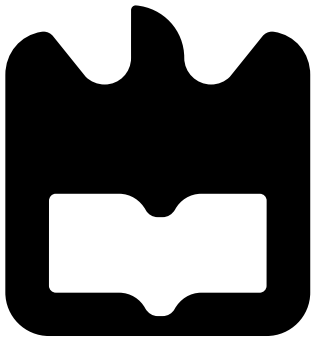
Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2019

Fábio Daniel Rodrigues Barros

Contributos para uma Abordagem Computacional ao Estudo do Desempenho de Oradores



Universidade de Aveiro Departamento de Electrónica, Telecomunicações e Informática,
2019

Fábio Daniel Rodrigues Barros

Contributos para uma Abordagem Computacional ao Estudo do Desempenho de Oradores

Dissertação apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Engenharia Informática, realizada sob a orientação científica de António José Ribeiro Neves, Professor Auxiliar do Departamento de Eletrónica, Telecomunicações e Informática da Universidade de Aveiro e coorientação de Samuel de Sousa Silva, Investigador no Instituto de Engenharia Eletrónica e Telemática de Aveiro da Universidade de Aveiro.

o júri / the jury

presidente / president	**Professor Doutor José Maria Amaral Fernandes** Professor Auxiliar da Universidade de Aveiro
vogais / examiners committee	**Professora Doutora Liliana da Silva Ferreira** Professora Catedrática Convidada do Departamento de Engenharia Informática da Faculdade de Engenharia da Universidade do Porto
	Doutor Samuel de Sousa Silva Investigador no Instituto de Engenharia Eletrónica e Telemática de Aveiro da Universidade de Aveiro (co-orientador)

agradecimentos /

acknowledgements

Ao professor António Neves, pela irrepreensível orientação e acompanhamento que mostrou ao longo desta dissertação, mostrando sempre vontade de ajudar e com quem tive a oportunidade de adquirir novos conhecimentos.

Ao Samuel Silva, também pela irrepreensível orientação e acompanhamento, que proporcionou ao longo desta dissertação, manifestando sempre toda a disponibilidade para ajudar a resolver a mais diversas adversidades. Foi um dos elementos com quem tive a oportunidade de adquirir e aprofundar inúmeros conhecimentos a diversos níveis e trocar longas conversas.

À professora Sandra Soares, que de forma direta ou indireta, contribuiu como o seu conhecimento e ajuda numa área na qual era totalmente desconhecida para mim.

Aos meus pais e irmãos, pelo exemplo de dedicação e esforço que foram, por todo o apoio incondicional e incentivo. Foram excecionais, sem eles nada nunca teria conseguido chegar a este ponto.

À minha família, que desde os primeiros dias estiveram sempre presentes e contribuíram de diferentes formas, mas todas elas importantes, para que me torna-se a pessoa que sou hoje.

À Rosa, por todos os momentos bons e menos bons que teve de tolerar, mostrando sempre um apoio sem fim. Sem este apoio teria sido, todo o percurso, muito mais difícil.

Aos amigos de sempre, Luísa, Kelly, Sandrina, Ilda, João Carlos, Miguel, Hugo, Amorim, Tiago e Cristóvão que estiveram, de uma forma ou de outra, presentes em todo o meu percurso, sendo um exemplo de companheirismo, força e apoio em todos os momentos.

Aos novos amigos que Aveiro me presenteou, Bárbara, Elisabete, Cláudia, Raquel, Mimi, Bruno, Pintor, José, Carlos, Tiago que foram fundamentais durante todo o meu percurso universitário. A forma como me receberam, acompanharam e suportaram, aos mais diversos níveis, foi fundamental tornando tudo mais simples.

Ao Sérgio, o companheiro de todas as horas, com quem tive a oportunidade de partilhar diversos conhecimentos, ideologias, mas também uma amizade sincera.

Ao Ricardo e ao Daniel, com que eu tive a oportunidade de trocar longas conversas e inúmeros conhecimentos na área da investigação, mostrando sempre prontidão a ajudar.

Palavras-Chave Comunicação verbal e não verbal; Postura; Expressões Faciais; Voz

Resumo

A capacidade de comunicar bem em público é uma competência muito importante a nível profissional, académico e pessoal. A caracterização do que é um bom comunicador tem sido amplamente abordada em diferentes áreas, particularmente na Educação, no sentido de, por exemplo, contribuir para melhorar a prestação de professores. Neste contexto, a literatura sustenta que a competência em comunicar não se define apenas tendo em consideração a componente verbal dado que muitos dos aspetos não verbais fornecem informação redundante e/ou complementar e são, eles próprios, parte da mensagem.

Cada um de nós consegue avaliar a qualidade da prestação de um orador e, em contexto formativo, por exemplo, esta é usualmente realizada por um especialista que aprecia a prestação com base num conjunto de critérios resultantes do conhecimento da área e da sua experiência pessoal. Ainda assim, pouco de sabe, de forma objetiva, sobre que aspetos – e com que importância – definem a qualidade da prestação. O avanço do conhecimento neste âmbito permitiria compreender mais sobre o fenómeno da comunicação e sustentaria trabalho no sentido de propor métodos automáticos de avaliação da qualidade da comunicação em público.

Este trabalho, pretende ajudar a estudar, de forma mais sistemática e objetiva, as características definidoras de um bom comunicador, contribuindo com uma abordagem computacional para caracterizar, a partir de vídeos de oradores, os diferentes elementos envolvidos na comunicação, tais como os movimentos dos braços, a postura corporal, a expressão facial e a voz.

Nesse sentido, começa-se por realizar uma contextualização do problema, seguindo-se um levantamento do estado de arte relativamente à caracterização da atividade presente em diferentes canais com um papel na comunicação. Com base nesse levantamento, são selecionados e aplicados métodos computacionais para extração dessas características.

Tendo em consideração que o estudo da comunicação em público é relevante para uma comunidade científica alargada foi considerada essencial uma anotação dos dados extraídos com informação de eventos significativos (por exemplo, "elevou os braços", "sorriu", "silêncio") que torna mais interpretáveis – atribuindo-lhe significado – os dados extraídos anteriormente.

Finalmente, e no sentido de ilustrar, de forma simples, como o trabalho desenvolvido abre novas perspetivas – e lança novas questões – no estudo da comunicação em público, são apresentados alguns exemplos ilustrativos de métodos computacionais que podem suportar o estudo exploratório da informação agora tornada disponível.

Keywords Verbal and non-verbal communication; Posture; Facial expressions; Voice

Abstract

The ability to communicate in public is a relevant competence at the professional, academic, and personal levels. The characterization of what is a good communicator has been addressed by several areas, particularly in Education, in the sense of, for example, contributing to improve the performance of teachers. In this context, the literature suggests that the ability to communicate is not only defined by the verbal component, but also by a set of non-verbal components, since many non-verbal aspects provide redundant and/or complementary information, sometimes being the message itself.

Each one of us can evaluate the performance of a speaker. In a formative context, for example, this evaluation is usually accomplished by a specialist who evaluates a performance based on a set of criteria resulting from knowledge regarding the phenomenon and personal experience. Yet, objectively, little is knwow about what aspects - and how important they are – define the quality of a presentation. The advancement of knowledge in this context would enable a greater understanding about the communication phenomenon and could support the proposal of automatic forms of evaluation of the quality of communication in public.

The goal of this project is to support the study of the defining characteristics of good communicator in a more systematic and objective form. This contribution will be performed with a computational approach to characterize the different elements that are involved in comunication, such as the movement of the arms, body posture, facial expressions and voice.

To this end, it begins with the contextualization of the problem, followed by a survey of the state of the art relating to the characterization of the activity for different channels deemed relevant for communication. Based on this survey, computational methods are selected and applied to extract these characteristics.

Considering that the study of public communication is relevant to an extended scientific community, an annotation of the extracted data with events of activities deemed relevant (eg, "raised arms", "smile", "silence") was performed, adding to the interpretability of the extracted data.

In order to illustrate how the work carried out opens new perspectives - and raises new questions - in the study of public communication, some illustrative examples of computational methods that can support the exploratory study of the information now made available are presented.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AUs	Action Units
FACS	Facial Action Coding System
HNR	Harmonics-to-Noise Ratio
JSON	JavaScript Object Notation
KNN	k-Nearest Neighbors
MFCCs	Mel-frequency cepstral coefficients
RGB	Red-Green-Blue
SVM	Support-vector machine
VAD	Voice Activity Detection
ZCR	Zero-crossing Rate

Capítulo 1

Introdução

Neste capítulo é feita uma introdução à temática envolvente abordada na presente dissertação. Nesse sentido, será apresentada a contextualização do tema e a respetiva motivação. Para completar, serão ainda identificados os principiais desafios, os objetivos que se pretendem atingir e os contributos que este estudo traz para a comunidade. Por último, é explicada a organização e a estrutura do documento.

1.1 Contexto

A comunicação é inerente à vida do ser humano, e é através dela que conseguimos interagir uns com os outros, trocar ideias e experiências.

Hoje em dia, e cada vez mais, a capacidade de comunicar adequadamente em público é uma competência muito importante a nível profissional, académico e até mesmo pessoal. Em apresentações ao público é necessário que exista uma boa prestação por parte do orador para que as ideias expressas por ele sejam compreendidas e aceites. No entanto, a estes momentos estão associados diferentes fatores que podem influenciar a prestação dos oradores e a forma com a sua mensagem é recebida.

A comunicação entre humanos não se restringe à componente verbal e muitos dos aspetos não verbais fornecem informação redundante e/ou complementar. Posto isto, afirma-se que a comunicação é multimodal. Alguma literatura propõe que através do movimento do corpo, gestos, expressões faciais e entoações da voz o público identifica um conjunto de informações socialmente relevantes, como atribuições de dominância, confiabilidade, competência e outros

traços de personalidade.

1.2 Motivação

No processo de comunicação, existe um conjunto de boas práticas que podem ser adotadas por um orador, nomeadamente a maneira como este se move, fala, gesticula, encara a audiência e até na forma como estes diferentes aspetos se articulam entre si. A maior parte dos métodos utilizados na avaliação de oradores é baseado no *feedback* recebido por especialistas na área. No entanto, estes métodos são subjetivos e de difícil entendimento uma vez que não só derivam de conhecimentos gerais sobre a comunicação em público, mas também da própria experiência.

Neste contexto, seria interessante estudar o desempenho de um orador de uma maneira mais sistemática e controlada para melhorar o conhecimento sobre o fenómeno e seja possível obter medidas mais objetivas no que diz respeito a competências de comunicação, que poderiam ser aproveitadas, por exemplo, recorrendo a métodos automatizados ou semi-automatizados de avaliação utilizando sistemas computacionais.

Com uma abordagem computacional a esta temática é relevante para diversas áreas de aplicação que podem tirar partido do treino de competências de comunicação como é o caso do ensino.

