui o número de movimentos com aumento de energia, o número de rejeições aumenta e o passo diminui. A temperatura mais baixa, e o consequente passo mais pequeno, reduzem o espaço de procura e concentram a procura.

O algoritmo inclui também critérios de paragem. É verificado se não foram efectuados movimentos significativos nas ultimas N_ε temperaturas. Assumindo que o valor óptimo obtido na temperatura T_k foi ε_k^*, e sendo ε_{opt} o valor óptimo actual, à temperatura T_{k+1}, se

$$|\varepsilon_k^* - \varepsilon_{k-m}^*| \leq \eta, \quad m = 1, \ldots, N_\varepsilon \tag{4.49}$$

$$|\varepsilon_k^* - \varepsilon_{opt}| \leq \eta \ , \tag{4.50}$$

acaba a procura. η é uma constante com um valor baixo. Outro critério de paragem utilizado é o de parar a procura quando o número total de vezes que foi calculada a função de custo atinge um valor máximo previamente fixado, N_{tot}.

Resumindo, o processo começa a uma temperatura elevada especificada pelo utilizador. É gerada uma sequência de pontos até se atingir o equilíbrio. Durante este processo aleatório, o comprimento do passo é periodicamente ajustado por forma a seguir melhor a função de custo. Depois de atingido o equilíbrio, é reduzida a temperatura e o processo anterior repetido. O processo termina numa temperatura baixa, quando já não é possível efectuar melhorias úteis.

4.5.2.6 Aplicação do algoritmo de Corana *et al.* (1987) na inversão de vogais orais

Em termos gerais, a relação entre a forma do tracto e o seu resultado acústico pode ser representado por uma função multidimensional de um argumento multidimensional, $\vec{y} = f(\vec{x})$, onde $\vec{x}$ é formado pelos parâmetros articulatórios e $\vec{y}$ é o vector formado por propriedades acústicas que lhe correspondem, sendo $f()$ a função que relaciona os dois vectores. Sendo dado $\vec{y}_d$, o problema é o de obter os parâmetros articulatórios $\vec{x}_o$ tal que $f(\vec{x}_o)$ seja a melhor aproximação para $\vec{y}_d$.

Os parâmetros articulatórios, já descritos neste capítulo, constituem o vector $\vec{x}$. Na implementação efectuada o utilizador pode decidir quais os parâmetros a optimizar. Os outros

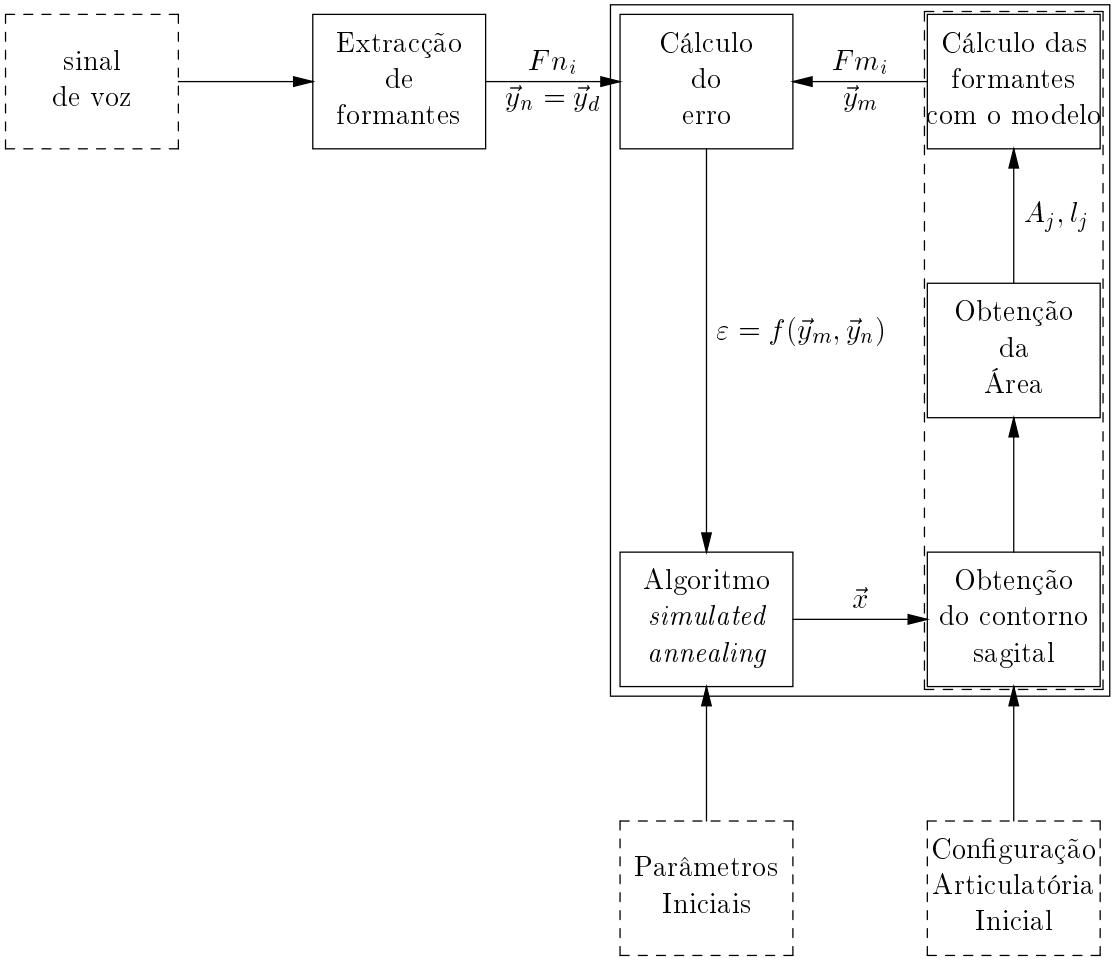


Figura 4.14: Diagrama de blocos do processo de inversão utilizando *simulated annealing*. O bloco a tracejado agrupa as fases do processo relacionadas com o chamado problema directo, a obtenção das formantes dada a posição dos articuladores.

mantêm o seu valor durante a optimização. No máximo teremos 11 componentes [14], ficando,

$$\vec{x} = [\text{tbodx tbody ttx tty jaw lipp lipo hyoid wh hk1 g1k}] \ . \tag{4.51}$$

O valor inicial para $\vec{x}$, constituindo a configuração inicial do tracto, pode ser a configuração neutra, uma configuração aleatória ou outra configuração definida pelo utilizador, recorrendo à interface do sintetizador.

O vector acústico é composto pelas primeiras N formantes

$$\vec{y} = [F_1 \ F_2 \ \ldots \ F_N] \ . \tag{4.52}$$

Geralmente, apenas se utiliza até $N = 4$, ou seja quatro ou menos formantes.

[14] O velo mantém-se sempre a zero, por ainda só estar desenvolvido o processo de inversão para vogais orais.

	Parâmetro	Valor por defeito
T	temperatura (parâmetro de controlo)	$0.1 - 0.2$ graus
r_T	coeficiente de redução da temperatura	0.85
N_S	número de repetições até ajustar o passo	20
N_T	número de ajustamentos a cada temperatura	5
N_ε	número de temperaturas usadas para critério paragem	4
η	critério de paragem	0.005
N_{tot}	número total avaliações da medida de erro	5001
v_i	comprimento do passo, $i = 1, 2, \ldots, M$	3.0

Tabela 4.3: Valores por defeito para os parâmetros do algoritmo de *simulated annealing*. Valores propostos por Hsieh (1994).

A função custo (medida de erro) é obtida através da comparação efectuada entre as formantes obtidas através do modelo articulatório com as formantes obtidas de voz natural. Foram implementadas várias formas de cálculo da medida de erro, todas com base nas formantes:

Distância Euclidiana pesada utilizada por Hsieh (1994)

$$\varepsilon = \sum_{i=1}^{N} \frac{w_i |Fm_i(\vec{x}) - Fn_i|}{Fn_i} \ \% \ ; \tag{4.53}$$

Diferença na escala Bark utilizada por Båvegård e Fant (1995)

$$\varepsilon = \sqrt{\sum_{i=1}^{N} \left[Bark(Fm_i(\vec{x})) - Bark(Fn_i) \right]^2} \ ; \tag{4.54}$$

Métricas uniformes propostas por Sorokin (1992)

$$\varepsilon = \max \frac{Fm_i(\vec{x}) - Fn_i}{Fn_i} \ , \tag{4.55}$$

onde Fm_i é a formante índice i obtida com base no modelo e Fn_i o valor, para a formante i, estimado por análise de sinal de voz natural e w_i é um factor multiplicativo. O número de formantes utilizadas pode ser escolhido entre 1 e 4.

As restrições, referidas aquando da descrição do algoritmo, são garantidas, utilizando-se o modelo articulatório para eliminar configurações articulatórias impossíveis.

O mínimo ideal para o erro, ε, é de 0 %, mas devido às várias aproximações utilizadas nos modelos articulatório e acústico um valor de cerca de 1 % é adequado.

Na Tabela 4.3 apresentam-se os parâmetros e os seus valores por defeito, utilizados pelo algoritmo. Através da interface gráfica é possível alterar quase a totalidade destes valores.

4.5.3 Obtenção das formantes do sinal de voz natural

Foram utilizadas duas técnicas distintas para obtenção das formantes de sinal natural, utilizadas no processo de inversão.

A primeira técnica, baseada em análise LPC, consistiu na utilização do programa **formant**, comercializado pela Entropics (Entropic, 1993b,a,c). Foi utilizada pre-ênfase, janela rectangular com 250 amostras, 100 janelas por segundo, método de covariância com ordem 12, seguido da determinação das raízes do polinómio e programação dinâmica.

Utilizou-se também o algoritmo *Weighted RLS with Variable Forgetting Factor* (WRLS-VFF) (Childers *et al.*, 1995; Ting, 1994; Lee, 1992). De referir que este método permite a inclusão de zeros, possibilitando a extracção de antiformantes. Oferece portanto capacidades para utilização em futuras extensões do processo de inversão aos sons nasais.

Os valores obtidos pelas duas técnicas são semelhantes. Exemplos de valores obtidos para 9 vogais orais do Português e a sua utilização na inversão de vogais foram apresentados em Teixeira *et al.* (1997c).

4.5.4 Obtenção das singularidades do som sintético

Para o processo de inversão torna-se necessário obter as singularidades, formantes e antiformantes. Apesar de, no estado actual, a inversão se aplicar apenas a sons não nasais, optámos por utilizar um processo que permitisse futuras extensões da inversão. Para tal torna-se necessário obter não só informação dos pólos (frequência e largura de banda), como também dos zeros. A utilização de uma forma alternativa, à anteriormente apresentada utilizando matrizes, de cálculo da resposta em frequência dos tractos, permite separar os pólos dos zeros. É esta técnica e a forma de obtenção a partir da resposta em frequência dos pólos e zeros que se apresenta de seguida.

4.5.4.1 Função de transferência de uma secção

Como base para o cálculo da resposta total, torna-se necessário saber como tratar o caso elementar de apenas um tubo.

Considere-se primeiro uma secção cilíndrica de comprimento l_i e área A_i. O circuito em T da Figura 4.15 é definido pelos seus elementos série e paralelo:

$$Z_a = Z_i \tanh\left(\frac{\Gamma_i}{2}\right) , \tag{4.56}$$

$$Z_b = \frac{Z_i}{\sinh\left(\Gamma_i\right)} . \tag{4.57}$$

A constante de transferência, $\Gamma_i = \gamma_i l_i$, e a impedância característica, Z_i, da secção estão relacionadas com os parâmetros distribuidor $RLGC$ e o comprimento da secção pelas expressões

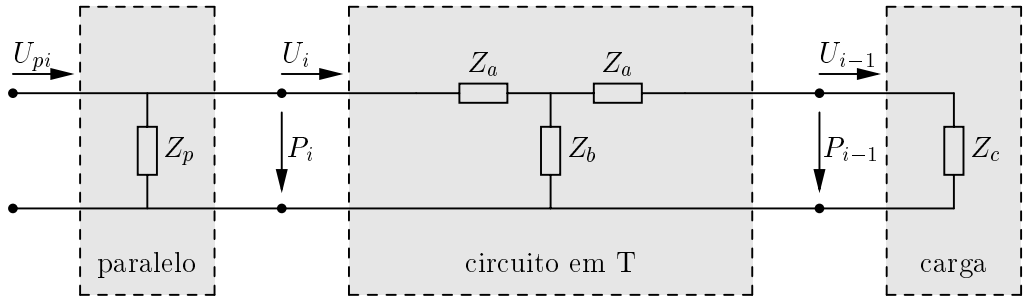


Figura 4.15: Circuito em T com carga Z_c e impedância em paralelo Z_p na entrada. Adaptado de (Lin, 1990, Figura 2.3)

apresentadas na secção 3.4.3.3 do capítulo anterior, com início na página 62.

As relações entre a entrada e saída de uma secção elementar podem ser escritas da seguinte forma:

$$P_i = P_{i-1}\cosh(\Gamma_i) + U_{i-1}Z_i\sinh(\Gamma_i) \ , \tag{4.58}$$

$$U_i = \frac{P_{i-1}}{Z_i}\sinh(\Gamma_i) + U_{i-1}\cosh(\Gamma_i) \ , \tag{4.59}$$

onde P_i e P_{i-1} representam, respectivamente, as pressões à entrada e saída da secção. U_i e U_{i-1} representam os fluxos correspondentes.

Evitando usar matrizes, podem usar-se as seguintes relações (Badin e Fant, 1984; Lin, 1990):

$$\frac{U_i}{U_{i-1}} = \cosh(\Gamma_i) + \frac{Z_{e_{i-1}}}{Z_i}\sinh(\Gamma_i) \ . \tag{4.60}$$

O quociente U_{i-1}/U_i é denominado de função de transferência da secção i. O termo $Z_{e_{i-1}}$ representa a carga vista na direcção dos lábios que é a impedância de entrada da secção $i-1$.

A impedância vista da entrada de cada secção, Z_{e_i}, é dada por

$$Z_{e_i} = \frac{P_i}{U_i} = Z_i\frac{\cosh\Gamma_i - \frac{U_{i-1}}{U_i}}{\sinh\Gamma_i}. \tag{4.61}$$

Dois casos especiais podem aparecer no que concerne a $Z_{e_{i-1}}$. No caso de considerar a impedância de radiação nula teremos Z_a em série como o paralelo de Z_a com Z_b, a impedância de entrada da secção será

$$Z_{e_i} = Z_a + \frac{Z_a Z_b}{Z_a + Z_b} = Z_i\tanh(\Gamma_i)\,. \tag{4.62}$$

Outro caso especial resulta da impedância de radiação ser infinita, facto que ocorre quando existe oclusão, como no caso das consoantes nasais. Teremos Z_a em série com Z_b, a impedância

de entrada da secção será

$$Z_{e_i} = Z_a + Z_b = Z_i \coth(\Gamma_i)\,. \tag{4.63}$$

4.5.4.2 Obtenção da função de transferência global

Baseando-nos na função de transferência de uma secção uniforme, a função de transferência global é facilmente determinada por passos sucessivos. Analisemos primeiro o caso mais simples das vogais. O processo começa com a impedância de radiação Z_r. O recíproco da função de transferência da primeira secção é (ver eq. 4.60)

$$\frac{U_1}{U_0} = \left(\frac{U_0}{U_1}\right)^{-1} = \cosh(\Gamma_1) + \frac{Z_r}{Z_1}\sinh(\Gamma_1)\ , \tag{4.64}$$

em que U_o é o fluxo nos lábios ou outro ponto de radiação, como as narinas. A seguir obtém-se a impedância de entrada da secção:

$$Z_{e_1} = Z_1 \frac{\cosh\Gamma_1 - \frac{U_0}{U_1}}{\sinh\Gamma_1} = \frac{P_1}{U_1}. \tag{4.65}$$

A impedância é utilizada para o cálculo da segunda secção:

$$\frac{U_2}{U_1} = \cosh(\Gamma_2) + \frac{Z_{e_1}}{Z_2}\sinh(\Gamma_2)\ . \tag{4.66}$$

Este processo é repetido até se obter U_g. A Função de transferência obtém-se multiplicando a função de transferência de todas as secções:

$$H(w) = \frac{U_o}{U_g} = \frac{U_{g-1}}{U_g}\cdots\frac{U_{i-1}}{U_i}\cdots\frac{U_1}{U_2}\times\frac{U_o}{U_1}\ . \tag{4.67}$$

Neste caso, $H(w)$ apenas contém pólos, dado que U_i/U_{i-1} apenas tem zeros.

4.5.4.3 Separação dos pólos e zeros em circuitos paralelos

Quando se considera o sistema subglotal ou o tracto nasal, a função de transferência passa a conter não só pólos como também zeros. Neste caso, as duas partes devem ser cuidadosamente separadas se se pretender calcular as frequências das formantes e respectivas larguras de bandas. Caso contrário, os pólos podem ser contaminados e tornar-se impossível a sua determinação. Através de um algoritmo simples pode-se efectuar a decomposição.

Assumindo que a impedância a colocar em paralelo à entrada da secção i é $Z_p = Z_{z,p}/Z_{p,p}$, isto é, Z_p tem numerador $Z_{z,p}$ e denominador $Z_{p,p}$. O primeiro índice, z e p, representa os zeros e os pólos respectivamente.

A função de transferência é dada por

$$\frac{U_i}{U_{pi}} = \frac{Z_p}{Z_p + Z_{e_i}} = \frac{Z_{z,p}}{Z_{z,p} + Z_{p,p} \times Z_{e_i}} \ . \tag{4.68}$$

A impedância de entrada correspondente, Z_{pi}, é o paralelo de Z_p e Z_{e_i}

$$Z_{pi} = \frac{Z_p \times Z_{e_i}}{Z_p + Z_{e_i}} \ . \tag{4.69}$$

A cada passagem por uma impedância em paralelo resulta uma contribuição para o numerador de $Z_{z,p}$, o numerador da impedância, e uma contribuição para o denominador de $Z_{z,p} + Z_{p,p} \times Z_{e_i}$.

A função de transferência completa, com separação dos pólos e zeros, será:

$$H(w) = \frac{H_z(w)}{H_p(w)} \ , \tag{4.70}$$

em que

$$H_p(w) = \frac{U_g}{U_{g-1}} \ldots (Z_{z,p} + Z_{p,p} Z_{e_i}) \frac{U_i}{U_{i-1}} \ldots \frac{U_1}{U_0} \tag{4.71}$$

$$H_z(w) = \prod_{m=1}^{m=M} Z_{p,p_m} \ , \tag{4.72}$$

sendo M o número de impedâncias em paralelo.

Claro que é necessário decompor inicialmente a impedância Z_p em numerador e denominador.

4.5.4.4 Separação de pólos e zeros nos casos de radiação em vários pontos

A discussão anterior pressupôs que a radiação ocorria apenas nos lábios. Em tal sistema, os zeros são apenas causados pelos circuitos adicionais em paralelo. Se existirem vários pontos de radiação, outros zeros são criados ao serem adicionadas as várias radiações. Sabemos que a radiação pode ter lugar nos lábios, nas narinas ou em ambos. A energia pode ainda ser radiada pelas paredes do tracto. A seguir, descreve-se o procedimento para separar os pólos e zeros no caso de existir radiação, simultaneamente, pelas narinas e lábios. Considera-se o tracto nasal representado por um único tubo. Para o caso de se considerarem os dois tubos e as duas narinas, temos um caso semelhante na zona em que as duas cavidades se unem na região da nasofaringe.

Seja U_f o fluxo antes da divisão, proveniente da faringe, o fluxo que entra para a cavidade oral U_{fo}, e o fluxo que entra na cavidade nasal U_{fn}. A figura 4.16 representa esquematicamente a bifurcação. Seguindo o processo já descrito, as funções de transferência para as cavidades

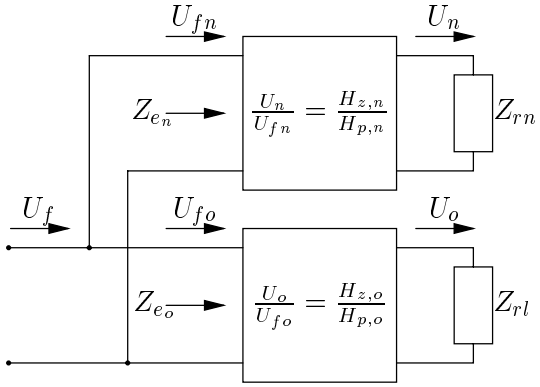


Figura 4.16: Soma da energia radiada em diferentes pontos.

orais e nasais, com separação de pólos e zeros, são calculadas como se segue:

$$\frac{U_o}{U_{fo}} = \frac{H_{z,o}}{H_{p,o}} , \tag{4.73}$$

$$\frac{U_n}{U_{fn}} = \frac{H_{z,n}}{H_{p,n}} . \tag{4.74}$$

Se não existirem impedâncias em paralelo, nas cavidades oral e nasal, as funções $H_{z,o}$ e $H_{p,o}$ reduzem-se à unidade, caso contrário são funções complexas da frequência. Caminhando em direcção à glote obtém-se

$$\frac{U_f}{U_{fo}} = \frac{Z_{e_{oral}} + Z_{e_{nasal}}}{Z_{e_{nasal}}} , \tag{4.75}$$

em que $Z_{e_{oral}} = Z_{z,o}/Z_{p,o}$ é a impedância de entrada da cavidade oral, e $Z_{e_{nasal}} = Z_{z,n}/Z_{p,n}$ é a impedância de entrada da cavidade nasal. Substituindo em 4.75 obtém-se

$$\frac{U_f}{U_{fo}} = \frac{Z_{p,n}Z_{z,o} + Z_{p,o}Z_{z,n}}{Z_{z,n}Z_{p,o}} , \tag{4.76}$$

Portanto

$$\frac{U_o}{U_f} = \left(\frac{U_o}{U_{fo}}\right)\left(\frac{U_{fo}}{U_f}\right) = \frac{Z_{z,n}H_{z,o}}{Z_{p,n}Z_{z,o} + Z_{p,o}Z_{z,n}} . \tag{4.77}$$

Porque o denominador da função de transferência é igual ao da impedância de entrada $H_{p,o} = Z_{p,o}$, estes cancelam-se na equação 4.77.

De uma forma semelhante se obtém

$$\frac{U_n}{U_f} = \left(\frac{U_n}{U_{fn}}\right)\left(\frac{U_{fn}}{U_f}\right) = \frac{Z_{z,o}H_{z,n}}{Z_{p,n}Z_{z,o} + Z_{p,o}Z_{z,n}} . \tag{4.78}$$

Note-se que ambas as equações 4.77 e 4.78 têm o mesmo denominador, como previsto. Finalmente, somando a energia radiada nos dois pontos teremos

$$\frac{U_n + U_o}{U_f} = \frac{Z_{z,o}H_{z,n} + Z_{z,n}H_{z,o}}{Z_{p,n}Z_{z,o} + Z_{p,o}Z_{z,n}} \ . \tag{4.79}$$

Depois de passar pela junção o processo iterativo já descrito é prosseguido até termos a função de transferência completa.

4.5.4.5 Obtenção das singularidades

Aplicando o exposto nas secções anteriores obtém-se

$$H(\omega) = \frac{A_b(\omega) + jA_a(\omega)}{N_b(\omega) + jN_a(\omega)}, \tag{4.80}$$

em que,

$$H_z(\omega) = A_b(\omega) + jA_a(\omega), \tag{4.81}$$

$$H_p(\omega) = N_b(\omega) + jN_a(\omega), \tag{4.82}$$

são o numerador e denominador de $H(s)$, respectivamente. Em geral ambos são funções complexas da frequência. O que se pretende é obter as frequências em que $|1/H_p(\omega)|$ tem um máximo, pólo, e frequências em que $|H_z(\omega)|$ tem um mínimo, zero. Os pólos e zeros podem não coincidir com os picos e vales da função composta $|H(\omega)|$ quando pólos e zeros se encontram próximos.

Existem dois processos iterativos no cálculo da resposta do tracto vocal. Para cada frequência, a função de transferência é obtida iterando desde o ponto, ou pontos, de radiação para trás até se atingir a fonte de excitação. Fonte que no caso dos sons sonoros se encontra situada na glote. A frequência é incrementada, de um valor dado pelo incremento na frequência, e a função de transferência é calculada para essa nova frequência. O processo repete-se até se ter varrido toda a gama de frequência de interesse. A maioria dos cálculos envolvem números complexos, tornando-os demorados.

De entre os vários métodos existentes para obter as singularidades (Lin, 1990), optamos pelo método rápido proposto por Lin (1992). Este método explora o facto de que para perdas reduzidas os resultados, usando um modelo com perdas, não diferem muito dos obtidos não considerando as perdas. Os cálculos tornam-se muito mais rápidos, segundo Lin pelo menos 100%, devido a passarmos a ter números reais. Uma vantagem adicional, de não incluirmos as perdas, é a de que se torna mais fácil detectar formantes bastante atenuadas.

A função de transferência $H(\omega)$ é, em geral, uma quantidade complexa. Depois de remover todos os elementos dissipativos, torna-se real. $H_z(\omega)$ e $H_p(\omega)$ também se reduzem a números reais A_b e N_b, respectivamente. Fazendo $N_b = 0$, é fácil calcular os pólos de $H_p(\omega)$ como

se segue. Primeiro, os intervalos, onde N_b muda de sinal, são determinados. O incremento deve ser tal que apenas exista uma raiz em cada intervalo. Para vogais o passo pode ser até 100 Hz. A raiz sem perdas pode ser determinada usando o método de Newton, no intervalo onde foi detectada mudança de sinal.

Sejam as raízes sem perdas representadas por $\overline{\omega}_n$. Quando a raiz $\overline{\omega}_1$ é obtida, passa-se a fazer os cálculos incluindo as perdas, calculando-se a função com perdas em $\overline{\omega}_1$. Agora $H(\omega)$, $H_z(\omega)$ e $H_p(\omega)$ são complexos. Assumindo que $H_p(\omega) = N_b + jN_a$, a raiz ω_1 da função complexa é obtida por interpolação de $\overline{\omega}_1$. Note-se que N_b e N_a são números reais e que N_b, no caso com perdas, difere do valor obtido para o caso sem perdas.

Devido às perdas, as ressonâncias do sistema com perda desviam-se das do sistema sem perdas. O desvio na frequência pode ser obtido usando (Lin, 1990):

$$\Delta\omega_1 = -\frac{N_b \cdot N_b^{'} + N_a \cdot N_a^{'}}{N_a^{'2} + N_b^{'2}}, \tag{4.83}$$

obtendo-se a nova frequência

$$\omega_1^{(i+1)} = \omega_1^{(i)} + \Delta\omega_1^{(i)}, \tag{4.84}$$

em que $\omega_1^{(1)} = \overline{\omega}_1$.

A equação 4.83 é recalculada para a nova frequência $\omega_1^{(i+1)}$ até que $\Delta\omega_1^{(i+1)} < 2\pi f_{resol}$, obtendo-se uma resolução de f_{resol} Hz. Usualmente é utilizada uma resolução de 1 Hz. O algoritmo converge em poucas iterações. Os pólos de ordem mais elevada, assim como os zeros, são calculados de forma semelhante. Mais detalhes podem ser encontrados em Lin (1992) e Branco (1997).

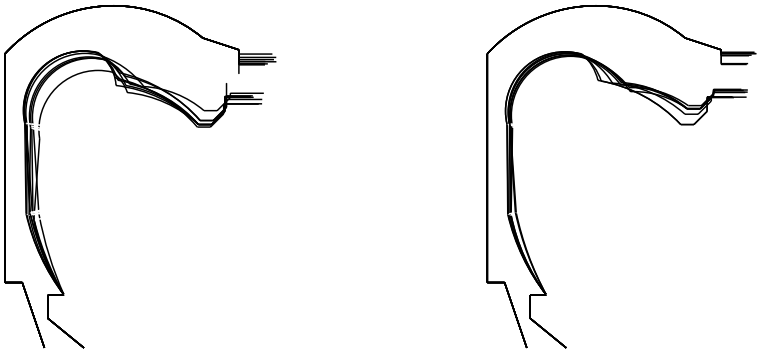
A parte real do pólo pode ser obtida empregando (Lin, 1995)

$$\sigma_n = -B_n \cdot \pi = \frac{N_a \cdot N_b^{'} + N_b \cdot N_a^{'}}{N_a^{'2} + N_b^{'2}}\ , \tag{4.85}$$

em que B_n é a largura de banda da formante n.

4.5.5 Resultados do processo de inversão

Para aferir as capacidades do processo de inversão utilizando o algoritmo de *simulated annealing* realizamos dois tipos de testes. Primeiro, testámos as capacidades do método com formantes geradas pelo próprio modelo. A segunda fase contemplou a situação com interesse prático, testando-se o método com formantes obtidas de sinal de voz natural.



(a) Resultados para várias repetições do processo de inversão.

(b) Utilizando as três métricas.

Figura 4.17: Configurações para a vogal [a], obtidas pelo processo de inversão, com $F_1 = 624$, $F_2 = 1316$, $F_3 = 2432$ e $F_4 = 3475$.

4.5.5.1 Testes com formantes geradas pelo modelo acústico

Devido aos múltiplos graus de liberdade do sistema (medida de erro, modelo acústico, modelo de radiação, inicialização, parâmetros do algoritmo de *simulated annealing*, ...) não foi realizada uma avaliação exaustiva do processo de optimização. Apenas se pretendeu aferir se usando parâmetros recomendados por outros investigadores (Hsieh, 1994) se consegue obter configurações próximas das utilizadas na obtenção dos parâmetros acústicos. Pretendia-se também ter uma ideia do desempenho das diferentes medidas de erro e métodos de inicialização. Foi também testado se o método de inversão implementado dá geralmente os mesmos resultados finais.

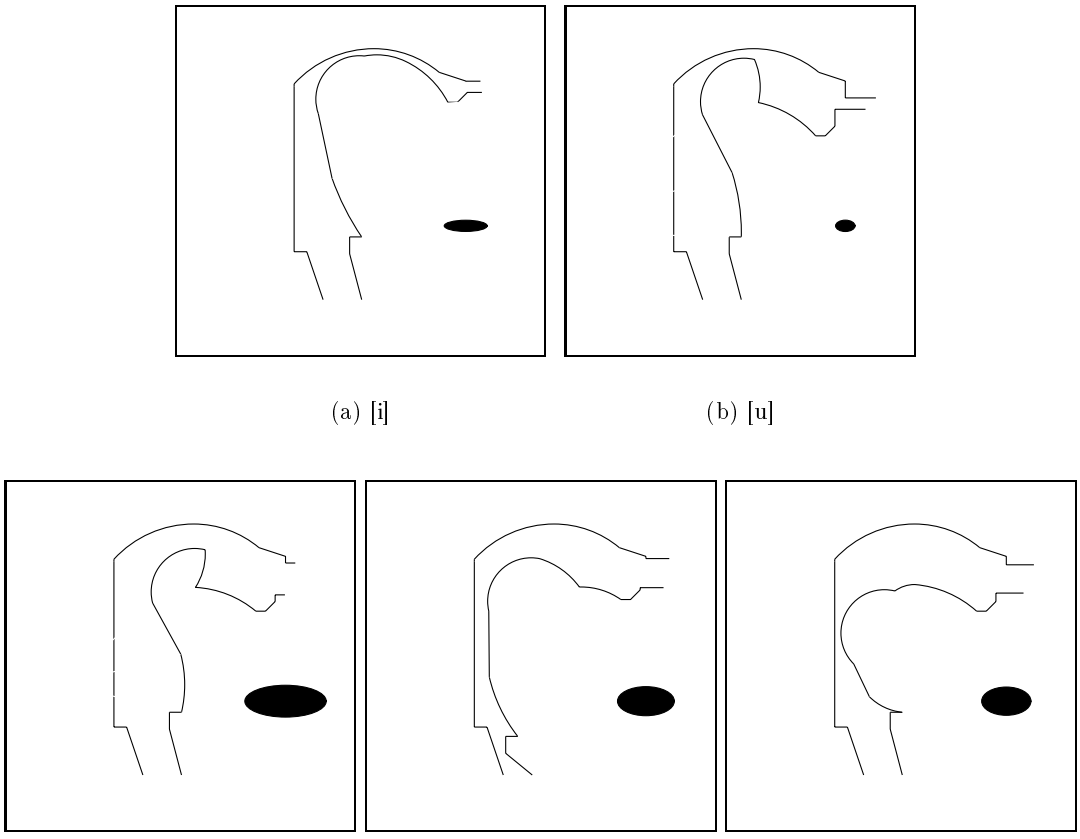
Os resultados obtidos, com 5000 iterações, mostraram que o processo é capaz de chegar a configurações muito próximas das originais. A diferença entre as formantes da configuração obtida por inversão e a configuração original consegue ser inferior a 1 %, ficando todas as diferenças entre formantes individuais abaixo da *Just Noticeable Difference* (JND). Também a posição e a área de máxima constrição obtidas são muito próximas das originais. Testes com subconjuntos de parâmetros também deram bons resultados. Por exemplo, sabendo-se a configuração dos lábios apenas se optimizaram os outros parâmetros.

O método chega a resultados similares em várias repetições do processo. Exemplificando este facto apresentam-se, na Figura 4.17(a), diversas configurações obtidas para a vogal [a]. Para a obtenção das configurações foram utilizadas 10000 iterações e uma métrica Euclidiana de erro. A configuração obtida está, também, de acordo com as descrições da articulação de vogal [a] em Português. O maxilar encontra-se abaixado, e a língua baixada e central.

Resultados obtidos com as três métricas são comparados na Figura 4.17(b) para o caso da vogal [a]. As configurações obtidas são muito semelhantes no que concerne à posição da língua, abertura do maxilar, e lábios. Os resultados desta, e outras simulações, não apontaram claramente para uma métrica como sendo capaz de permitir a obtenção de melhores resultados.

4.5.5.2 Testes com formantes obtidas de voz natural

Os resultados obtidos na inversão, utilizando formantes naturais, foram muito promissores. Exemplos foram apresentados em Teixeira *et al.* (1997c). A título de exemplo apresentam-se, na Figura 4.18, configurações obtidas para cinco vogais orais do Português, utilizando a medida de erro Euclidiana.



(a) [i] (b) [u]

(c) [ɛ] (d) [a] (e) [ɔ]

Figura 4.18: Configurações obtidas por inversão para cinco vogais orais do Português Europeu. Foi utilizada como medida de erro uma distância Euclidiana.

Analisando as configurações obtidas conclui-se que se aproximam das descrições da Fonética Articulatória (veja-se a secção 2.5.2 na página 26). A vogal [i] apresenta uma posição da língua elevada e frontal. O [u] apresenta-se também com a língua elevada mas agora recuada.

É, também, visível o arredondamento dos lábios. Para o [a], como era esperado, existe uma faringe bastante estreita, o maxilar encontra-se na sua abertura máxima (permitida pelo modelo). Permitiu-se, neste caso, a optimização da região baixa da faringe, para se poder obter o valor elevado de F_1 desejado. A vogal [ɔ] tem também a língua claramente recuada.

4.6 Detalhes de implementação

O sintetizador foi desenvolvido em C/C++ e Tcl/Tk (Welch, 1995) para utilização num computador pessoal, com placa de som, correndo o sistema operativo Linux. Esta combinação de hardware com software do domínio público torna o sistema de custo muito reduzido.

Os vários modelos apresentados neste capítulo (modelo articulatório, modelo acústico e fonte glotal) foram implementados como objectos, para tornar mais fácil a adição futura de novos modelos ao sistema.

A utilização de Tcl/Tk facilita o desenvolvimento de uma interface com o utilizador, assim como a sua reconfiguração. A interface com o utilizador permite:

1. A definição de uma configuração do tracto tendo, ao mesmo tempo, informação acerca da função de área, comprimento total do tracto, posição da zona de menor área, relação entre a área de entrada do tracto nasal e a área da passagem oral na zona do velo, configuração dos lábios, resposta em frequência e respectivas formantes. Esta facilidade é muito útil para a obtenção manual de uma configuração. O utilizador pode obter um ficheiro em PostScript da configuração para inserção directa em publicações [15];

2. Sintetizar o som correspondente a essa configuração, e de seguida ouvi-lo, gravá-lo, ver uma zona em detalhe e, mesmo, o espectrograma. A onda de excitação é também apresentada e é possível guardá-la em ficheiro;

3. Guardar em ficheiro, para posterior análise, a resposta em frequência de diversas zonas do tracto. É, por exemplo, possível obter a impedância de entrada do tracto nasal;

4. Controlar o processo de interacção entre a fonte e o tracto. É possível escolher entre: incluir a carga completa do tracto, apenas a carga do tracto oral, ou não considerar a interacção;

5. Escolher o sinal radiado que se pretende. Pode ter-se o sinal radiado apenas pelos lábios, apenas pelas narinas ou a soma dos dois;

6. Processar um ficheiro *batch* para sintetizar um conjunto alargado de sons. Esta facilidade é da máxima importância para a obtenção dos estímulos necessários para testes

[15] As figuras de configurações utilizadas neste trabalho foram obtidas desta forma. Um exemplo é a Figura 4.18 da página 123.

perceptuais. O ficheiro de entrada pode ser obtido manualmente ou ser o resultado de um programa;

7. Obter a configuração do tracto, usando o processo de inversão descrito, para valores de formantes fornecidos. O utilizador pode controlar os parâmetros do algoritmo de *simulated annealing*, escolher a medida de erro, o número de formantes a utilizar, etc.

4.7 Resumo

The whole is more than the sum of the parts.

ARISTOTLE
Metaphysica

Neste capítulo foi apresentado, com algum detalhe, o sintetizador articulatório desenvolvido para servir de suporte aos estudos de produção e percepção dos sons nasais do Português Europeu. Procurou-se descrever os modelos utilizados e apresentar as razões para a sua escolha, de entre as várias opções existentes. Não foi dedicada especial atenção a nenhum dos modelos. O objectivo não foi inovar no modelamento de cada parte constituinte do sintetizador, mas sim, pela avaliação, escolher o modelo mais adequado. Pretendeu-se, pela integração de todos os modelos num sintetizador desenvolvido com ferramentas de baixo custo, implementar uma ferramenta de trabalho versátil e com as capacidades adequadas ao fim em vista.

O modelamento da forma do tracto vocal foi efectuado utilizando um modelo articulatório sagital (Hsieh, 1994; Prado, 1991; Mermelstein, 1973). O nosso trabalho em relação a este modelo centrou-se mais na implementação de interfaces com o utilizador do que propriamente na evolução do modelo.

O modelo acústico desenvolvido, baseado no proposto por Sondhi e Schroeter (1987), teve como preocupação principal o modelamento de sons nasais. As principais características relacionadas com o modelamento do tracto nasal são: a possibilidade de utilizar configurações do tracto nasal facilmente definidas pelo utilizador; a capacidade do modelo nasal ser constituído por mais do que um tubo acústico; permitir a obtenção do fluxo radiado nos lábios e narinas em separado; e no processo de cálculo da impedância de entrada do tracto vocal decidir-se pela inclusão ou não da impedância de entrada do tracto nasal, capacidade esta importantíssima para o estudo dos efeitos de interacção entre a fonte e o tracto.

O modelo acústico foi desenvolvido de forma a permitir a simulação de sons resultantes da variação no tempo dos articuladores. Ao contrário de outros sintetizadores, como o CASY (Rubin *et al.*, 1981) e o desenvolvido por Story (1995), em que sempre que se define um valor para um dos parâmetros articulatórios, num determinado instante no tempo, se tem de definir o valor para todos os outros parâmetros, o sintetizador desenvolvido permite a definição dos valores de cada um dos parâmetros articulatórios de forma completamente independente.

Ainda em relação ao modelo acústico, e para reduzir o tempo de cálculo, apenas se calcula a função de transferência e impedância de entrada, no início de cada novo período da onda glotal.

Apresentou-se também um modelo interactivo para a fonte glotal do sintetizador. O modelo baseia-se no trabalho de Allen e Strong (1985) com os seguintes melhoramentos: permite sintetizar sons com configurações variáveis no tempo, quer dos articuladores que definem o tracto, quer dos parâmetros da fonte; a resistência dos pulmões foi corrigida para um valor mais realista; as cavidades subglotais são simuladas por três secções representadas por circuitos RLC, podendo, cada um deles, ser desligado; a síntese de sons nasais é possível; pode incluir-se ou não, consoante as necessidades, o efeito de carga do tracto; foram incluídos os efeitos de *jitter*, *shimmer* e da aspiração.

O sintetizador inclui, também, facilidades de obtenção dos parâmetros articulatórios com base nos valores das primeiras 4 formantes. Este processo encontra-se apenas desenvolvido para vogais orais. Nas nossas experiências de inversão das vogais orais do Português, obtivemos configurações plausíveis. No entanto, este trabalho tem de ser continuado, sendo necessário dados de produção obtidos por métodos directos para aferição de resultados.

Capítulo 5

Interacção fonte-tracto em sons nasais

> Future high-quality speech synthesizers may also have to forego the fiction that the vocal cords and vocal tract are completely decoupled mechanical systems.
>
> Manfred Schroeder
> (Schroeder, 1999, pág. 87)

É conhecido que a forma da onda de excitação glotal é de grande importância para a produção de som sintético de qualidade. Diversos estudos abordaram este assunto (Båvegård, 1995a; Båvegård e Fant, 1994; Wong, 1991; Allen e Strong, 1985; Koizumi *et al.*, 1987; Fant e Lin, 1987; Fant *et al.*, 1985a; Koizumi *et al.*, 1985; Childers *et al.*, 1983; Ananthapadmanabha e Fant, 1982; Rothemberg, 1981; Ishizaka e Flanagan, 1972), sem no entanto terem incluído sons nasais. Uma excepção é o estudo de Titze e Story (1997).

Como os efeitos da nasalização se fazem sentir na região da primeira formante (Hawkins e Stevens, 1985) (Sampson, 1999, pág. 8) e os estudos de interacção mostram ser a região da primeira formante a principal responsável pelo efeito da interacção (Ananthapadmanabha e Fant, 1982), justifica-se o estudo dos efeitos de interacção fonte-tracto para os sons nasais.

É também conhecido que os sons nasais são mais atenuados a altas-frequências, o que pode dever-se, em parte, às características espectrais da fonte.

Para estudar os efeitos de interacção, no caso de vogais nasais, assunto que ocupa este capítulo, começamos por efectuar simulações das alterações da impedância de entrada, onda glotal e suas características espectrais. Estas simulações constituem a primeira secção do capítulo.

Para saber se as alterações são perceptíveis ao ouvido humano, foram efectuados testes perceptuais, apresentados na segunda secção.

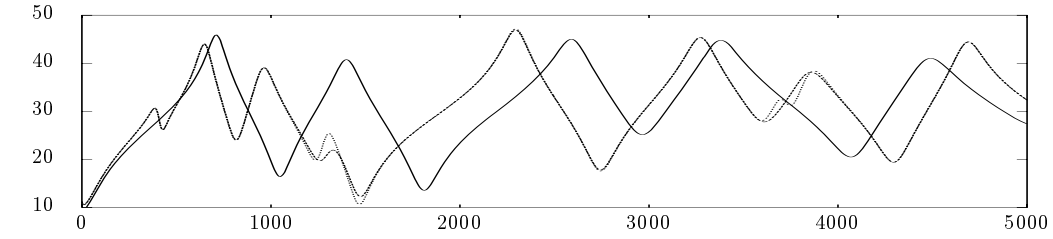
Na terceira secção apresentam-se simulações realizadas após os testes perceptuais, investigando possíveis explicações para os resultados obtidos.

O capítulo encerra resumindo o trabalho efectuado e os resultados principais obtidos.

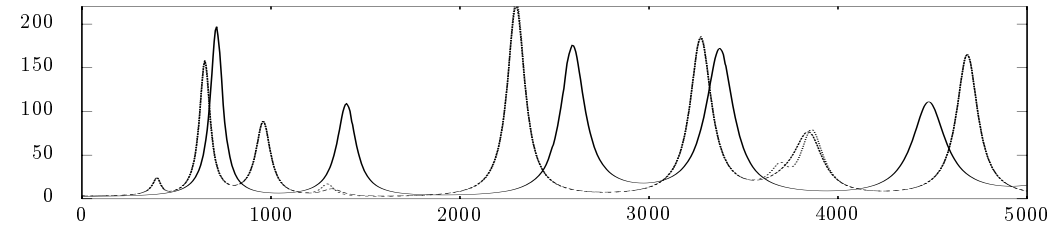
5.1 Simulações

Foram efectuadas diversas simulações para estudar o efeito do acoplamento do tracto nasal. Este acoplamento altera a função de transferência e a impedância de entrada. As alterações nesta última alteram as condições de carga da laringe, modificando por consequência a onda glotal.

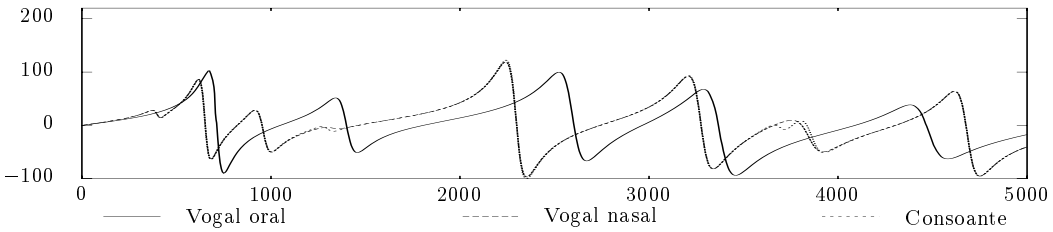
O estudo começou pela análise de configurações estáticas e continuou, devido a resultados obtidos indicarem a importância da variação no tempo do velo (Teixeira *et al.*, 1999b), estudando também o comportamento da onda glotal quando o velo varia a sua abertura ao longo do tempo em contextos CṼC.



(a) $20\log(|Z_e(\omega)|)$



(b) Parte real da impedância de entrada.



(c) Parte imaginária da impedância de entrada.

Figura 5.1: Comparação da impedância de entrada de uma vogal oral, um [ɐ], a correspondente vogal nasal e a consoante nasal bilabial, produzida fechando a passagem oral na zona dos lábios, mas mantendo a configuração do tracto da vogal.

5.1.1 Configurações estáticas

As simulações iniciaram-se pelo estudo do efeito do acoplamento adicional do tracto nasal na impedância de entrada. De seguida investigou-se o efeito dessa variação (da impedância de entrada) na onda de excitação glotal. Foram efectuadas simulações para a totalidade das vogais nasais portuguesas, apresentando-se, no entanto, de seguida apenas alguns exemplos que consideramos significativos.

5.1.1.1 Efeitos na impedância de entrada do tracto

Na Figura 5.1 compara-se a impedância de entrada de uma vogal oral, um [ɐ], com a impedância de entrada de uma vogal nasal produzida com a mesma configuração do tracto oral. Compara-se também com a impedância de entrada da consoante nasal produzida fechando a passagem oral na região dos lábios, mantendo a posição dos restantes articuladores. Não se pode considerar que exista uma alteração profunda da impedância. Existem no entanto alterações. Existem picos adicionais, em especial na zona da primeira e segunda formante. A impedância da consoante nasal é muito próxima da impedância obtida para a vogal nasal.

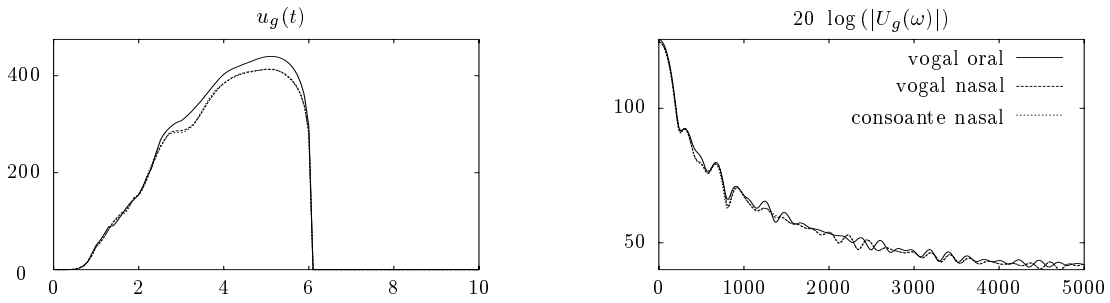


Figura 5.2: Onda glotal e módulo da respectiva transformada de Fourier para a vogal [ɐ]/[ɐ̃]. São apresentados três simulações: velo fechado (vogal oral); velo aberto (vogal nasal); velo aberto e passagem oral fechada na zona dos lábios (consoante nasal bilabial com configuração do tracto característica de [ɐ]). Em todos os casos utilizou-se a carga total do tracto.

5.1.1.2 Efeitos na onda glotal

Foi calculada a onda glotal para diversas vogais para estudar a influência do acoplamento adicional do tracto nasal.

Primeiro, comparamos a onda glotal de uma vogal oral com a vogal nasal obtida da anterior apenas alterando a abertura do velo. Apresenta-se também a consoante nasal bilabial diferindo apenas da vogal nasal na oclusão do tracto oral na zona dos lábios. Apresenta-se nas Figuras 5.2 e 5.3 exemplos para duas vogais. Para a onda glotal, a diferença entre o caso

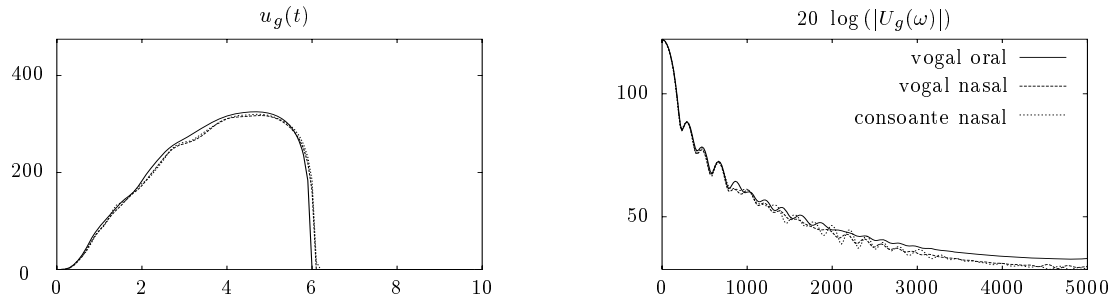


Figura 5.3: Onda glotal e módulo da respectiva transformada de Fourier para [ɛ]/[ɛ̃]. São apresentados três simulações: velo fechado (vogal oral); velo aberto (vogal nasal); velo aberto e passagem oral fechada na zona dos lábios (consoante nasal bilabial com configuração do tracto característica de [ɐ]). Em todos os casos utilizou-se a carga total do tracto.

oral e o nasal não é muito significativa. No domínio da frequência, a diferença é ainda mais reduzida. Comparando a consoante nasal com a vogal nasal, nota-se que o efeito de fechar os lábios é dificilmente perceptível.

Interessa também saber qual a contribuição da impedância de entrada do tracto nasal para a interacção. Para tal, calculou-se a onda glotal não incluindo a impedância de entrada do tracto nasal comparando-a com o caso em que se inclui a totalidade das cavidades (orais e nasais) no cálculo da impedância. Na Figura 5.4 apresenta-se o resultado deste tipo de simulação para a vogal [ĩ]. Na figura inclui-se também a forma da onda glotal para o caso de não existir interacção fonte-tracto. É notório que, pelo menos nesta vogal, o tracto nasal tem muita influência na forma de onda glotal. De forma alguma se poderá calcular a interacção desprezando o tracto nasal.

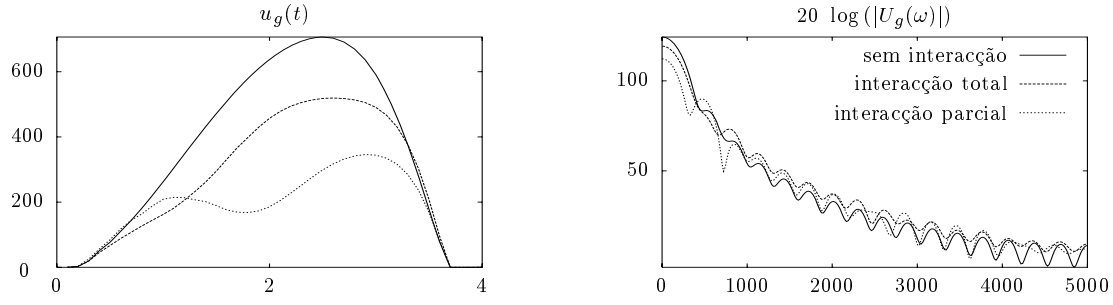
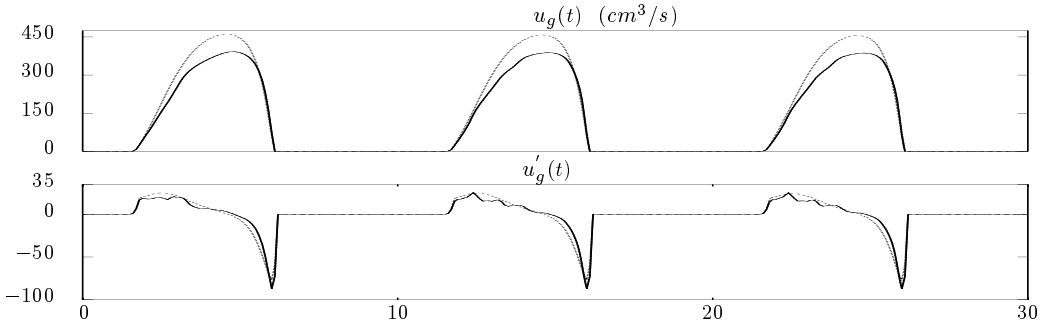
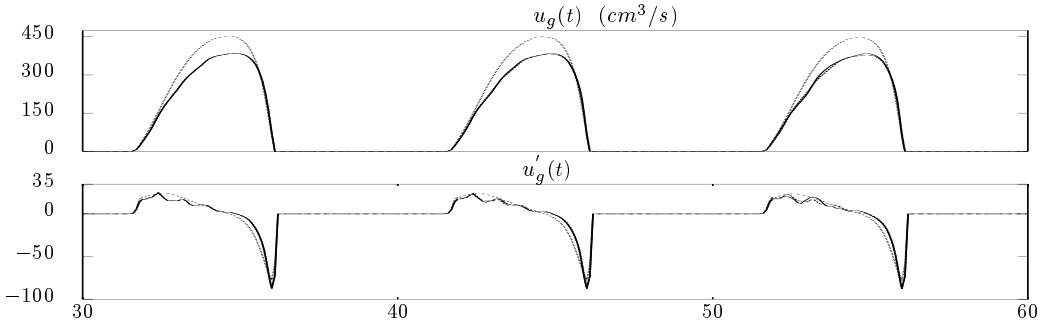


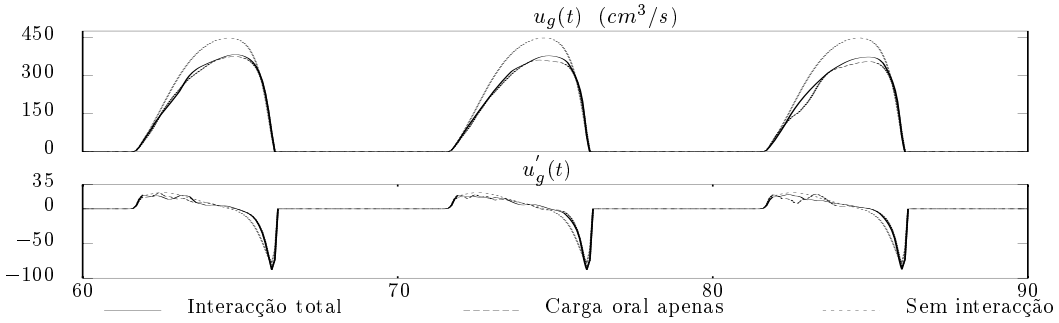
Figura 5.4: Onda glotal e módulo da transformada de Fourier para a vogal [ĩ]. Apresentam-se três situações: onda glotal calculada sem considerar o efeito de interacção (sem interacção); utilização da impedância total de entrada do tracto no cálculo da interacção (interacção total); e considerar como impedância apenas a carga devida ao tracto oral, não incluindo a impedância de entrada do tracto nasal (interacção parcial).



(a) Sinal entre 0 e 30 ms, zona com velo fechado.



(b) Sinal entre 30 e 60 ms, zona com radiação pela boca e narinas.



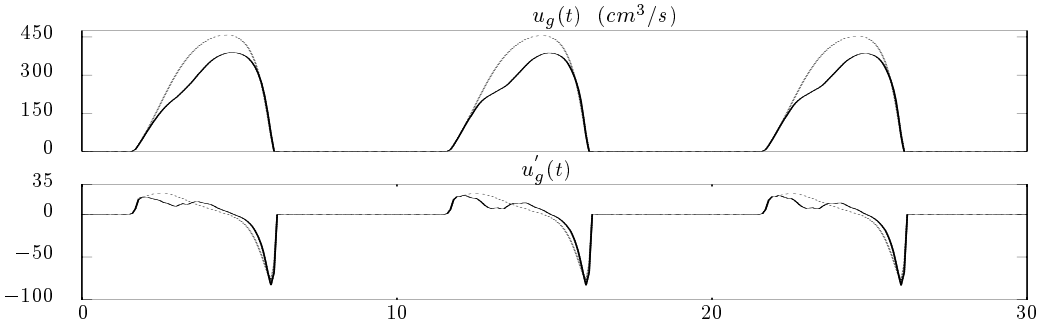
(c) Sinal entre 60 e 90 ms, fecha a passagem oral produzindo-se uma consoante nasal.

Figura 5.5: Efeito da interacção fonte-tracto para a vogal [ẽ] entre duas consoantes oclusivas (não nasais).

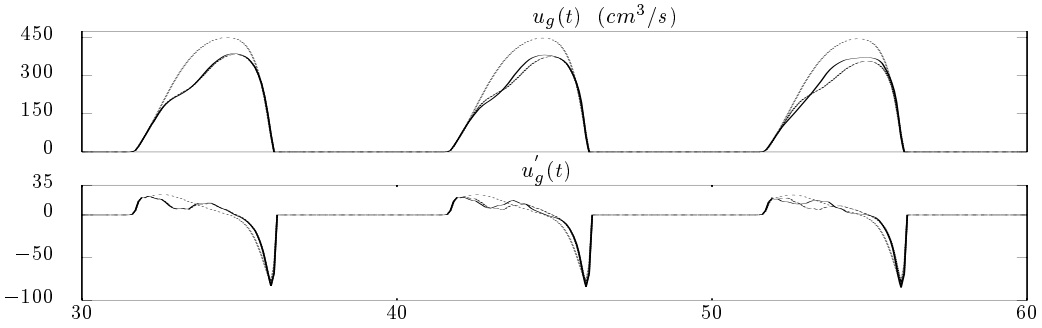
5.1.2 Interacção fonte-tracto em sequências CṼC

Os nossos resultados no estudo da influência da dinâmica na percepção de nasalidade, que apresentaremos no capítulo 6, indicam que é necessário ter em conta a variação no tempo do velo. Nada mais natural que investigar o efeito dessa variação, no tempo, em termos de interacção. Os estudos efectuados apenas abordaram o caso de vogais nasais entre duas

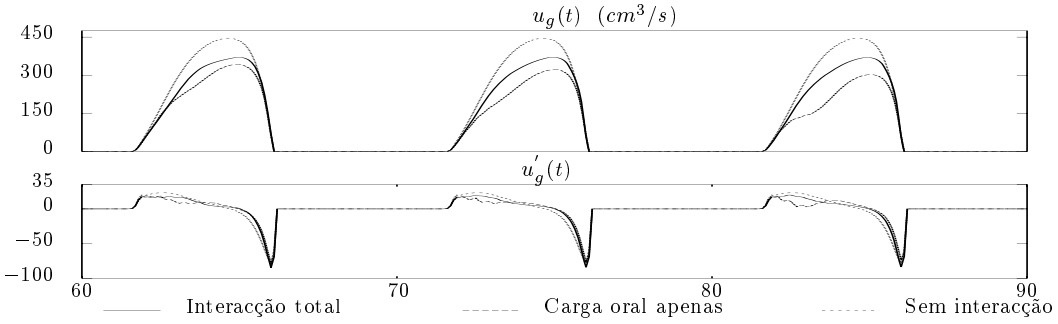
consoantes oclusivas não nasais, como, por exemplo, na palavra *canto*. Este contexto implica que na fase inicial exista uma zona oral, devido à inércia de abertura do velo. Segue-se uma fase em que existe radiação simultânea pela boca e nariz. Na parte final, o fecho da passagem oral, para a produção da oclusiva seguinte, antes do velo ter fechado, dá origem a uma consoante nasal.



(a) Sinal entre 0 e 30 ms, zona com velo fechado.



(b) Sinal entre 30 e 60 ms, zona com radiação pela boca e narinas.



(c) Sinal entre 60 e 90 ms, fecha a passagem oral produzindo-se uma consoante nasal.

Figura 5.6: Efeito da interacção fonte-tracto para a vogal [ĩ] entre duas consoantes oclusivas (não nasais).

Nas Figuras 5.5 e 5.6 apresenta-se o resultado da simulação para duas vogais, [ɐ̃] e [ĩ]. Para

facilitar a inclusão da figura cada uma das três fases tem a duração de 30 ms, o que não é muito natural. Para facilitar a visualização das diferenças inclui-se não só $u_g(t)$ mas também a sua derivada. Representam-se nos gráficos, simultaneamente, as ondas glotais obtidas para três situações distintas: (1) não existência de interacção; (2) interacção utilizando a impedância de entrada calculada incluindo a totalidade das cavidades supraglotais; (3) interacção calculada não incluindo a impedância de entrada do tracto nasal.

Para a vogal [ɐ̃] o efeito do tracto nasal não é muito significativo, a onda glotal calculada com interacção total não é muito diferente da onda glotal calculada utilizando a impedância de entrada apenas das cavidades orais.

O mesmo já não se verifica para a vogal [ĩ] onde é notória a influência da impedância nasal. Esta influência é particularmente notória na parte final, em que existe oclusão oral. É também notória, neste caso, a variação das características da onda glotal ao longo do tempo, causada pela variação da abertura do velo e da passagem oral.

O efeito em termos da frequência é menos notório. Apresenta-se na Figura 5.7 o módulo da transformada de Fourier da onda glotal e da sua derivada para as zonas média e final de realização da vogal nasal [ĩ]. Verifica-se que a inclusão da interacção provoca o decréscimo do valor do máximo nos gráficos do módulo da transformada de Fourier de $u_g^{'}(t)$. Este comportamento corresponde a uma variação da relação entre as durações das fases de abertura e fecho da glote (Fant, 1995, Figura 9, pág. 131). A inclusão da interacção aumenta o quociente de velocidade, em especial na fase final. Caso não seja incluído o efeito do tracto nasal, obtém-se um valor para o quociente de velocidade algo superior ao valor correcto.

Existe também uma ligeira variação no declive (em Inglês *tilt*) espectral, para frequências mais elevadas. Como o declive está relacionado com a velocidade de fecho da glote (Fant, 1995, Fig. 8, pág. 130), existem portanto alterações na fase próxima do fecho da glote. No entanto, estas alterações não dependem grandemente do acoplamento adicional do tracto nasal.

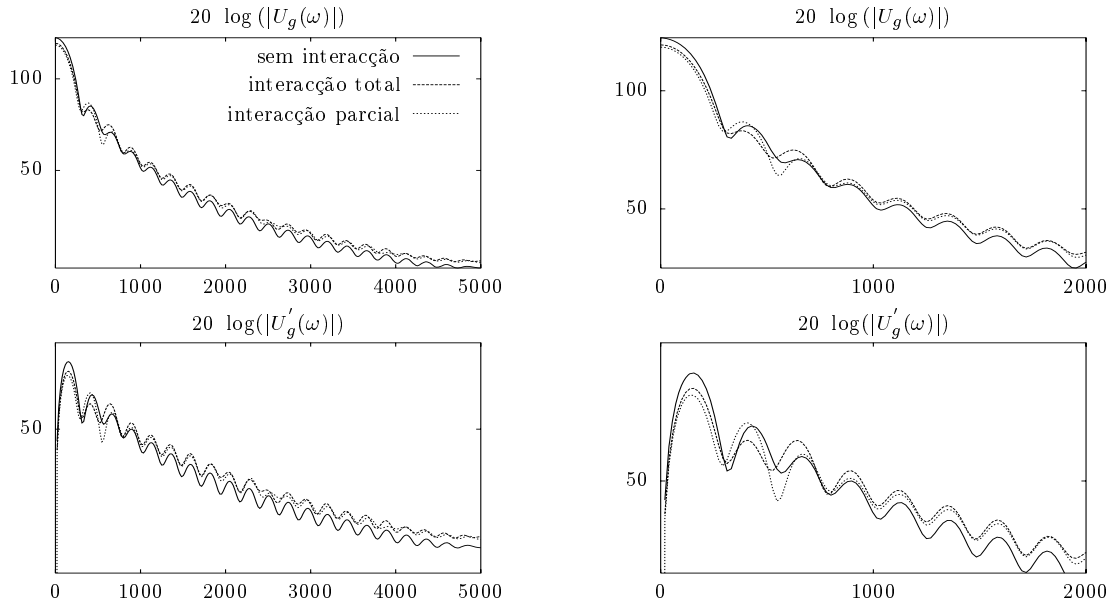
Tem-se portanto um efeito directo do tracto nasal no quociente de velocidade, mas não na fase de retorno da derivada de fluxo, característica considerada muito importante no modelamento da excitação glotal [1].

5.2 Testes perceptuais

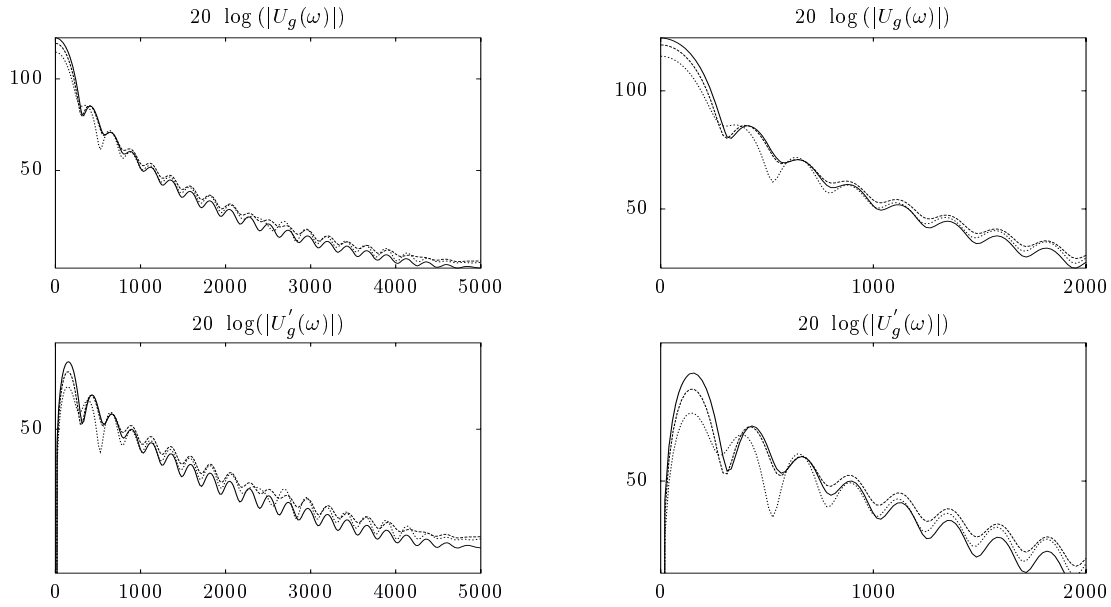
Mostrou-se, na secção anterior, a existência de alterações na onda glotal, causada pelas variações na impedância de entrada do tracto, resultantes do acoplamento do tracto nasal. Resta saber se essas alterações são perceptíveis para um ouvinte. Apenas sendo perceptíveis necessitarão os sistemas de síntese incluir essas variações.

Investigou-se se é detectável, pelos ouvintes, a inclusão do efeito da cavidade nasal e se a inclusão do tracto nasal, no cálculo da interacção, melhora a qualidade.

[1] O parâmetro Ta que a representa é um dos principais contributos do modelo LF proposto por Fant *et al.* (1985b).



(a) Zona média, radiação simultânea pelos lábios e narinas.



(b) Zona final, radiação apenas pelas narinas devido a oclusão da passagem oral.

Figura 5.7: Módulo da transformada de Fourier para a vogal [ĩ] na zona média e final. Para cada zona apresenta-se a transformada de $u_g(t)$, na primeira linha, e $u_g^{'}(t)$ na segunda linha. Em cada linha, apresenta-se primeiro o gráfico para frequências até 5 kHz e, do lado direito, apenas frequências até 2 kHz, para facilitar a visualização das diferenças nesta gama de frequência, com mais interesse no estudo da interacção fonte-tracto. Em todos os casos, apresentam-se simulações: sem interacção; interacção total; e interacção considerando apenas a carga do tracto oral.

Para tentar obter uma resposta, realizaram-se testes perceptuais com estímulos gerados pelo sintetizador articulatório. Para a primeira questão, foi escolhido um teste de discriminação e, para a segunda, um teste de preferência.

A utilização de um programa, por nós desenvolvido, tornou a realização dos testes completamente automática. O programa cria os pares de estímulos, as repetições necessárias, baralha a ordem de apresentação, apresenta os estímulos e guarda os resultados para processamento posterior.

Os estímulos foram apresentados aos ouvintes utilizando auscultadores, modelo Sennheiser Headmax HD 470, em salas com ruído ambiente baixo a moderado. Os ouvintes responderam aos testes, escolhendo a opção que consideravam mais adequada com a ajuda de um rato e uma interface gráfica.

5.2.1 Características gerais dos estímulos

As características comuns a todos os estímulos utilizados nos dois testes perceptuais são descritas de seguida.

As configurações orais para as diversas vogais utilizadas foram obtidas usando o método de inversão descrito na secção 4.5 na página 106. As configurações obtidas foram verificadas com base em configurações e descrições publicadas. A abertura do velo foi ajustada manualmente por forma a se obter uma relação entre a área da passagem oral e a abertura do velo de 1 para 10.

A variação no tempo da abertura do velo utilizada, devido à falta de dados detalhados, consistiu numa aproximação bastante grosseira. Baseia-se em valores médios para as transições do velo e análise de sequências CṼC naturais. Além disso, foi utilizada a mesma variação temporal para todas as vogais sintetizadas. Durante os 100 ms iniciais o velo mantém-se fechado, abrindo depois durante 60 ms até atingir a máxima abertura. Mantém-se na posição de abertura máxima até ao fim da vogal.

Os estímulos acabam numa consoante nasal bilabial, [m], produzida fechando os lábios. O movimento de fecho estende-se por 50 ms com início em 200 ms. Existe em consequência deste movimento dos lábios uma consoante final com duração de 50 ms, devido a todos os estímulos terem duração de 300 ms.

Relativamente à fonte, foi utilizado o modelo com interacção fonte-tracto com F_0 variável no tempo. F_0 começa em 100 Hz, aumenta 20 Hz nos primeiros 100 ms e decresce depois gradualmente até 100 Hz . Utilizaram-se valores de 60 % e 2 para o quociente de abertura e de velocidade, respectivamente (valores usados por Allen e Strong (1985)). Utilizou-se ainda *jitter* e *shimmer* para melhoria da naturalidade.

5.2.2 Ouvintes

Um total de 14 ouvintes, 11 do sexo masculino e 3 do sexo feminino, tendo o Português por língua materna, participaram nos dois testes perceptuais. As suas idades estavam compreendidas entre os 13 e 53 anos. Nenhum deles tinha conhecimento de possuir qualquer tipo de problema auditivo.

5.2.3 Teste de discriminação

Um teste de discriminação entre estímulos, do tipo 4IAX, foi utilizado para indagar se os ouvintes conseguem detectar no sinal de voz sintético as diferenças provocadas pela inclusão adicional do efeito de carga das cavidades nasais.

A escolha do teste 4IAX, em detrimento do método ABX geralmente utilizado, deveu-se a este último ter sido questionado, por obrigar a um esforço desnecessário de armazenamento em memória, pelo ouvinte, de informação acústica (Mais detalhes podem ser encontrados em (Garman, 1990, pág. 200)). No método 4IAX os ouvintes ouvem dois pares de estímulos. Os membros de um par são iguais (AA), sendo os membros do outro par diferentes (AB). Aos ouvintes é perguntado qual o par constituído por estímulos diferentes. A vantagem deste teste reside no facto de que assim que é ouvido um par, o ouvinte decide se é constituído pelos mesmos estímulos ou estímulos diferentes. Apenas é necessário guardar o resultado dessa decisão em memória, aliviando o ouvinte da necessidade de reter a informação acústica dos dois estímulos. Vários investigadores obtiveram melhores resultados de discriminação com este tipo de teste (Borden *et al.*, 1994, pág. 213).

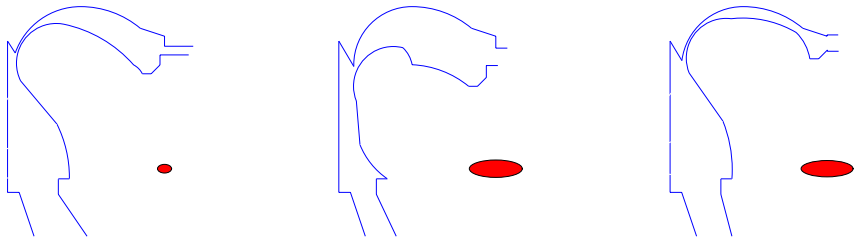
5.2.3.1 Procedimento

Cada uma das quatro combinações de dois estímulos (ABAA, ABBB, AAAB, BBAB) foi apresentada aos ouvintes três vezes. Desta forma, cada par de estímulos foi ouvido 12 vezes. A apresentação dos conjuntos de estímulos aos ouvintes seguiu uma ordem aleatória e diferente para cada um. Utilizou-se um intervalo entre estímulos de 400 *ms* e um intervalo entre os dois pares de 700 *ms*. Não foi imposto limite de tempo para decidir, podendo os ouvintes repetir a audição dos estímulos as vezes que desejassem.

5.2.3.2 Estímulos

Foram produzidos estímulos para 3 vogais nasais, [ɐ̃], [ĩ] e [ũ]. A utilização de apenas três vogais visou reduzir o tempo necessário para a realização dos testes [2]. Apresenta-se na Figura 5.8 as configurações do tracto, para o instante 190 *ms* onde já se atingiu o máximo de abertura do velo.

[2]Outro facto que motivou esta escolha é o de, por vezes, se referir, na literatura da área da Fonética, a realização como ditongos das outras duas vogais nasais portuguesas.



(a) [ũ] (b) [ẽ] (c) [ĩ]

Figura 5.8: Configuração do tracto para as três vogais nasais utilizadas no teste de discriminação. Apresenta-se a configuração no instante 190 *ms* em que a abertura do velo é máxima.

Apenas se utilizaram estímulos com velo dinâmico e consoante nasal final, considerados como mais naturais em trabalho anterior (Teixeira *et al.*, 1999b).

O único factor considerado foi a impedância de entrada do tracto, utilizada no cálculo da onda glotal. Este factor podia tomar dois valores: (1) a impedância de entrada inclui a totalidade das cavidades supraglotais; ou (2) no cálculo da impedância do tracto não se inclui a impedância de entrada das cavidades nasais. Produziram-se assim 6 estímulos, 2 para cada vogal.

	Ouvinte	Sexo	[ẽ]	[ĩ]	[ũ]	Média
1	EDU	M	50.0	33.3	41.7	41.7
2	CSST	M	58.3	100.0	50.0	69.4
3	ENE	F	50.0	41.7	50.0	47.2
4	MJT	F	33.3	83.3	66.7	61.0
5	FVAZ	M	16.7	58.3	33.3	36.1
6	AJST	M	66.7	66.7	66.7	66.7
7	VIT	M	50.0	50.0	41.7	47.2
8	RIC	M	58.3	58.3	41.7	52.8
9	PAT	F	41.7	50.0	66.7	52.7
10	TOS	M	58.3	50.0	58.3	55.6
11	NFF	M	33.3	83.3	58.3	58.3
12	NMF	M	75.0	58.3	58.3	63.9
13	RSM	M	50.0	41.7	33.3	41.4
14	AP	M	83.3	50.0	58.3	63.8
	Média		51.8	58.9	51.8	54.1
	Desvio padrão		17.3	18.6	11.9	10.3

Tabela 5.1: Resultados do teste de discriminação, 4IAX, do efeito do tracto nasal na interacção entre a fonte glotal e o tracto vocal.

5.2.3.3 Resultados

Na Tabela 5.1 apresentam-se a percentagens de respostas certas no teste de discriminação. São apresentados os resultados para cada ouvinte em cada vogal, assim como a média para as três vogais. Nas linhas finais da tabela apresenta-se a média e desvio padrão para cada vogal e para o conjunto das três vogais. A representação gráfica dos valores da mediana e quartis (*boxplot*), utilizando o diagrama extremos-e-quartis (Bryman e Cramer, 1993, pág. 113), é apresentada na Figura 5.9.

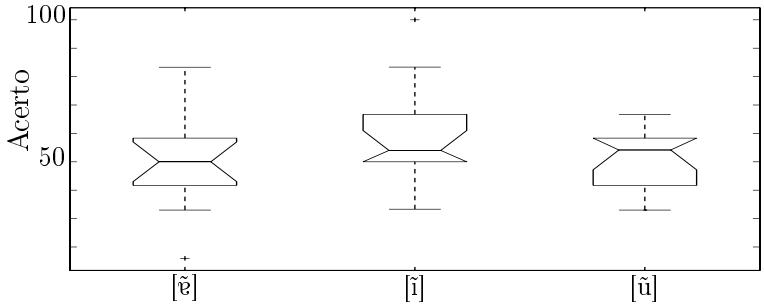


Figura 5.9: Resultados do teste de discriminação.

Em termos gerais ressalta que os valores obtidos são próximos dos 50%, indicando que a diferença é difícil de detectar.

A vogal [ĩ] apresenta um comportamento ligeiramente diferente. Em termos médios a taxa de acerto para o [ĩ] é um pouco superior a 50%. No diagrama extremos-e-quartis nota-se que cerca de 75 % das observações se encontram acima dos 50 %. Note-se que houve mesmo um ouvinte que conseguiu acertar 100 % para o [ĩ] e apenas conseguiu valores próximos dos 50 % nas outras duas vogais.

Testes estatísticos, tendo como hipótese nula $H_0 : \mu = 50$ e hipótese alternativa $H_1 : \mu > 50$, apenas deram diferença significativa, com um nível de significância de 5%, para o [ĩ] ($p < 0.05$). Para esta vogal o intervalo de confiança para a média é de 50.08 a 67.71. Para o [ɐ̃] obteve-se $p = 0.3599$ e para o [ũ] $p = 0.2939$. Para todos os resultados em conjunto a média também não é significativamente superior a 50% ($p = 0.0804$).

5.2.4 Teste de preferência

Estando interessados na qualidade do som sintetizado decidiu-se, também, pela realização de um teste de qualidade. Refira-se que este teste foi realizado em simultâneo com o teste de discriminação para aproveitar a disponibilidade dos ouvintes. O teste de qualidade escolhido foi o teste AB. Apesar de exigir mais decisões para cada ouvinte, aumentando a duração dos testes, é mais preciso que testes baseados na avaliação de um estímulo individual.

5.2.4.1 Estímulos

Para este segundo teste foram produzidos estímulos para as cinco vogais nasais do Português Europeu. O teste AB, para uma duração similar e o mesmo número de repetições, permite aumentar o número de estímulos em relação ao teste 4IAX. Apenas foi utilizado um valor para a abertura do velo para cada uma das vogais. Variou-se apenas a impedância de entrada, considerando-se três situações: não existência de interacção; interacção utilizando a impedância de entrada da totalidade das cavidades supralaríngeas; e interacção não incluindo o efeito da impedância de entrada do tracto nasal. Obtiveram-se, desta forma, um total de 15 estímulos.

5.2.4.2 Procedimento

A questão colocada aos ouvintes foi: "Qual dos dois estímulos prefere como uma vogal nasal do Português ?". O ouvinte podia optar por uma de quatro resposta: "o primeiro"; "o segundo"; "ambos"; e "nenhum". A inclusão das duas últimas, hipóteses de resposta, teve por objectivo evitar que o ouvinte tivesse que optar por um estímulo, nos casos em que eles são semelhantes e/ou ambos de má qualidade.

Cada estímulo foi apresentado, de forma aleatória, 10 vezes. Em metade dessas repetições o estímulo aparecia em primeiro lugar no par (ordem AB) na outra metade em segundo (ordem BA). Entre cada estímulo de um par foi inserido um intervalo de silêncio com duração de 600 ms. Foi permitida a audição do par de estímulos tantas vezes quantas o ouvinte desejasse.

5.2.4.3 Resultados

Os ouvintes tiveram muitas dificuldades em escolher. Em mais de metade dos pares a consistência dos ouvintes foi inferior a 60 %. Nas condições em que o teste foi realizado, claramente, os estímulos são muito semelhantes perceptualmente. Uma análise das respostas consistentes também não mostrou uma tendência geral. Alguns ouvintes preferem estímulos sem inclusão do efeito do tracto, outros preferem os estímulos com a impedância total. Diversos factores podem justificar estes resultados. Primeiro, a área glotal já é enviesada (em Inglês *skewed*). Também os sons nasais podem ser percebidos mais claramente sem o efeito de carga adicional do tracto nasal. Os ouvintes consideram-nos melhores por serem mais nítidos, o que está errado, pois os melhores sons nasais não podem ser, obviamente, os mais nítidos. Optou-se, devido aos factos referidos, por não efectuar qualquer outra análise dos resultados e pela sua não inclusão em detalhe neste trabalho.