1.3 Desafios

Um dos aspetos mais desafiantes neste estudo prende-se com a falta de conhecimento sobre que aspetos da atuação do orador estão a influenciar a forma como a mensagem é percebida e apreendida. A fim de compreender quais os aspetos que influenciam a maneira como um orador é avaliado pelo público é necessário que exista um conhecimento prévio e aprofundado das formas (verbais e não verbais) utilizadas para veicular a mensagem, e qual o seu impacto no público, tais como, a postura adotada, expressões faciais e voz.

Ainda que os métodos automáticos de classificação aplicados sobre um conjunto de características indiferenciadas pudessem parecer, à partida, uma solução viável per si, neste contexto, a sua aplicação, ainda que podendo traduzir-se em sistemas de medição automática de níveis de prestação em público, torna-se pouco interessante se não contribuírem para uma perceção um pouco melhor do fenómeno em questão. Assim, numa primeira fase, para que

se possa estudar o fenómeno da comunicação, os métodos de avaliação automáticos devem basear-se, tanto quanto possível, em características que revelem significado para os seres humanos. O grande desafio passa por, numa primeira fase, construir um conjunto de métodos que permitam a avaliação do desempenho de oradores de forma não cega. Isto é, pretende-se entender o fenómeno: Este orador é mau, **mas porquê?**

1.4 Objetivos

Falar em público é uma temática ainda pouco explorada e não existe um conhecimento objetivo e uniforme de quais os aspetos que influenciam a prestação de um orador. No sentido de contribuir para um melhor conhecimento, nesta área, o grande objetivo desta dissertação é contribuir para tornar o estudo da comunicação em público mais objetiva, concretamente através de:

- Selecionar, com base na literatura, quais os aspetos mais relevantes na comunicação humano-humano (verbal e não verbal), de um modo particular aqueles que têm um impacto relevante na comunicação em público.
- Propor um conjunto de métodos que descrevam as ações/conteúdos presentes nos diferentes canais de comunicação identificados.
- Considerar os contributos, no contexto dos objetivos acima referidos, para complementar/anotar uma base de dados audiovisual existente focada no estudo do desempenho de oradores.
- Testar de que forma os diferentes dados e informações anotadas podem suportar a proposta de sistemas (semi-) automáticos para avaliação do desempenho de um orador.

1.5 Contribuições

Como já referido, a forma como a temática da comunicação em público tem sido estudada carece de métodos mais objetivos de análise. Alinhado com os objetivos propostos, este trabalho contribui com um conjunto de métodos que permitem descrever, de forma sistemática e objetiva, várias dimensões da comunicação humano-humano apontadas pela literatura como caracterizadoras do desempenho de um orador.

Outra contribuição deste trabalho é o de disponibilizar, à comunidade, uma base de dados audiovisual de comunicação espontânea enriquecida com um conjunto de características descritoras e anotações de atividades consideradas relevantes para cada um dos canais de comunicação analisados.

Por fim, esta dissertação contribui, ainda, com alguns exemplos ilustrativos de métodos computacionais que podem ser aplicados aos diferentes conjuntos de informação, extraídos de cada canal de comunicação, de forma a permitir uma exploração mais objetiva da influência destes para a caracterização do desempenho de um orador. Estes, lançando já algumas questões sobre futuras direções da investigação, podem servir de base para trabalhos futuros de análise.

1.6 Organização do Documento

Este documento está organizado em 6 Capítulos. No Capítulo 2, é apresentado o levantamento do estado da arte relativamente a modos de comunicação humano-humano, mais especificamente as características por estes apresentadas e a identificação das bibliotecas de software existentes para a extração de características presentes na comunicação verbal e não verbal. São apresentados também, alguns datasets diferentes e com anotações e alguns métodos automatizados para a avaliação dos oradores. De seguida, no Capítulo, 3 são expostos e explicados métodos para a extração de características descritoras dos diferentes canais de comunicação. Estes métodos são posteriormente complementados com um conjunto de métodos para anotação de atividades relevantes durante a comunicação, tal como descrito no Capítulo 4. Para demonstrar como os diferentes métodos propostos podem ser aplicados, num caso concreto, o Capítulo 5 ilustra alguns resultados da sua aplicação a uma base de dados de vídeos de oradores. São ainda apresentados, no Capítulo 6 alguns exemplos de aplicação que podem ser usados através do uso de todas as características extraídas na comunicação humano-humano. Por fim, no Capítulo 7, é feita uma conclusão sobre os resultados obtidos, incluindo sugestões de trabalho futuro.

Capítulo 2

Estado da Arte

Neste capítulo será apresentado o estado da arte relativamente à comunicação humano-humano e as principais características presentes nos canais de comunicação para o aumento do conhecimento na dinâmica de comunicação em público. De seguida, será apresentado um conjunto de métodos computacionais para a extração dessas características. São expostos, também, alguns *datasets* que podem ser usados para avaliação de resultados. São ainda apresentados alguns métodos computacionais para uma exploração posterior dos diferentes canais de comunicação. O capítulo termina com uma breve discussão sobre os diferentes desafios que ainda subsistem e de como eles devem orientar a investigação, na área.do.

2.1 Comunicação Humano-Humano

A comunicação humano-humano não é feita apenas por palavras. Embora a comunicação verbal seja o principal meio de comunicação entre os seres humanos, a comunicação não-verbal (e.g., expressões faciais, gestos e postura corporal) desempenha um papel muito importante na comunicação [1] [2]. As diferentes formas de comunicação afetam a formação de impressões por parte do público [3]. Dessa forma, esta secção aborda um conjunto de características presentes na comunicação verbal e não verbal julgadas relevantes no âmbito deste trabalho.

2.1.1 Postura Corporal

O movimento do corpo humano é uma forma de comunicação não-verbal presente na vida quotidiana. É possível reconhecer não só emoções mas também traços de personalidade e

fazer julgamentos sociais nos movimentos do corpo, gestos e até na postura adotada durante fenómenos de comunicação [4]. Segundo a literatura relacionada com a expansividade das posturas (e.g. [5]), os seres humanos parecem usar posturas expansivas e abertas (tornar-se maior e ocupar mais espaço) para projetar sinais de poder, confiança e assertividade. Por outro lado, posturas contrativas e fechadas (minimização do espaço ocupado e corpo encolhido) projetam sinais de impotência e baixa confiança. A literatura refere ainda que, durante a comunicação, os seres humanos utilizam gestos amplos e posturas corporais expansivas para projetar domínio [4].

A Figura 2.1 retrata como a postura corporal pode ter um impacto no julgamento das capacidades de comunicação de um orador: inseguro à esquerda e confiante à direita [6].

Figura 2.1: Diferença na linguagem corporal: inseguro à esquerda e confiante à direita [6].

2.1.2 Gestos

Os seres humanos geralmente produzem gestos enquanto falam. De acordo com a literatura (e.g. [7] e [8]), tais gestos são ações que estão diretamente relacionadas com o conteúdo lexical e semântico e são particularmente adequados para reforçar ou complementar a mensagem que está a ser veiculada.

A literatura refere, ainda, que os gestos desempenham um papel crucial tanto para os oradores como para o público [9]. Relativamente aos oradores, os gestos são utilizados com o intuito de auxiliar a exposição das ideias e de recuperar conteúdos difíceis de memorizar. Por outro lado, para o público, a gesticulação utilizada pelos oradores podem relevar informações não disponíveis na fala, dando destaque a partes importantes do discurso ou desambiguando

o seu significado, tornando-o mais claro [10].

2.1.3 Expressão Facial

A face do ser humano tem despertado cada vez mais a atenção de investigadores em diversas áreas, nomeadamente extração de emoções, reconhecimento facial e comportamentos sociais [11] [12] [13]. A face humana é extremamente expressiva dado que consegue transmitir inúmeras emoções sem dizer uma única palavra, e que ao contrário de algumas formas de comunicação não-verbal, são universais. Desse modo, as expressões faciais são um dos aspetos mais importantes na comunicação humana dado que estas podem transmitir o estado emocional do orador, mas também intenções, através dos movimentos musculares faciais, como por exemplo, enrugamento das sobrancelhas ou levantamento dos cantos labiais.

Contudo, as expressões faciais não são as únicas com um papel importante na comunicação não-verbal através da face humana. A postura da cabeça e a direção do olhar são igualmente importantes indicadores da intenção comunicativa, uma vez que influenciam o nível de naturalidade e competência percebidos [14] [15].

2.1.4 Comunicação verbal

A capacidade de um orador falar bem, em diversas situações, pode contribuir significativamente para o seu fracasso ou sucesso, dado que a voz transporta em si inúmeras informações. Além da mensagem verbal, quando se refere a comunicação em público, a literatura afirma que o ser humano não infere apenas o significado transmitido mas também a forma como isso é feito (e.g. [16]). Nesse sentido, as pistas prosódicas são parte integrante da comunicação humana.

A literatura sustenta, ainda, que existem um conjunto de características presentes nos recursos áudio que tem sido amplamente utilizadas na investigação sobre como as vozes são ouvidas e interpretadas. Essas características passam, por exemplo, pelo volume e respetivas variações, duração da fala, duração das pausas, consideração de um campo lexical restrito (utilização de um grupo de palavras restritas), entre outras [17].

2.2 Métodos para extração dos canais de comunicação

Nesta secção serão apresentadas algumas ferramentas computacionais para a extração características presentes nos modos de comunicação verbal e não verbal e a respetiva justificação de escolha.

2.2.1 Extração de características Corporais

Para a extração dos dados necessários a análise de postura corporal, movimentos, e gestos, observados durante o processo de comunicação, existem algumas alternativas como *Kinect Skeletal Tracking* [18], ArtTrack [19] e DeeperCut [20]. No entanto, para fazer face a algumas deficiências que estas apresentam, foi apresentada em abril de 2017 a biblioteca OpenPose [21] que veio revolucionar área da computação visual. Através do uso de uma imagem Red-Green-Blue (RGB) como *input* possibilita a deteção e extração de valores a duas dimensões das principias partes do corpo humano num total de 130 *keypoints*, 15 ou 18 para o corpo, 21 para cada mão e 70 para a face.

O OpenPose usa algoritmos de *deep learning* para a deteção dos keypoints tendo por base modelos treinados a partir de dois conjuntos de dados (*Common Objects in Context (COCO)* ou *MPII Human Pose Dataset*) que contém imagens de pessoas anotadas com o esqueleto humano.

A Figura 2.2 exemplifica o resultado da utilização da biblioteca OpenPose aplicada a uma imagem RGB para a obtenção do esqueleto humano.