5.3 Simulações pós-testes perceptuais

Para tentar obter uma explicação para as diferenças do efeito entre vogais altas, como o [ĩ], e baixas, como o [ɐ̃], foram efectuadas novas simulações. Sendo a alteração em $u_g(t)$ devida directamente à impedância de entrada, começou-se por aferir o efeito na impedância quando se inclui o tracto nasal. Investigou-se também o comportamento da impedância de carga da zona faríngea do tracto que é o paralelo das impedâncias de entrada do tracto nasal e da cavidade bucal.

5.3.1 Vogal [ɐ̃]

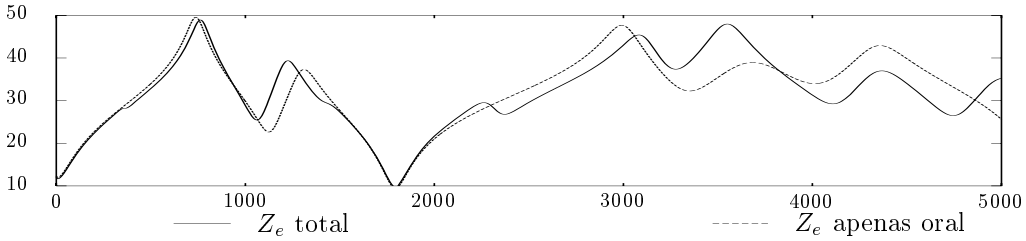
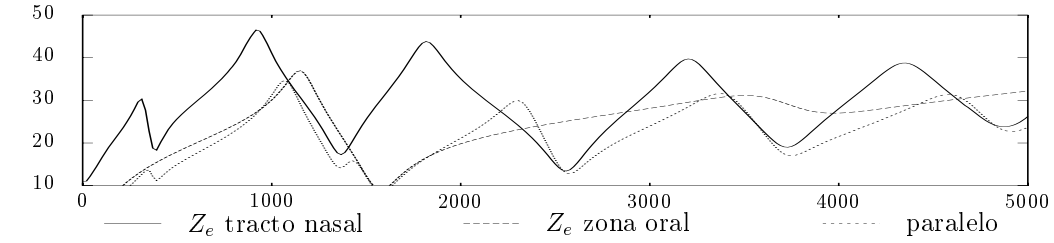


Figura 5.10: Impedância de entrada calculada para a vogal [ɐ̃] com e sem a inclusão da impedância de entrada do tracto nasal.

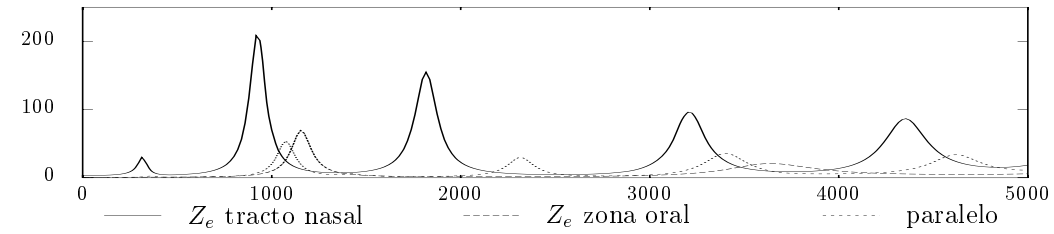
Começou-se por ver qual o efeito da impedância de entrada do tracto nasal na impedância de entrada do tracto vista da glote. Na Figura 5.10 apresenta-se o módulo da impedância em função da frequência. É notório que a inclusão da carga nasal não afecta grandemente a impedância de entrada. Isto é válido em especial para as frequências mais baixas (até cerca de 1 kHz).

Como a alteração é devida, essencialmente, à inclusão da impedância de entrada do tracto nasal, foi investigado com mais detalhe o que se passa na zona do velo em termos de impedâncias. Para tal, foram calculadas as impedâncias de entrada do tracto nasal, para a abertura máxima do velo utilizada, assim como a impedância de entrada da secção oral do tracto entre o velo e os lábios. Sendo a carga da secção faríngea do tracto o paralelo destas duas impedâncias, foi também calculado o valor da impedância resultante deste paralelo. Os resultados, para a vogal [ɐ̃], apresentam-se nas Figuras 5.11(a) a 5.11(c). A primeira figura representa o módulo das impedâncias, a segunda a parte real e a terceira a parte imaginária. É notório que o valor do paralelo é essencialmente dado, para frequências até 1.5 kHz, pelo valor da impedância da cavidade oral entre a zona do velo e os lábios.

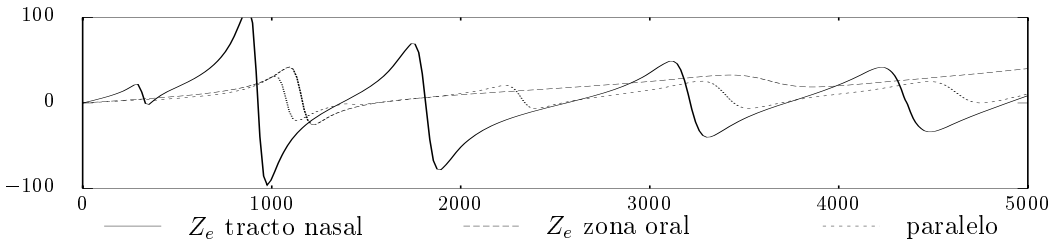
A oclusão labial na fase final da vogal nasal, cujos resultados se apresentam na Figura 5.12, provoca alguma alteração na impedância de entrada, em especial para frequências mais elevadas. Esta variação deve-se à variação da impedância de entrada da cavidade bucal, pois não existiu variação da impedância de entrada do tracto nasal. A configuração bucal, bastante aberta, leva a que, nesta vogal, o efeito da oclusão labial seja mais acentuado.



(a) Módulo



(b) Parte real



(c) Parte imaginária

Figura 5.11: Impedâncias na zona do velo para a vogal [ɐ̃]. Apresenta-se a impedância da cavidade oral, a impedância de entrada do tracto nasal e o paralelo das duas impedâncias. O paralelo das impedâncias é a impedância de carga da zona da faringe

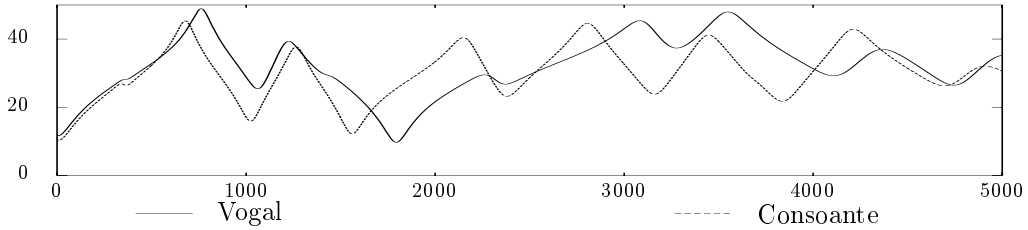


Figura 5.12: Comparação da impedância de entrada da vogal [ɐ̃] com a da consoante nasal bilabial diferindo de [ɐ̃] apenas pela oclusão.

5.3.2 Vogal [ĩ]

Para a vogal [ĩ] foi efectuado um procedimento similar. Primeiro, calculou-se a impedância de entrada do tracto com e sem a inclusão da carga do tracto nasal. Representados na Figura 5.13, os resultados mostram alguma diferença dos dois casos mesmo nas baixas frequências.

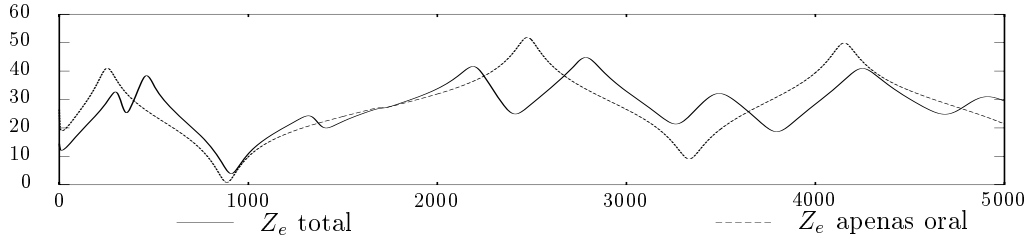


Figura 5.13: Impedância de entrada para a vogal [ĩ], com e sem a inclusão da impedância de entrada do tracto nasal.

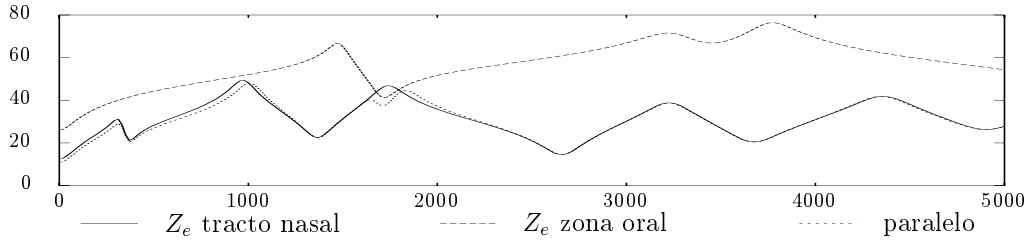
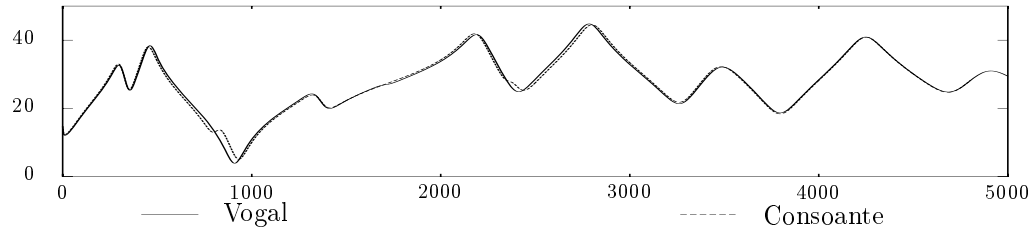


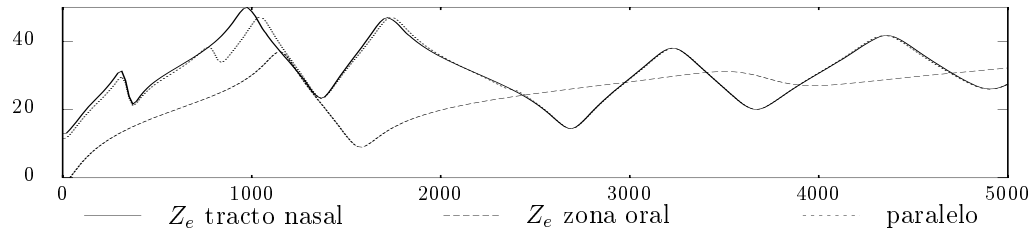
Figura 5.14: Módulos das impedâncias na zona do velo para a vogal [ĩ]. Apresenta-se a impedância da cavidade oral, a impedância de entrada do tracto nasal e o paralelo das duas impedâncias.

Os cálculos para a zona do velo, apresentados na Figura 5.14, mostram que a impedância de carga da faringe, é aproximadamente igual à impedância de entrada do tracto nasal. Este resultado não é de estranhar devido à reduzida área da cavidade bucal da vogal. Esta área reduzida é causadora de uma impedância de entrada da cavidade bucal elevada, sendo o paralelo quase igual à impedância do tracto nasal.

O efeito de fechar a passagem oral, na terceira fase da vogal nasal em contextos C$\tilde{\text{V}}$C, é apresentado na Figura 5.15. A oclusão labial não provoca quase alteração na impedância de entrada. A configuração bucal bastante fechada leva a que, nesta vogal, o efeito da oclusão labial seja muito menos acentuado do que no caso do [ɐ̃], apresentado na Figura 5.12. A oclusão da passagem oral pode ser efectuada noutros pontos da cavidade bucal. Compara-se a impedância de entrada total para diferentes pontos de articulação da consoante nasal na Figura 5.16. O efeito de alteração da zona de oclusão, devido à área reduzida na zona entre o velo e os lábios, é muito reduzido.



(a) Impedâncias (totais) de entrada.



(b) Impedâncias na zona do velo.

Figura 5.15: Impedâncias para a consoante nasal bilabial produzida com os outros articuladores na configuração usada para [ĩ].

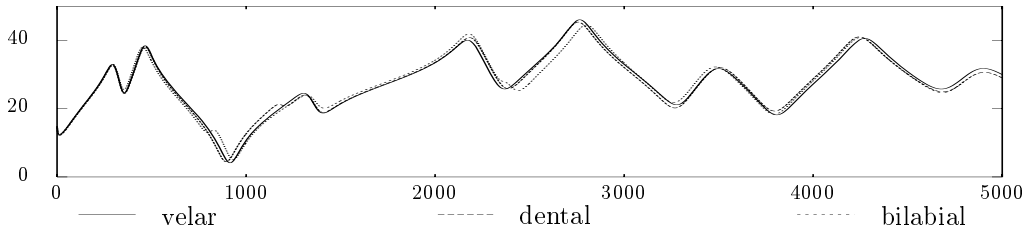


Figura 5.16: Efeito da zona de oclusão oral na impedância de entrada. Apresentam-se três pontos de oclusão: labial, dental, e velar.

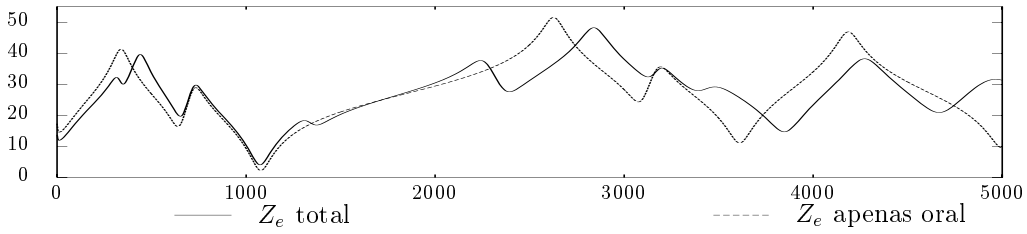


Figura 5.17: Impedância de entrada para a vogal [ũ], com e sem a inclusão da impedância de entrada do tracto nasal.

5.3.3 Vogal [ũ]

Para a vogal [ũ], apresenta-se a impedância de entrada do tracto, com e sem a inclusão da carga do tracto nasal, na Figura 5.17. O tracto nasal é responsável por alguma alteração na impedância de entrada, sem contudo ser tão significativo o seu contributo como no caso do [ĩ], anteriormente analisado.

As impedâncias na zona do velo, na Figura 5.18, mostram que a impedância de carga da zona faríngea é, para as frequências até cerca de 1 kHz, muito mais próxima do valor da impedância de entrada da cavidade bucal do que da impedância de entrada do tracto nasal. Tem-se portanto, um caso mais próximo do da vogal [ɐ̃] do que da vogal [ĩ].

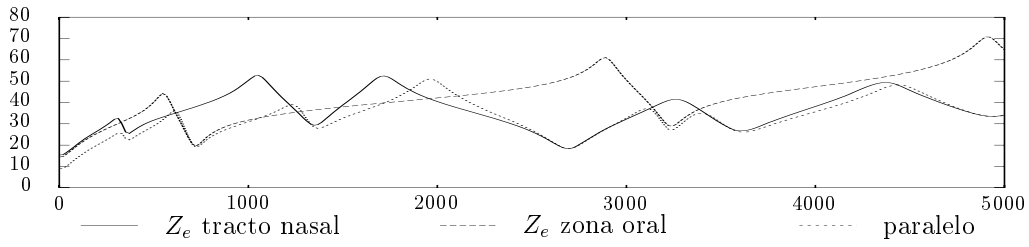


Figura 5.18: Módulos das impedâncias na zona do velo para a vogal [ũ]. Apresenta-se a impedância da cavidade oral, a impedância de entrada do tracto nasal e o paralelo das duas impedâncias.

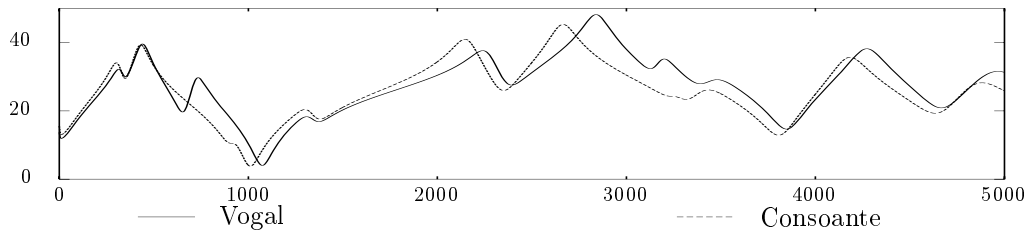


Figura 5.19: Comparação da impedância de entrada da vogal [ũ] com a da consoante nasal bilabial diferindo de [ũ] apenas pela oclusão.

Encerrou-se este conjunto de simulações analisando o efeito da oclusão labial para o [ũ]. As impedâncias de entrada, apresentadas na Figura 5.19, mostram que existe alguma alteração, sem no entanto ser tão acentuada como no caso do [ɐ̃]. A oclusão não é uma grande alteração à configuração dos lábios, que para a vogal nasal já têm uma abertura reduzida.

5.4 Resumo

Neste capítulo foram apresentados resultados, de simulações e testes perceptuais, acerca da interacção entre a fonte e as cavidades supra-laríngeas durante a produção de sons nasais. Abordou-se essencialmente as alterações adicionais causadas pelo acoplamento do tracto nasal. Pretendia-se averiguar da necessidade da inclusão do efeito da carga de entrada do tracto nasal num modelo de interacção entre fonte glotal e o tracto.

As primeiras simulações por nós efectuadas mostraram algum efeito nas características da onda de excitação glotal, devidas ao acoplamento do tracto nasal. Este efeito é muito mais visível no domínio do tempo do que no domínio da frequência. As simulações também mostraram que o efeito não tem a mesma relevância para todas as vogais. Nas simulações efectuadas, para vogais nasais no contexto C$\tilde{\text{V}}$C, é também notório que o efeito da carga nasal adicional varia ao longo da vogal nasal, dependendo da posição do velo e de outros articuladores.

Os resultados destas primeiras simulações motivaram a realização de testes perceptuais para saber se as alterações são perceptíveis pelos ouvintes. Os resultados obtidos mostram que as alterações não são de fácil identificação. Este resultado está de acordo com o obtido por Titze e Story (1997), em que o acoplamento nasal não apresentou qualquer efeito digno de registo no fluxo glotal. Existe, no entanto, uma tendência para o efeito ser mais facilmente detectável, portanto mais relevante, para a vogal elevada [ĩ], produzida com uma cavidade bucal de área bastante reduzida.

Simulações efectuadas depois dos testes sugerem como explicação para esta diferença, do efeito de interacção entre as vogais, a relação entre a impedância de entrada do tracto nasal e a impedância de entrada da cavidade oral na zona de acoplamento nasal. Para vogais baixas, como o [ɐ̃], a impedância de entrada da cavidade bucal é baixa o que torna difícil a carga de entrada do tracto nasal afectar a impedância de carga da zona da faringe. Para vogais com passagem oral bastante reduzida, com o [ĩ], a impedância de carga da faringe é essencialmente a impedância de entrada nasal. A oclusão oral, criada no final de uma vogal nasal quando se encontra antes de uma consoante oclusiva, não provoca grandes alterações na impedância de entrada, e por conseguinte na onda glotal. O efeito desta oclusão aumenta com o aumento da área da cavidade bucal e da área de abertura dos lábios. Vogais como o [ɐ̃] são as que sofrem maior variação devido à sua configuração bucal e dos lábios, bastante abertos.

Os resultados apontam para a necessidade de, pelo menos para as vogais com passagem oral reduzida entre o velo e os lábios, considerar a impedância de entrada do tracto, ao modelar-se a interacção entre a fonte e o filtro.

Devido aos múltiplos factores envolvidos, apenas se pode considerar o trabalho apresentado neste capítulo como primeiras experiências, inseridas num conjunto mais alargado que é necessário efectuar. Muitos parâmetros podem influenciar a interacção: a frequência fundamental; área do acoplamento nasal; sistema subglotal, etc. Também são possíveis, e necessárias, muitas melhorias nos modelos e parâmetros utilizados. É necessária informação das configurações

articulatórias e parâmetros relacionados com a fonte para sons nasais naturais da língua portuguesa.

As simulações por nós efectuadas não incluem efeitos da variação dos articuladores e da glote durante cada período de excitação glotal. Ultrapassada esta limitação [3] poderá chegar-se a outros resultados.

[3]Foram já realizados estudos deste tipo para vogais orais (Laine, 1999). No entanto, a técnica utilizada não é de fácil adaptação a sons nasais.

Capítulo 6

Efeito da variação dos articuladores na percepção de nasalidade

Eppur si muove.

Galileo Galilei

An understanding of ... timing constraints in speech production (as well as **the response of listeners to time varying signal of the kind that occur in speech) is of basic importance in the study of language**, since this knowledge may eventually provide some **basis for explaining the constraints that exist in the sequential patterns of features that are allowed in various languages**, and in some of the modifications that occur in utterances that are produced in a rapid and casual manner." [a]"

[a] Negrito adicionado pelo autor.

Kenneth Stevens
(citado em Hardcastle e Laver, 1996, pág. 355)

A qualidade das vogais nasais produzidas pelos sintetizadores actuais, especialmente os que não utilizam concatenação de sinal natural, precisa de ser melhorada. Uma forma de contribuir para a solução deste problema é o aprofundamento do nosso conhecimento sobre como estes sons são produzidos e percebidos.

Grande parte dos conhecimentos sobre nasais relaciona-se com características estáticas. No entanto, diversos estudos apontam para a necessidade de considerar a voz como um fenómeno dinâmico (Olive *et al.*, 1993; Kühnert e Nolan, 1997, por exemplo). A influência da informação dinâmica na percepção de vogais orais tem sido bastante estudada (Strange, 1989, para uma panorâmica sobre o assunto). Também foi referida a possibilidade de informação dinâmica desempenhar um papel relevante na percepção de vogais nasais.

Um dos primeiros a levantar essa hipótese, da possível dependência da percepção de nasalidade da forma de variação no tempo do velo, foi Clumeck (1976).

Feng e Castelli (1996), baseando-se em simulações de vogais nasais do Francês, propuseram que as vogais nasais deviam ser consideradas como começando numa configuração simples, a configuração de uma vogal oral, e tendendo para outra configuração simples, a configuração faringo-nasal, em que é eliminada a passagem oral, sendo a passagem do ar feita pela faringe e tracto nasal [1]. Para eles a nasalização está relacionada com a dinâmica da variação entre estas duas configurações, inicial e final.

Ohala e Ohala (1993) também referem a possibilidade da utilização de propriedades dinâmicas na percepção de vogais nasais utilizadas no inventário fonológico de uma língua.

Para o Português, a manifestação da nasalidade na dimensão temporal foi primeiro mencionada por Lacerda e Strevens em 1956 (facto referido em Almeida, 1976). De acordo com vários estudos na área da Fonética, as vogais nasais do Português diferem das nasais vogais de outras línguas, como o Francês, por serem apenas fortemente nasais no final (Trigo, 1993). Alguns estudos consideram mesmo que as vogais nasais portuguesas apresentam contornos de nasalidade, isto é, a sua parte inicial pode ser considerada como oral e a sua parte final como nasal [2]. Um estudo acústico das vogais nasais do Português do Brasil (de Sousa, 1994) refere também a existência das três fases: início oral, transição, e murmúrio nasal final. É proposto que o murmúrio nasal é co-articulado com a vogal e não com a consoante seguinte, como é geralmente referido noutros trabalhos. Esta hipótese, segundo a autora, necessita de experiências com estímulos sintéticos que a suportem ou rejeitem.

A naturalidade das vogais nasais depende da língua nativa do ouvinte. Stevens *et al.* (1987) obtiveram, em testes de classificação de naturalidade e qualidade, usando ouvintes falantes nativos de Francês, Português, e Inglês, diferenças dependentes da língua. Os falantes de Português preferem algum murmúrio nasal e parecem ter uma ligeira preferência para mais nasalização da vogal, mas não necessariamente durante toda a vogal.

Na tentativa de produzir vogais nasais de qualidade natural, usando síntese articulatória, chegámos à conclusão que, utilizando uma abertura do velo adequada, as vogais produzidas eram de facto percebidas como nasais mas de qualidade nada natural. Claramente, apenas a abertura do velo, com o consequente acoplamento das cavidades nasais, não era suficiente para obter qualidade natural. Decidimos pela introdução de mais informação no processo de síntese. Quando fizemos o velo variar no tempo produzindo uma vogal nasal consistindo em três fases (início oral, transição causada pela abertura do velo e fase final com velo aberto), característica geralmente mencionada acerca das vogais nasais do Português Europeu, a qualidade melhorou acentuadamente. Estes primeiros testes, preliminares, foram efectuados para

[1] No original, "...the conception of the nasal vowel as a trend beginning with a simple configuration (the oral one), which is terminated in the same manner (the pharyngonasal one)..." (Feng e Castelli, 1996, pág. 3704).

[2] Estes estudos referem-se quase sempre a vogais nasais entre oclusivas, contexto que se representa neste trabalho por C$\tilde{\text{V}}$C.

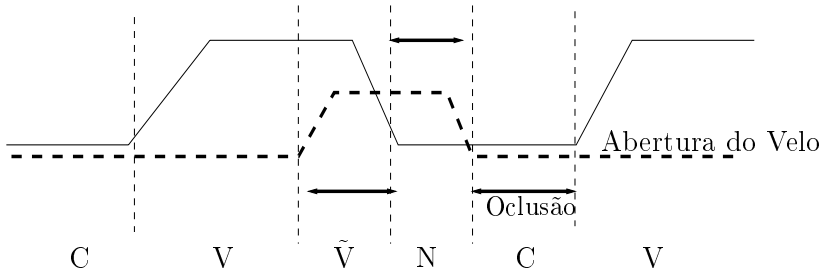


Figura 6.1: Variação do velo (linha a tracejado) e articuladores orais numa sequência CṼC.

uma vogal (Teixeira *et al.*, 1998b). No entanto, os resultados motivaram a realização de testes perceptuais mais cuidados para ter um melhor conhecimento da influência da variação do velo no tempo, a dinâmica do velo, na qualidade das vogais nasais portuguesas.

Os estudos apresentados de seguida têm como objectivo tentar obter uma resposta para a seguinte questão: É importante a dinâmica do velo para a produção de vogais nasais, do Português Europeu, com qualidade natural ?

Primeiro, estudamos a vogal nasal entre duas oclusivas orais, secção 6.1, e depois, vogais noutros contextos fonéticos. Na secção 6.2, estudam-se vogais nasais depois de uma consoante nasal, incluindo o caso, raro, de vogais nasais entre duas consoantes nasais. Na secção 6.3 aborda-se o caso das vogais nasais isoladas. Encerra o capítulo uma breve secção, resumindo o conteúdo e os resultados principais.

6.1 Vogal nasal entre consoantes não nasais

Nesta primeira experiência estudamos o caso em que a vogal nasal se encontra entre duas consoantes oclusivas não nasais.

Na Figura 6.1 apresenta-se, esquematicamente, a variação do velo e da abertura oral neste contexto. Durante a consoante oclusiva antes da vogal nasal o velo e a passagem oral encontram-se fechadas. O início da vogal nasal coincide com a abertura da passagem oral. Como o velo é um articulador lento vai existir um período de tempo em que existe passagem oral desobstruída e velo fechado, ou quase, tendo-se por consequência uma zona com características de uma vogal oral, representada na figura por V. A abertura do velo provoca uma zona com passagem de fluxo simultaneamente pelo tracto oral e nasal, zona representada por Ṽ. Como a passagem oral tem de fechar para a consoante seguinte, e ao fechar antes do velo, cria uma zona com radiação apenas nasal, representada por N. O ponto de articulação desta consoante nasal, criada devido a coarticulação, é igual ao da consoante seguinte. Refira-se que a cavidade oral tem de fechar antes do velo para evitar a criação de uma zona com passagem apenas oral.

6.1.1 Teste perceptual

Estando interessados na qualidade decidimos pela utilização de um teste de qualidade: o teste AB. Apesar de obrigar a um maior número de decisões, aumentando a duração do teste, o método AB é preciso.

6.1.1.1 Procedimento

Antes da realização do teste AB, e para saber se de facto os ouvintes são capazes de discriminar os estímulos, foi efectuado um teste de discriminação 4IAX por 2 ouvintes. Os 2 ouvintes foram capazes de distinguir os estímulos mais de 95% das vezes.

Os sinais foram apresentados aos ouvintes através de auscultadores Sennheiser Headmax HD 470 numa sala com baixo (mas não desprezável) ruído ambiente. O teste foi efectuado utilizando um programa de computador, usando uma interface gráfica, em que os ouvintes indicam a resposta utilizando o rato.

A pergunta feita aos ouvintes foi a seguinte: "Qual dos dois sons prefere como uma vogal nasal do Português ?".

Durante a preparação do teste notou-se a dificuldade por parte dos ouvintes em escolher em diversas situações. As causas para esta dificuldade eram de dois tipos: as vogais serem demasiado iguais ou ambas serem de baixa qualidade. Para obviar a esta situação foram adicionadas duas novas possibilidades de resposta ao teste. Ficaram assim as seguintes 4 hipóteses de resposta disponíveis: "o primeiro", "o segundo", "ambos", e "nenhum".

Os diversos estímulos foram repetidos 5 vezes em ambas as ordens AB e BA. O intervalo entre os estímulos (ISI do Inglês *Inter Stimuli Interval*) utilizado foi de 600 ms.

O teste foi dividido em 2 partes. Na primeira foram comparados estímulos produzidos com o velo estático com estímulos produzidos com velo dinâmico. Na segunda parte efectuou-se a comparação entre estímulos com e sem uma consoante nasal bilabial no final. Na segunda parte apenas se utilizaram estímulos com velo variável.

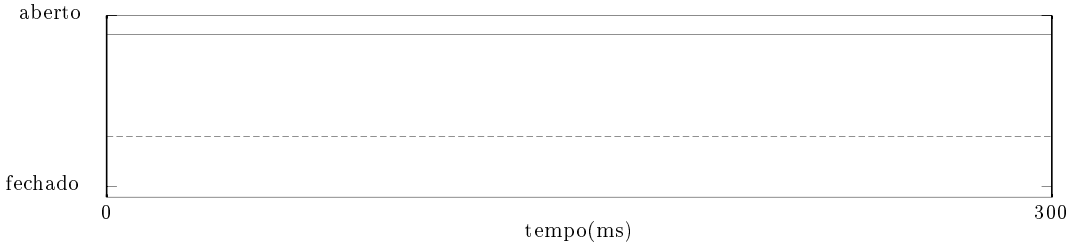
6.1.1.2 Estímulos

Foram produzidos estímulos para as 5 vogais nasais do Português padrão, usando o sintetizador articulatório.

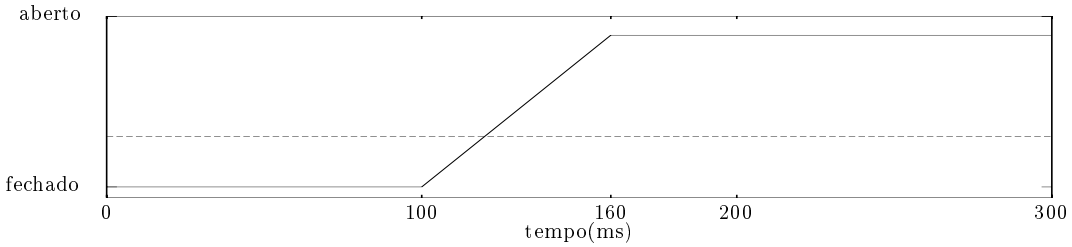
As configurações do tracto oral foram obtidas pelo processo de inversão, utilizando optimização. As configurações resultantes foram verificadas utilizando configurações publicadas. A abertura do velo foi obtida de uma forma manual ajustando o quociente entre a área de entrada do tracto nasal e a área da passagem oral, na zona do velo, para um valor de 10.

Os tempos utilizados para o velo, nos casos variáveis, não foram muito precisos, pois não possuímos dados detalhados sobre o movimento do velo. Usamos informação da literatura e

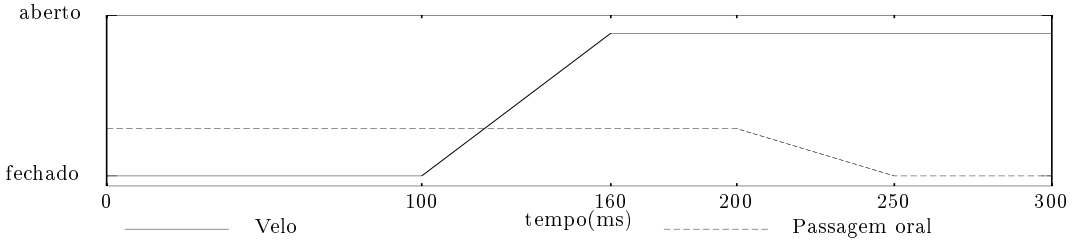
análise de vogais naturais em contextos CṼC. Utilizaram-se os mesmos tempos para todas as vogais. Nos primeiros 100 *ms* o velo mantém-se fechado, faz depois uma transição para a máxima abertura durante 60 *ms*, mantendo-se depois na posição de máxima abertura.



(a) Estímulo com abertura constante do velo.



(b) Estímulo com abertura do velo variável.



(c) Estímulo como velo variável e consoante nasal no final.

Figura 6.2: Variação da abertura do velo e da passagem oral para os três estímulos utilizados, para cada vogal, no estudo do contexto CṼC.

Os estímulos com a consoante nasal, bilabial ([m]), no final foram produzidos fechando os lábios. O movimento de fecho dos lábios começa em 200 *ms* e dura 50 *ms*.

A duração dos estímulos, de 300 *ms*, foi mantida constante para todas as vogais.

Para cada uma das vogais, cinco no total, gerou-se estímulos com velo dinâmico e velo estático. No caso do velo dinâmico foram gerados dois estímulos: um com uma consoante nasal bilabial no final, o outro sem consoante no final. Obtiveram-se assim 3 estímulos para cada vogal, apresentados na Figura 6.2 para o caso do [ɐ̃].

Na produção dos estímulos foi utilizado o modelo de fonte glotal incluindo interacção entre

a fonte e o tracto. Foi utilizada frequência fundamental variável no tempo. F_0 começa em 100 Hz, sobe até 120 Hz aos 100 ms e depois decresce até 100 Hz novamente no final do estímulo. Outros valores para parâmetros da fonte utilizados durante esta experiência encontram-se resumidos na Tabela 6.1.

Parâmetro	Símbolo	Valor	Unidade
Pressão Pulmonar	P_l	10000	$dine/cm^2$
Quociente de Abertura	OQ	0.6	
Quociente de Velocidade	SQ	2	
Área glotal mínima	Ag_0	0	cm^2
Área glotal máxima	Ag_{max}	0.3	cm^2
Constante de *steepness*	Sk=A2-A1	0.03	
Jitter		2	%
Shimmer		5	%

Tabela 6.1: Valores para os parâmetros da fonte glotal do sintetizador articulatório utilizados para a síntese dos estímulos utilizados no teste da influência da dinâmica em contextos CṼC.

A utilização de *jitter* e *shimmer* foi motivada por testes com vogais orais. Esses testes mostraram que a utilização de F_0 variável conjuntamente com *jitter* melhora a qualidade obtida. O ganho de qualidade com a adição de *shimmer* não é tão notório.

Os estímulos foram gerados utilizando uma frequência de amostragem de 10 kHz.

6.1.1.3 Ouvintes

Um total de 11 ouvintes, 9 do sexo masculino, falantes nativos do Português Europeu participaram nos testes. Nenhum sofria de problemas auditivos. As suas idades variavam entre os 23 e os 53 anos. Em termos de local de nascimento e local onde tinham residido mais tempo, os ouvintes que participaram nos testes provinham, na sua maioria, da região Norte do país. Em termos de escolaridade, encontravam-se representados diversos níveis, desde o ensino básico até a ouvintes com Doutoramento, com predominância de universitários.

Os resultados para cada par de estímulos foi verificado em termos de consistência da resposta. Apenas os pares em que o ouvinte preferia um dos estímulos em 70% dos casos ou mais foram retidos para análise. Desta formai, pares em que os ouvintes não tinham a certeza da resposta são rejeitados. Na primeira parte do teste, 9 num total de 66 casos, 13.7 %, foram rejeitados. Refira-se que dois dos ouvintes foram responsáveis por 7 dos casos. Na segunda parte, os ouvintes tiveram muito mais dificuldades em serem coerentes. Em 41 dos 66 casos, aproximadamente 62 % dos casos, os ouvintes obtiveram coerência inferior a 70 %. No caso da vogal [ũ], nenhum dos ouvintes conseguiu a coerência necessária. Dois ouvintes não conseguiram ser coerentes em nenhum dos casos. Tendo sido utilizadas as mesmas condições e os mesmos ouvintes nas duas partes do teste, ressalta a muito maior dificuldade da segunda parte dos testes.

A confiança (em Inglês *reliability*) (Rosenthal e Rosnow, 1991) dos ouvintes foi examinada calculando a correlação entre as respostas dos vários ouvintes. O valor obtido foi baixo.

Julgamos poder atribuir esta baixa confiança às condições, longe de óptimas, de realização dos testes. De referir que alguns ouvintes responderam claramente na direcção oposta à tendência geral. Deste resultado, conjuntamente com resultados de outros testes por nós realizados, este facto pode dever-se às vogais nasais serem de mais difícil percepção. Ao se pedir ao ouvinte para escolher a melhor nasal, este tem a tendência para escolher o som mais nítido (fácil de perceber). No entanto, o som mais nítido que ele prefere, claramente, não pode ser a melhor vogal nasal. Mantiveram-se, mesmo assim, os resultados de todos os ouvintes que participaram nos testes.

6.1.1.4 Resultados

Os resultados da primeira parte do teste, comparação entre velo fixo e velo variável, encontram-se na Tabela 6.2(a). Claramente, os ouvintes preferem os estímulos com velo variável. A preferência atinge, incluindo todas as vogais e ouvintes, uma média de 7.1 vezes em 10 possíveis. A preferência mantém-se se analisarmos os resultados para cada vogal, sendo no entanto menos notória no caso da vogal [õ]. Três dos ouvintes revelaram preferir o caso estático, indo claramente contra a tendência geral. Pelo menos para dois deles, esta tendência pode estar relacionada com a baixa consistência das suas escolhas. De facto, dois dos ouvintes que responderam de forma contrária à tendência geral são os mesmos que tiveram problemas notórios de consistência. Mesmo com estes casos, o efeito da variação do velo é muito notório.

Os resultados de análise de variância (ANOVA) de medida-repetida mostraram como significativo o efeito de variação do velo $[F(1,10) = 5.67,\ p < 0.05]$ e não significativo $(p > 0.05)$ o efeito da vogal e da interacção entre a vogal e a variação do velo.

Na Tabela 6.2(b) apresentam-se os resultados da segunda parte do teste. Como já foi referido, na segunda parte testou-se a influência na qualidade da existência ou não de uma consoante nasal na fase final da vogal. Em termos gerais, os ouvintes preferiram os estímulos com consoante nasal. Esta preferência é também patente nos resultados individuais para cada vogal. Apenas na primeira versão da vogal [õ] a preferência é menos notória.

Os resultados de análise de variância, com dois efeitos principais, revelam como significativo o efeito da consoante nasal no final $[F(1,8) = 9.5,\ p < 0.05]$ e não significativo $(p > 0.05)$ o efeito da vogal e da interacção entre os efeitos principais.

6.1.2 Discussão

A hipótese de as vogais nasais produzidas com velo fixo ou variável serem percebidas como de qualidade semelhante foi rejeitada, com um nível de significância de 5%. Claramente, as vogais nasais com o velo variável no tempo foram consideradas de maior naturalidade. Este resultado indica que é necessário incluir a informação sobre a forma de variação da abertura do velo ao longo do tempo para se obterem vogais sintéticas de elevada qualidade. Os resultados destes testes devem ser, e foram, encarados como motivação para continuar este tipo de estudos.

Ouvinte	[ẽ]		[õ]		[ɐ̃]		[ĩ]		[ũ]		[ɔ̃]		Total	
	V	F	V	F	V	F	V	F	V	F	V	F	V	F
1	3	7	-	-	-	-	1	9	-	-	-	-	2.0	8.0
2	10	0	10	0	10	0	10	0	10	0	10	0	10.0	0.0
3	10	0	10	0	9	1	10	0	10	0	10	0	9.8	0.2
4	-	-	3	7	-	-	-	-	8	1	0	9	3.7	5.7
5	2	8	1	9	0	10	10	0	3	7	0	10	2.7	7.3
6	1	9	3	7	10	0	10	0	10	0	0	10	5.7	4.3
7	10	0	9	1	9	1	10	0	9	1	8	2	9.1	0.8
8	10	0	9	1	10	0	10	0	8	1	9	1	9.3	0.5
9	10	0	1	9	0	9	0	10	9	1	0	10	3.3	6.5
10	-	-	10	0	9	0	7	0	10	0	10	0	9.2	0.0
11	9	1	9	1	10	0	10	0	10	0	-	-	9.6	0.4
Média	7.2	2.8	6.5	3.5	7.4	2.3	7.8	1.9	8.7	1.1	5.2	4.7	7.1	2.7
Var	15.7	15.7	15.6	15.6	18.0	16.8	15.8	16.1	4.7	4.6	30.0	23.8	15.4	15.2

(a) Resultados do teste comparativo de estímulos com velo variável (V) e fixo (F).