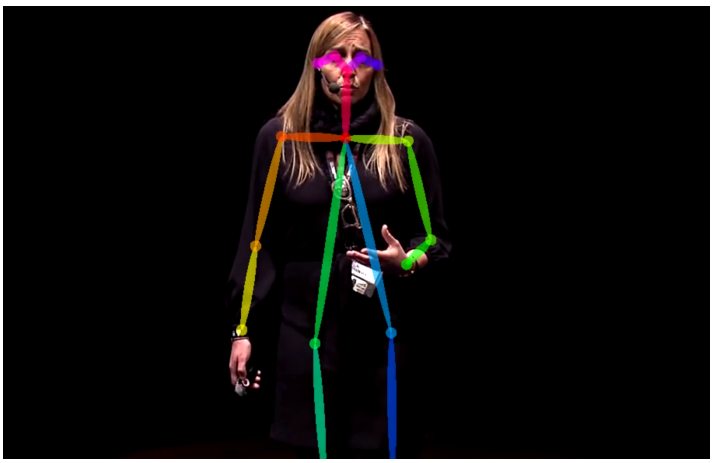


Figura 2.2: Aplicação do OpenPose a uma imagem RGB.

2.2.2 Extração de Características Faciais

Para a extração de marcadores faciais encontram-se algumas bibliotecas como Menpo [22], LEAR [23] ou o OpenFace 2.0 [24]. Porém, o OpenFace 2.0, uma biblioteca *open source* destinada aos investigadores de computação visual, tem como objetivo detetar pontos de referência facial, estimar a posição da cabeça, reconhecer *Action Units* (AUs) (ativação de músculos faciais) e estimar a direção do olhar. Por esse facto, torna-se assim mais vantajosa relativamente às anteriormente referidas. Esta biblioteca, com base numa imagem RGB, é capaz de detetar e extrair valores a duas dimensões da face, com um total de 68 *keypoints* para a deteção das *landmarks*, posição da cabeça, direção do olhar e reconhecimento da ativação e intensidade das AUs.

A Figura 2.3 apresenta a aplicação do OpenFace 2.0 a um conjunto de imagens onde são mostrados os pontos de referência faciais.

Figura 2.3: Aplicação do OpenFace 2.0 a um conjunto de imagens RGB. [24]

2.2.3 Extração de Características do Áudio

As ferramentas OpenEAR [25], SPAC [26] e Praat [27] são alguns exemplos do que pode ser considerado para processamento e análise de recursos áudio. Em alternativa a estas surge o OpenSMILE (*The Munich open-Source Media Interpretation by Large feature-space Extraction*) [28] que é fortemente usado pela comunidade de investigadores das áreas de reconhecimento de voz, reconhecimento de emoções e MIR (*Music Information Retrieval*). É uma biblioteca flexível e modular destinada ao processamento de sinais e aplicações de *Machine Learning*. O seu principal foco é a extração de características presentes no sinal de áudio, em tempo real ou sobre conjuntos de dados.

Esta biblioteca possibilita um conjunto de métodos diferenciados pelas seguintes categorias: *data input/output*, processamento geral do sinal áudio, extração de recursos relacionados com a fala, funcionalidades estatísticas, classificadores. Relativamente ao recursos relacionados com a voz, o OpenSMILE permite a extração de *Mel Frequency Cepstral Coefficient*, *Pitch*, *Jitter*, energia, intensidade, *Zero crossing rate*, entre outras. Todas a funcionalidades podem ser vistas com mais detalhe em [29].

2.3 Base de Dados Anotadas

Para a realização deste trabalho é útil considerar conjuntos de dados validados para que seja possível, por exemplo, uma análise dos métodos desenvolvidos. Nesse sentido, nesta secção são apresentados alguns exemplos de bases de dados anotadas que ilustram possíveis alternativas a considerar.

2.3.1 Ryerson Audio-Visual Database of Emotional Speech and Song

A RAVDESS (*Ryerson Audio-Visual Database of Emotional Speech and Song*) [30] é uma base de dados audiovisual validada, dinâmica e anotada de expressões faciais e vocais que conta com a participação de 24 atores (12 do sexo masculino e 12 do sexo feminino) profissionais.

Os atores proferiram duas frases distintas: *"Dogs are sitting by the door"* e *"Kids are talking by the door"*, onde cada uma das declarações foi expressa através de oito emoções, que foram exportados em três tipos de arquivos distintos: áudio - vídeo (rosto e voz), somente áudio (voz, mas sem rosto) e vídeo (rosto, mas sem voz).

A esta base de dados estão associadas emoções como: felicidade, tristeza, raiva, medo, surpresa, repugnância, tranquilidade e emoções neutras (i.e., sem emoção expressa). A Figura 2.4 ilustra dois exemplos de emoções da qual este dataset dispõe: Raiva e Felicidade.

Figura 2.4: Exemplo de duas emoções disponíveis na base de dados RAVDESS [30].

Contudo, este dataset não é o mais adequado, no contexto deste trabalho, uma vez que não aborda a temática da comunicação em público, sendo apenas um conjunto de dados que dispõe de discursos emocionais segmentados com expressões faciais.

2.3.2 Extended Cohn-Kanade Dataset

Extended Cohn-Kanade Dataset (CK+) [31] é uma base de dados de emoções associadas as expressões faciais que contém 5876 imagens de emoções faciais anotadas com *Action Units* (AUs) segmentadas pelo *Facial Action Coding System* (FACS). Esta base de dados foi composta com o auxílio de 123 participantes, onde cada um exibiu um conjunto de 23 expressões faciais que começavam e acabavam com um rosto neutro. As imagens foram adquiridas em vistas frontais e vistas com 30 graus.

A base de dados CK+ apresenta um conjunto de 7 emoções: raiva, desprezo, repugnância, felicidade, tristeza e surpresa. Estão apresentados alguns exemplos na Figura 2.5.

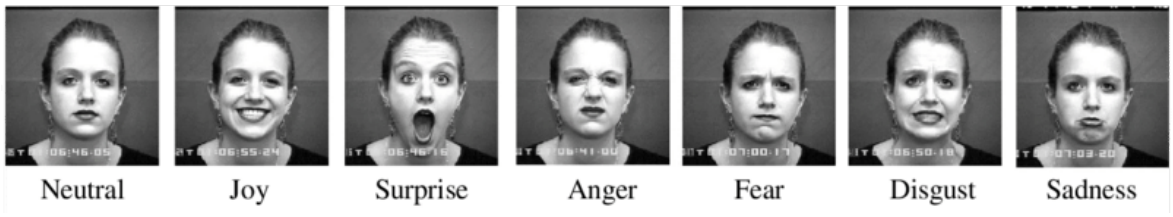


Figura 2.5: Exemplo das 7 disponíveis na base de dados CK+ [31].

Porém, este dataset é centralizado nas emoções faciais e, por esse motivo, apresenta uma relevância limitada para o trabalho em curso, uma vez que não cobre outros aspetos relacionados com a comunicação espontânea.

2.3.3 Interactive Emotional Dyadic Motion Capture

A base de dados *Interactive Emotional Dyadic Motion Capture* (IEMOCAP) [32] foi produzida com o auxílio de 10 atores que continham marcadores na face, cabeça e mãos que fornecem informações detalhadas sobre as expressões faciais e movimentos da mão durante cenários de comunicação espontânea. A Figura 2.6 ilustra a disposição dos marcadores que foram colocados nos atores para registar movimentos da cabeça e mãos e expressões faciais.

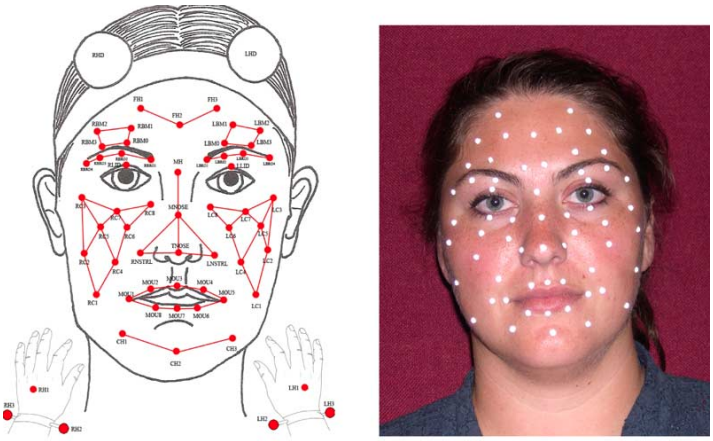


Figura 2.6: Ilustração da disposição dos marcadores nos atores [33].

Esta base de dados contém aproximadamente 12 horas de dados audiovisuais onde foi registado a fala e o movimento de rosto e mãos para as emoções de raiva, felicidade, tristeza e neutra, a que foram associados julgamentos sociais como valência, ativação e dominância.

2.3.4 DEP-UA TEDx Talks Dataset

Foi desenvolvida, pelo Departamento de Educação e Psicologia da Universidade de Aveiro (DEP-UA) a anotação de um dataset, com objetivo geral de estudar a comunicação em público de oradores. Em particular, pretende-se aferir os melhores preditores de inferência social e de *performance* de um orador com o intuito de melhorar a sua comunicação, tendo por objetivo final o seu uso na formação de professores.

Este insere-se num estudo no âmbito do Programa Doutoral em Multimédia da Educação (PDMMEdu) de Ângelo Silva Conde, resultante de uma colaboração entre o Departamento de Educação e Psicologia (DEP-UA) e o Departamento de Eletrónica Telecomunicações e Informática (DETI-UA), que com uma abordagem multidisciplinar pretende investigar qual a influência das posturas corporais dos professores como geradoras de índices de confiança dos alunos, para melhorar a comunicação destes em sala de aula.

O dataset é constituído por um conjunto de 36 vídeos selecionados do TEDx Portugal (18 oradoras e 18 oradores) do TEDx Portugal que relatam temas de diversas áreas. Os vídeos contêm uma duração de 30 a 60 segundos, obtidos de ângulos frontais ou $\frac{3}{4}$, dois planos (americano e médio). A Figura 2.7 apresenta a diferença entre dois planos (plano Americano

e médio) do qual dataset é composto.

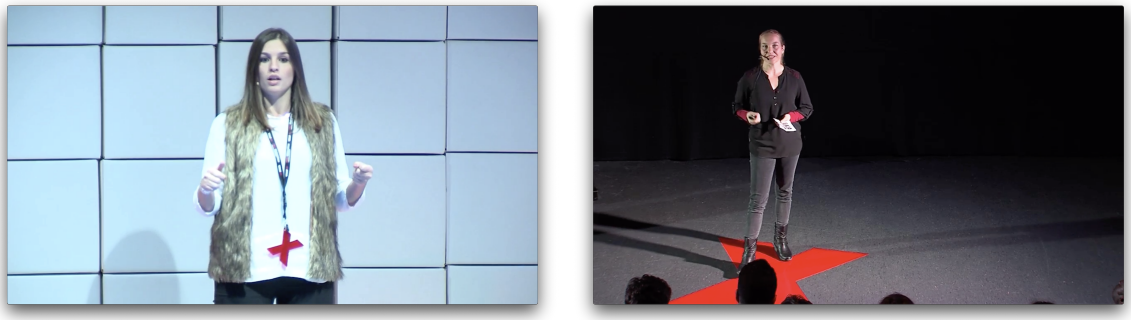


Figura 2.7: Diferença entre dois planos dos vídeos que compõe o dataset dispõe.