Ouvinte	[ẽ]		[õ]		[ɐ̃]		[ĩ]		[ũ]		[ɔ̃]		Média	
	S	C	S	C	S	C	S	C	S	C	S	C	S	C
1	0	10	0	10	-	-	-	-	-	-	0	10	0.0	10.0
2	7	3	-	-	-	-	-	-	-	-	0	9	3.5	6
3	9	0	-	-	7	3	-	-	-	-	7	2	7.7	1.7
4	-	-	-	-	-	-	-	-	-	-	-	-	-	-
5	0	9	0	10	0	9	0	8	-	-	0	10	0.0	9.2
6	0	10	0	10	1	9	0	8	-	-	0	10	0.2	9.4
7	-	-	-	-	0	10	-	-	-	-	1	7	0.5	8.5
8	-	-	-	-	-	-	-	-	-	-	-	-	-	-
9	-	-	-	-	1	7	-	-	-	-	-	-	1.0	7.0
10	8	2	-	-	-	-	-	-	-	-	-	-	8.0	2.0
11	-	-	0	9	0	10	-	-	-	-	0	10	0.0	9.7
Média	4.0	5.7	0.0	9.8	1.5	8.0	0.0	8.0	-	-	1.1	8.3	1.6	7.8
Var	19.6	20.3	0.0	0.3	7.5	7.2	0.0	0.0	-	-	6.8	8.9	9.5	9.8

(b) Resultados do teste comparativo de estímulos com (C) e sem (S) consoante nasal no final.

Tabela 6.2: Resultados do teste perceptual da influência da variação dos articuladores na qualidade de vogais nasais entre oclusivas.

6.2 Vogais nasais depois de uma consoante nasal

As vogais nasais do Português Europeu não surgem apenas entre consoantes oclusivas (orais). Existem também, embora em menor número, vogais nasais depois de consoantes nasais. Este contexto coloca novos problemas. Contrariamente ao caso analisado na secção anterior o velo não se encontra fechado na fase inicial da vogal.

Para as vogais nasais depois de consoante nasal os nossos estudos consistiram de três partes: (1) análise de vogais nasais naturais no contexto em causa; (2) simulações utilizando o sintetizador articulatório; e (3) testes perceptuais.

6.2.1 Análise de vogais nasais naturais

A análise de vogais naturais centrou-se na determinação de semelhanças para as vogais depois de uma oclusiva e depois de uma consoante nasal. Analisaram-se os vários contextos [3] em que a vogal nasal se pode encontrar.

Na Figura 6.3 apresenta-se um exemplo de vogal nasal antes de uma oclusiva, contexto que designamos por NṼOcl. Na figura representa-se em simultâneo o sinal de voz e a energia ao longo do tempo. É visível o aumento de energia na transição da consoante para o início da vogal, o decréscimo gradual, e a reduzida energia no final e no segmento seguinte, o [t].

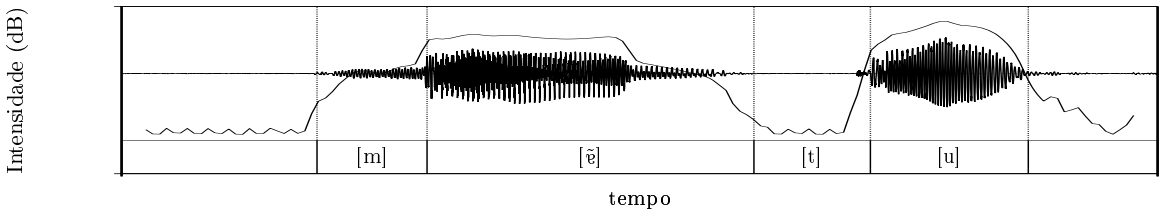


Figura 6.3: Exemplo de uma vogal nasal natural, depois de uma consoante nasal e antes de uma oclusiva (NṼOcl). Apresenta-se o sinal de voz e a energia.

A comparação entre uma vogal oral e a vogal nasal correspondente, antes de uma fricativa, é apresentada na Figura 6.4. A transição de energia elevada para uma energia baixa é muito mais rápida para a vogal oral. A energia é muito mais estável durante a realização da vogal oral do que durante a vogal nasal.

A vogal nasal no final de palavra, outro dos contextos, é analisada na Figura 6.5. Para comparação é apresentada a vogal oral correspondente no mesmo contexto. A energia aumenta na transição entre a consoante nasal anterior e o início de ambas as vogais. O decréscimo de energia no final é muito mais gradual para a vogal nasal.

[3] A representação adoptada para o contexto é a seguinte: uma consonante nasal é representada por N; uma vogal oral por V; vogal nasal por Ṽ; consoante oclusiva Ocl; consoante fricativa Fric; o símbolo # indica que o segmento que se encontra à sua esquerda se encontra no final.

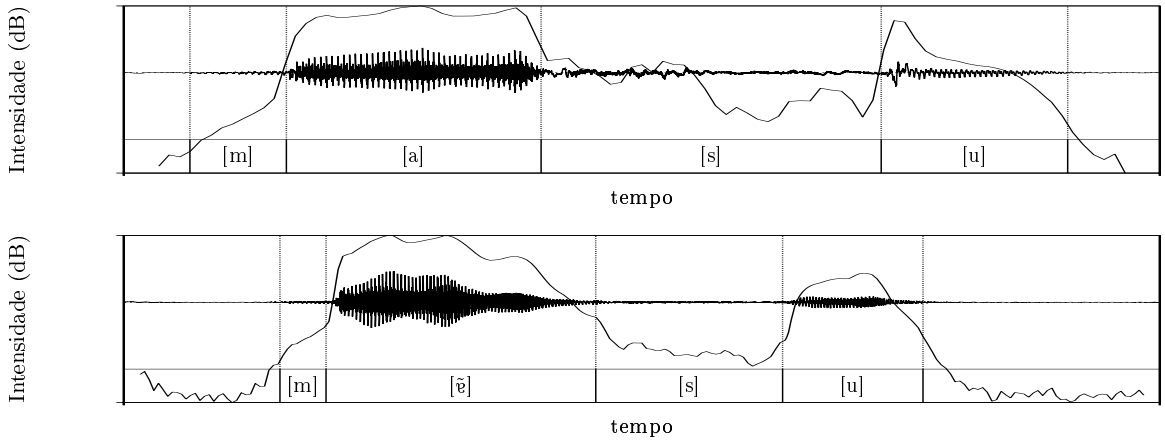


Figura 6.4: Exemplo de uma vogal nasal natural depois de uma consoante nasal e antes de uma fricativa (NṼFric). Para comparação, apresenta-se a vogal oral correspondente no mesmo contexto. Apresenta-se o sinal de voz e a energia.

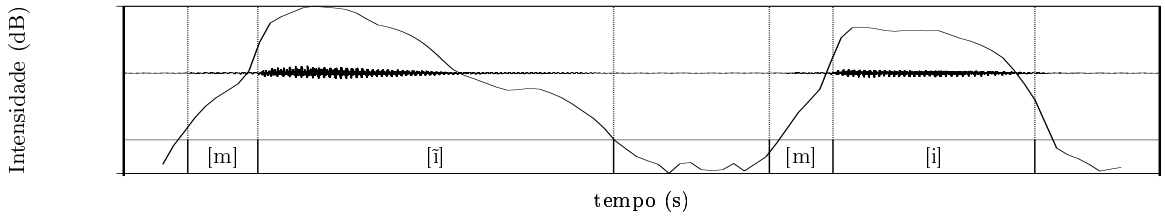


Figura 6.5: Exemplo de uma vogal nasal depois de uma consoante nasal em posição final (NṼ#). Apresenta-se o sinal de voz e a energia.

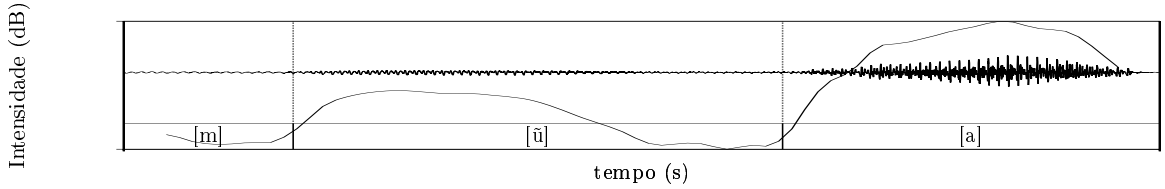


Figura 6.6: Exemplo de uma vogal nasal natural depois de uma consoante nasal e antes de uma vogal. (NṼV). Apresenta-se o sinal de voz e a energia.

Outros contextos apenas ocorrem em sequências de palavras. Na Figura 6.6 é apresentado o caso de vogal nasal antes de uma vogal oral. A vogal nasal constitui o fim de uma palavra e a oral o início de outra. Mais uma vez o final da vogal nasal apresenta uma baixa energia. Apenas depois de um segmento de baixa energia, a energia aumenta, devido à ocorrência da vogal oral.

Resumindo, em todos os contextos, é visível um aumento de energia no início da vogal nasal,

e o decréscimo gradual da energia durante a vogal. Também nunca se segue, imediatamente à vogal nasal, um segmento de energia elevada.

6.2.2 Simulações

A variação da energia no início e durante a vogal nasal foi investigada através de simulações. Na Figura 6.7 apresenta-se o sinal radiado, separadamente, pelos lábios e pelas narinas, para a sequência [mɐ̃]. É notório que a abertura da passagem oral faz com que a radiação oral se torne dominante no início da vogal nasal. Este aumento ocorre mesmo para vogais com uma passagem oral bastante reduzida, como é o caso da vogal [ĩ], na Figura 6.8. Em relação à radiação nasal, esta mantém-se quase inalterada no caso de vogais baixas, como o [ɐ̃], ou decresce, como no [ĩ]. Estas alterações levam a que se tenha uma predominância da radiação oral logo no início da vogal nasal. Esta dominância deve-se à menor resistência oferecida pela passagem oral à propagação do sinal sonoro. Considerando o efeito conjunto das duas radiações, assiste-se a um aumento da energia total no início da vogal.

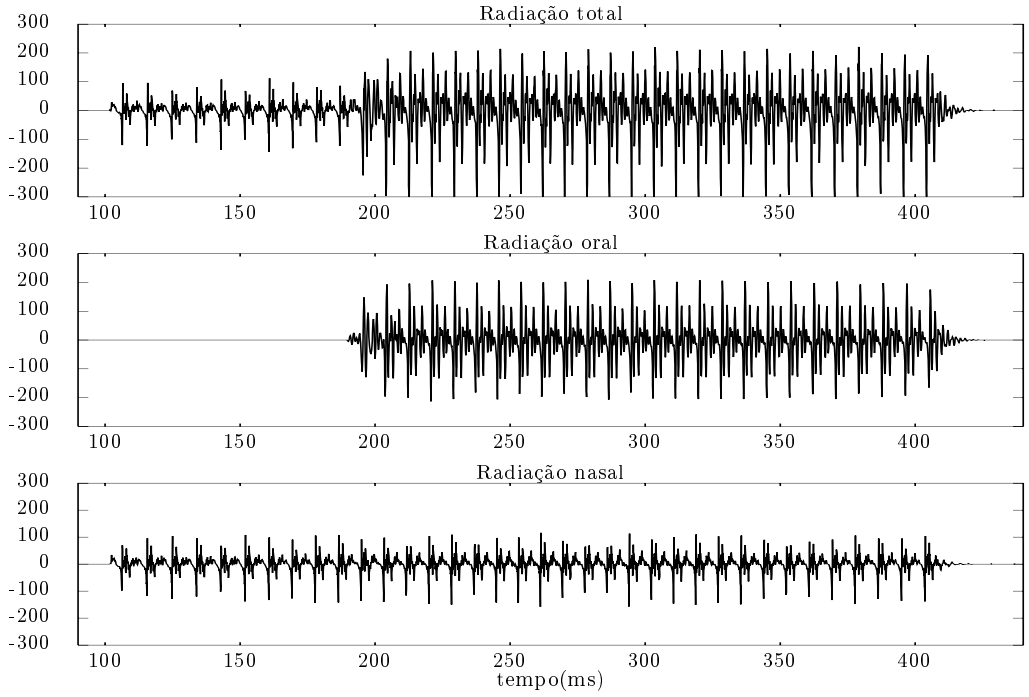


Figura 6.7: Resultados da simulação para a sequência [mɐ̃].

Na Figura 6.8 apresenta-se o resultado da simulação para uma vogal nasal antes de uma oclusiva. A variação temporal da posição do velo, típica deste contexto, provoca um decréscimo progressivo da radiação oral. Na fase final, a oclusão da passagem oral, imposta pela oclusiva seguinte, antes do fecho do velo, origina um segmento apenas com radiação nasal. Obtêm-se desta forma uma consoante nasal no final. O efeito na energia total radiada é uma diminuição

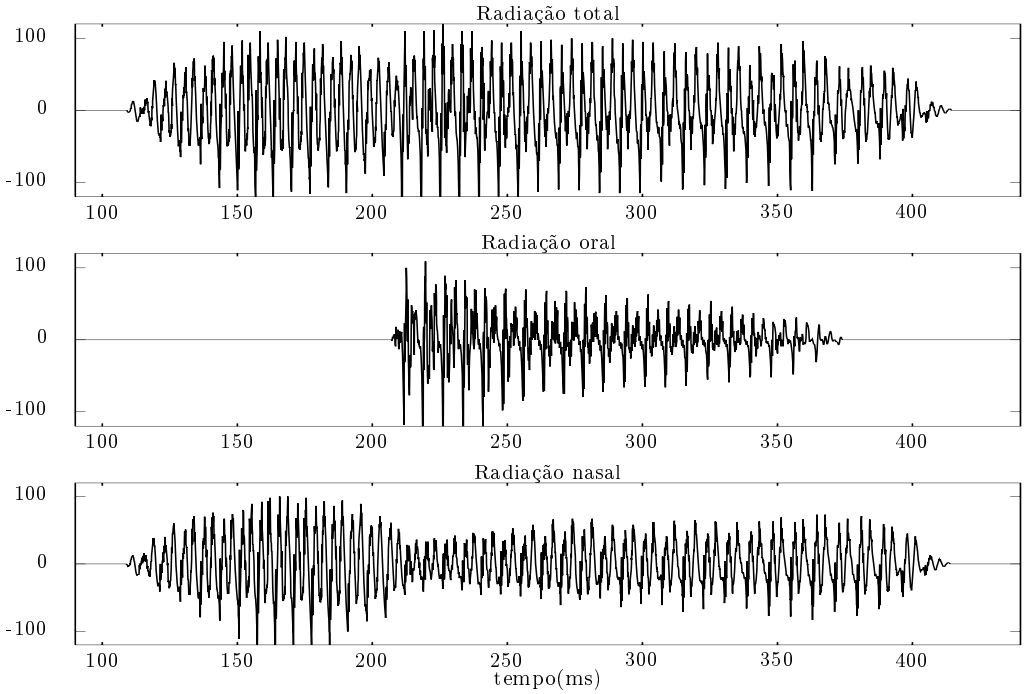


Figura 6.8: Simulação da sequência [mĩOcl].

gradual, terminando num segmento de baixa energia com radiação nasal. No caso de o segmento seguinte ser uma fricativa, o segmento final terá dominância da radiação nasal causada pela reduzida passagem oral.

6.2.3 Testes perceptuais

Para estudar a influência na percepção da variação no tempo da posição do velo para este contexto, foram efectuados dois testes perceptuais. A qualidade obtida, devido à falta de informação detalhada sobre a posição dos articuladores ao longo do tempo, motivou a realização de um teste de identificação. A facilidade de identificação pode ser utilizada como medida da qualidade dos estímulos. Apesar das reservas, em relação à qualidade, foi também efectuado um teste de preferência.

6.2.3.1 Descrição dos estímulos utilizados nos testes

Para a realização dos testes perceptuais foram sintetizados estímulos para várias vogais. Foram obtidos estímulos com variação do velo e restantes articuladores, característicos de vários contextos, diferindo entre si pelo segmento que se segue à vogal nasal.

Os estímulos são apresentados na Tabela 6.3. A primeira coluna da tabela apresenta o contexto. São apresentados, além da vogal nasal, o segmento anterior e o segmento seguinte.

Contexto	Factor	Níveis	Estímulo
N$\tilde{V}$V	duração da	0	1
	vogal oral	50	2
	seguinte	-40	3
N$\tilde{V}$Ocl	consoante nasal final	40	4
	criada por	-40	5
	coarticulação	10	6
N$\tilde{V}$#	consoante	0	7
	nasal final	40	8
	criada por	-1	9
	coarticulação	fric 40	10
N$\tilde{V}$	abertura	N=$\tilde{V}$	11
abertura do velo		N<>$\tilde{V}$	12
constante		$\tilde{V}$ = 0	13
N$\tilde{V}$Fric	duração da fase com	40	14
	passagem oral reduzida	10	15
N$\tilde{V}$N	duração	0	A
	da segunda	25	B
	consoante	50	C
	nasal	100	D
NVN		100	E

Tabela 6.3: Estímulos utilizados nos testes de identificação e preferência realizados para vogais nasais depois de uma consoante nasal.

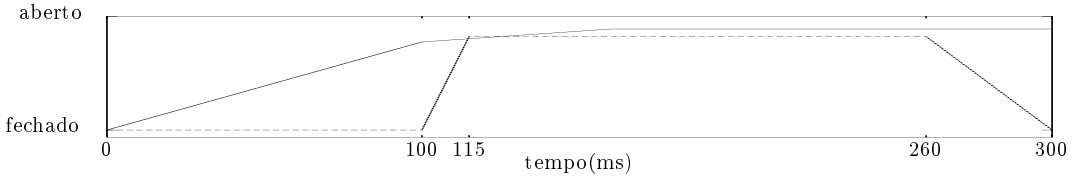
Na segunda coluna indica-se o parâmetro variável utilizado para cada contexto. Os diversos níveis utilizados para o parâmetro apresentam-se na terceira coluna. De forma a facilitar a referência a um tipo de estímulo, apresenta-se na última coluna um identificador que será utilizado nas descrições seguintes.

Na descrição, bastante resumida, dos diversos estímulos, seguiremos a ordem pela qual os contextos aparecem na Tabela 6.3.

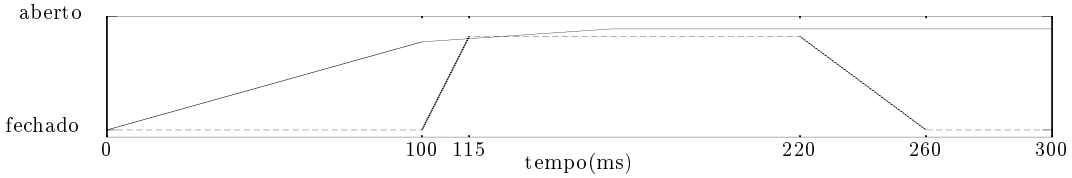
No primeiro contexto, em que a vogal nasal se encontra antes da vogal oral produzida apenas pela alteração da posição do velo, pretendeu-se estudar o efeito da duração dessa vogal oral. Realizaram-se três estímulos. O primeiro estímulo consiste no caso em que a vogal tem duração zero, isto é, o velo apenas fecha no final do estímulo. No segundo estímulo a vogal oral tem 50 ms de duração. No terceiro caso foi retirada também a fase de transição. Como nos estímulos gerados se considerou necessários 40 ms para o velo realizar o movimento de fecho, este caso é indicado com -40 para representar a inexistência da vogal e da transição.

Para o segundo contexto, vogal nasal seguida de oclusiva, o parâmetro variável representa a duração da consoante nasal, criada por coarticulação, no final da vogal nasal. Três casos foram considerados: consoante nasal de duração 40 ms; consoante de 10 ms; e não existência

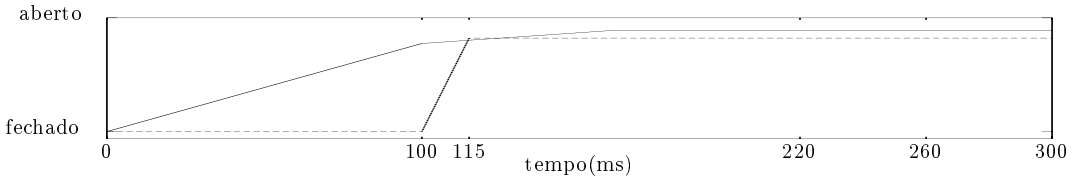
da consoante nem da zona de transição para a formação dessa consoante. Este último caso é representado como tendo duração $-40\ ms$, devido a ser precisamente $40\ ms$ o tempo utilizado para a passagem do tracto vocal de aberto a fechado. Os valores de $10\ ms$ e $40\ ms$ baseiam-se nos resultados de Stevens *et al.* (1987). Segundo estes, ouvintes portugueses preferem estímulos com consoante nasal antes da oclusiva, com duração de cerca de $40\ ms$ e ouvintes franceses preferem durações de $10\ ms$.



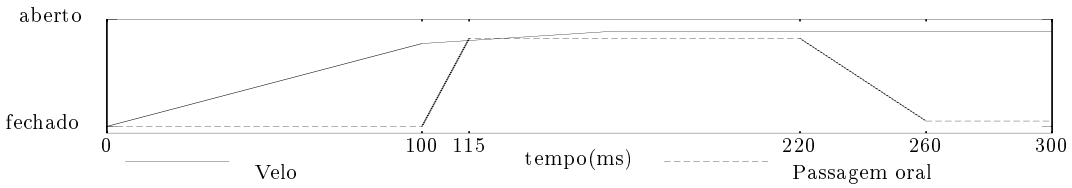
(a) Estímulo 7: A passagem oral fecha mesmo no final.



(b) Estímulo 8: A passagem oral é fechada $40\ ms$ antes do final.



(c) Estímulo 9: A passagem oral mantém-se aberta no final.



(d) Estímulo 10: A passagem oral é reduzida $40\ ms$ antes do final.

Figura 6.9: Variação do velo e abertura oral para os estímulos do contexto NṼ#.

Quando a vogal nasal se encontra em posição final, terceiro contexto apresentado na tabela, interessa-nos investigar as diferentes formas de realização da fase final. Podem ocorrer três situações: a passagem oral manter-se aberta até ao final; a passagem oral fechar antes de acabar o fluxo nasal; ou ainda, a passagem oral tender para o fecho, sem o atingir. A primeira situação resulta no estímulo com o nível -1, indicando que não existe formação de consoante

nasal no final. Para a segunda situação, podem ter-se várias durações da consoante nasal final. Optou-se por utilizar apenas dois casos: duração igual a 40 *ms*; e duração 0, isto é, o fecho da passagem oral coincide com o fim da produção da vogal. Para a terceira situação apenas se considerou um único caso, duração da fase final com passagem oral reduzida de 40 *ms*. A indicação do nível como *fric* 40 deve-se ao facto de a área oral reduzida é semelhante à utilizada na produção de fricativas. O valor de 40 *ms* baseou-se no trabalho de Stevens *et al.* (1987). A variação no tempo do velo e abertura oral, para os quatro estímulos representativos deste contexto, encontra-se na Figura 6.9.

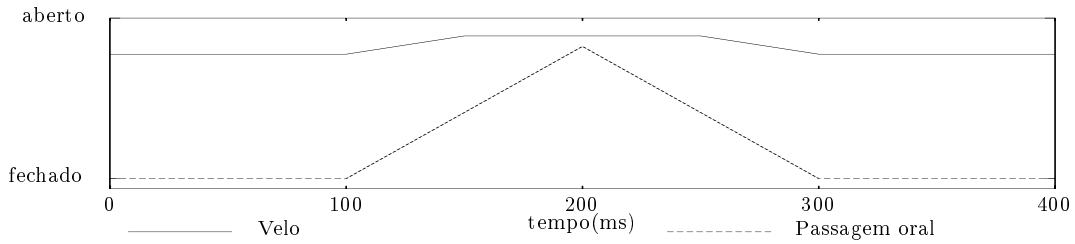
O conjunto seguinte de estímulos não se referem propriamente a um contexto. Os estímulos produzidos servem de termo de comparação ao não incluírem variação no tempo, da posição do velo. Produziram-se três estímulos para o caso em que se considera o velo constante: um com o velo constante durante toda a realização do estímulo (representado por N=$\tilde{\text{V}}$ para indicar que abertura do velo durante a consoante N e a vogal nasal $\tilde{\text{V}}$ é igual); um segundo caso com diferentes aberturas para a consoante e para a vogal nasal (representado por N<>$\tilde{\text{V}}$); um último em que o velo se mantém fechado durante toda a vogal nasal ($\tilde{\text{V}}$=0 para indicar velo fechado durante a vogal nasal).

Para o caso de vogal nasal antes de uma fricativa, o parâmetro variável utilizado foi a duração da fase final da vogal em que existe passagem oral com área reduzida, mantendo-se o velo aberto. Duas durações, desta fase final, foram consideradas: 10 e 40 *ms*. Note-se que não é produzida a fricativa que se segue. Também não é considerada a possibilidade de durante a fase em que a passagem oral se encontra bastante reduzida, com valores típicos de uma fricativa, se produzir um som fricativo nasal.

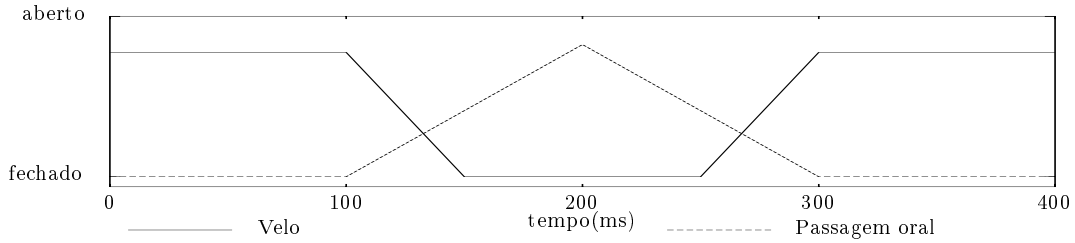
Para o caso de vogal nasal entre duas consoantes nasais, no nosso caso, duas consoantes nasais bilabiais, variou-se a duração da segunda consoante. Pretendia-se saber qual o efeito desta segunda consoante na percepção da vogal nasal. Gerou-se também um estímulo em que o velo fecha e volta a abrir durante a vogal, para simular o caso de uma vogal oral no mesmo contexto. Obtiveram-se 5 estímulos, identificados na tabela pelas letras A, B, C, D e E. Nos primeiros 4, o velo mantém-se sempre aberto e varia-se a duração da segunda consoante. Utilizaram-se durações de 0, 25, 50 e 100 *ms*. No quinto estímulo, o velo fecha e abre durante a vogal e a consoante tem duração 100 *ms*. Os estímulos D e E são representados na Figura 6.10.

É óbvio que se poderia ter variado outros factores em cada um dos contextos, podendo-se também experimentar com muitos outros níveis para cada parâmetro. No entanto, a necessidade de ter um número reduzido de estímulos levou a restringir os casos possíveis.

Para os vários contextos referidos, o movimento do velo durante a consoante nasal inicial e início da vogal nasal foi modelado da forma seguinte: para vogais elevadas, o velo desce durante a consoante e depois sobe durante a parte final da consoante e fase inicial da vogal; para as vogais baixas o velo continua a baixar durante a fase inicial da vogal. Considera-se que o velo se encontra, no início, na posição de fechado. Este modelamento baseou-se nos resultados de (Clumeck, 1976, pág. 345).



(a) NṼN



(b) NVN

Figura 6.10: Variação do velo e abertura oral para dois estímulos representativos de vogais orais e nasais, entre consoantes nasais.

A configuração do tracto utilizada para o [m] foi obtida por ajuste manual da posição dos articuladores, com base em configurações publicadas em (Delattre, 1968, pág. 69), (Stevens, 1998, pág. 488) e (Fant, 1960). As configurações do tracto oral para as vogais foram obtidas por inversão a partir de valores de formantes para vogais orais.

Os parâmetros utilizados para a excitação glotal foram semelhantes aos utilizados no estudo do contexto CṼC, descritos na secção 6.1.1.2, página 150.

6.2.4 Teste de identificação

6.2.4.1 Estímulos

Os estímulos utilizados para este teste foram obtidos da forma já referida. Para não tornar o teste demasiado extenso foram utilizadas apenas três vogais em alguns dos contextos. O número de vogais utilizadas em cada contexto encontra-se indicado na Tabela 6.5, página 165.

6.2.4.2 Procedimento

Foi efectuado um teste de identificação. A tarefa dos ouvintes consistiu em identificar qual a vogal ouvida, escolhendo de entre um número fixo de hipóteses incluindo nove vogais orais ([a],[ɐ],[ɛ],[e],[ɪ],[ɔ],[o],[u] e [ə]) e cinco vogais nasais ([ɐ̃],[ẽ],[ĩ],[õ] e [ũ]). Quando nenhuma era

considerada adequada, os ouvintes podiam usar outras duas hipóteses: uma para representar outra vogal oral diferente das nove existentes em Português, a outra para representar um som nasal diferente das cinco vogais nasais do Português.

Os estímulos foram apresentados aos ouvintes através de auscultadores em salas com ruído ambiente baixo. O teste foi realizado individualmente. Foi permitido repetir o estímulo o número de vezes que o ouvinte pretendesse.

6.2.4.3 Ouvintes

Participaram no teste 8 ouvintes, 7 do sexo masculino, todos falantes nativos do Português Europeu. Nenhum sofria de problemas auditivos. As suas idades variavam ente os 23 e os 53 anos. Em termos de local de nascimento e local onde tinham residido mais tempo os ouvintes que participaram nos testes provinham da zona litoral entre Porto e Aveiro. Em termos de escolaridade, encontravam-se representados diversos níveis, desde o ensino básico até a ouvintes com Doutoramento, com grande predominância de ouvintes com frequência universitária.

Para aferir a consistência de cada um dos ouvintes, calculou-se o número médio de identificadores diferentes utilizados para cada um dos estímulos. Um ouvinte com consistência óptima apenas utilizaria um único identificador, enquanto um ouvinte que utilizasse sistematicamente três identificadores diferentes seria inconsistente. Os vários ouvintes utilizaram entre 1.30 e 1.94 identificadores diferentes. Em termos dos vários tipos de estímulos (1 a 15 e A,B,C,D,E) os valores situaram-se entre 1.38 e 1.88, com média 1.61 e desvio padrão 0.14.

A concordância entre os ouvintes foi calculada usando um método bastante simples. O número de vezes que dois ouvintes obtiveram o mesmo número de identificações, dividido pelo total de possibilidades, multiplicado por cem, dá a taxa de concordância (Schweigert, 1994, pág. 87). Os valores obtidos encontram-se na Tabela 6.4(a). Os valores médios para cada ouvinte, apresentados na última coluna da tabela, variam entre 47.62 e 68.57. Como existem 3 repetições, existem 4 resultados possíveis. Os casos em que dois ouvintes não obtiveram o mesmo resultado não podem ser considerados semelhantes. É muito diferente não estarem de acordo, quando um identificou três e o outro duas ou quando um identificou duas vezes e o outro nenhuma. Na Tabela 6.4(b), apresenta-se a concordância dos ouvintes quando se consideram apenas dois resultados possíveis: identificou pelo menos um; não identificou nenhum.

6.2.4.4 Resultados

Os resultados do teste de identificação permitem várias análises. Pode analisar-se o efeito do contexto e, para cada contexto, analisar a relação entre os vários tipos de estímulos utilizados. Pode também estudar-se o efeito da vogal nos resultados. Os resultados dos testes foram processados automaticamente, de forma a se obter o número de identificações da vogal nasal pretendida. São estes valores que consideramos os resultados do teste e cuja análise se apre-

Ouvinte	1	2	3	4	5	6	7	8	média (%)
1	—	60.00	66.67	53.33	73.33	73.33	80.00	73.33	68.57
2	60.00	—	33.33	60.00	46.67	33.33	53.33	46.67	47.62
3	66.67	33.33	—	53.33	66.67	53.33	60.00	73.33	58.10
4	53.33	60.00	53.33	—	46.67	46.67	46.67	46.67	50.48
5	73.33	46.67	66.67	46.67	—	73.33	80.00	80.00	66.67
6	73.33	33.33	53.33	46.67	73.33	—	73.33	66.67	60.00
7	80.00	53.33	60.00	46.67	80.00	73.33	—	80.00	67.62
8	73.33	46.67	73.33	46.67	80.00	66.67	80.00	—	66.67
								média global	60.71

(a) Cálculo efectuado considerando todos os 4 resultados possíveis como diferentes.

Ouvinte	1	2	3	4	5	6	7	8	média (%)
1	—	69.70	62.12	62.12	71.21	75.76	77.27	72.73	70.13
2	69.70	—	62.12	53.03	56.06	66.67	62.12	57.58	61.04
3	62.12	62.12	—	60.61	63.64	68.18	60.61	71.21	64.07
4	62.12	53.03	60.61	—	51.52	65.15	48.48	71.21	58.87
5	71.21	56.06	63.64	51.52	—	59.09	90.91	77.27	67.10
6	75.76	66.67	68.18	65.15	59.09	—	62.12	72.73	67.10
7	77.27	62.12	60.61	48.48	90.91	62.12	—	71.21	67.53
8	72.73	57.58	71.21	71.21	77.27	72.73	71.21	—	70.56
								média global	65.80

(b) Cálculo considerando apenas dois resultados possíveis: identificou uma ou mais vezes; não identificou.

Tabela 6.4: Concordância entre os ouvintes no teste de identificação de vogais nasais depois de consoante nasal.

senta de seguida. Na análise detalhada de cada contexto, apresenta-se também a distribuição das respostas.

Efeito do contexto

A soma, para todas as vogais e ouvintes, do número de vezes que um tipo de estímulo é identificado como a vogal nasal pretendida é apresentada na Tabela 6.5. Como o número de vogais não foi igual em todos os contextos, o valor anterior expresso em percentagem do máximo de identificações possíveis, também indicado na tabela, pode ser utilizado como medida de qualidade de cada estímulo. Para facilitar a análise apresenta-se também a ordenação dos estímulos dentro de cada contexto (correspondendo 1 ao mais identificado) e os primeiros 6 classificados de entre todos os estímulos, com excepção do contexto N$\tilde{\text{V}}$N.

Numa primeira análise da tabela ressalta a existência de contextos com taxas de identificação bastante baixas, o contexto com velo constante e N$\tilde{\text{V}}$V, e outros com taxas na ordem dos 50 % para alguns dos estímulos que os constituem. Estímulos com variação do velo ao longo do tempo obtêm taxas de identificação razoáveis. A presença de uma vogal oral a seguir à vogal nasal torna muito baixa a taxa de identificação.

O resultado de análise de variância confirma como significativo o efeito do contexto [$F(5, 35) =$

Contexto	Vogais	Número do Estímulo	Identificação Total	Identificação %	Ordenação Contexto	Ordenação Geral
N$\tilde{V}$V	5	1	14	11.7	2	
		2	6	5.0	3	
		3	20	16.7	1	
N$\tilde{V}$Ocl	5	4	63	52.5	1	1
		5	15	12.5	3	
		6	40	33.3	2	6
N$\tilde{V}$#	3	7	22	30.6	3	
		8	34	47.2	1	2
		9	13	18.1	4	
		10	31	43.1	2	3
Velo	3	11	3	4.2	2	
Constante		12	3	4.2	2	
		13	4	5.5	1	
N$\tilde{V}$Fric	5	14	47	39.2	1	4
		15	42	35.0	2	5
N$\tilde{V}$N	1	A	12	50.0	1	
		B	9	37.5	4	
		C	11	45.8	2	
		D	9	37.5	4	
NVN	1	E	11	45.8	2	

Tabela 6.5: Resultados dos testes de identificação com 8 ouvintes e 3 repetições.

$17.26,\ p < 0.001]$, da vogal $[F(4,28) = 7.34,\ p < 0.001]$, e da interacção contexto vogal $[F(12,84) = 8.50,\ p < 0.001]$.

Apenas foram efectuados testes *post hoc* (teste de Newman-Keuls usando nível de significância de 0.05) para o contexto. Os resultados revelam como significativamente ($p = 0.05$) pior o contexto N$\tilde{V}$V do que todos os restantes, com excepção do contexto em que o velo é constante. Também o contexto com velo constante obtém significativamente piores resultados do que todos os restantes, com excepção do contexto N$\tilde{V}$V. Todas as outras diferenças entre contextos revelaram-se não significativas.

Análise de cada contexto

Para cada contexto, foi analisada a existência de diferença entre os vários tipos de estímulos que o compõem.

Para o contexto N$\tilde{V}$V, as taxas médias de identificação são algo diferentes para os três casos. Os estímulos do tipo 2 apresentam a taxa mais baixa, seguidos dos de tipo 1, sendo os melhores os de tipo 3. Se olharmos para a forma como os estímulos foram identificados, apresentada na Tabela 6.6, nota-se que os ouvintes identificaram maioritariamente os estímulos como vogais orais. Para duas das vogais, [ẽ] e [ĩ], quando consideraram os estímulo nasal, não foram capazes de identificar a vogal como uma das 5 vogais nasais portuguesas. Análise de variância confirma como significativa a diferença com $F(2,14) = 4.71,\ p < 0.05$. O efeito da vogal não é

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
1	a	**15**	4						3			2					
2	a	**17**	1	1					2			3					
3	a	**17**	3						2			2					
1	e			5	**12**	1					1		5				1
2	e			7	**10**	2					1		2				2
3	e	1	1	4	2	2					1	2	5				**6**
1	o			1			**16**	1	3					3			
2	o		1	1			**13**	4	3		2						
3	o		1				**15**	1	1		1			5			
1	u								**21**						3		
2	u							1	**20**					1	1		1
3	u						1		**17**		1				4		1
1	i			**9**		1				4			2			1	7
2	i			6		2				**8**					2		6
3	i			6		1			1	2			1		3	4	**7**

Tabela 6.6: Resultados do teste de identificação para o contexto NṼV (vogal nasal depois de consoante nasal e antes de vogal oral).

significativo ($p > 0.05$). Testes *post hoc* revelaram como significativa apenas a diferença entre os estímulos 2 e 3. O 3 é significativamente mais identificado como vogal nasal do que o 2. A existência de uma vogal oral (estímulo 2) com 50 *ms* influencia negativamente a percepção da vogal nasal.

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
4	a	2	6								2	**14**					
5	a	**20**	3									1					
6	a	**11**	8									5					
4	e	1	1		8	2						2	**9**				1
5	e		1	1	**21**						1						
6	e				**14**	3						1	6				
4	o		3				**8**	1				3		7			2
5	o						**19**	1			1			3			
6	o		1				**17**	2				1		3			
4	u						1		2						**21**		
5	u						1	1	**13**		1			1	7		1
6	u								5						**19**		
4	i			5							1		3			**12**	3
5	i			5	1					**7**	1		4			4	3
6	i			**9**		1				2			2			7	3

Tabela 6.7: Resultados do teste de identificação para o contexto NṼOcl (vogal nasal depois de consoante nasal e antes de oclusiva).

No contexto NṼOcl, as taxas de identificação são também diferentes para os três casos. Os estímulos tipo 5 apresentam a taxa mais baixa, seguidos dos de tipo 6, sendo os melhores os de tipo 4. Se olharmos para a forma como os estímulos foram identificados, apresentada na Tabela 6.7, nota-se que o estímulo 4 foi identificado maioritariamente como nasal para quatro das cinco vogais, o estímulo 5 apenas para uma vogal e o outro nenhuma. Para todas as vogais, o número de identificações como vogal nasal foi sempre maior para o 4 que para o 6.

Para o [ĩ], a nasalização nalguns casos levou os ouvintes a ouvirem um som nasal diferente das cinco nasais do Português. Análise de variância confirma como significativa a diferença com $F(2,14) = 19.31,\ p < 0.001$. O efeito da vogal é significativo $[F(4,28) = 5.06,\ p < 0.01]$. Os testes *post hoc* revelaram como significativa a diferença entre os três estímulos. Pode concluir-se que a duração da consoante nasal tem influência significativa na qualidade dos estímulos. Os estímulos com consoante nasal de 40 ms são os mais identificados, seguidos dos que possuem uma consoante nasal de 10 ms.

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
7	a	**13**	7								1	3					
8	a	3	7						1			**12**		1			
9	a	**19**	3									2					
10	a	2	8								1	**13**					1
7	u								5						**19**		
8	u								2						**22**		
9	u							3	**11**						10		
10	u								6						**18**		
7	i			4									1		**14**		5
8	i			4					1		3		2		**8**		6
9	i			6					1				1		**11**	1	5
10	i			3		1			4					1	**11**		5

Tabela 6.8: Resultados do teste de identificação para o contexto NṼ# (vogal nasal depois de consoante nasal no final de palavra).

No contexto NṼ#, as taxas de identificação são diferentes para os quatro casos. Os estímulos tipo 8 e 10 apresentam as taxas mais elevadas de identificação, seguidos do tipo 7. O tipo 9 obtém, claramente, menos identificações. A forma como os estímulos foram identificados, apresentada na Tabela 6.8, mostra que os estímulos 8 e 10 são os únicos maioritariamente identificados como a vogal nasal pretendida, para a três vogais. Para duas das vogais, o estímulo 9 foi maioritariamente identificado como vogal oral. Para a vogal [ĩ], uma parte dos estímulos foi considerado nasal, mas não sendo uma das cinco vogais nasais do Português. Análise de variância confirma como significativa a diferença entre os estímulos com $F(3,21) = 14.40,\ p < 0.001$. O efeito da vogal é significativo $[F(2,14) = 24.71,\ p < 0.001]$. Testes *post hoc* revelaram como não significativa a diferença entre os tipos 8 e 10. Este grupo é significativamente melhor que o tipo 7. Todos os outros tipos são significativamente melhores que o tipo 9, em que não existe movimento de fecho da cavidade oral. Os resultados apontam como melhorando a identificação dos estímulos a existência de fecho, ou redução acentuada, da passagem oral no final da vogal nasal.

Quando se considera o velo constante, as taxas de identificação são bastante baixas e semelhantes para os três casos considerados. Se olharmos para a forma como os estímulos foram identificados, na Tabela 6.9, para duas das vogais, os estímulos são identificados maioritariamente como vogal oral e na outra como uma vogal nasal, mas sem ser nenhuma das vogais do Português. Análise de variância confirma como não significativa a diferença entre os estímulos $[F(2,14) = 0.57,\ p > 0.05]$ e não significativo o efeito da vogal $[F(2,14) = 1.11,\ p > 0.05]$.

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
11	a	**23**															1
12	a	**20**	3						1								
13	a	**22**	1									1					
11	u								**22**						2		
12	u						1		**21**						3		
13	u							1	**22**						1		
11	i			6		1			1	2			5		1	1	**8**
12	i			6		1			1		2		3		4		**7**
13	i			**9**		2				7			4			2	

Tabela 6.9: Resultados do teste de identificação para estímulos com velo constante.

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
14	a	6	**9**	1							2	4		1			1
15	a	**9**	8								2	5					
14	e		2	1	**7**	2						4	7				2
15	e		1	1	**14**						1	1	5				1
14	o	1	3				**11**	2			2			3			2
15	o	1	1				**14**	1			1	1		6			
14	u							1	1						**22**		
15	u								5						**19**		
14	i			6						1			1			**11**	5
15	i			**7**						4			2			7	4

Tabela 6.10: Resultados do teste de identificação para o contexto NṼFric (vogal nasal depois de consoante nasal e antes de fricativa).

No contexto NṼFric, ambos os tipos de estímulos apresentam taxas de identificação entre os 30 e 40 %, não se podendo dizer que difiram grandemente. A forma como os estímulos foram identificados, na Tabela 6.10, é bastante variável com a vogal. Existe uma vogal, o [ũ], em que ambos os estímulos são maioritariamente identificados como a vogal nasal pretendida; vogais, como o [õ], a maioria dos estímulos é identificada como [ɔ]. Torna-se muito difícil apontar a preferência entre os dois tipos de estímulos. Análise de variância confirma como não significativa a diferença entre os estímulos $[F(1,7) = 0.30,\ p > 0.05]$. O efeito da vogal é significativo $[F(4,28) = 9.29,\ p < 0.001]$.

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
A	a	2	7	1							1	**12**					1
B	a	5	8								1	**9**					1
C	a	4	8								1	**11**					
D	a	7	8								1	**9**					
E	a	4	9									**11**					1

Tabela 6.11: Resultados do teste de identificação para o contexto NṼN e NVN (vogal nasal, e oral, entre duas consoantes nasais).

Para vogais situadas entre duas consoantes nasais, contextos NṼN e NVN, as taxas de iden-

tificação são algo diferentes. Dois dos estímulos apresentam taxas de 37.5 % e os outros três taxas próximas dos 50 %. A forma como os estímulos foram identificados, Tabela 6.11, neste caso, não traz grande informação adicional devido ter sido utilizada apenas uma vogal. Refira-se, no entanto, a distribuição das respostas, quando consideram os estímulos como vogal oral, entre as vogais [ɐ] e [a], com certa preponderância para a primeira. Como na obtenção dos estímulos se utilizou uma configuração do tracto a que corresponde em termos perceptuais um [a], a abertura do velo tem influência na altura da vogal. A nasalização torna a vogal, perceptualmente, como tendo sido produzida por uma configuração oral com a língua mais elevada. Análise de variância não confirma como significativa a diferença entre os estímulos $[F(4, 28) = 0.50,\ p > 0.05]$. Em termos de identificação não parece ser detectado o movimento de fecho do velo que ocorre no estímulo E. A presença da consoante nasal a seguir transforma este estímulo numa vogal nasal, com taxas de identificação semelhantes aos outros casos.

6.2.5 Teste de preferência

O objectivo do teste é descobrir os estímulos preferidos, quanto á sua qualidade, de entre um conjunto alargado de estímulos. A utilização directa de um teste AB tornaria necessário comparar cada estímulo com todos os restantes. Sendo N o número total de estímulos este método obrigaria a $N(N-1)$ comparações [4]. Naturalmente, durante as comparações existem estímulos que são considerados sistematicamente de qualidade inferior. Tendo em vista o objectivo, de saber qual os melhores, é ineficiente tê-los até ao fim.

Uma forma de obviar aos problemas referidos é realizar o teste AB usando um método em que se vão eliminando, sucessivamente, os piores estímulos (Rosemberg, 1971, pág. 588). Desta forma reduz-se o número de comparações necessárias e as realizadas são utilizadas com os pares de melhor qualidade. Este método é semelhante aos torneios, realizados na idade média, ou actualmente, em modalidades desportivas. O método de eliminação pode ser simples, em que se elimina um estímulo após uma única derrota ou, mais complexo, eliminando-se os piores classificados ao fim de um conjunto de comparações. A forma como se classifica pode também variar. A técnica mais simples é ordenar pelo número de vitórias. Outra técnica consiste em atribuir um ponto mais a pontuação actual do oponente derrotado ao estímulo vencedor. Esta técnica é designada por pontuação geométrica, pois as pontuações aumentam de uma forma aproximadamente geométrica, reflectindo, a dificuldade que aumenta de uma forma, também, aproximadamente geométrica. Está implícita a transitividade: se A derrota B e B derrota C, então A derrotaria C. No início de cada fase os estímulos podem ser agrupados de forma aleatória ou utilizando os resultados anteriores. Um exemplo desta última hipótese é utilizada em torneios de futebol. Após uma eliminatória, os primeiros classificados de cada grupo jogam com os segundos de outros grupos. Rosemberg (1971) comparou várias destas variantes e concluiu que "...os resultados não são dependentes do tipo de torneio usado para

[4] No nosso caso, estamos interessados em ter as duas combinações de cada par, isto é, AB e BA.

os obter.". Por outras palavras, os resultados por ele obtidos pelos vários métodos não foram significativamente diferentes.

6.2.5.1 Procedimento

No teste desenvolvido optou-se por: pontuação geométrica, com uma ligeira alteração; realização do teste uma única vez; pares construídos de forma aleatória; eliminação de metade dos estímulos após cada fase (definida pelo número de comparações a efectuar para cada estímulo).

Para cada comparação de um par de estímulos, os ouvintes tiveram quatro opções de resposta: escolher o primeiro estímulo como o melhor; escolher o segundo; considerar ambos como de boa qualidade e perceptualmente indistintos; ou considerar que nenhum dos estímulos é um exemplar de qualidade do estímulo em causa. A última opção não é geralmente incluída em torneios, mas consideramos que, com a sua utilização, aumenta a capacidade do teste de identificar os melhores estímulos [5]. Como já foi referido, a vitória implica a atribuição ao estímulo vencedor da pontuação do perdedor adicionada de um. A escolha da última opção implica a perda de uma percentagem da pontuação por parte de cada estímulo. Na implementação efectuada esta percentagem é de 50 %, o que é bastante penalizador. Como estamos interessados nos melhores, a má sorte dos outros é considerada vantajosa.

Cada vogal foi testada em separado. Para não tornar o teste demasiado demorado, cada fase consistiu em 4 comparações para as vogais [ĩ] e [ũ], em que se tinha 16 estímulos [6], e de 3 comparações para [ɐ̃] devido aos 20 estímulos.

Os testes foram realizados individualmente, numa sala com ruído ambiente baixo. Os estímulos foram ouvidos através de auscultadores.

O resultado retido dos testes é apenas a ordenação final dos estímulos, obtida com base nos pontos obtidos. Não foi utilizado qualquer critério para desempate. Estímulos com os mesmos pontos ficam empatados.

6.2.5.2 Estímulos

Neste teste foram apenas utilizados exemplares dos estímulos para três das vogais. Foram escolhidas as vogais [ĩ], [ũ] e [ɐ̃]. As duas outras vogais nasais são muitas vezes realizadas como ditongos, e ocupam posições intermédias, sendo, por isso, as candidatas naturais a serem excluídas.

[5] A utilização desta opção em certos torneios desportivos seria certamente do agrado do público, permitindo penalizar equipas que entraram em jogo para empatar!

[6] Um dos 15 estímulos foi repetido para se obter um número par. Pode usar-se os resultados obtidos para os dois casos do estímulo repetido como medida da validade do teste.

6.2.5.3 Ouvintes

Participaram no teste 8 ouvintes, sendo 7 do sexo masculino, todos falantes nativos do Português Europeu. Nenhum sofria de problemas auditivos. As suas idades variavam ente os 23 e os 53 anos. Em termos de local de nascimento e local onde tinham residido mais tempo, os ouvintes que participaram nos testes provinham da zona litoral entre Porto e Aveiro. Em termos de escolaridade, encontravam-se representados diversos níveis, desde o ensino básico até a ouvintes com Doutoramento, com predominância de universitários.

Ouvinte	Vogal	correlação
5	[ũ]	0.5756
7	[ɐ̃]	0.8463
7	[ĩ]	0.8856
6	[ĩ]	0.8070
6	[ũ]	0.9002

Tabela 6.12: Correlação entre as classificações dos estímulos obtidas par duas repetições do teste de preferência para vogais depois de consoante nasal.

A consistência dos ouvintes foi aferida através da repetição do teste (Rosenthal e Rosnow, 1991, pág. 47) por alguns ouvintes para uma ou duas das vogais, escolhidas aleatoriamente. A correlação entre as classificações nas duas repetições é apresentada na tabela 6.12.

Para avaliar se os ouvintes estavam de acordo entre si, foi calculada a correlação entre a classificação atribuída aos estímulos pelos vários ouvintes. A correlação ordinal (em Inglês *rank correlation*) foi obtida utilizando o método apresentado em (Sachs, 1984, pág. 402) para o cálculo do coeficiente de correlação ordinal de Spearman (Bryman e Cramer, 1993, pág. 213). Esta escolha deveu-se às classificações serem variáveis ordinais e existirem muitos empates. Os valores obtidos para as três vogais encontram-se na Tabela 6.13. O valor médio, apresentado na última coluna, é a confiança r de cada ouvinte. O valor de r variou entre 0.35 e 0.80, excluindo-se os casos negativos. Apenas um dos ouvintes, para a vogal [ũ], respondeu claramente de forma contrária a todos os outros [7]. O valor de r apenas nos dá a confiança de um dos ouvintes. O valor para a confiança do conjunto dos n ouvintes, designada por confiança efectiva, pode ser calculada utilizando a fórmula de Spearman-Brown (Rosenthal e Rosnow, 1991, pág. 51),

$$R = \frac{nr}{1 + (n-1)r}. \tag{6.1}$$

A média conjunta, incluindo todos os resultados, das correlações para as três vogais é de $r = 0.56$. Se excluirmos o caso com correlação negativa obtém-se $r = 0.61$. Estes valores médios resultam em $R = 0.91$ e $R = 0.93$ respectivamente. Devido ao número de ouvintes

[7] Esta situação pode ter sido motivada pela utilização do sistema de torneio. Neste sistema, caso o ouvinte no início do teste, por inexperiência, por exemplo, cometa alguns erros de avaliação, pode levar à eliminação de estímulos de boa qualidade, o que provoca, na fase final, dificuldade em escolher o melhor e consequentemente a resultados estranhos.

Ouvinte	1	2	3	4	5	6	7	8	r
1	—	0.62	0.54	0.41	-0.68	0.34	0.54	0.64	0.35
2	0.62	—	0.72	0.80	-0.68	0.69	0.85	0.79	0.54
3	0.54	0.72	—	0.74	-0.74	0.68	0.53	0.76	0.46
4	0.41	0.80	0.74	—	-0.61	0.89	0.87	0.86	0.57
5	-0.68	-0.68	-0.74	-0.61	—	-0.54	-0.61	-0.74	-0.66
6	0.34	0.69	0.68	0.89	-0.54	—	0.82	0.79	0.52
7	0.54	0.85	0.53	0.87	-0.61	0.82	—	0.88	0.56
8	0.64	0.79	0.76	0.86	-0.74	0.79	0.88	—	0.57
								média	0.36
								R	0.82
								média	0.51
								R	0.89

(a) Resultados para [ũ]

Ouvinte	1	2	3	4	5	6	7	8	r
1	—	0.39	0.68	0.58	0.41	0.65	0.25	0.61	0.51
2	0.39	—	0.82	0.72	1.00	0.78	0.82	0.83	0.77
3	0.68	0.82	—	0.91	0.82	0.77	0.66	0.90	0.79
4	0.58	0.72	0.91	—	0.72	0.76	0.65	0.89	0.75
5	0.41	1.00	0.82	0.72	—	0.78	0.80	0.84	0.77
6	0.65	0.78	0.77	0.76	0.78	—	0.66	0.81	0.75
7	0.25	0.82	0.66	0.65	0.80	0.66	—	0.71	0.65
8	0.61	0.83	0.90	0.89	0.84	0.81	0.71	—	0.80
								média	0.72
								R	0.95

(b) Resultados para [ĩ]

Ouvinte	1	2	3	4	5	6	7	8	r
1	—	0.46	0.35	0.58	0.75	0.82	0.66	0.19	0.54
2	0.46	—	0.55	0.59	0.62	0.39	0.76	0.63	0.57
3	0.35	0.55	—	0.48	0.35	0.35	0.50	0.84	0.49
4	0.58	0.59	0.48	—	0.69	0.68	0.77	0.58	0.62
5	0.75	0.62	0.35	0.69	—	0.87	0.91	0.49	0.67
6	0.82	0.39	0.35	0.68	0.87	—	0.75	0.32	0.60
7	0.66	0.76	0.50	0.77	0.91	0.75	—	0.66	0.71
8	0.19	0.63	0.84	0.58	0.49	0.32	0.66	—	0.53
								média	0.59
								R	0.92

(c) Resultados para [ẽ]

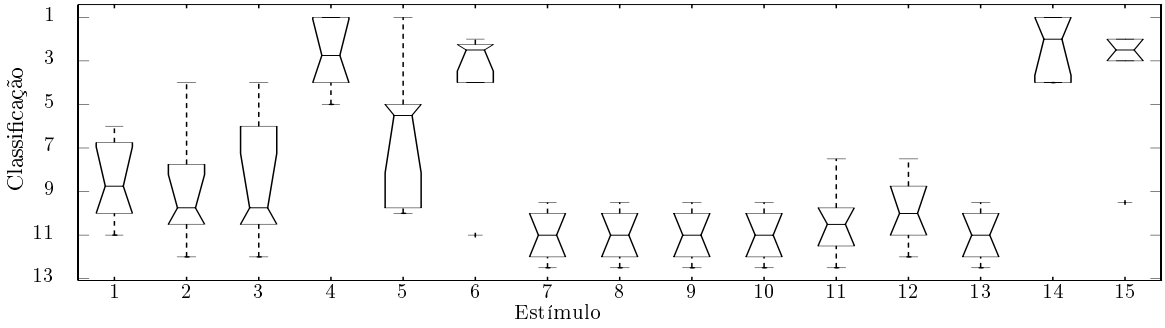
Tabela 6.13: Correlação entre as classificações atribuídas pelos vários ouvintes no teste de preferência.

este valor é muito superior ao valor médio de r. O valor obtido é geralmente considerado como bom. Por exemplo (Chen, 1996, pág. 70), obteve um valor de 0.93 com 8 ouvintes.

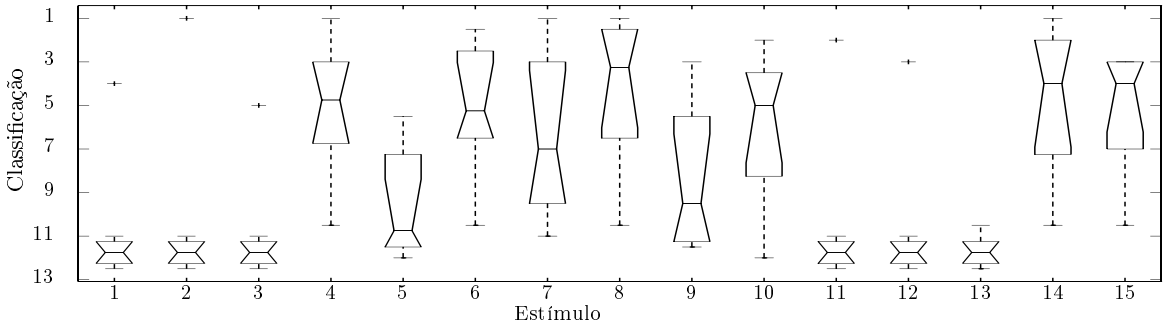
Os resultados obtidos permitem considerar os resultados dos testes efectuados como válidos.

6.2.5.4 Resultados

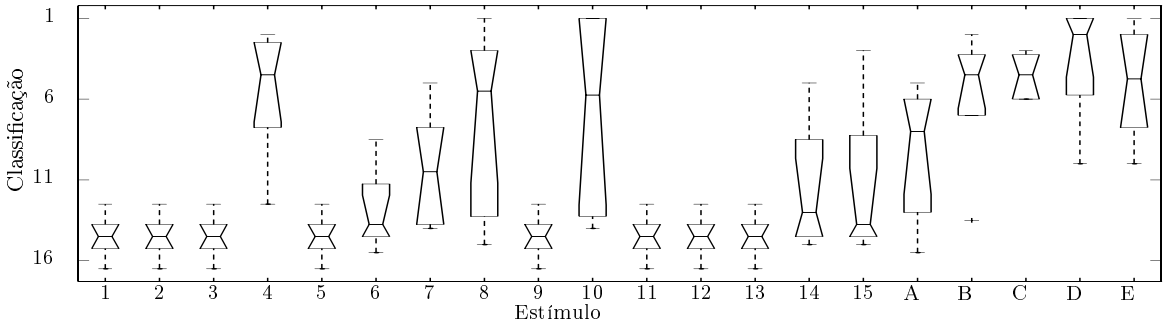
Os resultados encontram-se representados graficamente na Figura 6.11. No eixo horizontal encontra-se o estímulo a que se referem os resultados. O eixo vertical representa a classificação obtida indicando 1 a primeira posição, ou seja, o melhor.



(a) Vogal [ĩ]



(b) Vogal [ũ]



(c) Vogal [ɐ̃]

Figura 6.11: Resultado do teste de preferência para vogais nasais depois de consoante nasal. No eixo vertical apresenta-se a classificação obtida, representando 1 o melhor classificado. O número, ou letra, no eixo horizontal identifica o estímulo. Os estímulos identificados por letras referem-se ao contexto NṼN, apenas estudado para o [ɐ̃].

Efeito do contexto

As figuras mostram claramente que existem estímulos de maior qualidade do que outros. Para as três vogais, os estímulos correspondentes aos contextos NṼV (números 1, 2 e 3) e com velo constante (11, 12 e 13) ficam posicionados nos últimos lugares. Dois dos estímulos referentes

ao contexto NṼOcl (4 e 6) atingem posições cimeiras. Outro contexto que atinge as posições cimeiras é o NṼFric (14 e 15). As classificações dos estímulos do contexto NṼ# (7 a 10) varia bastante entre as vogais. Para a vogal [ĩ] ficam pelos últimos lugares, enquanto para as outras duas vogais, alguns deles, atingem posições de liderança. Conclui-se que a percepção de nasalidade depende do contexto.

Para cada vogal, foi realizada análise de variância para verificar se o efeito do contexto é significativo [8]. Os resultados confirmam o efeito muito notório do contexto. Obteve-se $F(4,28) = 60.64,\ p < 0.001$ para o [ĩ], $F(4,28) = 8.12,\ p < 0.001$ para a vogal [ũ], e $F(5,35) = 16.55,\ p < 0.001$ para [ɐ̃].

Foram efectuados testes estatísticos *post hoc* (Schweigert, 1994, pág. 218 e 261) (Bryman e Cramer, 1993, pág. 181), usando o teste de Newman-Keuls (Montgomery, 1991, pág. 76), para saber quais os contextos que resultam em qualidade significativamente diferente. Utilizou-se um nível de significância de 0.05.

Para a vogal [ũ] os contextos NṼV e velo constante não diferem significativamente. Os outros contextos também não diferem significativamente entre si. Tem-se assim os contextos divididos em duas classes. A primeira, constituída pelos casos com velo constante e vogal nasal antes de vogal oral, obteve classificações inferiores aos outros contextos.

Para o [ĩ] os resultados são algo diferentes. A grande diferença é a última posição em termos de preferência do contexto NṼ#. Este resultado pode dever-se a uma deficiência na geração deste tipo de estímulos [9]. Os restantes estímulos têm um comportamento semelhante ao da vogal anterior. Os primeiros lugares são ocupados pelos contextos NṼFric e NṼOcl, não sendo significativa a diferença entre estes. O contexto NṼV classifica-se a seguir, diferindo significativamente dos últimos classificados. Novamente o contexto em que o velo se mantém constante ocupa as últimas posições, sendo significativa a sua diferença para todos os contextos com excepção do NṼ#.

Para o [ɐ̃], o contexto NṼN obtém significativamente melhores posições do que os restantes. Os outros contextos têm um comportamento semelhante ao verificado para a vogal [ũ], o grupo constituído pelos contextos NṼOcl, NṼFric e NṼ# não difere entre si, mas é significativamente melhor que os outros dois contextos (NṼV e velo constante).

Claramente, para as três vogais, os contextos NṼV e velo constante obtêm as piores classificações. Não incluir a variação do velo no tempo ou ter um segmento oral logo a seguir são as causas.

[8] Não se realizou uma análise conjunta de todos os dados, devido a uma das vogais ter um número diferente de estímulos. Também não faz muito sentido estudar a influência da vogal com duas ou três vogais apenas.

[9] De facto, a configuração bastante elevada e recuada da língua causa alguns problemas no modelo articulatório ao movimento de abertura do velo. A baixa qualidade tinha também sido notada pelo autor antes da execução dos testes perceptuais.

Análise de cada contexto

Tem, também, interesse analisar o que acontece para cada um dos contextos. Foram efectuadas análises de variância para cada contexto, utilizando-se os resultados das três vogais. Apenas foi considerado um factor, o tipo de estimulo, nas análises. Para os contextos em que existe uma diferença significativa foram usados testes *post hoc* de Newman-Keuls (nível de significância de 0.05).

Para os contextos N$\tilde{V}$V, N$\tilde{V}$Fric e velo constante a diferença entre os vários tipos de estímulos não é significativa ($p > 0.05$).

A diferença entre os três tipos de estímulos utilizados para o contexto N$\tilde{V}$Ocl é clara nos gráficos. Os resultados de ANOVA confirmam como significativa essa diferença, com $F(2, 14) = 15.36,\ p < 0.001$. Todos os três diferem significativamente entre si. Os estímulos do tipo 4 são claramente os melhores e os de tipo 5 os piores. Os ouvintes preferem os casos em que existe uma consoante nasal no final. A consoante nasal de duração 40 *ms* é também preferida a uma com 10 *ms*i de duração.

É também significativa $[F(3, 21) = 7.30,\ p < 0.01]$ a diferença entre os quatro tipos de estímulos, no contexto N$\tilde{V}$#. Os testes *post hoc*, revelam que os estímulos 7, 8 e 10, semelhantes entre si, são significativamente melhores que o 9. Os ouvintes preferem os estímulos em que existe um movimento de fecho da passagem oral na fase final da vogal.

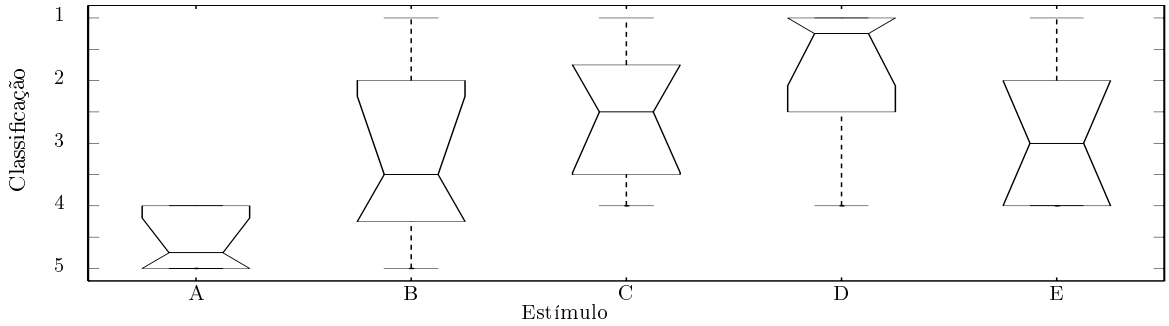


Figura 6.12: Resultado do teste de preferência para vogal oral e nasal entre consoantes nasais (N[ɐ̃]N e N[a]N). Os resultados apresentados foram obtidos considerando que apenas estes estímulos tinham participado no teste. No eixo vertical, apresenta-se a classificação obtida, representando 1 o melhor classificado. O número no eixo horizontal identifica o estímulo.

Analisemos agora o contexto N$\tilde{V}$N, cujos resultados se apresentam na Figura 6.12. Destaca-se a pior classificação do estímulo A e a qualidade um pouco superior do D. Os resultados de análise de variância revelam como significativo o efeito de variar a duração da consonante nasal que se segue à vogal nasal $[F(4, 28) = 5.02,\ p < 0.01]$. Apenas é significativa a diferença entre o estímulo A e os restantes. Este resultado é muito interessante, pois não se obteve significativamente qualidade nasal inferior, mesmo com a manobra de fecho do velo no caso do estímulo E. Em contraste, se não existir a consoante nasal, estímulo A, a qualidade é

inferior. Estes resultados estão de acordo com a dificuldade de utilizar a distinção entre vogal oral e vogal nasal neste contexto.

6.2.6 Discussão dos resultados dos testes de identificação e preferência

Conjugando os resultados dos dois testes pode concluir-se que:

- O efeito do contexto é notório: os contextos N$\tilde{V}$Ocl, N$\tilde{V}$Fric e N$\tilde{V}$# resultam em vogais nasais de qualidade e os contextos N$\tilde{V}$V e velo constante em vogais de baixa qualidade. Os contextos com baixa qualidade resultam da não inclusão da variação do velo ao longo da realização da vogal nasal, e da presença de um segmento de elevada energia logo a seguir à vogal;
- Para o contexto N$\tilde{V}$Ocl a qualidade depende da duração da consoante nasal final, preferindo os ouvintes o estímulo em que a duração é de 40 ms;
- No contexto N$\tilde{V}$# o movimento de fecho, mesmo que não exista, de facto, oclusão oral, resulta em melhores estímulos;
- No contexto N$\tilde{V}$N a consoante nasal seguinte é utilizada na percepção da vogal nasal. É também difícil produzir uma vogal oral neste contexto.

6.3 Vogais nasais isoladas

Foi realizado um teste de identificação de vogais para o caso isolado. A realização de apenas um tipo de teste deveu-se: à menor relevância das vogais nasais isoladas, quando se está interessado em produzir discurso contínuo; à escassez de dados disponíveis acerca do comportamento do velo neste caso; a limitações temporais.

6.3.1 Teste de identificação

6.3.1.1 Procedimento

Este teste foi realizado conjuntamente com o teste de vogais nasais depois de consoante nasal, descrito na secção 6.2.4.2, página 162.

6.3.1.2 Estímulos

Para cada uma das cinco vogais foram gerados três tipos de estímulos: o primeiro com o velo fechado, o segundo com o velo aberto e constante durante toda a realização da vogal, e o terceiro com abertura do velo variável no tempo. A abertura do velo para o segundo tipo de estímulo foi ajustada manualmente por forma a se ter uma relação entre a área de acoplamento

nasal e área oral na zona do velo de aproximadamente 10. A variação ao longo do tempo, para o terceiro tipo, seguiu uma variação aproximando os gráficos publicados em (Benguerel *et al.*, 1977, pág. 152), em que o velo começa fechado e vai abrindo progressivamente ao longo da vogal. Segundo o trabalho referido, o velo atinge a sua posição mais elevada, fechando a passagem nasal, 100 a 200 ms antes do início da vogal nasal. O velo posiciona-se numa espécie de posição preparatória para a vogal nasal. O movimento de abertura é feito numa fase inicial de forma rápida sendo muito mais suave na fase final. Aproximámos esta variação por uma variação contínua de inclinação constante entre velo fechado e o valor de abertura utilizado nos estímulos do segundo tipo, valor atingido no final da vogal apenas.

A duração foi mantida igual a 250 ms para todos os estímulos.

6.3.1.3 Ouvintes

Os ouvintes foram os mesmos utilizados no teste de identificação para o caso de vogais nasais depois de consoante nasal, descritos na secção 6.2.4.3 na página 163.

O número médio de identificadores utilizados, indicador da consistência de cada ouvinte, variou entre 1.0 e 1.87. Esta tendência manteve-se também para os três tipos de estímulos, com médias de 1.30, 1.63, e 1.43 respectivamente.

A concordância entre os ouvintes foi calculada da mesma forma que no caso de vogais depois de consoante nasal. Os valores médios para cada ouvinte encontram-se entre 47.6 % e 68.6 %, com valor médio global de 60.7 %. Se se considerar apenas se identificam ou não os valores situam-se entre 54.3 % e 79.1 %, sendo o novo valor médio global de 70.5 %.

6.3.1.4 Resultados

Para as cinco vogais, os estímulos produzidos com velo fechado foram quase na totalidade identificados como vogal oral. Em relação aos estímulos em que existe abertura do velo, o seu comportamento foi bastante dependente da vogal. Para o [ɐ̃], apenas com o velo variável se conseguiu obter a identificação como nasal, e para um número muito reduzido de casos. Para o [õ], quase todos foram considerados [ɔ]. Para o [ẽ], o estímulo com velo variável obteve maioritariamente identificação como a vogal nasal pretendida; o estímulo com velo fixo foi identificado como nasal, mas sendo outra vogal não pertencente às cinco vogais nasais do Português. Para o [ũ], com velo fixo ou variável obtiveram-se resultados semelhantes: identificação maioritariamente como a vogal nasal pretendida. Para o [ĩ], o caso com velo variável é o que obtém mais identificações como [ĩ]. Para esta vogal, parte substancial é identificada como som nasal não pertencendo às cinco vogais nasais do Português.

Os resultados são bastante confusos, mas pode ver-se que em nenhum caso a utilização de velo variável levou a uma diminuição do número de estímulos identificados como a vogal nasal pretendida, havendo dois casos, o [ẽ] e [ĩ], em que a melhoria é nítida.

Estímulo		Vogal oral										Vogal nasal					
ident	V	a	ɐ	ə	ɛ	e	ɔ	o	u	i	?	ɐ̃	ẽ	õ	ũ	ĩ	?
1	a	**22**										2					
2	a	**24**															
3	a	**21**										3					
1	e		1		**21**								2				
2	e					2			1			1	6		4		**10**
3	e				3	3			1			4	**12**		1		
1	o						**21**	1						2			
2	o						**20**	1				2		1			
3	o						**20**	1					1	2			
1	u								**24**								
2	u								11					1	**12**		1
3	u							1	11						**12**		
1	i			5	1	4				**12**	1		1				
2	i			5					2	1			4		1	3	**8**
3	i			3		1				3			2		2	6	**7**

Tabela 6.14: Resultados do teste de identificação para vogais nasais isoladas.

Somando as vezes que os estímulos foram "correctamente" identificados como a vogal pretendida, obtêm-se percentagens de 5.0, 18.3 e 29.1 % para os estímulos com velo fechado, velo aberto fixo e velo aberto com variação no tempo, respectivamente.

Os resultados de análise de variância revelam que, quer o tipo dos estímulos $[F(2,14) = 6.60,\ p < 0.01]$, quer o efeito da vogal $[F(4,28) = 4.56,\ p < 0.01]$ são significativos. Os testes *post hoc* revelam que apenas é significativa a diferença entre os estímulos do tipo 1 (velo fechado) e tipo 3 (velo aberto e variável). Apenas abrindo o velo e fazendo-o variar no tempo se obtiveram maior número de identificações das vogais como a vogal nasal pretendida.

6.4 Resumo e comentários finais

Neste capítulo foi explorada a influência da variação no tempo do velo e outros articuladores, na qualidade de vogais nasais. O sintetizador articulatório foi utilizado para a realização de simulações e para obtenção dos estímulos utilizados nos testes perceptuais. Estudaram-se vogais nasais em vários contextos: entre consoantes oclusivas orais, depois de consoante nasal, e isoladas.

Dos resultados das experiências apresentadas neste capítulo, ressalta que a variação do velo, e mesmo de outros articuladores, influencia a percepção de nasalidade. Incluindo a forma como o velo e os outros articuladores variam nos vários contextos em que as vogais nasais do Português aparecem, a qualidade aumenta.

Analisando o que é comum a todas as situações em que se obtêm vogais nasais de qualidade, conclui-se que, em todas as situações a vogal nasal tem um início com energia elevada e radiação oral dominante (ou mesmo sendo a única radiação) e tende, de uma forma gradual, para uma configuração de baixa energia, com radiação nasal dominante. A forma como se obtém esta dominância nasal é variável, tanto podendo haver uma consoante nasal no final,

devida a necessidade de coarticulação com uma consoante seguinte, como ser o resultado de uma entrada para a cavidade bucal bastante reduzida pelo abaixamento do velo.

Com base nos resultados deste capítulo, consideramos que as vogais nasais do Português Europeu devem ser vistas como ditongos [10], já propostos para análise fonológica das vogais nasais do Português por Parkinson (1983), sendo também já referidos anteriormente por Louro (1954-1955) nos seus estudos. O ditongo inicia-se com radiação pelos lábios dominante e acaba com radiação nasal dominante. A obtenção de dominância da radiação nasal no final pode ser conseguida por passagem oral reduzida ou obstruída, como continuação nasal da primeira parte do ditongo ou como resultado de coarticulação com o segmento seguinte. A transição entre as duas configurações, inicial e final, tem de ser gradual.

[10]Entendendo-se aqui ditongo como composto por dois sons, o que está de acordo com o significado da origem do termo, a palavra grega *diphthoggos*, que significa "que tem dois sons".

Capítulo 7

Avaliação da qualidade

> People in both areas [speech production and applied research in speech synthesis] need to collaborate more to solve the **naturalness** problem in speech synthesis that I consider the next important milestone to reach.
>
> J. Schroeter em (van Santen *et al.*, 1996, pág. 183)

Tendo este trabalho como motivação inicial principal a obtenção de síntese de elevada qualidade, em especial para a língua portuguesa, não poderíamos deixar de avaliar a qualidade que o sistema desenvolvido, conjuntamente com os conhecimentos obtidos, permite obter. Pretendeu-se, não só avaliar a qualidade do sistema desenvolvido, mas também ter uma avaliação relativa a outras técnicas de síntese desenvolvidas para o Português Europeu. A avaliação incide quase exclusivamente nos sons nasais.

Na próxima secção descreve-se o teste realizado para avaliação de qualidade, encerrando o capítulo com um resumo das principais conclusões.

7.1 Teste perceptual para avaliação de qualidade

Para obter uma medida da qualidade com significado esta tem de basear-se na resposta de ouvintes (Quackenbush *et al.*, 1988, pág. 15). Para a avaliação de sinais com inteligibilidade elevada, que consideramos acontecer para os sinais produzidos pelo nosso sistema [1], existem variados testes, sendo o mais utilizado o método de classificação dos estímulos em cinco categorias, entre mau e excelente, que corresponde a uma pontuação de um a cinco e do qual resulta uma opinião média. Este teste é habitualmente designado por teste *Mean Opinion Score* (MOS) (Quackenbush *et al.*, 1988; Rothauser, 1969).

7.1.1 Procedimento

O teste MOS foi realizado recorrendo a um programa de computador, por nós desenvolvido. Além da audição do estímulo, foi fornecida ao ouvinte informação acerca do som, sequência de sons ou palavra que era suposto ser percebido pelo ouvinte. A interface com o utilizador do programa permitia aos ouvintes atribuir a classificação utilizando o rato. Permitia também a audição do estímulo as vezes que o ouvinte desejasse.

Na classificação dos estímulos os ouvintes utilizaram a seguinte escala :

1. Não percebo ou percebo outra coisa;
2. Pobre (mas é o que se pretende);
3. Razoável;
4. Bom;
5. Muito bom (natural).

Como proposto por Rothauser (1969), o teste teve duas fases: a primeira de treino, a segunda de avaliação propriamente dita. Apenas se retiveram para análise os resultados da segunda fase. A fase de treino, em que se utilizaram dez estímulos, diferentes dos avaliados, serviu para habituação dos ouvintes ao teste, à escala e também para estes tomarem contacto com os limites superior e inferior da escala. Deu-se particular atenção ao limite superior da escala, incluindo-se, por essa razão, vários exemplos naturais de vogais nasais, ou palavras contendo vogais nasais, para habituar o ouvinte a estímulos a que deverá corresponder a classificação de Muito Bom, pontuada com cinco. Na fase de avaliação, foram também incluídos exemplos de estímulos naturais para "refrescar" a informação acerca do nível cinco da escala (aconselhado no ponto 5.2.2, pág. 232, de Rothauser, 1969).

[1] Realizamos diversos testes informais de identificação de algumas palavras, tendo os ouvintes identificado geralmente a palavra correctamente.

Para permitir obter informação acerca da consistência das respostas de cada ouvinte e para se obter um número maior de avaliações, cada um dos estímulos foi apresentado quatro vezes para avaliação. A ordem de apresentação foi aleatória e diferente para cada um dos ouvintes.

O teste foi realizado individualmente, em salas com ruído baixo a moderado, sendo os estímulos apresentados aos ouvintes através de auscultadores.

7.1.2 Estímulos

				Estim.
vogais isoladas	velo fixo		[ɐ̃] [ĩ] [ũ]	6
	velo variável		[ɐ̃] [ĩ] [ũ]	
entre oclusivas (CṼC)	velo fixo		[ɐ̃] [ĩ] [ũ] [ẽ] [õ]	10
	velo variável		[ɐ̃] [ĩ] [ũ] [ẽ] [õ]	
depois de consoante nasal (NṼ)	antes de oclusiva		[ɐ̃] [ĩ] [ũ]	15
	no final	0 *ms*	[ɐ̃] [ĩ] [ũ]	
	no final	40 *ms*	[ɐ̃] [ĩ] [ũ]	
	velo fixo	igual N e Ṽ	[ɐ̃] [ĩ] [ũ]	
	velo fixo	diferente N e Ṽ	[ɐ̃] [ĩ] [ũ]	
palavras	mão	F_0 sintético	sem [m] no final	2
		F_0 sintético	com [m] no final	
	mãe	F_0 natural	[a], [ɲ] no final	6
		F_0 sintético	[a], [ɲ] no final	
		idem	[ɐ], [ɲ] no final	
		idem	[a], sem [ɲ] no final	
		F_0 sintético, sem interacção	[ɐ], [ɲ] no final	
		F_0 sintético, sem interacção	[a], [ɲ] no final	
	António	F_0 natural		2
		F_0 sintético		

(a) Estímulos gerados usando síntese articulatória (SAP).

Palavra	tipo de síntese	técnica
mãe	concatenação	PSOLA
mãe	formantes	
mãe	concatenação	MBROLA
canto	concatenação	MBROLA
manto	concatenação	MBROLA
mim	concatenação	MBROLA
mão	concatenação	MBROLA
nem	concatenação	MBROLA
não	concatenação	MBROLA

(b) Estímulos gerados pelo sistema DIXI.

Descrição	Qualidade gravação	Sexo locutor	obs.
[ũ]	elevada	feminino	extraída de *rum*
[mɐ̃]	média	masculino	
[mũ]	média	masculino	
[ɐ̃]	média	masculino	
[ĩ]	média	masculino	
[ũ]	média	masculino	
António	média	masculino	

(c) Estímulos naturais.

Tabela 7.1: Estímulos utilizados no teste perceptual de avaliação da qualidade.

No teste incluíram-se vários tipos de estímulos. Pretendeu-se avaliar a qualidade dos estímulos utilizados no estudo da influência da dinâmica, a qualidade de algumas palavras e tentar comparar a qualidade obtida, usando síntese articulatória com outras técnicas de síntese. O tipo de teste utilizado, necessitando de estímulos representativos do máximo da escala, motivou a inclusão de gravações de vários sons naturais.

Estímulos já utilizados nos estudos do efeito da dinâmica

Para avaliar o efeito na qualidade da inclusão da variação dos articuladores no tempo, foram incluídos alguns estímulos utilizados nos testes descritos no Capítulo 6.