Cada um dos vídeos que compõe o dataset foi anotado por 40 participantes quanto a diferentes componentes da prestação do orador: uma medida do nível de confiança percebida, recolhida, em tempo real, ao longo do vídeo, e várias medidas de julgamento social (e.g., competência, assertividade) recolhidas no final de cada vídeo. Adicionalmente, cada participante identificou, para cada vídeo, que aspeto da comunicação mais influenciou o seu julgamento. As Tabelas 2.1 e 2.2 apresentam uma lista dos aspetos avaliados, para cada vídeo. Cada um dos aspetos mencionados foi avaliado usando uma escala analógica visual traduzida para um valor de 0 a 100.

Julgamentos Sociais	**Escala analógica visual**
Confiança	0 a 100
Assertividade	0 a 100
Dominância	0 a 100
Competência	0 a 100
Atratividade	0 a 100

Tabela 2.1: Anotação de Julgamentos Sociais e respetiva escala de anotação analógica.

Aspetos de Comunicação	Escala analógica visual
Voz	0 a 100
Postura	0 a 100
Gestos	0 a 100
Expressão Facial	0 a 100
Vestuário	0 a 100

Tabela 2.2: Anotação de aspetos de comunicação e respetiva escala de anotação analógica.

As Figuras 2.8 e 2.9 ilustram de que forma são obtidas as anotações para o nível de confiança percebido (em tempo real) e os diferentes julgamentos sociais (no final de cada vídeo), respetivamente.

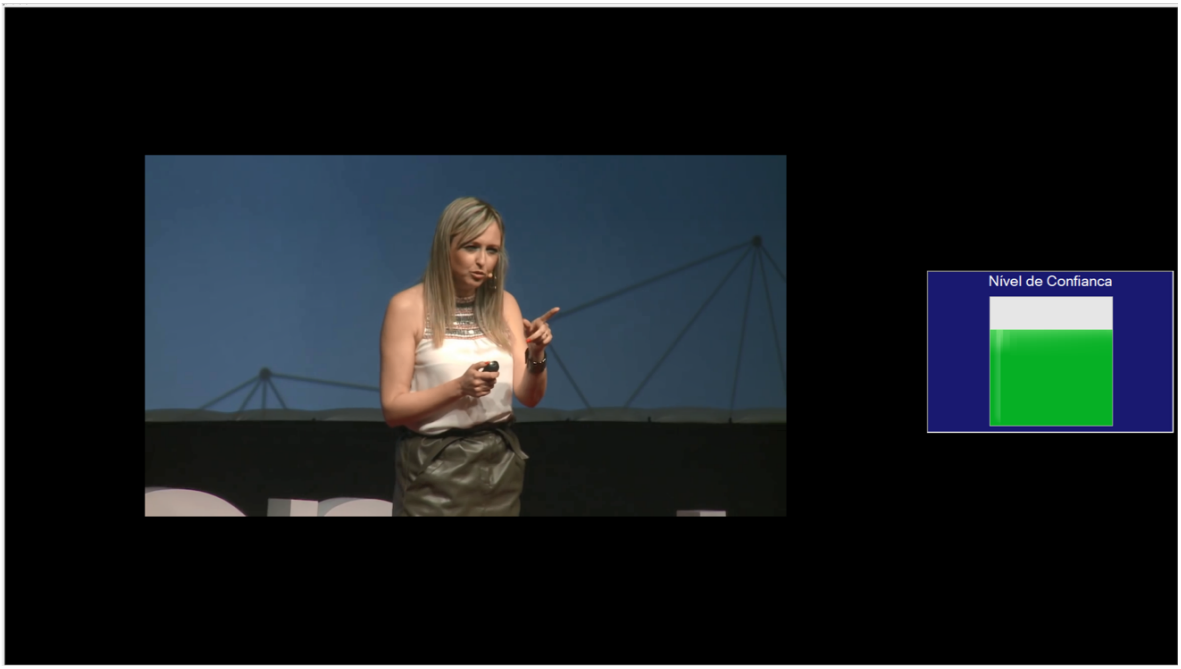


Figura 2.8: Ilustração da interface de aquisição da anotação, em tempo real, para o nível de confiança percebido.

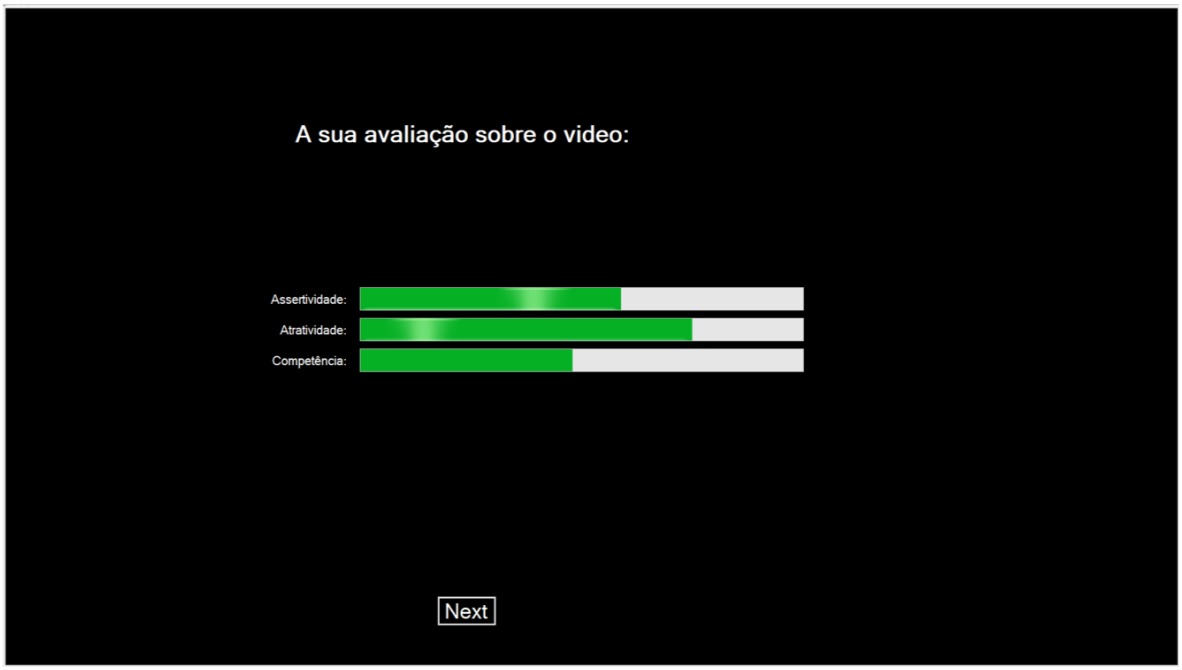


Figura 2.9: Ilustração da interface de aquisição da anotação, na prestação global, para os diferentes julgamentos sociais.

2.4 Conclusões

Neste capítulo é feito o levantamento do estado da arte relativamente a múltiplos aspetos presentes na comunicação verbal e não verbal, e de que for este podem ser extraídos para utilização de métodos computacionais. Falar em público é uma temática pouco explorada e o *feedback* existente recebido é apenas baseado na partilha, muitas vezes subjetiva, dos especialistas na área, o que dificulta a compreensão de que aspetos influenciam o público. Embora existam um conjunto de características na comunicação verbal e não verbal apontadas como relevantes, falta uma abordagem mais sistemática (e quantitativa) ao estudo do processo de comunicação.

Um dos aspetos dignos de nota é o de que algumas bases de dados existentes, eventualmente relevantes para o estudo da comunicação, carecem de informação complementar (anotações) que permitam interpretar o que ocorre e que permitam, a diferentes grupos de investigação, trabalhar os dados de forma mais independente, não havendo assim a necessi-

dade de conhecimentos técnicos aprofundados sobre técnicas de extração de dados para os diferentes canais de comunicação. Com uma maior importância, uma base de dados mais anotada pode permitir, mais facilmente, contributos multidisciplinares.

Por ultimo, os métodos de *Machine Learning*, para proposta de sistemas automáticos de avaliação da prestação dos comunicadores, ainda que sejam uma abordagem importante, carecem de uma abordagem que não afaste o investigador das diferentes vertentes em estudo. Não existe o interesse em utilizar um algoritmo de classificação com uma agregação de dados sem que se perceba o que são esses dados, o que representam e a sua relevância na comunicação humano-humano. É necessário a compreensão sobre o contributo de cada um e não simplesmente utilizar o métodos de *Machine Learning* de forma cega. O trabalho desenvolvido deverá criar as condições necessárias para que isso possa acontecer.

Capítulo 3

Características descritoras dos canais de comunicação

Tendo em conta os diferentes modos de comunicação identificados no capítulo anterior, neste capítulo é descrito um conjunto de descritores computacionais que permitem descrever as características do que se passa em cada um destes canais.

3.1 Influência das Propriedades da Imagem

Tendo em conta que estão a ser considerados vídeos, como fonte de dados, foi considerado importante caracterizar as propriedades gerais dos vídeos, antes de caracterizar a atividade dos oradores, uma vez que essas características, podem ser aspetos a ter em conta como potenciais influenciadoras da "agradabilidade" de um vídeo e, consequentemente, ter impacto, ainda que indireto, no julgamento feito sobre os oradores. Nesse sentido, foi desenvolvido um método que recebendo como *input* um ou mais vídeos, realiza a extração de duas propriedades: Luminância média e Contraste.

Para a obtenção dos valores representativos da luminância média e contraste de cada vídeo é feita uma conversão da escala cor (RGB) para uma escala de cinzentos. Após a realização da conversão, são aplicadas as Equações 3.1 e 3.2 para o cálculo de luminância e contraste, respetivamente, para cada *frame*, onde $P(x, y)$ identifica o valor do pixel na posição x, y e N

representa o número total de pixéis na *frame*.

$$L = \frac{\sum_{i=0}^{N} P(x,y)}{N} \tag{3.1}$$

$$C = \sqrt{\frac{\sum_{i=0}^{N} P(x,y) - L)^2}{N}} \tag{3.2}$$

Como resultado, é obtido para cada vídeo um vetor de dados de tamanho igual ao número de *frames* de cada vídeo, com valores que variam entre 0 e 255. A Figura 3.1 apresenta dois exemplos para valores de luminância média diferentes, isto é, a imagem a esquerda possui um valor de luminância mais elevado do que a imagem à esquerda.