Para o caso de vogais nasais entre oclusivas (secção 6.1.1.2, página 150) incluíram-se exemplos, para as cinco vogais, dos estímulos pelos quais os ouvintes revelaram preferência: os produzidos com velo variável e consoante nasal no final, com 40 ms de duração. Foram também incluídos, para termo de comparação, os estímulos com velo aberto, mas mantendo o valor durante toda a vogal nasal e sem consoante nasal.

Para vogais nasais depois de consoante nasal (secção 6.2.3.1, página 158), retiveram-se alguns exemplos dos estímulos preferidos e alguns exemplos com o velo fixo. Como exemplos de estímulos com articuladores variáveis incluíram-se: exemplos de vogais nasais antes de oclusiva, com consoante nasal produzida por coarticulação, com duração de 40 ms; exemplos de vogal nasal em final de palavra, com oclusão oral na parte final, ocorrendo em simultâneo com o final da vogal nasal ou 40 ms antes. Apenas se utilizaram estímulos de três vogais.

Foram, também, incluídos exemplos de vogais nasais isoladas. Além do caso com velo aberto e permanecendo na mesma posição durante toda a vogal, foram incluídos exemplos, para três vogais, com o velo variável no tempo (mais detalhes na secção 6.3.1.2, na página 176).

Palavras contendo sons nasais

Foram sintetizadas, usando o sintetizador articulatório, diversas versões de palavras compostas apenas por sons nasais, *mãe* e *mão*, e também da palavra *António*. O processo de obtenção das palavras foi o seguinte:

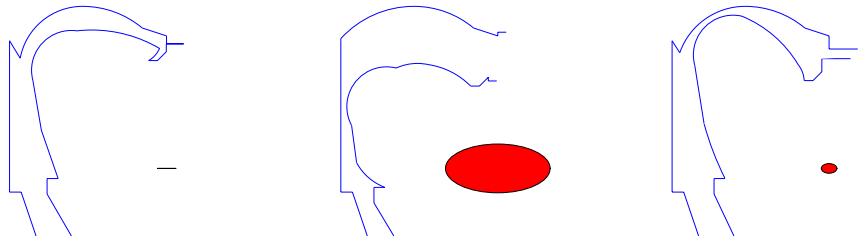
1. Obtenção da duração de cada segmento fonético que entra na composição da pronunciação da palavra, com base na análise de um exemplo de voz natural;

2. Obtenção de configurações para os articuladores orais para cada um dos segmentos fonéticos. Para o caso das vogais, são utilizados os valores obtidos aplicando inversão a valores médios de formantes. Não foi efectuada a inversão usando o sinal de voz natural para as palavras; limitamo-nos a usar configurações que tinham sido obtidas anteriormente. Para as consoantes, de que não temos um processo de inversão, foram obtidas, manualmente, configurações baseadas na descrição da Fonética Articulatória e imagens de raios X e MRI, publicadas na literatura;

3. Definição da trajectória do velo, usando valores adequados para cada vogal e para cada consoante;

4. Definição dos parâmetros da fonte glotal, em especial da frequência fundamental.

A título exemplificativo, apenas se apresenta a forma como foi obtida a palavra *mão*. Primeiro, da análise de um exemplo natural da palavra, obtiveram-se as durações de 100 ms para o [m] e 465 ms para o ditongo nasal. Obteve-se assim uma duração total de 565 ms. De seguida, obtiveram-se as configurações. Neste caso, o [m] foi obtido manualmente e usaram-se as configurações de [a] e [u]. As três configurações encontram-se na Figura 7.1. Em terceiro



(a) [m] (b) [a] (c) [u]

Figura 7.1: Configurações do tracto usadas na obtenção da palavra *mão*.

lugar, foi definida a variação do velo, representada na Figura 7.2. Este começa fechado, abre para a consoante nasal, abre ainda mais na parte inicial da vogal [ɐ̃], subindo de seguida para a vogal [ũ].

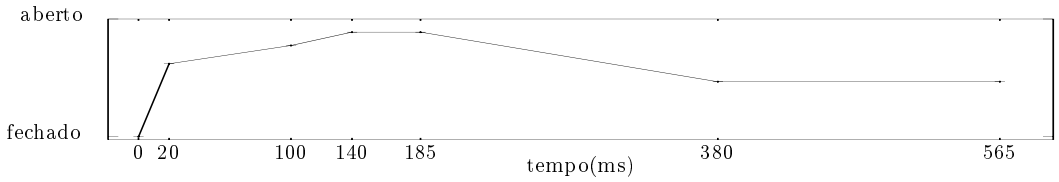
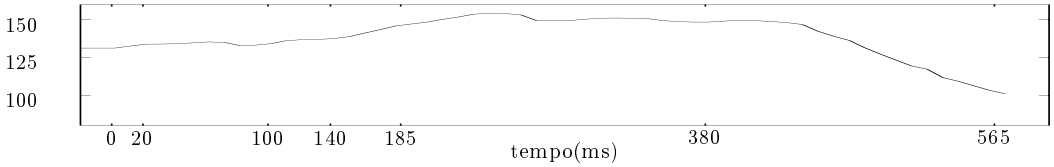
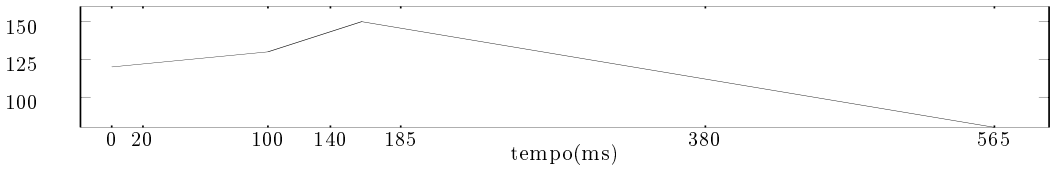


Figura 7.2: Variação do velo para a palavra *mão*.

O quarto passo consistiu na definição da frequência fundamental e outros parâmetros da fonte. A frequência fundamental apresentada na Figura 7.3(b), começa em 120 Hz, sobe 10 Hz até ao final da consoante nasal, sobe depois até aos 150 Hz, para acentuar a vogal [ɐ̃], descendo, de seguida, até 80 Hz no final da palavra. Esta trajectória baseou-se, em parte, na trajectória obtida do sinal natural, sendo, no entanto, muito mais simples. Na figura 7.3(a) apresenta-se a frequência fundamental obtida por análise de um exemplo natural da palavra *mão*. Para a excitação glotal foram utilizados: quociente de abertura (OQ) igual a 60 %, quociente de velocidade (SQ) igual a 2, *jitter*, *shimmer* e interacção entre a fonte glotal e o tracto. Foram



(a) Valores obtidos por análise de sinal de voz natural.



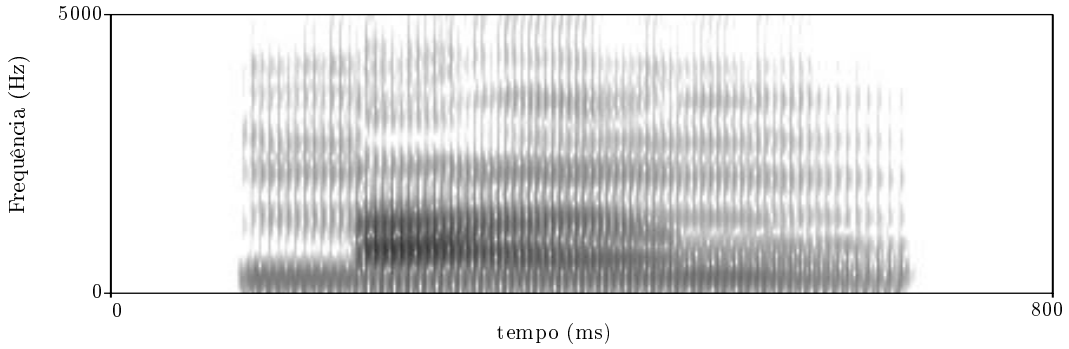
(b) Frequência fundamental não natural.

Figura 7.3: Variação da frequência fundamental para a palavra *mão*. Apresentam-se os dois casos: frequência fundamental obtida de sinal de voz natural e frequência fundamental sintética.

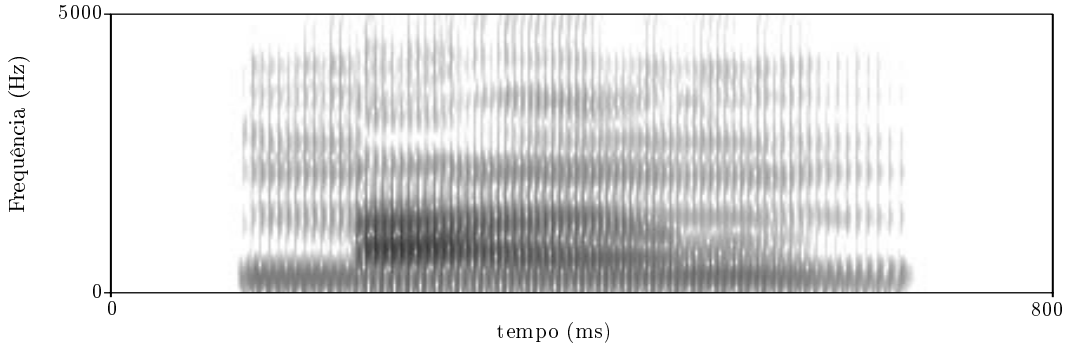
produzidas duas versões diferindo pela existência, ou não, de uma consoante bilabial ([m]) no final. Os espectrogramas das duas versões da palavra *mão* encontram-se na Figura 7.4.

Uma transcrição fonética possível para a palavra *mãe* é ['mɐ̃ĩɲ], incluindo-se no final uma consoante nasal palatal (Laver, 1994, pág. 292). Não existe informação precisa acerca da configuração das cavidades orais durante a produção da sequência [ɐ̃ĩ]. Devido às incertezas existentes, decidimos pela inclusão no teste de versões da palavra com e sem a consoante nasal final e com duas configurações do tracto para a vogal [ɐ̃], sendo uma adequada à realização da vogal oral [a], outra, mais elevada, característica da vogal [ɐ]. Foram realizadas várias versões da palavra *mãe*, variando-se os seguintes factores: configuração do tracto oral para a vogal [ɐ̃], para a qual se usaram as configurações da vogal [a] e da vogal mais elevada [ɐ]; a formação ou não na fase final de uma oclusão na zona palatal, resultando na consoante nasal palatal [ɲ]; a forma de obtenção da frequência fundamental, obtida de análise de sinal de voz natural ou criada artificialmente; utilização ou não de interacção entre a fonte e o tracto (apenas para o caso de frequência fundamental não natural). Na Figura 7.5 apresenta-se o espectrograma para o estímulo produzido com frequência fundamental não natural, interacção entre o tracto e a fonte glotal, a configuração [a] para a vogal [ɐ̃] e oclusão palatal no final.

Para a palavra *António* sintetizaram-se dois exemplos, um com F_0 natural, outro com F_0 sintético. A oclusiva [t] foi sintetizada fechando e abrindo a passagem oral sem se terem modelado diversos fenómenos importantes para a percepção de uma oclusiva, como o ruído existente na fase inicial de abertura da oclusão.



(a) Sem [m] no final.



(b) Com [m] no final.

Figura 7.4: Espectrogramas das duas versões da palavra *mão* produzidas pelo sintetizador articulatório.

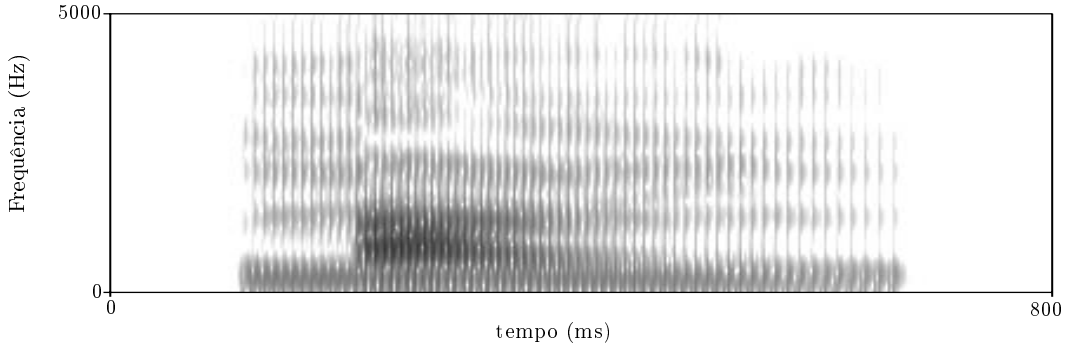


Figura 7.5: Espectrograma de uma das versões da palavra *mãe* produzidas pelo sintetizador articulatório.

Estímulos produzidos por outras técnicas de síntese

Além dos sons gerados por síntese articulatória foram incluídos estímulos criados pelo sistema DIXI de síntese de fala a partir de texto [2] (Oliveira *et al.*, 1991; Oliveira, 1996), apresentados na Tabela 7.1(b), para permitir uma análise comparativa da qualidade que é actualmente possível obter com diferentes técnicas de síntese existentes para o Português Europeu. Todas as versões usaram o mesmo módulo de análise de texto variando o módulo de geração da forma de onda: síntese de formantes, e concatenação (Carvalho *et al.*, 1998) *Pitch Synchronous Overlap-Add* (PSOLA) e *Multi-Band Re-synthesis Overlap-Add* (MBROLA) (Dutoit, 1993, 1997). Para a palavra *mãe* foram incluídos exemplos das três técnicas de síntese do sistema DIXI. Foram utilizados sete estímulos, usando síntese MBROLA, correspondentes às palavras *canto*, *manto*, *mim*, *mão*, *nem*, *não* e *mãe*.

Sons naturais

Para permitir aos ouvintes "refrescar" informação acerca do limite máximo da escala foram incluídos no teste estímulos naturais. Quatro dos estímulos usados consistiram em vogais nasais isoladas, dois estímulos em consoante nasal seguida de vogal nasal, sendo o último estímulo a palavra *António*. Informação adicional acerca da forma de obtenção destes estímulos encontra-se na Tabela 7.1(c). Os sons indicados na tabela, como tendo sido obtidos com elevada qualidade de gravação, foram gravados numa sala insonorizada, utilizando um microfone de condensador de elevada qualidade, no MMIRC, da University of Florida. Os outros casos foram gravados numa sala comum, com nível de ruído baixo, usando equipamento corrente (computador pessoal com placa de som Sound Blaster 128 e microfone de baixo custo da Creative Labs).

7.1.3 Ouvintes

Participaram nos testes 8 ouvintes, sendo 7 do sexo masculino, todos falantes nativos do Português Europeu, naturais e residentes na maioria na zona norte litoral do país. As idades variavam entre os 20 e 32 anos. Em termos de escolaridade, encontravam-se representados diversos níveis, desde o ensino básico até a estudantes de Doutoramento.

Para aferir a consistência das respostas de cada ouvinte foi calculada a correlação, usando o r de Pearson, entre cada uma das quatro repetições da totalidade dos estímulos. O valor médio da correlação para cada um dos ouvintes, o valor médio da correlação para todos os ouvintes, assim como o valor médio efectivo, R, calculado usando a fórmula de Spearman-Brown (Rosenthal e Rosnow, 1991, pág. 51), encontram-se na Tabela 7.2(a). Os valores médios de $r = 0.76$ e $R = 0.96$ mostram que os ouvintes foram bastante consistentes nas suas classificações.

[2] Agradece-se à equipa que desenvolveu o sistema DIXI, na pessoa do L. C. Oliveira, a colaboração dada, ao disponibilizar exemplos de palavras contendo sons nasais, geradas por várias versões do sistema.

Ouvinte	1	2	3	4	5	6	7	8	Média	R
r Pearson	0.82	0.68	0.88	0.87	0.88	0.51	0.73	0.72	0.76	0.96

(a) Informação acerca da consistência de cada ouvinte. Foi calculado o valor médio da correlação entre cada uma das quatro repetições dos estímulos.

Ouvinte	1	2	3	4	5	6	7	8	Média
1		0.63	0.70	0.50	0.59	0.56	0.56	0.49	
2	0.63		0.68	0.37	0.54	0.54	0.46	0.56	
3	0.70	0.68		0.68	0.82	0.73	0.83	0.79	
4	0.50	0.37	0.68		0.65	0.49	0.64	0.49	
5	0.59	0.54	0.82	0.65		0.62	0.89	0.80	
6	0.56	0.54	0.73	0.49	0.62		0.59	0.69	
7	0.56	0.46	0.83	0.64	0.89	0.59		0.81	
8	0.49	0.56	0.79	0.49	0.80	0.69	0.81		
Média	0.50	0.47	0.65	0.48	0.61	0.53	0.60	0.58	0.55
								R=	0.91

(b) Correlação entre as classificações atribuídas pelos vários ouvintes.

	[ũ]	[mɐ̃]	[mũ]	[ɐ̃]	[ĩ]	[ũ]	António	todos
Média	4.06	4.41	4.53	4.72	4.62	4.66	4.97	4.57
Desvio padrão	1.32	0.61	1.05	0.52	1.01	0.83	0.18	0.89

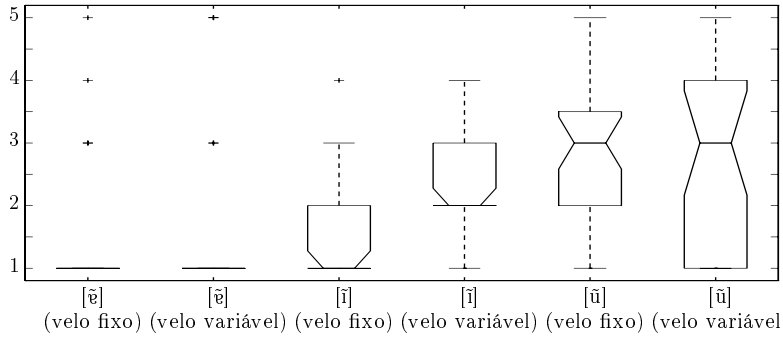
(c) Classificações médias atribuídas aos sons naturais.

Tabela 7.2: Dados acerca do desempenho dos ouvintes no teste MOS.

A correlação entre as classificações dos vários ouvintes foi também calculada, usando também o r de Pearson, para avaliar se os ouvintes concordaram uns com os outros nas classificações atribuídas. Os valores obtidos encontram-se na Tabela 7.2(b). O valor médio, considerando todos os ouvintes, é de 0.55 o que resulta num valor para a confiança conjunta dos n ouvintes, R, igual a 0.91. Os ouvintes concordam, razoavelmente, uns com os outros. Também em nenhum caso se obteve uma correlação negativa.

As classificações atribuídas pelos ouvintes aos estímulos naturais pode também fornecer informação acerca do desempenho dos ouvintes no teste efectuado. Os valores médios e desvio padrão, para os sete estímulos naturais, são apresentados na Tabela 7.2(c). Apresentam-se também os valores da média e desvio padrão, considerando todos os estímulos. Os valores médios, com excepção do primeiro estímulo, aproximam-se bastante do valor máximo da escala, como pretendido. A pior classificação do primeiro estímulo deve-se ao facto de este estímulo ter um volume muito baixo e ter sido extraído de uma palavra.

Do atrás exposto, pode concluir-se pela validade dos resultados obtidos.



(a) Diagrama extremos-e-quartis (*boxplot*)

Vogal	Velo fixo			Velo variável			diferença
	Número	Média	desvio padrão	Número	Média	desvio padrão	significativa ?
[ɐ̃]	1	1.50	1.11	2	1.38	1.07	$p > 0.05$ (ns)
[ĩ]	3	1.62	1.01	4	2.38	0.94	$p < 0.01$
[ũ]	5	2.84	0.95	6	2.72	1.35	$p > 0.05$ (ns)
Todas		1.99	1.18		2.16	1.26	$p > 0.05$ (ns)

(b) Classificações médias.

Figura 7.6: Resultados do tese de qualidade para as vogais nasais isoladas. Estímulos ímpares, com velo fixo; pares, com velo variável.

7.1.4 Resultados

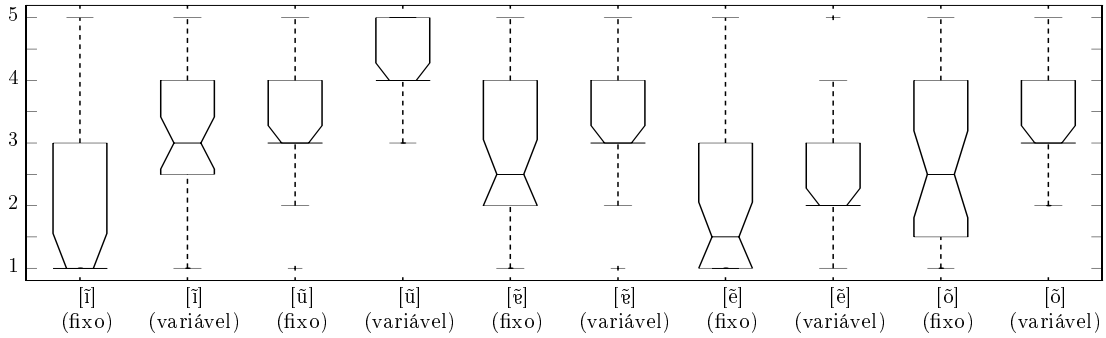
As classificações atribuídas pelos ouvintes aos estímulos foram analisadas, tendo por objectivo obter respostas para as seguintes questões:

1. Qual a melhoria na qualidade ao utilizar-se a variação dos articuladores?
2. Qual a qualidade dos sons nasais produzidos pelo sintetizador?
3. Em termos de qualidade qual é a posição relativamente a outras técnicas de síntese do sintetizador articulatório desenvolvido?

Efeito na qualidade da utilização de parâmetros articulatórios variáveis no tempo

Para o caso de vogais isoladas, cujos resultados se apresentam na Figura 7.6, o efeito de variação do velo depende da vogal. Apenas houve melhoria para a segunda vogal, o [ĩ]. Os resultados são em geral maus, o que pode, em parte, dever-se à pouca naturalidade da situação.

Os resultados para vogal nasal entre oclusivas encontram-se na Figura 7.7. A inclusão de velo variável no tempo, característico de vogal nasal entre oclusivas, bem como a existência na fase final de uma consoante nasal, criada devido à coarticulação com a oclusiva que se segue



(a) Diagrama extremos-e-quartis (*boxplot*)

	Velo fixo			Velo variável			diferença
Vogal	Número	Média	desvio padrão	Número	Média	desvio padrão	significativa ?
[ĩ]	1	1.91	1.23	2	3.28	1.05	Sim $p < 0.0001$
[ũ]	3	3.19	1.15	4	4.16	0.68	Sim $p < 0.0005$
[ɐ̃]	5	2.72	1.33	6	3.25	1.22	Não $p > 0.05$ (ns)
[ẽ]	7	2.06	1.29	8	2.47	1.24	Não $p > 0.05$ (ns)
[õ]	9	2.75	1.39	10	3.31	0.97	Não $p > 0.05$ (ns)
Todas		2.53	1.34		3.30	1.18	Sim $p < 0.0001$

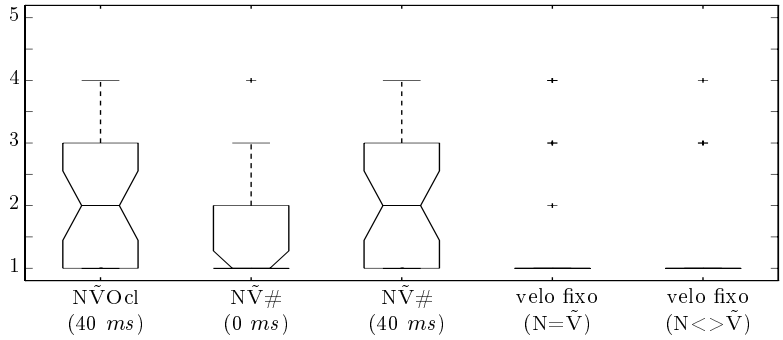
(b) Classificações médias.

Figura 7.7: Resultados do teste de qualidade para as vogais nasais entre oclusivas. Estímulos ímpares com velo fixo, estímulos pares com velo variável e consoante nasal na fase final da vogal.

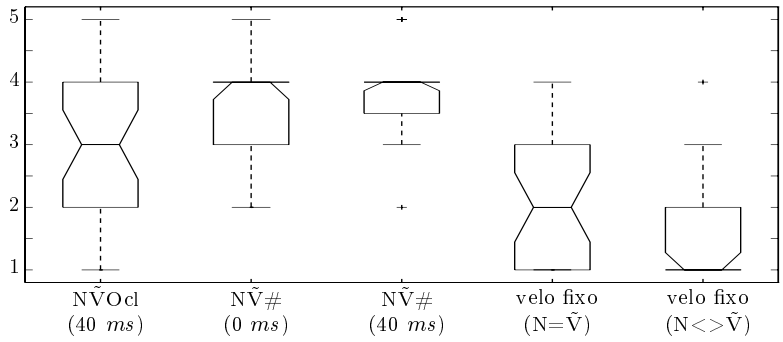
à vogal nasal, permitiu em termos médios, para as cinco vogais nasais do Português, um aumento da classificação de 0.77, obtendo-se 3.30 em vez de 2.53. Existe uma probabilidade de 0.95 de a melhoria de classificação se situar no intervalo entre 0.45 e 1.09. A diferença é estatisticamente bastante significativa ($p < 0.0001$). A inclusão do conhecimento obtido, acerca das várias fases de realização de uma vogal nasal e movimento do velo, permitiu passar de estímulos entre o pobre e razoável, para um pouco acima de razoável. Apesar de haver um aumento de qualidade para todas as vogais os ganhos foram particularmente notórios e estatisticamente significativos, para o [ĩ] e [ũ]. No caso do [ũ], atingiu-se a classificação média de 4.16, isto é, um pouco acima de bom. O intervalo de confiança [3] para a este caso situa-se entre os 3.92 e os 4.41. Apesar de inferior a 4, o limite inferior é suficientemente próximo desse valor para se poder considerar o estímulo como bom.

No caso de vogais nasais depois de consoante nasal, Figura 7.8, é notório que os três primeiros casos, para as três vogais, correspondentes à inclusão da variação dos articuladores, obtêm classificações superiores aos dois últimos, em que os articuladores se mantêm fixos e estando o velo aberto. Considerando os valores médios para o conjunto das três vogais, o efeito da

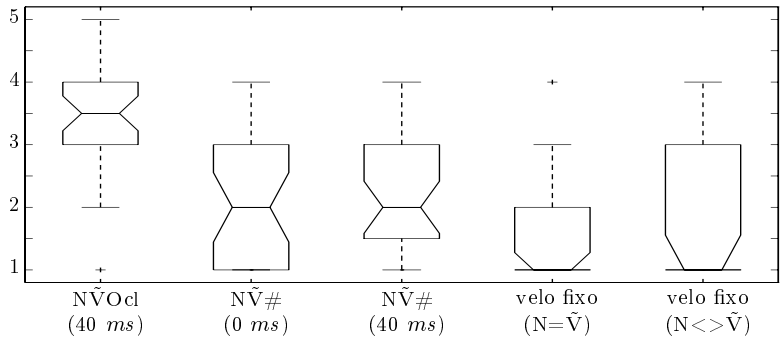
[3] Todos os intervalos de confiança apresentados neste capítulo utilizam uma probabilidade de 0.95.



(a) [ẽ]



(b) [ĩ]



(c) [ũ]

Contexto		Núm.	[ẽ]	[ĩ]	[ũ]	todas
NṼOcl	40 *ms*	1	2.12 (1.04)	2.97 (1.06)	3.50 (0.84)	2.86 (1.13)
NṼ#	0 *ms*	2	1.66 (0.97)	2.22 (1.04)	3.78 (0.75)	2.55 (1.29)
NṼ#	40 *ms*	3	2.16 (1.02)	3.91 (0.73)	2.38 (1.04)	2.81 (1.22)
Velo fixo	N=Ṽ	4	1.34 (0.87)	1.97 (1.03)	1.53 (0.88)	1.61 (0.96)
Velo fixo	N<>Ṽ	5	1.28 (0.77)	1.69 (0.90)	1.72 (1.08)	1.56 (0.93)

(d) Média e desvio padrão (entre parêntesis)

Figura 7.8: Resultados do teste de qualidade para as vogais nasais depois de consoante nasal. Os três primeiros estímulos, com velo variável; os dois últimos, com velo fixo. O número 1 corresponde a vogal nasal antes de oclusiva (NṼOcl); os 2 e 3 a vogal nasal em final de palavra (NṼ#).

variação dos articuladores é significativo ($p < 0.0001$), sendo estatisticamente significativas ($p < 0.001$), usando teste *post hoc* de Newman-Keuls, as diferenças entre os casos com velo fixo e cada um dos restantes. Não é significativa a diferença entre os dois casos com velo fixo, nem entre os três casos com articuladores variáveis. Os estímulos podem ser agrupados em duas classes: uma, com velo fixo com qualidade média próxima de 1.5; outra, compreendendo os estímulos produzidos utilizando variação dos articuladores, com qualidade média superior a 2.5. A diferença entre as duas classes é aproximadamente 1.

Devido ao desvios e ao reduzido número de ouvintes, torna-se necessária uma análise mais detalhada dos resultados. A média para o primeira classe de estímulos situa-se, com uma probabilidade de 0.95, entre 1.46 e 1.72. Usando a mesma probabilidade, o intervalo de confiança para a segunda classe de estímulos tem por limites 2.60 e 2.88 e o intervalo de confiança para a diferença situa-se entre 0.95 e 1.36. Mesmo com a dispersão existente nas classificações e reduzido número de ouvintes, é muito provável a melhoria de classificação ao incluir-se a variação dos articuladores.

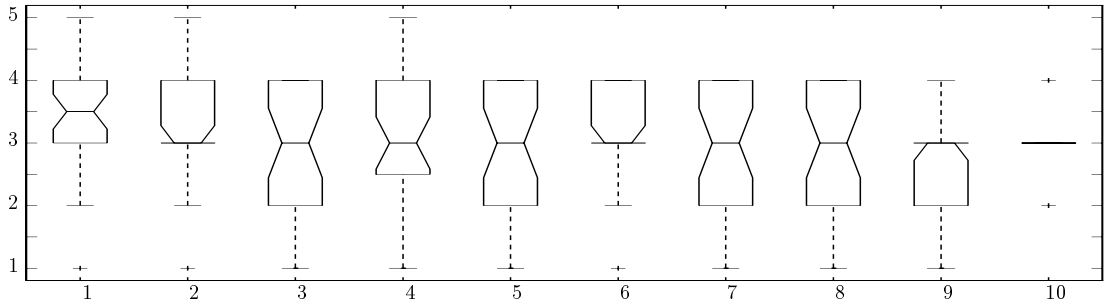
Em termos gerais, a qualidade melhorou ao incluir-se a variação do velo e de outros articuladores, contribuindo, portanto, os conhecimentos obtidos, para uma melhoria de qualidade.

Qual o nível de qualidade actual do sintetizador ?

A variação nas classificações atribuídas e o reduzido número de ouvintes utilizados no teste não permitiram a obtenção da qualidade com grande exactidão. Para facilitar a análise dos resultados, é indicado a seguir ao valor médio, entre parêntesis, o intervalo de confiança. Para estímulos individuais o intervalo de confiança é bastante grande.

Os estímulos, usando parâmetros articulatórios variáveis no tempo, obtiveram classificações médias entre os 2.16 ([1.90..2.42]), no caso de vogais isoladas e os 3.30 ([3.12..3.48]) no caso de vogal nasal entre oclusivas. Para vogal nasal depois de consoante nasal obteve-se uma média de 2.86 ([2.63..3.08]) quando o segmento seguinte é uma oclusiva e de 2.82 ([2.57..3.07]) quando em final de palavra.

A qualidade de algumas palavras geradas com o sintetizador articulatório, na Figura 7.9, obteve uma classificação média geral de 2.97 ([2.87..3.07]). O valor médio é muito próximo de 3 e, devido à média incluir vários estímulos, o intervalo de confiança é razoavelmente pequeno. O melhor resultado, para a palavra *mão*, foi de 3.38 ([3.04..3.72]). Refira-se que, em média, os estímulos naturais obtiveram uma classificação de aproximadamente 4.6 ([4.45..4.69]). Se os valores forem normalizados para os sons naturais terem em média 5, a média geral situa-se em 3.23 ([3.13..3.33]) e a melhor classificação média em 3.67 ([3.33..4.00]), valores entre o razoável e bom. Os melhores resultados para a palavra *mãe* foram obtidos com configuração oral de [a] e F_0 não natural, número 4; os piores para os casos 7 e 8, em que não se utilizou interacção entre a fonte e o tracto. Estas diferenças não se revelaram significativas estatisticamente, apontando para alguma liberdade na configuração oral para a produção do [ɐ̃] e também na



(a) Diagrama extremos-e-quartis (*boxplot*)

Palavra	mão		mãe						António		Todas
Número	1	2	3	4	5	6	7	8	9	10	
Média	3.38	3.03	2.94	3.06	2.94	3.00	2.88	2.84	2.66	3.00	2.97
Desvio padrão	0.94	0.97	0.98	0.95	0.88	0.98	1.04	0.99	0.75	0.67	

(b) Classificações médias

Figura 7.9: Qualidade das palavras geradas com o sintetizador articulatório. A ordem dos estímulos é a utilizada na Tabela 7.1(a).

forma de realização do final da palavra. Os melhores resultados com frequência fundamental não natural podem ser considerados como uma indicação das capacidades do modelo de fonte glotal utilizado, devidas à inclusão de fenómenos como o *jitter* e a interacção entre a fonte e o tracto. A palavra *António*, que contém oclusivas para as quais o sintetizador ainda não se encontra preparado, obteve 2.66 ([2.39..2.93]), com F_0 não natural e 3 ([2.76..3.24]), com F_0 natural. Em ambos os casos é inteligível [4].

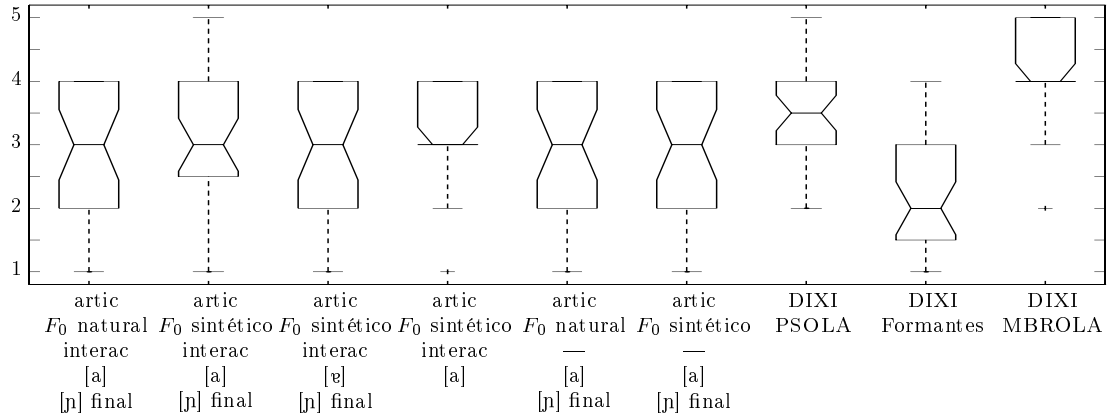
Qual o posicionamento relativamente a outras técnicas?

A ordenação das várias técnicas é a seguinte: MBROLA é a melhor seguida de PSOLA, duas técnicas usando voz natural; a seguir, todos os exemplares produzidos por síntese articulatória e, finalmente, com os piores resultados, a síntese de formantes.

Para a palavra *mãe*, na Figura 7.10, é estatisticamente significativa a diferença entre as 4 técnicas. Este resultado necessita de ser confirmado, pois foi obtido usando apenas uma palavra.

Para a palavra *mão*, cujos resultados se apresentam na Figura 7.11, a síntese articulatória obteve resultados semelhantes, ou mesmo ligeiramente superiores, à síntese por concatenação

[4]Esta afirmação não resulta do teste MOS efectuado, mas de testes informais de identificação realizados. Nestes testes não foi fornecida qualquer informação aos ouvintes, nem mesmo que se tratava de sons sintetizados.



(a) Diagrama extremos-e-quartis (*boxplot*)

	Sintetizador Articulatório						DIXI		
Síntese	articulatória						PSOLA	formantes	MBROLA
	1	2	3	4	5	6	7	8	9
Média	2.94	3.06	2.94	3.00	2.88	2.84	3.41	2.19	4.22
Desvio padrão	0.98	0.95	0.88	0.98	1.04	0.99	0.76	0.86	0.79

(b) Classificações médias

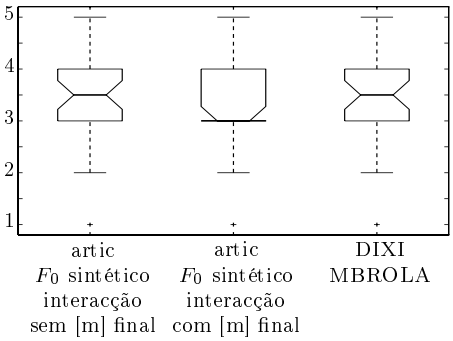
Figura 7.10: Relação entre a qualidade obtida pelas várias técnicas de síntese para a palavra *mãe*.

MBROLA, que utiliza sinal de voz natural. As diferenças não são estatisticamente significativas.

Os resultados obtidos pelos estímulos gerados pelo sistema DIXI encontram-se na Figura 7.12. Os piores resultados foram obtidos usando síntese de formantes, para a palavra *mãe*. Para a palavra *mão*, o sistema DIXI, mesmo usando síntese por concatenação MBROLA, não foi além de uma classificação média de 3.22, sendo o intervalo de confiança entre 2.83 e 3.61.

Os estímulos produzidos usando síntese MBROLA, de que existiam 7 exemplos no teste, obtiveram um valor médio de 4.10, atingindo para a palavra *não* 4.44. O valor médio para o conjunto dos 7 estímulos encontra-se, com uma probabilidade de 0.95, entre 3.98 e 4.22. A situação para a palavra *não* é muita mais imprecisa, sendo o intervalo de confiança entre 4.18 e 4.7.

Sendo, como já foi referido anteriormente, o intervalo de confiança da média para os estímulos produzidos usando síntese articulatória entre 2.87 e 3.07 e para os estímulos produzidos usando MBROLA entre 3.98 e 4.22, a diferença de qualidade destas duas técnicas situa-se entre os 0.97 e os 1.29, com média igual a 1.13.

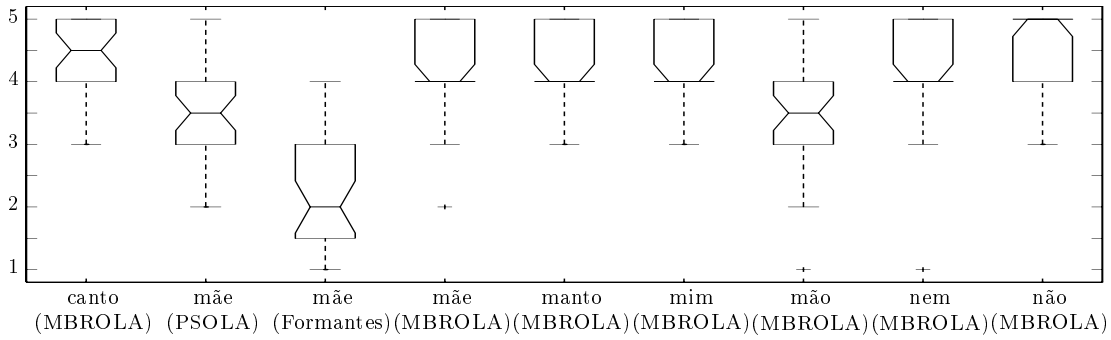


(a) Diagrama extremos-e-quartis (*boxplot*)

Técnica	Articulatório		DIXI MBROLA
Número	1	2	3
Média	3.38	3.03	3.22
Desvio padrão	0.94	0.97	1.07

(b) Média a desvio padrão.

Figura 7.11: Qualidade para a palavra *mão*. Os dois primeiros estímulos foram gerados por síntese articulatória, o terceiro pelo sistema DIXI, usando MBROLA.



(a) Diagrama extremos-e-quartis (*boxplot*)

Palavra	canto	mãe			manto	mim	mão	nem	não
Técnica	MBROLA	PSOLA	Formantes	MBROLA	MBROLA				
Média	4.34	3.41	2.19	4.22	4.31	4.16	3.22	4.03	4.44
Desvio padrão	0.75	0.76	0.86	0.79	0.64	0.72	1.07	1.28	0.72

(b) Média a desvio padrão.

Figura 7.12: Qualidade dos vários estímulos produzidos pelo sistema DIXI.

7.2 Resumo

Neste capítulo apresentou-se a avaliação da qualidade que actualmente é possível obter com o sintetizador articulatório desenvolvido. Como todo o trabalho apresentado, a avaliação incidiu sobre os sons nasais do Português.