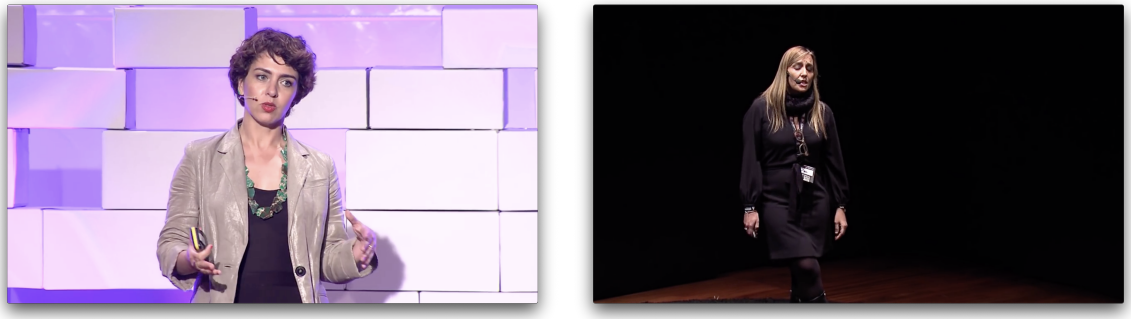


Figura 3.1: Diferença entre valores de luminância média: Mais elevado a esquerda e mais baixo a direita.

A Figura 3.2 apresenta dois exemplos para valores de contraste diferentes, isto é, a imagem a esquerda possui um valor de contraste mais elevado do que a imagem à esquerda.

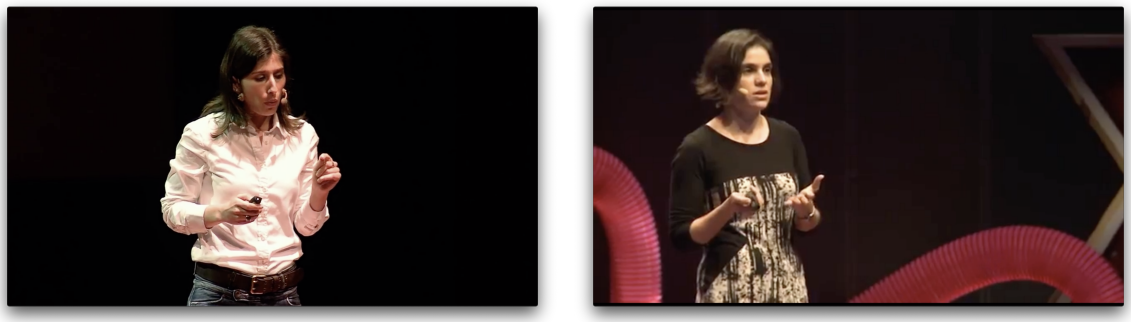


Figura 3.2: Diferença entre valores de contraste: Mais elevado a esquerda e mais baixo a direita.

3.2 Pontos de referência faciais, *Action Units* e Postura da Cabeça

Como referido no Capítulo 2, a face é um indicador de intenção comunicativa. Nesse sentido, pode-se avaliar a comunicação através da face em pelo menos duas vertentes: expressões faciais e posição da cabeça.

Para facilitar a descrição das propriedades das expressões faciais pode ser considerado o *Facial Action Coding System* (FACS). O FACS [34] baseia-se nas variações visíveis da face humana produzidas pela ativação individual dos músculos faciais, que são intituladas de AUs (*Action Units*). Assim, qualquer expressão facial pode ser representada através da combinação das diferentes AUs, como é possível observar na Figura 3.3. Cada uma das AUs pode ter duas informações associadas: (1) se está ativada e (2) a intensidade dessa ativação.

Figura 3.3: Múltiplas e distintas *Action Units* dois estados emocionais distintos [35].

Por outro lado para avaliar a posição da cabeça, e de acordo com [36], esta possui três graus de liberdade: *pitch, roll, yaw* (representado na Figura 3.4).

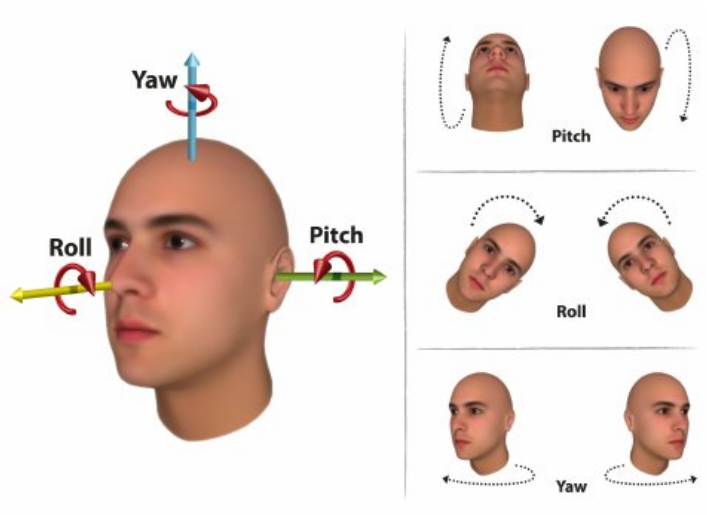


Figura 3.4: Posição da Cabeça em função de pitch, roll e yaw [37].

Para a obtenção das AUs (com respetiva intensidade), os pontos de referência faciais e os três graus de liberdade da cabeça foi desenvolvido um método que, recebendo com *input* um vídeo e usado a biblioteca OpenFace 2.0, procede a extração dos mesmos para cada *frame*. Foi ainda desenvolvido um método que permite uma visualização dos dados extraídos (pontos de referência faciais e intensidade das AUs), em tempo real, sobre o vídeo em questão com o intuito de perceber a dinâmica dos mesmos.

A Figura 3.5 ilustra a disposição dos pontos de referencia faciais e a intensidade medida para as diferentes AUs obtidas pelo método adotado. Note-se que na imagem da esquerda a face apresenta uma emoção neutra, enquanto que na imagem da direita apresenta uma expressão de felicidade.

Figura 3.5: Variação de pontos de referência faciais e *Action Units*.

3.3 Pontos de referência corporais

A comunicação através do corpo, como mencionado no Capítulo 2, é outro indicador comunicativo. Nesse sentido foi desenvolvido um método que tendo como input um vídeo e utilizando a biblioteca OpenPose, descrita no Capítulo 2, realiza a extração de cada ponto de referência corporal, num máximo de 18 *keypoints*, que representam o corpo. Além disso, e também de modo similar à secção anterior, foi também desenvolvido um método que utilizado os dados extraídos, permite a sua visualização sobre o vídeos de modo a permitir aferir a qualidade dos dados extraídos e perceber como estes variam. A Figura 3.6 retrata a disposição e identificação de cada ponto de referência corporal, de acordo com o método adotado.

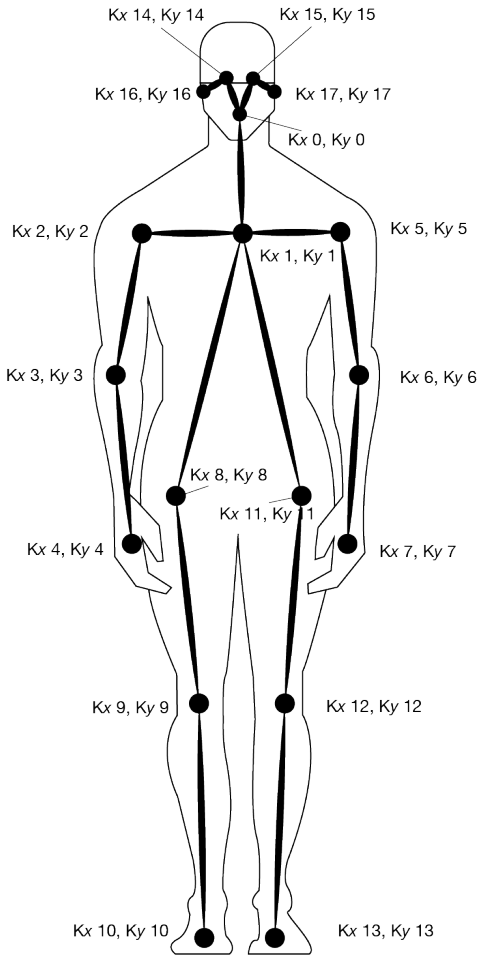


Figura 3.6: Extração de características da postura corporal: Ilustração da disposição e respetiva identificação dos pontos de referencia corporais consideradas no método adotado.

3.4 Expansividade, Velocidade dos Movimentos e Área ocupada

Segundo Koppensteiner et al. [4], os movimentos horizontais e verticais do corpo afetam a formação de impressões de um modo diferente. Da mesma maneira, e de acordo com Carney et al. [5], posturas expansivas e abertas projetam sinais de poder, domínio, confiança e assertividade, e por outro lado, posturas contrativas e fechadas projetam sinais de impotência e baixa autoestima.

Dessa forma foi desenvolvido um método que recebendo como input os pontos de referência corporais, realiza um processamento sobre os mesmos com o objetivo de calcular os valores relativamente aos movimentos horizontais e verticais do corpo, velocidade dos movimentos e a área ocupada pelo corpo humano.

3.4.1 Expansividade

Para o cálculo dos movimentos do corpo, estes foram decompostos em componentes: movimentos horizontais $H(x)$ e movimentos verticais $V(y)$. A fim de determinar a amplitude dos mesmos, foi utilizando algo semelhante ao que foi realizado em [4], onde para a obtenção dos movimentos horizontais das mãos foi feito o somatório da diferença entre coordenada x da garganta e a coordenada x da mão direita com a diferença da coordenada x da mão esquerda (x) e a coordenada x garganta. Por outro lado, para a amplitude dos movimentos verticais foi feita a soma da coordenada y púlpito menos a coordenada y da mão direita com a coordenada y do púlpito menos a coordenada y da mão esquerda.

Expansividade Horizontal

A amplitude dos movimentos horizontais $H(x)$ é obtida através da soma da distancia entre a coordenada x do *keypoint* do pescoço (Kx_1) e as coordenadas x dos *keypoints* do pulso esquerdo (Kx_7) e pulso direito (Kx_4), que é representada pela Equação 3.3. A Figura 3.7 ilustra a disposição dos pontos de referência corporais utilizado para o cálculo da expansividade horizontal.

$$H(x) = |Kx_1 - Kx_4| + |Kx_7 - Kx_1| \tag{3.3}$$

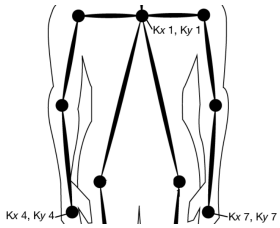


Figura 3.7: Ilustração da disposição dos pontos de referência corporais utilizado para o cálculo da expansividade horizontal.

Na Figura 3.8 está representada a variação dos valores da expansividade horizontal ao logo do tempo. De forma a ilustrar este processo, são apresentadas três *frames* que representam, visualmente, a variação do vetor da expansividade horizontal em estudo, segmentadas com um gráfico representativo da curva de variação da expansividade horizontal.