Os estudos de produção e percepção efectuados, usando o sintetizador articulatório como um informador versátil, permitiram aumentar a qualidade obtida. Para os casos de vogal nasal entre oclusivas e depois de consoante nasal, o ganho de qualidade, ao incluir-se no processo de síntese a variação dos articuladores no tempo, foi significativo. Entre oclusivas, o ganho médio atingiu os 0.7 e depois de consoante nasal, cerca de 1. Em alguns casos, atingiram-se classificações médias superiores a 4, indicação clara da boa qualidade obtida.

Os resultados revelam ser, no momento, possível sintetizar sons de qualidade entre o razoável e bom. Os resultados obtidos usando síntese articulatória revelaram-se superiores aos produzidos por síntese de formantes, para o caso testado. Em relação à síntese por concatenação MBROLA, a qualidade do sintetizador revelou-se, em média, inferior cerca de 1.1.

Não tendo a avaliação de qualidade sido exaustiva, os resultados apresentados não podem ser considerados como definitivos e absolutos. Possivelmente para outras palavras obter-se-iam resultados diferentes. Julgamos, no entanto, serem estes resultados suficientemente interessantes para que se continue o processo de melhoria do sintetizador, fazendo uso de um processo iterativo, com melhorias sucessivas do sintetizador e dos conhecimentos existentes acerca do processo de produção e percepção dos sons nasais, e não só, do Português.

Capítulo 8

Conclusões

> It is a mistake to believe that science consists of nothing but conclusively proven propositions, and it is unjust to demand that it should. It is a demand made by only those who feel a craving for authority in some form and a need to replace the religious catechism by something else, even if it be a scientific one.
>
> Sigmund Freud

Esta dissertação descreve o desenvolvimento de um sintetizador articulatório para os sons nasais e a sua aplicação, como informante versátil, em estudos de produção e percepção das vogais nasais do Português Europeu.

Neste último capítulo, apresenta-se um resumo do trabalho realizado, as principais conclusões e possíveis continuações para o trabalho.

8.1 Resumo do trabalho efectuado

Motivados pela necessidade de melhorar a qualidade da voz produzida por sistemas de síntese, seguiu-se, nesta tese, um caminho novo em relação ao Português: a produção de voz pela simulação dos fenómenos físicos e acústicos que estão na base da produção de voz pelos seres humanos. Este caminho é longo, apenas se podendo considerar este trabalho como um pequeno passo.

Uma parte essencial do trabalho desenvolvido consistiu no desenvolvimento de um sintetizador articulatório, especialmente vocacionado para os sons nasais.

O sintetizador constitui, não só um sistema capaz de gerar voz sintética, como também uma ferramenta valiosa para melhorar os nossos conhecimentos sobre a produção e percepção. Foram efectuados estudos sobre: o efeito de interacção acústica entre a fonte glotal e as cavidades supralaríngeas para sons nasais; a influência da variação no tempo do velo e outros articuladores, na percepção de vogais nasais.

Os conhecimentos obtidos foram utilizados na síntese de vogais isoladas, sequências constituídas por consoante nasal e vogal nasal e algumas palavras contendo sons nasais. A qualidade destes sons foi avaliada num teste de qualidade.

8.2 Conclusões

O primeiro resultado deste trabalho é o sintetizador articulatório. Especialmente desenvolvido para a síntese de sons nasais, constitui uma ferramenta muito útil na realização de estudos de produção e percepção acerca deste tipo de sons, tão importantes para a língua portuguesa. Uma escolha criteriosa dos modelos utilizados na sua implementação, com melhorias e adição de novas funcionalidades nalguns deles, conjuntamente com conhecimentos que foram obtidos em experiências realizadas com a sua ajuda, permitiram que se conseguisse obter vogais nasais e mesmo algumas palavras contendo sons nasais, com qualidade entre o razoável e bom.

O ponto mais importante do trabalho foi a constatação de que as vogais nasais, pelo menos em Português, são sons em que a variação temporal das suas características, causada pela dinâmica dos articuladores, tem especial relevância. Esta "descoberta", nas nossas experiências iniciais, tendo por objectivo a obtenção de vogais nasais de qualidade natural, motivou a realização de diversos estudos nos quais obtivemos vários resultados interessantes.

Do conjunto de simulações e testes perceptuais realizados sobre este tema, destacaria os seguintes resultados:

- É necessário incluir no processo de síntese a variação temporal da abertura do velo, conjuntamente com a variação de outros articuladores que controlam a passagem oral, para se obter vogais nasais de qualidade natural;

- Apesar do velo e os outros articuladores terem variações diferentes para os vários contextos, o resultado final é sempre caracterizado por uma predominância inicial do som radiado pelos lábios e por uma predominância, na fase final, da radiação nasal, consistindo a vogal nasal numa transição gradual entre estas duas fases;

- É muito importante a existência da fase final de baixa energia, provocada por predominância da radiação nasal. A existência de um segmento de elevada energia, imediatamente a seguir, reduz, drasticamente, a percepção de nasalidade da vogal;

- O segmento que se segue à vogal nasal influencia a percepção da nasalidade. No caso de se seguir, imediatamente, uma vogal oral, muito dificilmente se percebe a existência da vogal nasal. Os resultados estão de acordo com a evolução histórica das vogais nasais portuguesas, em que sequências de vogal nasal seguida de vogal oral desapareceram, dando lugar a uma única vogal nasal, a uma vogal oral ou a sequências de duas vogais orais com uma consoante nasal intermédia. No caso de se seguir à vogal nasal uma consoante nasal, é difícil a distinção entre vogais produzidas com velo aberto e vogais produzidas com velo fechado, dependendo a percepção de nasalidade também da duração da consoante nasal seguinte. Este resultado está de acordo com a reduzidíssima utilização que é feita, na língua portuguesa, do contraste entre oral e nasal para vogais entre consoantes nasais [1];

- A percepção da nasalidade não é só controlada pelo velo. Existe a possibilidade de utilização da oclusão, ou redução da passagem oral, para melhorar a qualidade de uma vogal nasal. Quando situada em final de palavra, o movimento na fase final de fecho dos lábios, mesmo sem oclusão, melhora a qualidade da vogal nasal, ao aumentar a preponderância da radiação nasal relativamente à radiação oral. A oclusão, devida a coarticulação, antes de oclusivas [2], também contribui para uma melhoria da qualidade. As durações preferidas pelos ouvintes para estas oclusões, estão de acordo com estudos anteriormente efectuados, usando síntese de formantes.

Com base nestes resultados, é proposta uma teoria para as vogais nasais portuguesas, em que estas são vistas como sons em que a variação temporal das suas propriedades é muito importante para a sua percepção. Como a variação é feita entre uma configuração inicial, com predominância da radiação pelos lábios e uma configuração final, com radiação nasal predominante, consideramos que as vogais nasais podem ser consideradas ditongos.

[1] Alguns dialectos do Português usam este contraste para distinguirem entre a primeira pessoa do plural do pretérito e do presente, tendo-se *am*[ẽ]*mos* (do Latim AMAMUS), e *am*[a]*mos* (do Latim AMMAMUS) (Hajek, 1997, pp. 77).

[2] Nos nossos estudos a oclusiva é bilabial.

Os resultados de testes perceptuais efectuados mostram existir uma dependência da perceptibilidade do efeito de interacção acústica entre a fonte glotal e o tracto com a altura da vogal. Para vogais produzidas com a língua numa posição elevada, o efeito é perceptível. Foi proposta uma explicação desta dependência, baseada em simulações realizadas com o auxílio do sintetizador articulatório.

Este estudo apenas foi possível pelo desenvolvimento de um modelo da fonte glotal, incluindo interacção acústica com as cavidades supraglotais e a implementação, no modelo acústico, da possibilidade de controlar a inclusão da impedância de entrada do tracto nasal, no cálculo da impedância de entrada das cavidades supraglotais.

Os estudos efectuados permitiram melhorar o conhecimento acerca das vogais nasais, contribuindo para a melhoria da qualidade de vogais nasais produzidas. Pode concluir-se que este tipo de técnica é útil como ferramenta em estudos dos fenómenos de produção e percepção, em especial, os relacionados com as características dinâmicas, como a coarticulação.

O trabalho desenvolvido baseou-se em duas ideias base: por um lado, utilização de mais conhecimento no sistema de síntese acerca do processo humano de produção de voz e por outro, no aumento desse conhecimento através da utilização do método científico. Em termos gerais, os resultados obtidos, em especial por terem sido obtidos por um único investigador com formação de base afastada das ciências da linguagem, como a Fonética, permitem afirmar que esta abordagem mais científica da síntese de voz, advogada por investigadores como Fant (1998) e Greenberg (1998), pode, e deve, ser utilizada como caminho alternativo para a resolução do problema da naturalidade de voz, produzida por sistemas de síntese.

8.3 Desenvolvimentos futuros

Muitos e variados, assuntos podem ser abordados como continuação do trabalho apresentado. Referem-se, de seguida, aqueles que pretendemos vir a abordar.

Obtenção de dados

É essencial para a continuação deste trabalho a obtenção de dados anatómicos e fisiológicos, para validação e melhoramento dos modelos utilizados. Torna-se necessário obter informação acerca da configuração do tracto vocal, durante a produção dos vários sons utilizados pela língua portuguesa, não só estática, mas também a forma como essa configuração varia ao longo do tempo, nas diversas transições entre sons. Este trabalho, já iniciado (Vaz e Teixeira, 1998), começará pela recolha de dados acerca da evolução no tempo de vários articuladores, incluindo o velo, usando EMMA. A utilização, prevista, de Ressonância Magnética permitirá a obtenção de informação acerca das cavidades nasais, posição da língua, e posição do velo.

Espera-se, num futuro mais distante, poder continuar as recolhas, usando novas técnicas como radar de baixa potência (Burnett *et al.*, 1999). Técnicas já antigas, como electroglotografia e videoendoscopia também têm interesse, para obtenção de dados acerca da excitação glotal.

A obtenção de dados directos pode ser complementada pela análise de sinal de voz natural, de mais fácil acesso, tendo sido dados os primeiros passos para a recolha e posterior análise de uma base de dados dos sons nasais do Português Europeu, contemplando as várias regiões do país. Para facilitar análises relacionadas com a excitação, está prevista a utilização de electroglotografia, além da gravação do sinal de voz.

Realização de mais experiências acerca de sons nasais

Em relação às vogais nasais existe a possibilidade de continuação dos testes perceptuais e simulações.

Relativamente ao efeito da dinâmica dos articuladores, seria interessante estudar, de forma exaustiva, quais as durações das várias fases de realização de uma vogal que os ouvintes portugueses preferem. Outro estudo interessante seria o da influência da forma como é feito o movimento de abertura e fecho do velo na qualidade de uma vogal nasal. Deveriam ser utilizadas variações da abertura mais aproximadas da realidade do que a actual aproximação, usando interpolação linear, entre os estados de fecho e abertura máxima. Parâmetros mantidos constantes nos estudos até agora efectuados, como é o caso da duração, deveriam ser incluídos como factores em novas experiências.

Os estudos de interacção entre a fonte glotal e o tracto devem ser continuados, utilizando-se valores para os parâmetros da fonte glotal adequados às vogais nasais do Português, obtidos por medição directa, complementada por análise de sinal de voz natural. A frequência fundamental e os parâmetros definindo a forma da área glotal podem ter influência nos resultados. Deverá, também, ser estudado o efeito nos resultados obtidos de diferentes configurações da faringe, em especial na zona imediatamente a seguir à laringe.

Seria, também, interessante estudar o efeito de alterações das cavidades nasais e paranasais, causadas por doenças, nos sons nasais produzidos.

Melhoramento do sintetizador

O sintetizador no seu estado actual apenas permite a síntese de vogais orais, de consoantes e vogais nasais. Torna-se necessário continuar o seu desenvolvimento para ser possível a síntese dos outros tipos de sons utilizados pela língua portuguesa: fricativas, oclusivas, laterais (Narayanan *et al.*, 1995) e vibrantes. A extensão implica a evolução do modelo representando as cavidades (modelo articulatório) e do modelamento da geração e propagação do ar nessas cavidades (modelo acústico). Para as fricativas, gostaríamos de explorar os resultados recentes de Sinder (1999), utilizando teorias da aerodinâmica.

Mesmo para os sons que o sintetizador é capaz de modelar, a evolução dos modelos articulatório e acústico é possível e desejável.

Relativamente à representação das cavidades supraglotais, a utilização dum modelo articulatório tridimensional (Engwall, 1999; Badin *et al.*, 1998; Wu e Wilhelms-Tricarico, 1994; Wilhelms-Tricarico, 1995) permitiria evitar a conversão, empírica, de distâncias sagitais para a função de área e incluir informação adicional acerca da forma das cavidades acústicas. No que concerne ao tracto nasal, será útil a utilização de dados anatómicos, considerando as duas passagens laterais, um melhor modelamento da zona do véu palatino e da área de abertura.

A obtenção da posição dos diversos articuladores, com base na activação dos vários músculos que os controlam, é uma capacidade que gostaríamos de ver incluída no sintetizador (Bouabana, 1995; Laboissiére *et al.*, 1999; Perrier e Ostry, 1994).

A utilização de novas técnicas de simulação do tracto, como por exemplo, *Transmission Line Model* ou *Transmission Line Matrix* (TLM) (Miki *et al.*, 1999), permitirá incluir informação sobre a forma das cavidades e a inclusão no cálculo de outros modos de propagação das ondas sonoras, para além do modo de propagação longitudinal actualmente utilizado.

Relativamente à fonte de excitação glotal, uma das áreas a abordar será a síntese amostra a amostra, para estudar o efeito da variação ao longo de um período de abertura-fecho das cordas vocais na onda de fluxo (Laine, 1999).

Extensão do processo de inversão

A extensão a outros tipos de sons do processo de obtenção dos parâmetros articulatórios, utilizando unicamente informação extraída de sinal de voz natural, deve ser estudada. O primeiro passo será a extensão da inversão a vogais nasais e consoantes nasais. Esta extensão implica: uma forma mais rápida e robusta de obtenção de pólos e zeros da função de transferência representativa do modelo do tracto, usando técnicas como as recentemente propostas por Jospa e Praag (1999); a obtenção de pólos e zeros do sinal acústico (Regalia, 1995; Miyanaga, 1990; Lee, 1992). Podem também tentar-se outros métodos de optimização, como os algoritmos genéticos (McGowan, 1994). Os dados obtidos por medição directa serão indispensáveis para validação do processo de inversão.

Divulgação/Disponibilização do sintetizador

O sintetizador desenvolvido foi até ao momento utilizado apenas pelo autor. Considera-se, no entanto, que poderá ser útil para o ensino e para trabalhos e investigação de outros. Tendo por objectivo a sua disponibilização, foi já iniciado o trabalho de adaptação do sintetizador existente a um ambiente computacional de utilização mais generalizada, Microsoft Windows, dotando-o de uma interface com o utilizador adequada aos utilizadores. Será aproveitada a oportunidade para evoluir alguns modelos e, se possível, incluir a síntese de novos tipos de sons, como as fricativas.

Aplicação em terapia da fala

A síntese articulatória pode ser usada na implementação de ajudas na terapia da fala. Um sistema que, utilizando métodos de inversão, permita transmitir visualmente informação sobre a forma correcta de posicionamento dos vários articuladores, será muito útil para ensino de surdos. Esta informação pode ser transmitida sobre a forma de um jogo, cativando mais facilmente crianças para a sua utilização.

Com a evolução do sintetizador, bem como dos processos de inversão, será interessante retomar o trabalho realizado por Branco (1997) nesta área.

Aplicações na área da música

A colaboração, recente, num trabalho de aplicação de processamento de voz na área da música (Sá Pinto, 2000) despertou o interesse do autor, e de outros, para a aplicação das técnicas de síntese articulatória em duas áreas relacionadas com a música: o canto (Sundberg, 1997; Cook, 1993) e o modelamento de instrumentos musicais (Scavone, 1997).

Integração num sistema de conversão de texto para fala

O nosso objectivo, a longo prazo, é a obtenção de um sistema de conversão de texto para fala, baseado no sintetizador em desenvolvimento.

8.4 Epílogo

> Para concluir, acho que só há um caminho para a ciência - ou para a filosofia: encontrar um problema, ver a sua beleza e apaixonarmo-nos por ele; casarmo-nos com ele, até que a morte nos separe - a não ser que encontremos outro problema ainda mais fascinante, ou a não ser que obtenhamos uma solução. Mas ainda que encontremos uma solução, poderemos descobrir, para nossa satisfação, a existência de toda uma família de encantadores, se bem que talvez difíceis, problemas-filhos, para cujo bem-estar poderemos trabalhar, com uma finalidade em vista, até ao fim dos nossos dias.
>
> Sir Karl Popper, (Popper, 1992, Prefácio de 1956, pág. 42)

Acrónimos

ANOVA *Analyis of Variance.*

ASY *Articulatory Synthesis.* Nome do sintetizador desenvolvido nos Laboratórios Haskins, inicialmente por Mermelstein (1973) e depois melhorado por Rubin *et al.* (1981).

CASY *Configurable Articulatory Synthesizer.* Nome da versão mais recente e melhorada do *Articulatory Synthesis* (ASY).

CVC *Consonant Vowel Consonant.* Sequências de consoante, vogal e consoante.

DANA *Dynamic Analog of the NAsal cavities.*

DRM *Distinctive Region Model.*

EGG *Electroglotography*

EMA *ElectroMagnetic Articulography.*

EMMA *ElectroMagnetic Midsagittal Articulography.*

FFT *Fast Fourier Transform.* Algoritmo rápido de cálculo da Transformada Discreta de Fourier.

FFTW *Fast Fourier Transform in the West.*

GNU *GNU is not UN*X.*

JND *Just Noticeable Difference.*

KTH *Kungl Tekniska Högskolan.* Instituto Real de Tecnologia de Estocolmo.

LIR *Lips Impulse Response.*

LPAT *Linear Prediction Acoustic Tube.*

LPC *Linear Predictive Coding.*

MBROLA *Multi-Band Re-synthesis Overlap-Add*

MIT *Massachussets Institute of Technology.*

MMIRC *Mind Machine Interaction Research Center.* Laboratório da University of Florida chefiado pelo Professor Donald Childers.

MOS *Mean Opinion Score*

MRI *Magnetic Ressonance Imaging.*

OQ *Open Quotient.* Veja-se a Figura 4.9, na página 98 para a sua definição.

PIS *Piston in Sphere* Modelo de radiação.

PSOLA *Pitch Synchronous Overlap-Add*

RLS *Recursive Least Squares*

SQ *Speed Quotient.* Veja-se a Figura 4.9, na página 98 para a sua definição.

TLM *Transmission Line Model* ou *Transmission Line Matrix.*

TTS *Text to Speech.* Sistemas de conversão de texto para fala.

WRLS-VFF *Weighted RLS with Variable Forgetting Factor.*

Bibliografia

> If I have seen further it's because I've stood on the shoulders of giants.
>
> *Sir Isaac Newton*

> If I haven't gone too far it's because giants stepped on me.
>
> *Anonymous*

P. Alku (1992). An automatic method to estimate the time-based parameters of the glottal pulseform. In *Proc. ICASSP*, páginas 29–32. San Francisco, USA.

P. Alku (1993). Estimation of the glottal excitation of speech with pitch-synchronous iterative adaptive inverse filtering. In Cooke *et al.* (1993), capítulo 9, páginas 140–146.

P. Alku, E. Vilkman, e A.-M. Laukkanen (1998). Estimation of amplitude features of the glottal flow by inverse filtering speech pressure signals. *Speech Communication*, 24:123–132.

D. Allen e W. Strong (1985). A model for the synthesis of natural sounding vowels. *Journal of the Acoustical Society of America*, 78:58–69.

A. Almeida (1976). The portuguese nasal vowels: Phonetics and phonemics. In J. Schmidt-Radefelt (editor), *Readings in Portuguese Linguistics*, páginas 348–396. North Holland, New York.

T. V. Ananthapadmanabha e G. Fant (1982). Calculation of true glottal flow and its components. *Speech Communication*, 1:167–187.

B. Atal, J. L. Miller, e R. D. Kent (editores) (1991). *Papers in Speech Communication: Speech Processing*. Acoustical Society of America. Ver também Kent *et al.* (1991); Miller *et al.* (1991).

B. S. Atal, J. J. Chang, M. V. Mathews, e J. W. Tukey (1978). Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique. *Journal of the Acoustical Society of America*, 63(5):1535–1555.

P. Badin, G. Bailly, M. Raybaudi, e C. Segebarth (1998). A three-dimensional linear articulatory model based on MRI data. In *Proc. ICSLP '98*. Sydney, Australia.

P. Badin e G. Fant (1984). Notes on vocal tract computation. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 2-3:53–108.

T. Baer, J. C. Gore, L. C. Gracco, e P. W. Nye (1991). Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels. *Journal of the Acoustical Society of America*, 90(2 (Pt 1)):799–828.

G. Bailly, R. Laboissière, e A. Galván (1997). Learning to speak: Speech production and sensori-motor representations. In P. Morasso e V. Sanguineti (editores), *Self-Organization, Computational Maps and Motor Control*, páginas 593–615. Elsevier, Amsterdam.

J. M. Barbosa (1961). Les voyelles nasales portugaises: Interpretation phonologique. In A. Sovijärvi e P. Aalto (editores), *Proc. Fourth International Congress of Phonetic Sciences*, páginas 691–709. Mouton & Co, The Hague.

J. M. Barbosa (1965). *Études de Phonologie Portugaise*. Junta de Investigações do Ultramar, Centro de Estudos Políticos e Sociais, Lisboa.

J. M. Barbosa (1994). *Introdução ao Estudo da Fonologia e Morfologia do Português*. Livraria Almedina, Coimbra.

M. Båvegård (1995a). Interactive voice source modelling. In *Proceedings ICPhS*, páginas 634–637. Stockolm.

M. Båvegård (1995b). Introducing a parametric consonantal model to the articulatory speech synthesizer. In *Proc. Eurospeech*, páginas III, 1857–1860. Madrid.

M. Båvegård (1996). Towards an articulatory speech synthesizer: Model development and simulations. *Speech, Music and Hearing, Quarterly Progress and Status Report*, TMH-QPSR 1:1–15.

M. Båvegård e G. Fant (1994). Audibility of interaction ripple. WP1.2 Delivery 14 bis, SPEECHMAPS (ESPRIT/BR Project No 6975).

M. Båvegård e G. Fant (1995). From formant frequencies to VT area fuction parameters. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4:55–66.

M. Båvegård, G. Fant, J. Gauffin, e J. Liljencrants (1993). Vocal tract sweeptone data and model simulations of vowels, laterals and nasals. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4:43–75.

D. Beautemps, P. Badin, e R. Laboissiére (1995). Deriving vocal-tract area functions from midagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data. *Speech Communication*, 16(1):27–47.

P. S. Beddor (1993). The perception of nasal vowels. In M. K. Huffman e R. A. Krakow (editores), *Nasals, Nasalization, and the Velum*, Phonetics and Phonology (vol. 5), páginas 171–196. Academic Press Inc.

P. S. Beddor e W. Strange (1982). Cross-language study of perception of the oral-nasal distinction. *Journal of the Acoustical Society of America*, 71:1551–1561.

F. Bell-Berti e T. Baer (1983). Velar position, port size, and vowel spectra. In *Proc. of the 11th International Congress of Acoustics*, páginas 4, 19–21. Paris.

F. Bell-Berti, R. A. Krakow, D. Ross, e S. Horiguchi (1993). The rise and fall of the soft palate: The Velotrace. In M. Stone (editor), *Measuring Speech Production: Abstracts and References*, páginas 21–22. Acoustical Society of America.

A. P. Benguerel, H. Hirose, M. Sawashima, e T. Ushijima (1977). Velar coarticulation in french: Fiberscopic study. *Journal of Phonetics*, 5:149–158.

A. P. Benguerel e A. Lafarge (1981). Perception of vowel nasalization in french. *Journal of Phonetics*, 9:309–321.

C. Benoît (1997). Speech synthesis: Present and future. In G. Bloothooft, W. van Dommelen, C. Espain, P. Green, V. Hazan, M. Huckvale, e E. Wigforss (editores), *The Landscape of Future Education in Speech Communication Sciences*, páginas 119–123. OTS Publications, Utrecht, The Netherlands.

L. Björk (1961). Velopharyngeal function in connected speech - studies using tomography and cineradiography synchronized with speech spectrography. *Acta Radiologica*, Supplement 202:1–94.

L. Bjork, B. Nylen, A. Möller, e G. Fant (1961). Velopharyngeal function in connected speech. *Speech Transm. Lab. - Q.P.S.R.*, 1:13–14.

G. Bjuggren e G. Fant (1964). The nasal cavity structure. *Speech Transm. Lab. - Q.P.S.R.*, 4:5–7.

E. L. Bocchieri e D. G. Childers (1984). Interactive graphics editor permits study of animated speech articulation. *Speech Technology*, páginas 10–14.

P. Boersma (1998). *Functional Phonology: Formalizing the interactions between articulatory and perceptual drives*. LOT International Series 11. Hollan Academic Graphics (HAG), The Hague. ISBN 90-5569-054-6. Doctoral Thesis University of Amsterdam.
URL `http://www.hagpub.com`

E. Bognar e H. Fujisaki (1986). Analysis, synthesis and perception of the french nasal vowels. In *Proc. ICASSP*, páginas Vol 2, 1601–1604.

G. J. Borden, K. S. Harris, e L. J. Raphael (1994). *Speech Science Primer - Physiology, Acoustics, and Perception of Speech*. William-Wilkins, segunda edição.

S. Bouabana (1995). *Modélisation des Mouvements Articulatoires de la Langue par la Méthode de la LPC Multi-impulsionnelle*. Mémoire de soutenance de thése, École Nationale Supérieure des Télécommunications, Paris.

A. Branco (1997). *Representação Visual do Modelo Articulatório para o estudo da Produção da Fala*. Tese de mestrado, Universidade de Aveiro.

A. Branco, A. Teixeira, A. Tomé, e F. Vaz (1997a). An Articulatory Speech Synthesizer. In H. Araújo e L. V. de Sá (editores), *9th Portuguese Conference on Pattern Recognition (RecPad'97)*, páginas 205–208. Coimbra.

A. Branco, A. Teixeira, A. Tomé, e F. Vaz (1997b). Um Sintetizador Articulatório para o Estudo da Produção da Voz. In *I Conferência Nacional de Telecomunicações*. Instituto de Telecomunicações, Aveiro.

A. Branco, A. Tomé, A. Teixeira, e F. Vaz (1997c). A method to extract articulatory parameters from the speech signal using Neural Networks. In *13th International Conference on Digital Signal Processing (DSP97)*. Santorini, Greece.

E. O. Brigham (1988). *The Fast Fourier Transform and its Applications*. Signal Processing Seies, Alan V. Oppenheim, Series Editor. Prentice Hall.

G. A. Brito (1975). The perception of nasal vowels in brazilian portuguese: A pilot study. In C. A. Ferguson, L. M. Hyman, e J. J. Ohala (editores), *Nasálfest - Papers from a Symposium on Nasals and Nasalization*, páginas 49–76. Language Universals Project, Department of Linguistics, Stanford University, Stanford, CA, USA.

C. P. Browman e L. Goldstein (1989). Articulatory gestures as phonological units. *Phonology*, 6:201–251.

C. P. Browman e L. Goldstein (1990). Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18:299–320.

C. P. Browman e L. Goldstein (1992). Articulatory phonology: An overview. *Phonetica*, 49:155–180.

C. P. Browman e L. Goldstein (1995). Gestural syllable position effects in american english. In F. Bell-Berti e L. J. Raphael (editores), *Producing Speech: Contemporary Issues, for Katherine Safford Harris*, AIP Series in Modern Acoustics and Signal Processing, capítulo 2, páginas 19–33. American Institute of Physics (AIP) Press, New York. ISBN 1-56396-286-1.

A. Bryman e D. Cramer (1993). *Análise de Dados em Ciências Sociais: Introdução às Técnicas usando o SPSS*. Métodos e Técnicas. CELTA, Oeiras, segunda edição.

G. C. Burnett, J. F. Holzrichter, T. J. Gable, e L. C. Ng (1999). Direct and indirect measures of speech articulator motions using low power EM sensors. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 2247–2249. University of California, Berkeley, San Francisco.

L. C. Cagliari (1977). *An experimental study of nasality with particular reference to Brazilian Portuguese*. Ph.d. thesis, University of Edinburgh.

R. Carlson (1994). Models of speech synthesis. In D. B. Roe e J. G. Wilpon (editores), *Voice Communication between Humans and Machines*, página 116. National Academy Press, Washington DC. ISBN 0-309-04988-1 (Hard back). Presented at the "Colloquium on Human-Machine Communication by Voice", Irvine, California, February 8-9, 1993, organized by the National Academy of Sciences, USA.

R. Carlson e B. Granström (1996). Speech synthesis. In Hardcastle e Laver (1996), capítulo 26, páginas 768–788.

P. M. Carvalho, L. C. Oliveira, I. M. Trancoso, e M. C. Viana (1998). Concatenative speech synthesis for european portuguese. In *Proc. of Third ESCA/COCOSDA Interaction Workshop on Speech Synthesis*. Jenolan Caves, Australia.

F. Charpentier (1984). Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic nonlinearities. *Speech Communication*, 3(4):291–308.

M. Y. Chen (1995). Acoustic parameters of nasalized vowels in hearing-impaired and normal-hearing speakers. *Journal of the Acoustical Society of America*, 98(5, Pt 1):2443–2453.

M. Y. Chen (1996). *Acoustic Correlates of Nasality in Speech*. Phd thesis, Massachusetts Institute of Technology, USA.

M. Y. Chen (1997). Acoustic correlates of english and french nasalized vowels. *Journal of the Acoustical Society of America*, 102(4):2360–2370.

S. Chennoukh, D. Sinder, G. Richard, e J. Flanagan (1997). Voice mimic system using articulatory codebook for estimation of vocal tract shape. In *Proc. Eurospeech '97*, páginas 429–432. Rhodes, Greece.

T. Chiba e M. Kajiyama (1958). *The Vowel, its nature and structure*. Phonetic Society of Japan, Tokyo. First published in 1941.

D. G. Childers (1985). Voice (as opposed to speech) synthesis. In *Voice I/O Systems Applications Conference*, páginas 349–361. Sponsored by American Voice I/O Society (AVIOS), San Francisco, CA.

D. G. Childers (2000). *Speech Processing and Synthesis Toolboxes*. John Wiley & Sons.

D. G. Childers e C. Ding (1991). Articulatory synthesis: nasal sounds and male female voices. *Journal of Phonetics*, 19:453–464.

D. G. Childers e J. N. Larar (1984). Electroglottography for laryngeal function assessment and speech analysis. *IEEE TRansactions on Biomedical Engineering*, BME-31(12):807–817.

D. G. Childers, J. C. Principe, e Y. T. Ting (1995). Adaptive WRLS-VFF for speech analysis. *IEEE Trans. Speech Audio Proc.*, 3(3):209–213.

D. G. Childers, J. J. Yea, e E. L. Bocchieri (1983). Source/vocal-tract interaction in speech and singing synthesis. In A. Askenfelt, S. Felicetti, E. Jansson, e J. Sundberg (editores), *Proceedings of the Stockolm Music Acoustics Conference (SMAC)*, páginas 125–141. Royal Swedish Academy of Music.

Chomsky e Halle (1968). *Sound Pattern of English*. Studies in Language. Harper & Row, Publishers, New York.

J. Clark e C. Yallop (1990). *An Introduction to Phonetics & Phonology*. Basil Blackwell, Inc., Cambridge, MA.

H. Clumeck (1976). Patterns of soft palate movements in six languages. *Journal of Phonetics*, 4:337–351.

C. H. Coker (1967). Synthesis by rule from articulatory parameters. In *Proc. 1967 Conference Speech Commun. Process.*, páginas 52–53. IEEE. Reimpresso em Flanagan e Rabiner (1973).

C. H. Coker (1976). A model of articulatory dynamics and control. *Proc. IEEE*, 64(4):452–460.

C. H. Coker (1997). Systems and methods for performing phonemic synthesis. United States Patent 5633983.

C. H. Coker, N. Umeda, e C. P. Browman (1973a). Automatic synthesis from ordinary english text. In Flanagan e Rabiner (1973), capítulo 36, páginas 400–411.

C. H. Coker, N. Umeda, e C. P. Browman (1973b). Automatic synthesis from ordinary English text. *IEEE Trans. Audio and Electr.*, AU-21(3):293–298.

P. R. Cook (1993). SPASM, a real-time vocal tract physical model controller; and singer, the companion software synthesis system. *Computer Music Journal*, 17(1):30–44.

M. Cooke, S. Beet, e M. Crawford (editores) (1993). *Visual Representations of Speech Signals*. Wiley professional Computing. John Wiley & Sons, Chichester, England.

F. S. Cooper (1961). Speech synthesizers. In A. Sovijärvi e P. Aalto (editores), *Proc. Fourth International Congress of Phonetic Sciences*, páginas 3–13. Mouton & Co, The Hague.

F. S. Cooper, P. C. Delattre, A. M. Liberman, e J. M. B. L. J. Gerstman (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, 24:597–606. Reimpresso em Flanagan e Rabiner (1973).

F. S. Cooper, A. M. Liberman, e J. M. Borst (1951). The interconversion of audible and visible patterns as a basis for research in the perception of speech. In 37 (editor), *Proc. Natl. Acad. Sci.*, páginas 318–325. Reimpresso em Flanagan e Rabiner (1973).

A. Corana, M. Marchesi, C. Martini, e S. Ridella (1987). Minimizing multimodal functions of continuous variables with the "simulated annealing" algorithm. *ACM Transactions on Mathematical Software*, 13(13).

K. E. Cummings e M. A. Clements (1995). Glottal models for digital speech processing: A historical survey and new results. *Digital Signal Processing*, 5:21–42.

C. F. Cunha (1982). *História da Língua Portuguesa*. Sá da Costa, Lisboa. Tradução de Teyssier (1980).

E. d'Andrade e A. Kihm (1988). Fonologia autosegmental e vogais nasais em português. In *Actas do III Encontro da Associação Portuguesa de Linguística*, páginas 51–60.

J. Dang e K. Honda (1994). MRI measurements and acoustic of the nasal and paranasal cavities. *Journal of the Acoustical Society of America*, 94(3, Pt 2):1765.

J. A. de Moraes (1997). Vowel nasalization in Brazilian Portuguese: An articulatory investigation. In *Proc. Eurospeech '97*. Rhodes, Greece.

E. M. G. de Sousa (1994). *Para a Caracterização Fonético-Acústica da Nasalidade no Português do Brasil*. Dissertação de mestrado em linguística, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Brasil.

P. Delattre (1954). Les atributs acoustiques de la nasalité vocalique et consonantique. *Studia Linguist.*, 8:103–109.

P. Delattre (1968). Divergences entre nasalités vocalique et consonantique en Français. *Word*, 24:64–73.

D. Demolin, V. Lecuit, T. Metens, B. Nazarian, e A. Soquet (1998). Magnetic resonance measurements of the velum port opening. In *Proc. ICSLP '98*. Sydney, Australia. Paper no. 532.

C. Ding (1990). *Articulatory Speech Synthesis*. Tese de Doutoramento, University of Florida.

M. H. Draper, P. Ladefoged, e D. Whitteridge (1959). Respiratory muscles in speech. *Journal of Speech and Hearing Research*, 2(1):16–27. Reimpresso em Kent *et al.* (1991).

H. Dudley e T. H. Tarnoczy (1950). The speaking machine of Wolfgang von Kempelen. *Journal of the Acoustical Society of America*, 22(2):151–166.

H. K. Dunn (1950). The calculation of vowel ressonances, and an electrical vocal tract. *Journal of the Acoustical Society of America*, 22:740–753. Reimpresso em Flanagan e Rabiner (1973).

T. Dutoit (1993). *High Quality Text-To-Speech Synthesis of the French Language*. Tese de Doutoramento, Faculté Polytechnique de Mons.

T. Dutoit (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publisher. ISBN 0792344987.

S. El-Masri, X. Pelorson, P. Saguet, e P. Badin (1996). Vocal tract acoustics using the transmission line matrix (TLM) method. In *Proc. ICSLP*.

W. Engelke, G. Hoch, T. Bruns, e M. Striebeck (1996). Velopharyngeal function under different dynamic conditions with EMA and Videoendoscopy. *Folia Phoniatr. Logop.*, 48:65–77.

O. Engwall (1999). Modelling of the vocal tract in three dimensions. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*, volume 1, páginas 113–116. Budapest, Hungary.

G. L. Entenman (1976). *The development of nasal vowels*. Ph.d. thesis, University Texas, Austin.

Entropic (1993a). *ESPS Library Functions and File Type (Version 5.0)*. Entropic Research Laboratory, Inc.

Entropic (1993b). *ESPS Programs (Version 5.0)*. Entropic Research Laboratory, Inc.

Entropic (1993c). *waves+ (Version 5.0)*. Entropic Research Laboratory, Inc.

J. H. Esling, J. Clayards, J. A. Edmondson, Q. Fuyuan, e J. G. Harris (1998). Quantification of pharyngeal articulations using measurements from laryngoscopic images. In *Proc. ICSLP '98*. Sydney, Australia.

D. S. Fagan (1988). Notes on diachronic nasalization in portuguese. *Diachronica*, 5:141–157.

G. Fant, , Q.-G. Lin, e C. Gobl (1985a). Notes on glottal flow interaction. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 2-3:21–45. Contribution to the Gotland Symposium on Voice Acoustics and Dysphonia, Aug. 1985.

G. Fant (1960). *Acoustic theory of speech production*. Mouton and Co., Gravenhage, The Netherlands. Second Edition in 1970.

G. Fant (1991). What can basic research contribute to speech synthesis. *Journal of Phonetics*, 19:75–90.

G. Fant (1995). The LF-model revisited. transformations and frequency domain analysis. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 2-3.

G. Fant (1998). Half a century with speech science. In P. K. Kuhl e L. A. Crum (editores), *16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America*, volume IV, páginas 2383–2384. Seattle, Washington, USA.
URL `http://www.apl.washington.edu/asa/asa.html`

G. Fant, J. Liljencrants, e Q.-G. Lin (1985b). A four-parameter model of glottal flow. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4:1–13.

G. Fant e Q.-G. Lin (1987). Glottal source-vocal tract acoustic interaction. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 1:13–27. Paper EE24, 113th Meeting of the Acoustical Societty of America, May, 1987.

I. H. Faria, E. R. Pedro, I. Duarte, e C. A. M. Gouveia (editores) (1996). *Introdução à linguística Geral e Portuguesa*. Caminho.

S. J. Farlow (1993). *Partial Differential Equations for Scientists and Engineers*. Dover Publications, Inc., New York.

G. Feng (1987). Modélisation acoustique et traitement du signal de parole: le cas des voyelles nasales et la simulation des poles et des zéros. *Bull. Lab. Commun. Parlée, Grenoble*, 16:1–102.

G. Feng e E. Castelli (1996). Some acoustic features of nasal and nasalized vowels: A target for vowel nasalization. *Journal of the Acoustical Society of America*, 99(6):3694–3706.

A. Fettweis (1986). Wave Digital Filters: Theory and Practice. *Proceedings of IEEE*, 74(2):270–327.

J. L. Flanagan (1972). *Speech Analysis, Synthesis and Perception*. Springer-Verlag, New York.

J. L. Flanagan e L. Cherry (1969). Excitation of vocal-tract synthesizers. *Journal of the Acoustical Society of America*, 45(3):764–769.

J. L. Flanagan e K. Ishizaka (1976). Automatic generation of voiceless excitation in a vocal cord - vocal tract speech synthesizer. *IEEE Trans. Ac. Sp. Sig. Proc*, ASSP-24(2):163–170.

J. L. Flanagan, K. Ishizaka, e K. L. Shipley (1975). Synthesis of speech from a dynamic model of the vocal cords and vocal tract. *The Bell System Technical Journal*, 54(3):485–506.

J. L. Flanagan, K. Ishizaka, e K. L. Shipley (1980). Signal models for low bit-rate coding of speech. *Journal of the Acoustical Society of America*, 68(3):780–791.

J. L. Flanagan e L. L. Landgraf (1968). Self-oscilating source for vocal-tract synthesizers. *IEEE Trans. Audio and Electr.*, AU-16(1):57–64. Reimpresso em Flanagan e Rabiner (1973).

J. L. Flanagan e L. Rabiner (editores) (1973). *Speech synthesis*. Benchmark Papers in Acoustics. Dowden Hutchinson & Ross.

M. Frigo (1997). *FFTW 1.1 User's Manual*. Massachusetts Institute of Technology.
URL `http://theory.lcs.mit.edu/~fftw`

M. Frigo e S. G. Johnson (1998). FFTW: An adaptive software architecture for the FFT. In *Proc. ICASSP'98*, volume 2, página 1381.

O. Fujimura (1960). Spectra of nasalized vowels. *Quarterly Progress Report, Research Laboratory of Technology, M. I. T.*, 58:214–218.

O. Fujimura e J. Ludqvist (1971). Sweep-tone measurements of vocal-tract characteristics. *Journal of the Acoustical Society of America*, 49(2 (Pt. 2)):541–558.