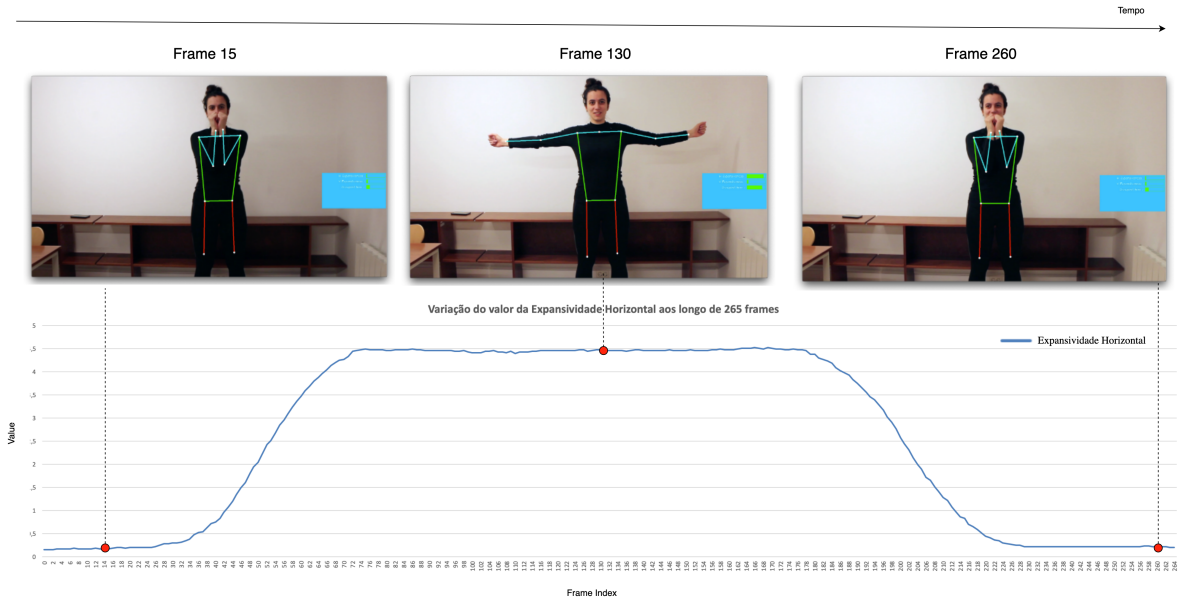


Figura 3.8: Variação dos valores para expansividade horizontal ao longo de um vídeo.

Expansividade Vertical

A amplitude dos movimentos verticais $V(y)$ é obtida através da soma da distancia entre a coordenada y do *keypoint* do pescoço (Ky_1) e as coordenadas y dos *keypoints* do pulso esquerdo (Ky_7) e pulso direito (Ky_4), que é representada pela Equação 3.4. A Figura 3.9 ilus-

tra a disposição dos pontos de referência corporais utilizado para o cálculo da expansividade vertical.

$$V(y) = (Ky_4 - Ky_1) + (Ky_7 - Ky_1) \tag{3.4}$$

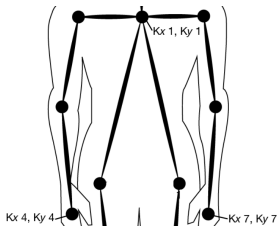


Figura 3.9: Ilustração da disposição dos pontos de referência corporais utilizado para o cálculo da expansividade vertical.

Na Figura 3.8 está representada a variação dos valores da expansividade vertical ao logo do tempo. De forma a representar este processo são apresentadas três *frames* que representam, visualmente, a variação do vetor da expansividade vertical em estudo, segmentadas com um gráfico representativo da curva de variação da expansividade vertical.

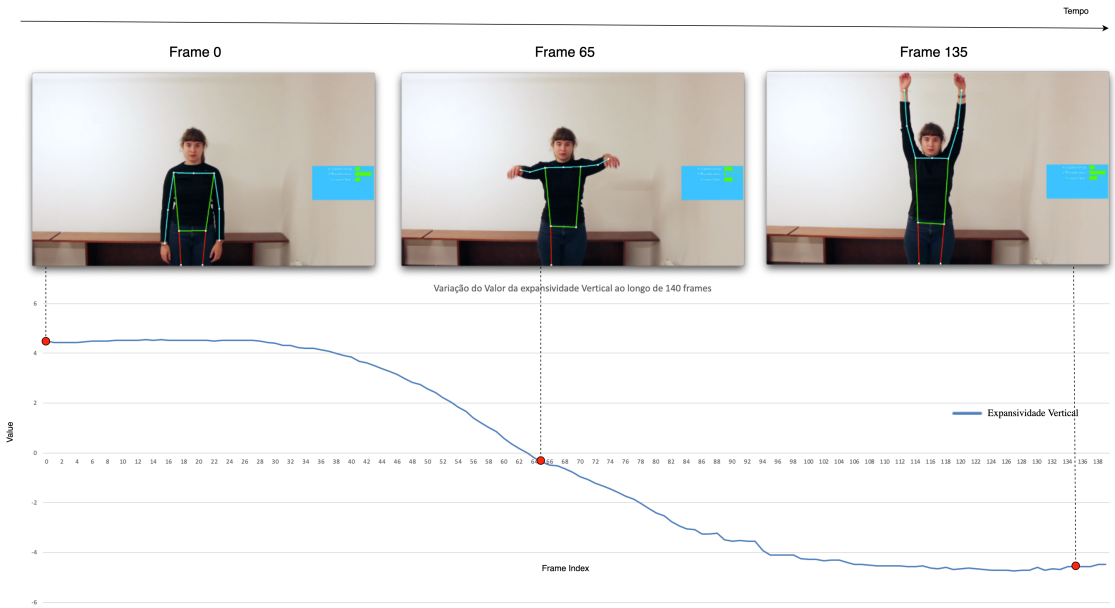


Figura 3.10: Variação dos valores para expansividade vertical ao longo de uma sequência de vídeo. A curva de variação passa por 0 quando os pulsos estão à altura do pescoço.

3.4.2 Velocidade dos Movimentos

Para além da amplitude dos movimentos, torna-se importante entender a velocidade dos mesmos, isto é, se são rápidos ou lentos. De forma a calcular a velocidade dos movimentos das mãos ao longo do tempo foi desenvolvido um método que calcula a velocidade de cada mão de forma individual. O cálculo é feito *frame* a *frame*, analisando a *frame* atual, as dez *frames* anteriores e as dez *frames* seguintes. Dessa forma é possível obter a velocidade de movimentos por unidade de tempo, para cada mão.

A Figura 3.11 ilustra a variação da velocidade do movimento do pulso esquerdo ao longo de um conjunto de *frames*. De forma a demonstrar este processo são apresentadas quatro *frames* que representam, visualmente, o *keypoint* (pulso esquerdo, identificado pelo círculo) em estudo, acompanhadas com um gráfico que identifica o valor do deslocamento para cada uma das 4 *frames*. Analisando a ilustração é possível realçar que entre as duas imagens mais à esquerda (100 *frames* de diferença) a variação da velocidade é menos acentuada que nas duas imagens mais à direita (35 *frames* de diferença), tal como mostra o gráfico.

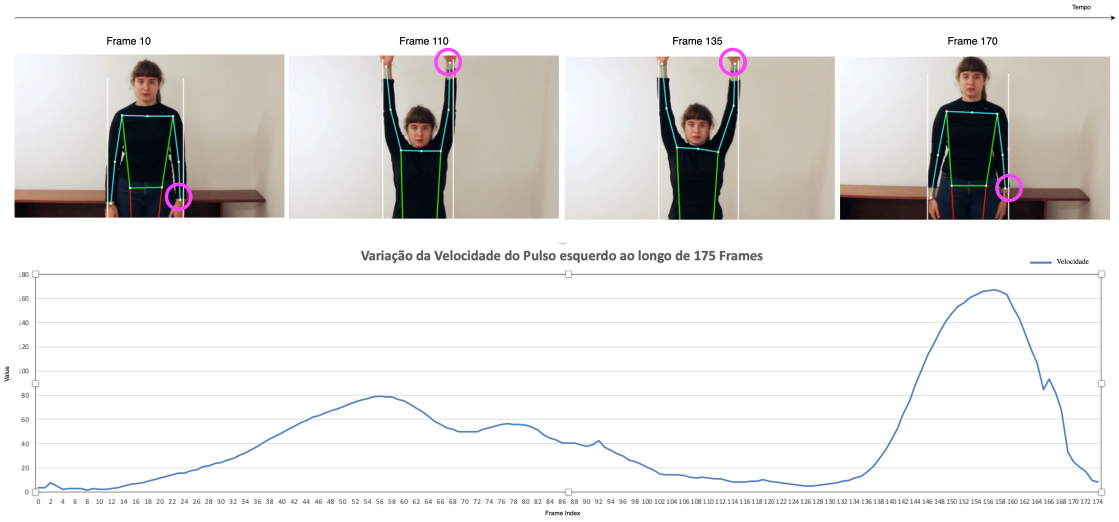


Figura 3.11: Ilustração da variação da velocidade de deslocamento para a mão esquerda. De realçar o movimento mais lento de elevação do braço do que o abaixamento.

3.4.3 Área Ocupada

A área ocupada por um orador consiste no espaço que este ocupa em relação a um plano (*frame*). Nesse sentido, para calcular a área ocupada em cada *frame* são obtidos os valores de

Kx_{max}, Kx_{min}, Ky_{max} e Ky_{min} do corpo humano. Para estes valores são verificados quais os *keypoints* que representam os valores mínimos e máximos para eixos X e Y em cada *frame*. A expressão 3.5 representa como é feito o respetivo cálculo da área ocupada. Na Figura 3.12 ilustra a disposição dos pontos de referência corporais máximos e mínimos para o cálculo da área ocupada.

$$A = |Kx_{max} - Kx_{min}| * |Ky_{max} - Kx_{min}| \tag{3.5}$$

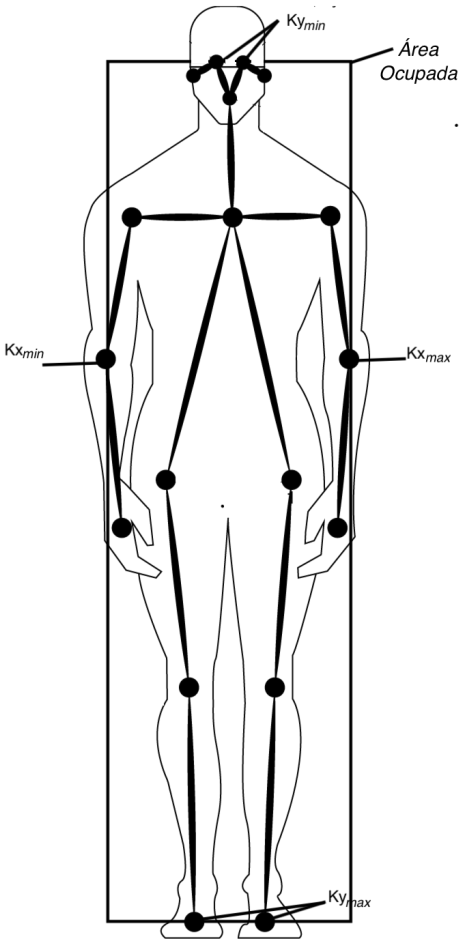


Figura 3.12: Ilustração da disposição dos pontos de referência corporais utilizados para o cálculo da área ocupada.

A Figura 3.13 exemplifica a diferença de da área ocupada usando duas posturas diferentes. À esquerda o Sujeito 1 está a usar uma postura (exageradamente) mais aberta do que o Sujeito

2, que se encontra à direita.

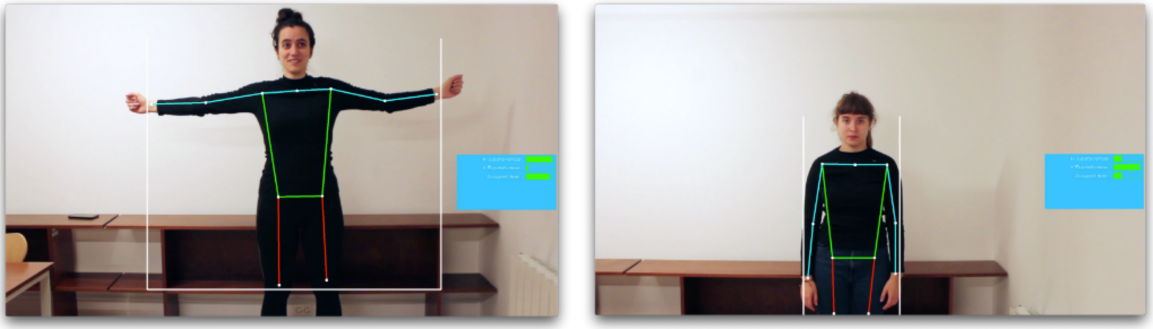


Figura 3.13: Ilustração da Área Ocupada por dois sujeitos destinos: à esquerda mais expansivo e à direita mais constrito.

3.5 Medidas Antropométricas

Com o objetivo de normalizar os valores de expansividade e área ocupada, visto 1que os vídeos podem conter planos diferentes, como é o caso do DEP-UA TEDx Talks Dataset evidenciado no Capítulo 2, foram obtidas as medidas antropométricas do corpo do humano.

A Figura 3.14 ilustra a diferença entre dois planos que o DEP-UA TEDx Talks Dataset dispõe, que no caso das medidas antropométricas (distancia entre ombros), possuem uma diferença entre valores muito acentuada.

Figura 3.14: Diferença entre planos e respetivas medidas antropométricas (distancia entre ombros).

Dessa forma, foi feito um processamento aos dados relativos aos pontos de referência

corporais com o intuito de obter o valor para a distancia entre ombros e o comprimento dos braços, dado que estes influenciam o cálculo das expansividades e da área ocupada.

Para a obtenção da distância entre ombros é calculada a distância entre o *keypoint* do ombro esquerdo e direito, para todas as *frames* do vídeo, atribuído assim a distância entre ombros o valor máximo encontrado.

De igual forma, o comprimento dos braços é obtido através da soma das distâncias entre o *keypoint* do ombro e cotovelo e entre o *keypoint* do cotovelo e pulso, para todas as *frames* do vídeo. Este processo é executado para cada braço, onde é designado o comprimento de cada braço com valor máximo encontrado.

Uma vez obtidas as medidas antropométricas, para a normalização dos valores das expansividades e área ocupada, para que esta passe a ter um valor máximo de 5, foi utilizada a Regra de três simples, cujo intuito é obter um novo valor normalizado, a partir de outros três: máximo possível em pixéis e normalizado e o valor a normalizar em pixéis.

3.6 Recursos Áudio

Como já referido no Capítulo 2, variação da intensidade, duração da fala, duração da pausa e o *pitch* têm sido alvo de investigação de forma a avaliar a maneira como as vozes são percebidas. Contudo, as características prosódicas tornam-se também um dos pilares do reconhecimento de traços para-linguísticos, ou seja, características prosódicas são conhecidas por predizer o carisma do orador [38]. Nesse sentido, estes são uma escolha natural para a caracterização do desempenho do orador. Contudo, e apesar do geralmente alto desempenho, estas características nem sempre permitem avaliar aspetos como emoções, por exemplo. Nesse contexto, os *Mel-frequency cepstrum coefficients* (MFCCs) têm sido frequentemente usados para detetar um conjunto de sinais sociais [38]. Dessa forma, foi desenvolvido um método que usando como *input* um ficheiro áudio procede a extração de recursos áudio através do uso da biblioteca OpenSMILE, considerando um conjunto de características análoga ao do desafio INTERSPEECH 2010 [39]. Após realizada a extração dos recursos é feito um processamento sobre os mesmos a fim de obter algumas métricas. A Tabela 3.1 detalha os recursos áudio extraídos e as respetivas métricas.

Recursos Áudio	Métricas
Intensity	Valor, Máximo, Mínimo, Intervalo, Média Aritmética, Media Quadrática, Desvio Padrão, Skewness, Kurtosis
Voice Quality	
RMS Energy	
Log Energy	
ZCR	
Loudness	
HNR	
Pitch	
Jitter	
Shimmer	
MCFF 1 - 12	

Tabela 3.1: Caracterização das propriedade da voz do orador: Recursos Áudio extraídos e métricas, tendo por base o desafio INTERSPEECH 2010 [39].

3.7 Aplicação

Neste capítulo são enunciados um elevando número de métodos que permitem explorar os diferentes canais de comunicação. Embora este possam ser extraídos de forma simples, foi desenvolvida uma aplicação para fazer uma aglomeração dos mesmos, possibilitando a utilizadores com diferentes níveis de competências técnicas, de forma ainda mais simples, procederem a extração dos descritores dos canais de comunicação.

A Figura 3.15 ilustra a interface da aplicação desenvolvida, onde é permitido ao utilizador especificar um conjunto de vídeos para análise, quais os conjuntos de dados a extrair e o respetivo diretório de destino, isto é, para onde serão gerados os ficheiros resultantes da extração. Como resultado são produzidos ficheiros no diretório de destino que podem facilmente ser usados para diversos fins como, por exemplo, análise estatística.

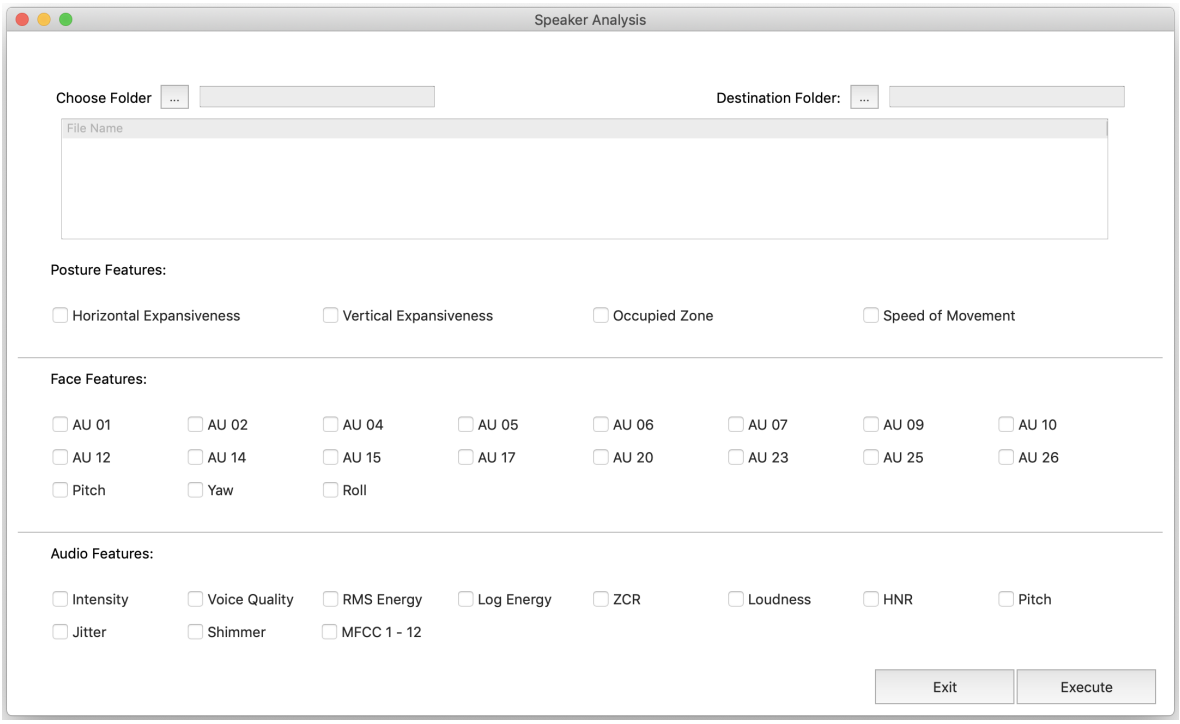


Figura 3.15: Aplicação desenvolvida para a extração dos conjuntos de informação dos diferentes canais de comunicação.

3.8 Conclusões

Durante este capítulo foram apresentados um conjunto de métodos que permitem extração de um conjunto de descritores computacionais que permitem descrever as características do que se passa em cada um destes canais, nomeadamente extração de expressões faciais, extração de postura e extração de recursos áudio. É também apresentado no final deste capítulo uma aplicação que permite de forma simples aos utilizadores procederem a extração dos diferentes conjuntos de dados.

Capítulo 4

Anotações de atividades relevantes

Com o intuito de permitir aos dados extraídos no capítulo anterior uma maior legibilidade, uma melhor identificação das atividades com relevância e de modo a que estes se adequem aos diferentes níveis de estudo, neste capítulo são expostos um conjunto de métodos para anotações de atividades relevantes desenvolvidos. Iniciando-se pela dedução de emoções e estimativa da posição da cabeça, até a anotação de gestos, momentos de silêncio e variação da intensidade da voz.

4.1 Dedução de Emoções

Depois de obtidos os vetores relativos à variação de intensidade das (*Action Units* (AUs) produzidos pelo método exposto na Secção 3.2, segue-se a análise dos mesmos, com o objetivo de realizar a dedução de emoções presentes. Em [40] foi feita uma abordagem em que, para cada emoção, foi verificado quais as AUs mais predominantes, que resultou numa tabela que apresentava quais as cinco AUs mais predominantes por emoção. Então, realizou-se uma abordagem similar, mas tendo em consideração outros aspetos. A abordagem usada baseou-se na aproximação sobre quais as AUs representam uma determinada emoção (segmentadas pelo FACS) e quais as mais predominantes segundo [40], que resultou numa seleção das três AUs que melhor caracterizam uma determinada emoção. A Tabela 4.1 mostra cada emoção versus as AUs ordenadas de forma crescente, da esquerda para a direita, em função do valor de predominância.