A. Galván-Rodríguez (1997). *Étude dans le cadre de l'inversion acoustico-articulatoire: Amélioration d'un modèle articulatoire, normalisation du locuteur et récuperation du lieu de constriction des plosives*. Thése, Institut National Polytechnique de Grenoble.

M. J. Galvão (1998). The nasal vowels in iberian portuguese. In P. K. Kuhl e L. A. Crum (editores), *16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America*, páginas 2949–2950. Seattle, Washington, USA.
URL `http://www.apl.washington.edu/asa/asa.html`

M. Garman (1990). *Psycholinguistics*. Cambridge Textbooks in Linguistics. Cambridge University Press.

Goffe, Ferrier, e Rogers (1994). Global optimization of statistical functions with simulated annealing. *Journal of Econometrics*, 60(1/2):65–100.

J. A. Goldsmith (1990). *Autosegmental & Metrical Phonology*. Blackwell, Oxford, UK.

B. Gopinath e M. M. Sondhi (1970). Determination of the shape of the human vocal tract from acoustical measurements. *The Bell System Technical Journal*, 49:1195–1214.

S. Greenberg (1998). Recognition in an new key - Towards a science of spoken language. In *Proc. ICASSP'98*, volume 2, página 1041.

S. K. Gupta e J. Schroeter (1991). Low update rate articulatory analysis/synthesis of speech. In *Proc. ICASSP*, páginas 481–484.

S. K. Gupta e J. Schroeter (1993). Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis. *Journal of the Acoustical Society of America*, 94(5):2517–2530.

J. Hajek (1997). *Universals of Sound Change in Nasalization*. Blackwell, Oxford. ISBN 0631204563.

G. Hammarström (1952). Le chromographe et le triangle tonométrique de Lacerda. *Revista do Laboratório de Fonética Experimental (de Coimbra)*, I:28–38.

W. J. Hardcastle (1976). *Physiology of Speech Production - An Introduction for Speech Scientists*. Academic Press, London.

W. J. Hardcastle e N. Hewlett (editores) (1999). *Coarticulation: Theoretical and Empirical Perspectives*. Cambridge University Press, Cambridge.

W. J. Hardcastle e J. Laver (editores) (1996). *Handbook of Phonetic Sciences*. Blackwell, Oxford.

W. J. Hardcastle e A. Marchal (editores) (1990). *Speech Production and Speech Modelling*. NATO Advanced Science Institute Series: D, Volume 55. Kluwer Academic Publishers, Dordrecht. ISBN 0-7923-0746-1.

M. Harrison e M. McLennan (1997). *Effective Tcl/Tk Programming*. Professional Computing Series. Addison-Wesley.

S. Hattori, K. Yamamoto, e O. Fujimura (1958). Nasalization of vowels in relation to nasals. *Journal of the Acoustical Society of America*, 30(4):267–274.

S. Hawkins e K. N. Stevens (1985). Acoustic and perceptual corelates of the non-nasal–nasal distinction for vowels. *Journal of the Acoustical Society of America*, 77(4):1560–1575.

S. Haykin (1996). *Adaptive Filter Theory*. Signal Processing Series. Prentice-Hall, New Jersey, terceira edição. ISBN 0 13 322760 X.

M. H. L. Hecker (1961). Dynamic analog of the nasal cavities. *Quarterly Progress Report, Research Laboratory of Technology, M. I. T.*, 62:196–197.

M. H. L. Hecker (1962). Studies of nasal consonants with an articulatory speech synthesizer. *Journal of the Acoustical Society of America*, 34(2):179–188.

G. C. Hegerl e H. Höge (1991). Numerical simulation of the glottal flow by model based on the compressible navier-stokes equations. In *Proc. ICASSP*, páginas 477–480.

J. M. Heinz e K. N. Stevens (1964). On the derivation of area functions and acoustic spectra from cineradiographic films of speech. *Journal of the Acoustical Society of America*, 36(A):1037–1038.

W. L. Henke (1966). *Dynamic articulatory model of speech production using computer simulation*. Phd thesis, MIT, Cambridge, MA.

D. Hermes (1993). Pitch analysis. In Cooke *et al.* (1993), capítulo 1, páginas 3–25.

P. Hoole e N. Nguyen (1997). Electromagnetic articulography in coarticulation research. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, 35:177–184. Também em Hardcastle e Hewlett (1999).
URL `http://www.phonetik.uni-muenchen.de/FIPKM/index.html`

S. Horiguchi e F. Bell-Berti (1987). The Velotrace: A device for monitoring velar position. *The Cleft Palate Journal*, 24(2):104–111.

Y. Horii (1979). Fundamental frequency perturbation observed in sustained phonation. *Journal of Speech and Hearing Research*, 22:5–19.

H. Horn, T. Scholl, R. Berndt, I. Hertich, H. Ackermann, e G. Goez (1999). Development and advantages of a new electromagnetic articulograhy instrument for the investigation of the orofacial structures. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*. University of California, Berkeley, San Francisco.

A. S. House e K. S. Stevens (1956). Analog studies of the nasalization of vowels. *Journal of Speech and Hearing Disorders*, 21(2):218–232.

Y.-F. Hsieh (1994). *A Flexible and High Quality Articulatory Speech Synthesizer*. Tese de Doutoramento, University of Florida.

L. Ingber (1993). Simulated annealing: Practice versus theory. *Journal Mathl. Comput. Modelling*, 18(11):29–57.
URL `http://www.ingber.com/asa93_sapvt.ps.Z`

K. Ishizaka e J. L. Flanagan (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *The Bell System Technical Journal*, 51:1233–1268. Reimpresso em Flanagan e Rabiner (1973).

K. Ishizaka, J. C. French, e J. L. Flanagan (1975). Direct determination of vocal tract wall impedance. *IEEE Trans. Ac. Sp. Sig. Proc*, ASSP-23(4):370–373.

K. Ishizaka, M. Matsudaira, e T. Kaneko (1976). Input acoustic-impedance measurement of the subglottal system. *Journal of the Acoustical Society of America*, 60(1):190–197.

S. W. Jacob, C. A. Francone, e W. J. Lossow (1990). *Anatomia e Fisiologia Humana*. Editora Guanabara, Rio de Janeiro, quinta edição. Tradução de Carlos Miguel Gomes Sequeira, de, Structure and Function in Men, 3rd edition, 1982, W. B. Saunders Company.

W. Jassem e F. Nolan (1984). Speech sounds and languages. In G. Bristow (editor), *Electronic Speech Synthesis : Techniques, Technology and Applications*, páginas 19–47. Granada, London.

P. Jospa e R. V. Praag (1999). Sound field computaion in a network of non uniform ducts. Application to the vocal tract. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 2141–2144. University of California, Berkeley, San Francisco.

H. Kawasaki (1986). Phonetic explanation for phonological universals: the case of distintive vowel nasalization. In J. J. Ohala e J. J. Jaeger (editores), *Experimental Phonology*, páginas 81–103. Academic Press, New York.

E. Keller (editor) (1994). *Fundamentals of Speech Synthesis and Speech Recognition - Basic Concepts, State-of-the-art and Future Challenges*. John Wiley & Sons. ISBN 0 471 94449 1.

J. L. Kelly Jr. e C. C. Lochbaum (1962). Speech synthesis. In *Proc. Fourth Intern. Congr. Acoust., Paper G42*, volume 22, páginas 1–4. Reimpresso em Flanagan e Rabiner (1973).

R. D. Kent, B. Atal, e J. L. Miller (editores) (1991). *Papers in Speech Communication: Speech Production*. Acoustical Society of America. Ver também Atal *et al.* (1991); Miller *et al.* (1991).

S. Kirkpatrick, C. D. Gelatt, e M. P. Vecchi (1983). Optimization by simmulated annealing. *Science*, 220(4598):671–680.

D. Klatt (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustic Society of America*, 67(3):971–995.

D. Klatt e L. Klatt (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2):820–857.

D. H. Klatt (1987). Review of text-to-speech conversion for english. *J. Acoust. Soc. Am.*, 82(3):737–793.

T. Koizumi, S. Taniguchi, e S. Hiromitsu (1987). Two mass models of the vocal cords for natural sounding voicei synthesis. *Journal of the Acoustical Society of America*, 82(4):1179–1192.

T. Koizumi, S. Tanigushi, e S. Hiromitsu (1985). Glottal source-vocal tract interaction. *Journal of the Acoustical Society of America*, 78(5):1541–1547.

R. A. Krakow, P. S. Beddor, L. M. Goldstein, e C. Fowler (1988). Coarticulatory influences on the perceived height of nasal vowels. *J. A. S. A.*, 83(3):1146–1158.

R. A. Krakow e M. K. Huffman (1993). Instrumentaion and techniques for investigationg nasalization and velopharyngeal function in the laboratory: An introduction. In M. K. Huffman e R. A. Krakow (editores), *Nasals, Nasalization, and the Velum*, Phonetics and Phonology (vol. 5), páginas 3–59. Academic Press Inc.

B. Kühnert e F. Nolan (1997). The origin of coarticulation. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München (FIPKM)*, 35:61–75. Também em Hardcastle e Hewlett (1999).
URL `http://www.phonetik.uni-muenchen.de/FIPKM/index.html`

R. Laboissiére, V. Sanguineti, e D. J. Ostry (1999). A model of biomechanics and neural control of the tongue, jaw, hyoid bone and larynx. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 1777–1780. University of California, Berkeley, San Francisco.

A. Lacerda e B. F. Head (1966). Análise de sons nasais e sons nasalizados do Português. *Revista do Laboratório de Fonética Experimental (de Coimbra)*, VI:5–70.

P. Ladefoged, J. Anthony, e C. Riley (1971). Direct measurement of the vocal tract. *Journal of the Acoustical Society of America*, 49(1 (Pt 1)):1971. Apresentado no 80th Meeting da ASA, 4 November 1970.

P. Ladefoged e I. Maddieson (editores) (1995). *The Sounds of the World's Languages*. Blackwell Publishers, Oxford.

U. K. Laine (1999). Modal synthesis and modeling of vowels. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*. Budapest, Hungary.

A. Lalwani (1991). *Flexible Formant Synthesizer: A Tool For Improving Speech Production Quality*. Tese de Doutoramento, University of Florida.

H. Lausberg (1981). *Linguística Românica*. Fundação Calouste Gulbenkian, Lisboa, segunda edição. Tradução do original alemão entitulado: Romanische Sprachwissenschaft.

J. Laver (1980). *The Phonetic Description of Voice Quality*. Cambridge Studies in Linguistics. Cambridge University Press, primeira edição.

J. Laver (1994). *Principles of Phonetics*. Cambridge Textbooks in Linguistics. Cambridge University Press, primeira edição.

S. Lawson e A. Mirzai (1990). *Wave Digital Filters*. Ellis Horwood Limited, Chichester, UK.

K. Lee (1992). *Pitch synchronous analysis/synthesis system using WRLS-VFF Algorithm*. Tese de Doutoramento, University of Florida.

S. E. Levinson e C. E. Schmidt (1983). Adaptive computation of articulatory parameters. *Journal of the Acoustical Society of America*, 74(4):1145–1154.

P. Lieberman e S. E. Blumstein (1988). *Speech Physiology, Speech Perception, and Acoustic Phonetics*. Cambridge Studies in Speech Science. Cambridge University Press.

J. Liljencrants (1985). *Speech Synthesis with a Reflection-Type Line Analog*. Ds dissertation, Dept. of Speech Comm. and Music Acoust., Royal Inst. of Tech., Stockolm, Sweden.

Q. Lin (1990). *Speech production theory and Articulatory Speech Synthesis*. Tese de Doutoramento, Dept. of Speech Comm. & Music Acoustics, Royal Institute of Technology (KTH), Stockolm, Sweden.

Q. Lin (1992). Vocal-tract computaion: How to make it more robust and faster. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4. Also in The Journal of the Acoustical Society of America , vol. 96, no. 4, pp. 2576-2579, 1994.

Q. Lin (1994). A three-channel model for nasals and nasalization. *Journal of the Acoustical Society of America*, 95(5, Pt 2):2922.

Q. Lin (1995). A fast algorithm for computing the vocal-tract impulse response from the transfer function. *IEEE Trans. Speech Audio Proc.*, 3(6):449–457.

B. E. F. Lindblom e J. E. F. Sundberg (1971). Acoustical consequences of lip, tongue, jaw, and larynx movement. *Journal of the Acoustical Society of America*, 50(4 (Pt 2)):1166–1179.

J. Lindqvist-Gauffin e J. Sundberg (1976). Acoustic properties of the nasal tract. *Phonetica*, 33:161–168.

R. Linggard (1985). *Electronic Synthesis of Speech*. Cambridge University Press.

J. I. Louro (1954-1955). Estudo e classificação das vogais. *Boletim de Filologia*, Tomo XV:215–248.

S. Maeda (1982a). A digital simulation method of vocal-tract system. *Speech Communications*, 1:199–229.

S. Maeda (1982b). The role of the sinus cavities in the production of nasal vowels. In *Proc. ICASSP*, páginas Vol. 2, 911–914.

S. Maeda (1990). Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In W. J. Hardcastle e A. Marchal (editores), *Speech Production and Speech Modelling*, NATO Advanced Science Institute Series: D, Volume 55. Kluwer Academic Publishers, Dordrecht. ISBN 0-7923-0746-1.

S. Maeda (1993). Acoustics of vowel nasalization and articulatory shifts in french nasal vowels. In M. K. Huffman e R. A. Krakow (editores), *Nasals, Nasalization, and the Velum*, Phonetics and Phonology (vol. 5), páginas 147–167. Academic Press Inc.

J. J. Mahshie (1993). Use of the electroglottograph in the laboratory and clinic. In M. Stone (editor), *Measuring Speech Production: Abstracts and References*, páginas 6–7. Acoustical Society of America.

M. H. M. Mateus (1975). *Aspectos de Fonologia Portuguesa*. Publicações do Centro de Estudos Filológicos, Lisboa.

M. H. M. Mateus, A. Andrade, M. do Céu Viana, e A. Villalva (1990). *Fonética, Fonologia e Morfologia do Português*. Universidade Aberta, Palácio Ceia, R. Escola Politécnica, 147, Lisboa.

R. S. McGowan (1994). Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 4(1).

P. Mermelstein (1967). Determination of the vocal-tract shape from measured formant frequencies. *Journal of the Acoustical Society of America*, 41(5):1283–1294.

P. Mermelstein (1973). Articulatory model for the study of speech production. *J. Acoust. Soc. Am.*, 53(4):1070–1082.

P. Mermelstein e S. Maeda (1971). Description of tongue and lip movement in a jaw-based coordinate system. *Journal of the Acoustical Society of America*, 49(1, (Pt. 1)):1971. Apresentado no 80th ASA Meeting, November 1970.

N. Metropolis, A. W. Rosembluth, marshall N. Rosembluth, A. H. Teller, e E. Teller (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6).

P. Meyer, J. Schroeter, e M. M. Sondhi (1991). Design and evaluation of optimal cepstral lifters for accessing articulatory codebooks. *IEEE Trans. Ac. Sp. Sig. Proc*, 39(7).

P. Meyer, R. Wilhelms, e H. W. Strube (1989). A quasiarticulatory speech synthesizer for german language running in real time. *Journal of the Acoustical Society of America*, 86(2):523–539.

W. L. Meyerhoff e S. D. Schaefer (1991). Physiology of the nose and paranasal sinuses. In Paparella e Shumrick (1991), capítulo 12, páginas 315–332.

N. Miki, T. Yokoyama, T. Ohtani, S. Masaki, I. Shimada, I. Fujimoto, e Y. Nakamura (1999). A vocal tract model using multi-line equivalent circuits. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*, volume 1, páginas 129–132. Budapest, Hungary.

J. L. Miller, R. D. Kent, e B. S. Atal (editores) (1991). *Papers in Speech Communication: Speech Perception*. Acoustical Society of America. Ver também Kent *et al.* (1991); Atal *et al.* (1991).

Y. Miyanaga (1990). Adaptive identification of ARMA model and model identification. In N. Nagai (editor), *Linear Circuits, Systems and Signal Processing*, capítulo 9, páginas 243–280. Marcel Dekker.

D. C. Montgomery (1991). *Design and Analysis of Experiments*. John Wiley & Sons, terceira edição.

P. M. Morse (1948). *Vibration and Sound*. McGraw Hill Book Co., New York, segunda edição.

P. M. Morse (1991). *Vibration and Sound*. Acoustical Society of America, segunda edição. Reedição de Morse (1948).

P. M. Morse e K. U. Ingard (1968). *Theorectical Acoustics*. McGraw Hill, New York.

N. Nagai (1990). Complex transmission-line circuit and complex wave digital filter. In N. Nagai (editor), *Linear Circuits, Systems and Signal Processing - Advanced Theory and Applications*, capítulo 12, páginas 343–388. Marcel Dekker.

S. Narayanan, A. Alwan, e K. Haky (1995). An articulatory study of fricative consonants using MRI. *Journal of the Acoustical Society of America*, 98(3):1325–1347.

S. Narayanan, A. Alwan, e K. Haky (1997a). Towards articulatory-acoustic models for liquid consonants based on MRI and EPG data - part i: the laterals. *Journal of the Acoustical Society of America*, 101(2):1064–1077.

S. Narayanan, A. Alwan, e K. Haky (1997b). Towards articulatory-acoustic models for liquid consonants based on MRI and EPG data - part ii: the rotics. *Journal of the Acoustical Society of America*, 101(2):1078–1089.

O. Nobiling (1974). As vogais nasais em português. *Littera*, 12:80–109.

J. Ohala e M. Ohala (1993). The phonetics of nasal phonology: Theorems and data. In M. K. Huffman e R. A. Krakow (editores), *Nasals, Nasalization, and the Velum*, Phonetics and Phonology (vol. 5), páginas 225–249. Academic Press Inc.

J. P. Olive, A. Greenwood, e J. Coleman (1993). *Acoustics of American English Speech*. Springer-Verlag.

L. C. Oliveira (1996). *Síntese de Fala a Partir do Texto*. Dissertação de doutoramento, Instituto Superior Técnico, Universidade Técnica de Lisboa, Lisboa.

L. C. Oliveira, M. C. Viana, e I. M. Trancoso (1991). DIXI: Portuguese text-to-speech system. In *Proc. EUROSPEECH*.

S. Ouni e Y. Laprie (1999). Design of hypercube codebooks for the acoustic-to-articulatory inversion respecting the non-linearities of the articulatory-to-acoustic mapping. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*, volume 1, páginas 141–144. Budapest, Hungary.

J. K. Ousterhout (1994). *Tcl and Tk Toolkit*. Addison-Wesley.

J. K. Ousterhout (1998). Scripting: Higher-level programming for the 21st century. *Computer*, 31(3):23–30.

M. M. Paparella e D. A. Shumrick (editores) (1991). *Otolaryngology*, volume I - Basic Sciences and Related Principles. W. B. Saunders Company, Philadelphia, PA, terceira edição.

S. Parkinson (1983). Portuguese nasal vowels as phonological diphthongs. *Lingua*, 61:157–177.

S. Parthasarathy e C. H. Coker (1990). Phoneme-level parametrization of speech using an articulatory model. In *Proc. ICASSP*, páginas 337–340.

S. Parthasarathy e C. H. Coker (1992). On automatic estimation of articulatory parameters in a text-to-speech system. *Computer Speech and Language*, 6:37–75.

J. S. Perkell (1974). *A physiologically-oriented model of tongue activity in speech production*. Phd thesis, MIT, Cambridge, MA.

J. S. Perkell, M. H. Cohen, M. A. Svirsky, M. L. Natthies, I. naki Garabieta, e M. T. T. Jackson (1992). Electromagnetic midsagital articulometer systems for transducing speech articulatory movements. *Journal of the Acoustical Society of America*, 92(6):3078–3096.

J. S. Perkell, M. A. Svirsky, M. L. Matthies, e J. Manzella (1993). Measuring articulatory movements with an electromagnetic midsagittal articulometer (EMMA) system. In M. Stone (editor), *Measuring Speech Production: Abstracts and References*, páginas 23–24. Acoustical Society of America.

P. Perrier e D. J. Ostry (1994). Dynamic modelling and control of speech articulators: Application to vowel reduction. In Keller (1994), capítulo 11, páginas 231–251. ISBN 0 471 94449 1.

N. B. Pinto, D. G. Childers, e A. L. Lalwani (1989). Formant synthesis: Improving production quality. *IEEE Trans. Ac. Sp. Sig. Proc*, 37(12).

Y. Ploysongsang e R. G. London (1991). Physiology of the lungs. In Paparella e Shumrick (1991), capítulo 14, páginas 343–359.

K. R. Popper (1992). *O Realismo e o Objectivo da Ciência*. Nova Enciclopédia. Publicações D. Quixote, Lisboa, segunda edição. Tradução de *Realism and the Aim of Science*, Hutchinson, London, 1956.

P. P. L. Prado (1991). *A Target-Based Articulatory Synthesizer*. Tese de Doutoramento, University of Florida.

P. P. L. Prado, E. H. Shiva, e D. G. Childers (1992). Optimization of acoustic-to-articulatory mapping. In *Proc. ICASSP*, volume 2, páginas 33–36. San Francisco, USA.

W. H. Press, B. P. Flannery, S. A. Teukolsky, e W. T. Vetterling (1992). *Numerical Recipes in C*. Cambridge University Press, segunda edição.

S. R. Quackenbush, T. P. B. III, e M. A. Clements (1988). *Objective Measures of Speech Quality*. Signal Processing Series. Prentice Hall.

L. R. Rabiner e R. W. Schafer (1978). *Digital processing of speech signals*. Prentice-Hall, Englewood Cliffs,NJ.

M. G. Rahim e C. C. Goodyear (1990). Estimation of vocal tract filter parameters using a neural net. *Speech Communication*, 9(1):49–55.

M. G. Rahim, C. C. Goodyear, W. B. Kleijn, J. Schroeter, e M. M. Sondhi (1993). On the use of neural networks in articulatory speech synthesis. *J. Acoust. Soc. Am.*, 93(2):1109–1121.

P. A. Regalia (1995). *Adaptive IIR Filtering in Signal Processing and Control*. Marcel Dekker Inc.

G. Richard, M. Liu, D. Sinder, H. Duncan, Q. Lin, J. Flanagan, S. Levinson, D. Davis, e S. Slimon (1995). Numerical simulations of fluid flow in the vocal tract. In *Proc. Eurospeech*. Madrid, Spain.

K. Richmond (1999). estimating velum height from acoustics during continuous speech. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*, volume 1, páginas 149–152. Budapest, Hungary.

E. L. Riegelsberger (1995). Acoustic-to-articulatory mapping of fricatives. Poster presented at the 129th Meeting ASA, 3 June 1995.

E. L. Riegelsberger (1997). *The Acoustic-to-Articulatory Mapping of Voiced and Fricated Speech*. Tese de Doutoramento, The Ohio State University.

A. E. Rosemberg (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49(2, Part 2):583–590.

G. Rosen (1958). Dynamic analog speech synthesizer. *Journal of the Acoustical Society of America*, 30(3):201–209.

R. Rosenthal e R. L. Rosnow (1991). *Essentials of Behavioral Research: Methods and Data Analysis*. McGraw-Hill Series in Psychology. McGraw-Hill, segunda edição. ISBN 0-07-053929-4.

S. Rossato e G. Feng (1999). Nasal vowel transfer functions: A statistical comparison between the nasal area and the area ratio. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 2101–2104. University of California, Berkeley, San Francisco.

S. Rossato, G. Feng, e R. Laboissière (1998). Recovering gestures from speech signals: A preliminary study for nasal vowels. In *Proc. ICSLP '98*. Sydney, Australia. Paper no. 540.

T. D. Rossing e N. H. Fletcher (1994). *Principles of Vibration and Sound*. Springer-Verlag.

E. H. Rothauser (1969). IEEE recommended practice for speech quality measurements. *IEEE Trans. Audio and Electr.*, AU-17(3):225–246. Standards Publications No. 297.

M. Rothemberg (1981). An interactive model for voice source. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4:1–17. Paper apresentado no Vocal Fold Physiology Conference, Univ. Wisconsin, 1981.

P. Rubin, T. Baer, e P. Mermelstein (1981). An articulatory synthesizer for perceptual research. *Journal of the Acoustical Society of America*, 70(2):321–328.

A. H. Sá Pinto (2000). *Detecção Automática da Frequência Fundamental de Sinais Musicais Produzidos por uma Fonte Vocal*. Tese de mestrado, Universidade de Aveiro.

L. Sachs (1984). *Applied Statistics: A Handbook of Techniques*. Springer Series in Statistics. Springer-Verlag, New York, segunda edição.

E. Saltzman e K. Munhall (1989). A dynamic approach to gestural patterning in speech production. *Ecological Psychology*, 1/3:333–382.

R. Sampson (editor) (1999). *Nasal Vowel Evolution in Romance*. Oxford University Press. ISBN 0-19823848-7.

G. P. Scavone (1997). *An Acoustic Analysis of Single-Reed Woodwind Instruments with an Emphasis on Design and Performance Issues and Digital Waveguide Modeling Techniques*. Phd thesis, Stanford University.

P. Schönle, K. Gräbe, P. Wenig, J. Hölme, J. Schrader, e B. Conrad (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, 31:26–35.

M. R. Schroeder (1967). Determination of the geometry of the human vocal tract by acoustic measurements. *Journal of the Acoustical Society of America*, 41(4, Pt. 2):1002–1010.

M. R. Schroeder (1999). *Computer Speech: Recognition, Compression, Synthesis*. Springer Series in Information Sciences, 35. Springer Verlag. ISBN 3540 643 974.

M. R. Schroeder, B. S. Atal, e J. L. Hall (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66(6):1647–1652.

J. Schroeter, P. Meyer, e S. Parthasarathy (1990). Evaluation of improved articulatory codebooks and codebook access distance measures. In *Proc. ICASSP*, volume 1, páginas 393–396. Albuquerque, USA.

J. Schroeter e M. M. Sondhi (1992). Speech coding based on physiological models of speech production. In S. Furui e M. M. Sondhi (editores), *Advances in Speech Siganl Processing*, páginas 231–268. Marcel Dekker,Inc., New York.

J. Schroeter e M. M. Sondhi (1994). Techniques for estimating vocal-tract shapes from speech signal. *IEEE Transactions on Speech and Audio Processing*, 2(1, Part II):133–150.

W. A. Schweigert (1994). *Research Methods & Statistics for Psychology*. Brooks/Cole Publishing Company, Pacific Grove, California.

C. Scully (1987). Linguistic units and units of speech production. *Speech Communication*, 6(2):77–142.

C. H. Shadle, M. Mohammad, J. N. Carter, e P. J. B. Jackson (1999). Multi-planar dynamic magnetic resonance imaging: New tools for speech research. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 623–626. University of California, Berkeley, San Francisco.

K. Shirai e T. Kobayashi (1986). Estimating articulatory motion from speech wave. *Speech Communication*, 5:159–170.

A. R. T. V. Silva (1995). *Análise de Fonemas Nasais da Língua Portuguesa*. Tese de mestrado, Universidade de Aveiro.

C. Silva, S. Chennoukh, e I. Trancoso (1999). On improving the decision algorithm for articulatory codebook search. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*, volume 1, páginas 153–156. Budapest, Hungary.

D. Sinder, R. G, H. D. adn J. Flanagan, M. Krane, S. Levinson, S. Slimon, e D. Davis (1997). Flow visualization in stylized vocal tracts. In *Proc. ASVA*. Tokyo, Japan.

D. J. Sinder (1999). *Speech Synthesis using an aeroacoustic fricative model*. Tese de Doutoramento, Rutgers, The State University of New Jersey.

S. Slimon, D. Davis, S. Levinson, M. Krane, G. Richard, D. Sinder, H. Duncan, Q. Lin, e J. Flanagan (1996). Numerical simulations of fluid flow in the vocal tract. In *17th American Institute of Aeronautics and Astronautics (AIAA) Aeroacoustics Conference*. State College, PA, USA.

M. M. Sondhi (1974). Model for wave propagation in a lossy vocal tract. *Journal of the Acoustical Society of America*, 55(5):1070–1075.

M. M. Sondhi (1979). Estimation of vocal-tract areas: The need for acoustical measurements. *IEEE Transactions on Acoustic, Speech, and Signal Processing*, ASSP-27(3):268–273.

M. M. Sondhi (1986). Resonances of a bent vocal tract. *Journal of the Acoustical Society of America*, 79(4):1113–1116.

M. M. Sondhi e J. R. Resnick (1983). The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis. *Journal of the Acoustical Society of America*, 73(3):985–1002.

M. M. Sondhi e J. Schroeter (1987). A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-35(7):955–967.

A. Soquet, V. Lecuit, T. Metens, B. Nazarian, e D. Demolin (1998). Segmentation of the airway from the surrounding tissues on magnetic resonance images: A comparative study. In *Proc. ICSLP '98*. Sydney, Australia.

V. N. Sorokin (1992). Determination of vocal tract shape for vowels. *Speech Communication*, 11(1):71–85.

V. N. Sorokin (1994). Inverse problem for fricatives. *Speech Communication*, 14(3):249–262.

R. Sproat (editor) (1998). *Multilingual Text-to-Speech Synthesis: the Bell Labs Approach*. Kluwer.

K. N. Stevens (1971). Airflow and turbulence noise for fricative and stop consonants: Static considerations. *Journal of the Acoustical Society of America*, 50(4, Part 2):1180–1192.

K. N. Stevens (1998). *Acoustic Phonetics*. Current Studies in Linguistics. MIT Press. ISBN 0-262-19404-X.

K. N. Stevens, A. Andrade, e M. C. Viana (1987). Perception of vowel nasalization in VC contexts: A cross-language study. *Journal of the Acoustical Society of America*, 82:S119.

K. N. Stevens, G. Fant, e S. Hawkins (1985). Some acoustical and perceptual correlates of nasal vowels. In *Festschrift für Ilse Lehiste*, páginas 241–254. Dordrecht, Holland.

K. N. Stevens e A. S. House (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America*, 27:484–493. Reimpresso em Stevens e House (1955).

K. N. Stevens, S. Kasowski, e C. G. Fant (1953). An electrical analog of the vocal tract. *Journal of the Acoustical Society of America*, 25:734–742.

J. Q. Steward (1922). An electrical analog of the vocal organs. *Nature*, 110:311–312.

B. H. Story (1995). *Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract*. Tese de Doutoramento, University of Iowa.

B. H. Story, I. R. Titze, e E. A. Hoffman (1995). Vocal tract shapes and area functions from magnetic resonance imaging (MRI). *Journal of the Acoustical Society of America*, 98(5, Pt. 2):2930. Abstract.

Strange (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America*, 85:2081–2087.

P. Strevens (1954). Some observations on the phonetics and pronounciation of modern Portuguese. *Revista do Laboratório de Fonética Experimental (de Coimbra)*, II:5–29.

H. W. Strube (1982). Time-varying wave digital filters and vocal-tract models. In *Proc. ICASSP*, páginas 923–926.

J. Sundberg (1997). Human singing voice. In M. J. Crocker (editor), *Encyclopedia of Acoustics*, volume Four, capítulo 139, páginas 1687–1695. John Wiley & Sons, Inc., New York. ISBN 0-471-80465-7.

J. Sundberg, C. Johansson, H. Wilbrand, e C. Ytterbergh (1987). From sagittal distance to area - A study of transverse, vocal tract cross-sectional area. *Phonetica*, 44:76–90. Also in STL-QPSR 4/1983.

H. Suzuki, T. Nakai, e H. Sakakibara (1995). 3-D FEM analysis of sound propagation in the nasal tract. In *EuroSpeech*, páginas 1301–1304.

S. Takeuchi, H. Kasuya, e K. Kido (1975). On the acoustic correlate of nasality. *Journal Acoustical Society of Japan*, 31(5):298–309.

P. Taylor, R. Caley, A. W. Black, e S. King (1999). *Edinburgh Speech Tools Library*. University of Edinburgh. System Documentation, Edition 1.2, for Speech Tools Library Version 1.2.0.
URL `http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/`

A. Teixeira e F. Vaz (1999). Fonte glotal. Relatório técnico do projecto PRAXIS/P/PLP/11222/1998 SAP 1/1999, IEETA, Universidade de Aveiro.

A. Teixeira e F. Vaz (2000a). Articulatory synthesis: The use of biological models in production of high quality speech. In *5 Congresso Português de Engenharia Biomédica (BioEng'2000)*. Coimbra, Portugal.

A. Teixeira e F. Vaz (2000b). Síntese articulatória de sons nasais do português. In *V Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada (PROPOR)*. Atibaia, São Paulo, Brasil. Submetido.

A. Teixeira, F. Vaz, e J. C. Príncipe (1997a). Síntese de Voz em Telecomunicações. In *I Conferência Nacional de Telecomunicações*. Instituto de Telecomunicações, Aveiro.

A. Teixeira, F. Vaz, e J. C. Príncipe (1997b). A Software Tool to Study Portuguese Vowels. In G. Kokkinakis, N. Fakotakis, e E. Dermatas (editores), *Proceedings Eurospeech'97*, volume 5, páginas 2543–2546. Rhodes, Greece.

A. Teixeira, F. Vaz, e J. C. Príncipe (1998a). A comprehensive nasal model for a frequency domain articulatory synthesizer. In *10th Portuguese Conference on Pattern Recognition (RecPad'98)*. IST, Lisbon.

A. Teixeira, F. Vaz, e J. C. Príncipe (1998b). Some studies of european portuguese nasal vowels using an articulatory synthesizer. In *5th IEEE International Conference on Electronics, Circuits and Systems*. Lisbon.

A. Teixeira, F. Vaz, e J. C. Príncipe (1999a). Effects of source-tract interaction in perception of nasality. In G. Gordos e G. Németh (editores), *Proceedings Eurospeech'99*, volume 1, páginas 161–164. Budapest, Hungary.

A. Teixeira, F. Vaz, e J. C. Príncipe (1999b). Influence of dynamics in the perceived naturalness of portuguese nasal vowels. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*. University of California, Berkeley, San Francisco.

A. Teixeira, F. Vaz, e J. C. Príncipe (2000). Nasal vowels following a nasal consonant. In *5th Seminar on Speech Production: Models and Data, CREST Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*. Kloster Seeon, Bavaria, Germany.

A. Teixeira, F. Vaz, J. C. Príncipe, e D. G. Childers (1997c). Articulatory Synthesis of Portuguese Vocoids. In H. Araújo e L. V. de Sá (editores), *9th Portuguese Conference on Pattern Recognition (RecPad'97)*, páginas 219–224. Coimbra.

P. Teyssier (1980). *Histoire de La Langue Portuguaise*. col. «Que Sais-je?». PUF, Paris.

G. Thimm e J. Luettin (1999). Extraction of articulators in X-Ray image sequences. In G. Gordos e G. Németh (editores), *Proc. Eurospeech*, volume 1, páginas 157–160. Budapest, Hungary.

T. J. Thomas (1986). A finite element model of fluid flow in the vocal tract. *Computer Speech and Language*, 1:131–151.

Y.-T. Ting (1994). *Adaptive Estimation of Time-varying Signal parameters with applications to speech*. Tese de Doutoramento, University of Florida.

I. R. Titze (1989). A four-parameter model of the glottis and vocal fold contact area. *Speech Communication*, 8(3):191–201.

I. R. Titze (1994). *Principles of Voice Production*. Prentice Hall, Englewood Cliffs.

I. R. Titze e B. H. Story (1997). Acoustic interactions of the voice source with the lower vocal tract. *Journal of the Acoustical Society of America*, 101(4):2234–2243.

R. L. Trigo (1993). The inherent structure of nasal segments. In M. K. Huffman e R. A. Krakow (editores), *Nasals, Nasalization, and the Velum*, Phonetics and Phonology (vol. 5), páginas 369–400. Academic Press Inc.

J.-P. Tubach (1996). Présentation générale. In H. Méloni (editor), *Fondements et perspectives en traitment automatique de la parole*. Universités Francophones.

V. Välimäki (1995). *Discrete-Time Modeling of Acoustic Tubes Using Fractional Delay Filters*. Phd thesis, Laboratory of Acoustics and Audio Signal Processing, Faculty of Electrical Engineering, Helsinky University of Technology, Espoo, Finland.

V. Välimäki, M. Karjalainen, e T. Kuisma (1994). Articulatory speech synthesis based on fractional delay waveguide filters. In *Proc. ICASSP*, páginas 585–588.

J. van den Berg (1958). Myoelastic-aerodynamic theory of voice production. *Journal of Speech and Hearing Research*, 1(3):227–244. Reimpresso em Kent *et al.* (1991).

J. van den Berg (1960). An electrical analog of the trachea, lungs and tissues. *Acta Physiol. Pharmacol. Neerl.*, 9:361–385.

J. van den Berg, J. T. Zantema, e P. Doornenbal Jr. (1957). On the air resistance and the Bernoulli Effect of the human larynx. *Journal of the Acoustical Society of America*, May(5):626–631.

R. van Praag (1997). *Formulation variationelle du champ sonore dans une arborescence de conduits non uniformes. Application à l'appareil vocal.* Dissertation docteur en sciences, Institut des Langues Viavantes et de Phonétique, Faculté des Sciences, Université Libre de Bruxelles.

J. P. H. van Santen, J. Olive, J. Hirschberg, e R. Sproat (editores) (1996). *Speech Synthesis*. Springer-Verlag.

F. Vaz e A. Teixeira (1998). Articulatory Synthesis of Portuguese, projecto PRAXIS/P/PLP/11222/1998, Fundação para a Ciência e Tecnologia.
URL `http://www.inesca.pt/~ajst/`

A. R. G. Viana (1883). Essai de phonétique et de phonologie de la langue portuguaise d'aprés le dialecte actuel de lisbonne. Paris.

A. R. G. Viana (1892). Exposição da pronúncia normal portuguesa para uso de nacionais e estrangeiros. Lisboa.

H. Wakita (1973). Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms. *IEEE Trans. Audio and Electr.*, AU-21(5).

H. Wakita (1979). Estimation of vocal-tract shapes from acoustical analysis of the speech wave: The state of the art. *IEEE Trans. Acoustics, Speech, and Signal Processing*, ASSP-27(3):281–285.

H. Wakita e G. Fant (1978). Toward a better vocal tract model. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 1:9–29.

H. Walter (1996). *A Aventura das Linguas do Ocidente: A sua origem, a sua história, a sua Geografia*. Terramar, Lisboa.

D. W. Warren (1967). Nasal emission of air and velopharyngeal function. *The Cleft Palate Journal*, 4:148–156.

A. G. Webster (1919). Acoustical impedance, and the theory of horns and the Phonograph. *Proceeding of the National Academy of Sciences of the United States of America*, 5:275–282.

B. Welch (1995). *Practical Programming in Tcl/Tk*. Prentice-Hall.

R. Wilhelms-Tricarico (1995). Physiological modeling of speech production: Methods for modeling soft-tissue articulators. *Journal of the Acoustical Society of America*, 97(5):3085–3098.

C.-F. Wong (1991). *The Incorporation of Glottal source-vocal tract interaction effects to improve the naturalness of synthetic speech*. Tese de Doutoramento, University of Florida.

A. A. Wrench (1999). An investigation of sagittal velar movement and its correlation with lip, tongue and jaw movement. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 435–438. University of California, Berkeley, San Francisco.

G. T. H. Wright e F. J. Owens (1993). An optimized multirate sampling technique for the dynamic variation of vocal tract length in the Kelly-Lochbaum speech synthesis model. *IEEE Trans. Speech Audio Proc.*, 1(1):109–113.

C.-M. Wu e R. Wilhelms-Tricarico (1994). Tongue structural model: Integrating MRI data and anatomical structure into a finite element model of the tongue. *Journal of the Acoustical Society of America*, 96(5, Pt. 2):3341–3342. Abstract.

Q. Xue, Y. H. Hu, e P. Milenkovic (1990). Analyses of the hidden units of the multi-layer perceptron and its application in acoustic-to-articulatory mapping. In *Proc. ICASSP*, páginas 869–872.

B. Yang (1999). Measurement and synthesis of the vocal tract of Korean monophthongs by MRI. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 2005–2008. University of California, Berkeley, San Francisco.

H. Yehia e F. Itakura (1994). Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal fourier analysis. In *Proc. ICASSP*, volume I, páginas 477–480.

Z. Yu (1993). A method to determine the area function of speech based on perturbation theory. *Speech Transmission Laboratory, Quarterly Progress ans Status Report*, STL-QPSR 4:77–95.

W. R. Zemlin (1988). *Speech and Hearing Science - Anatomy and Physiology*. Prentice Hall, terceira edição.

A. Zierdt, P. Hoole, e H. G. Tillmann (1999). Development of a system for three-dimensional fleshpoint measurement of speech movements. In J. J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, e A. C. Bailey (editores), *Proceedings of the XIVth International Congress of Phonetic Sciences*, páginas 73–75. University of California, Berkeley, San Francisco.

F. Zussa (1995). A new design for articulatory parametrization of speech: Application to low-bit rate coding and recognition. Industrial thesis report, CAIP, Rutgers University.

F. Zussa, Q. Lin, G. Richard, D. Sinder, e J. Flanagan (1995). Open-loop acoustic-to-articulatory mapping. *Journal of the Acoustical Society of America*, 98(5, Pt. 2):2931. Abstract.