Uma vez deduzida a tabela que enuncia as três AUs que melhor descrevem uma dada

Emoções	***Action Units***		
Raiva	AU 25	AU 4	AU 9
Medo	AU 1	AU 5	AU 25
Tristeza	AU 1	AU 4	AU 17
Felicidade	AU 12	AU 6	AU 25
Surpresa	AU 26	AU17	AU2
Desgosto	AU 9	AU 7	AU 4

Tabela 4.1: Emoções versus *Action Units*.

emoção, foi desenvolvido um método que, usando como *input* a intensidade das AUs, produz um *output* com a respetiva emoção. Para cada emoção é atribuído um valor numérico, que é obtido através da soma das intensidades das AUs (que descrevem a emoção) e para cada AU é atribuído um peso maior consoante o valor de predominância. Em seguida, é selecionada a emoção que possuir o valor numérico maior. No entanto se esse valor for inferior a um determinado *threshold*, é atribuída a emoção neutra.

Na Figura 4.1 está representada a variação dos valores da AUs 12, 6 e 15 ao logo do tempo. De forma a demonstrar este processo são apresentadas 3 *frames* que representam, visualmente, a variação do vetor das três AUs em estudo.

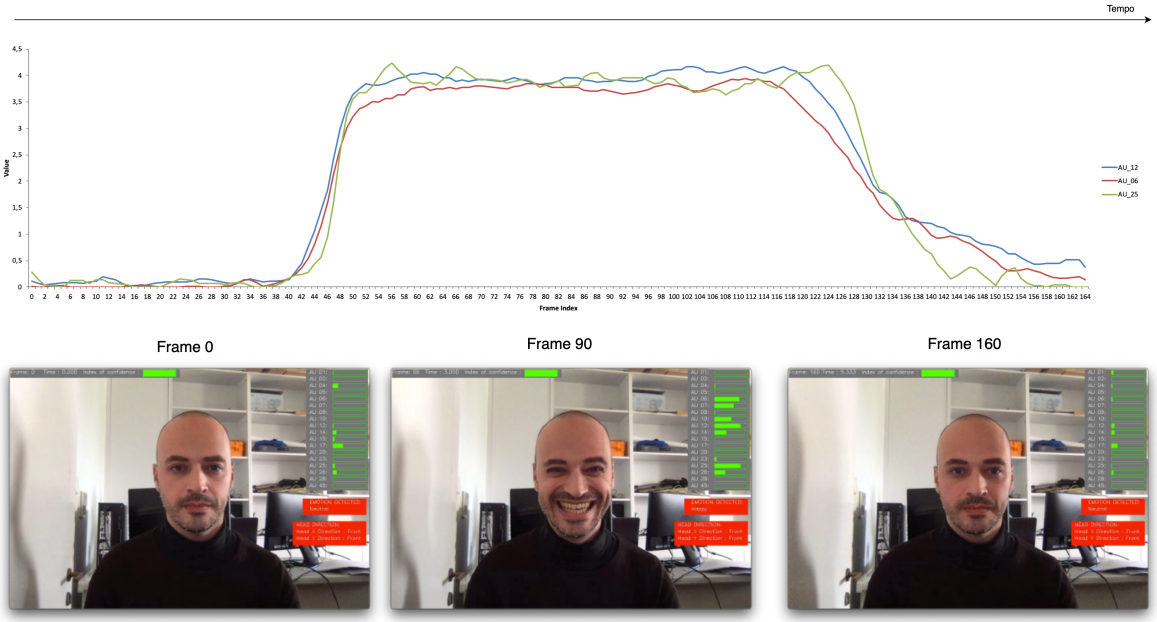


Figura 4.1: Variação da intensidade das *Action Units* para as emoções: Neutro-Feliz-Neutro.

4.2 Estimativa da Posição da Cabeça

Depois de obtidos os vetores relativos aos três graus de liberdade que definem a orientação da cabeça foi realizado o processamento apenas sobre os vetores de *pitch* e *yaw* a fim de anotar a variação da orientação em função do eixo vertical (movimento de cima para baixo e vice-versa) e em função do eixo horizontal (movimento da esquerda para a direita e vice-versa).

Deste modo, foi desenvolvido um método que usando como input os valores de *pitch* e *yaw*, produzidos pelo método exposto na Secção 3.2, produz como *output* a respetiva anotação do movimento. A anotação é feita através da variação dos valores, *pitch* e *yaw*, a cada combinação de dez *frames* consecutivas, isto é, se em cada dez *frames* consecutivas existir uma variação positiva ou negativa.

É feita a anotação do movimento da cabeça em função de do eixo vertical, de cima para baixo, se existir uma variação positiva do valor de *pitch* e vice-versa caso a variação seja negativa. Nas Figuras 4.2 e 4.3 é representado o processo de anotação do movimento da cabeça para cima e cabeça para baixo respetivamente, dado que na figura 4.2 existe uma variação negativa no valor do *pitch* enquanto que na figura 4.3 a variação é positiva. Estas apresentam um conjunto de quatro *frames*, correspondentes a uma janela de trinta *frames* complementadas com um gráfico que ilustra o valor do *pitch* para cada uma destas.

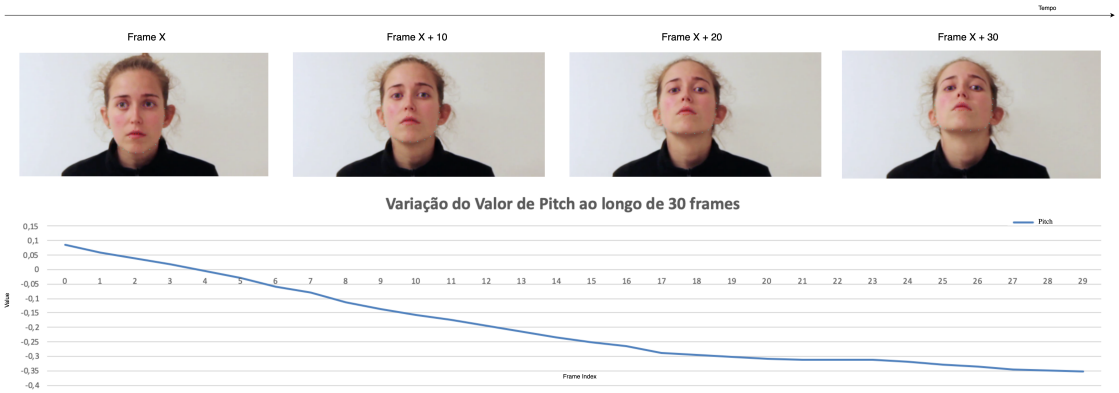


Figura 4.2: Processo de anotação para : *Head Moving up*.

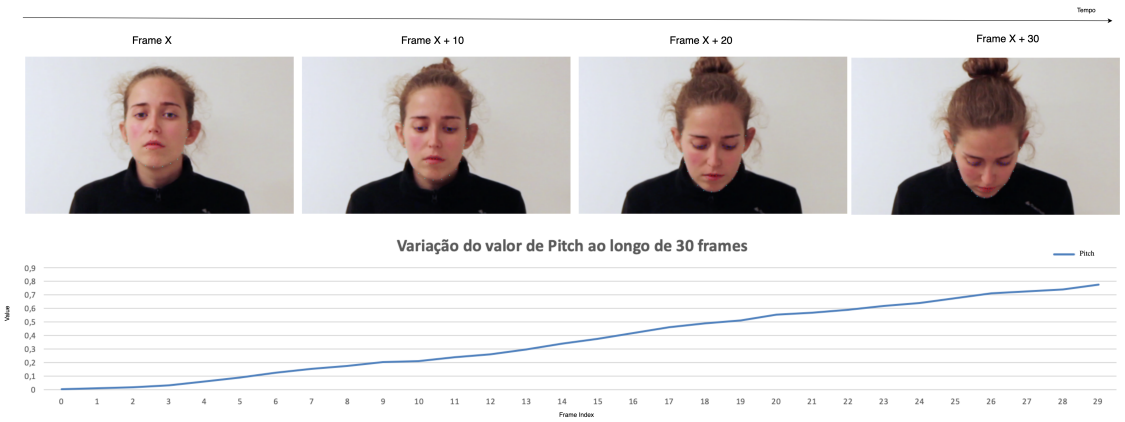


Figura 4.3: Processo de anotação para : *Head Moving Down*.

Do mesmo modo mas em função do eixo horizontal é feita a anotação do movimento da esquerda para a direita se a variação do valor de *Yaw* for uma variação positiva e vice-versa caso a variação seja negativa. É ilustrado nas Figuras 4.4 e 4.5 o processo de anotação do movimento da cabeça da esquerda para a direita e vice versa, respetivamente. Na Figura 4.4 existe uma variação positiva no valor de *Yaw*. Por outro lado na Figura 4.5 a variação do valor de *Yaw* é negativa. Estas apresentam um conjunto de quatro *frames*, correspondentes a uma janela de trinta *frames* complementadas com um gráfico que ilustra o valor do *Yaw* para cada uma destas.

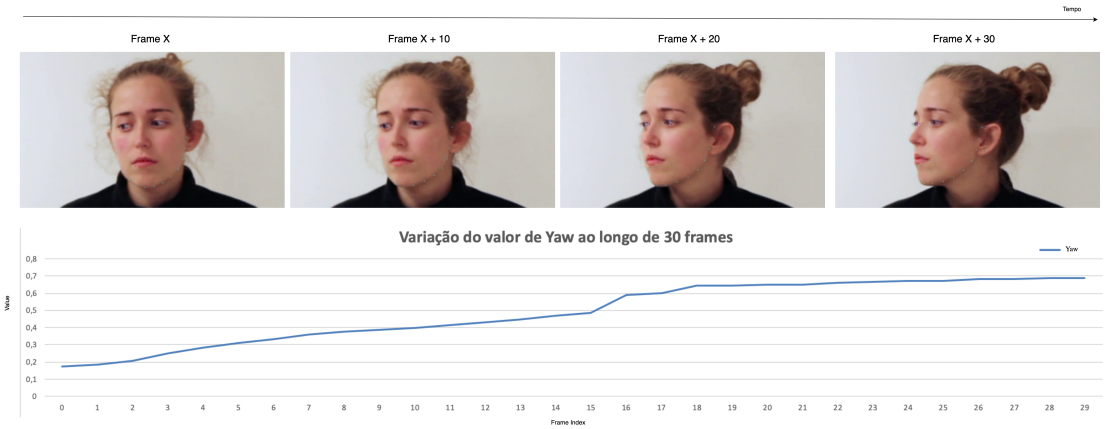


Figura 4.4: Processo de anotação para : *Head Moving Right*.

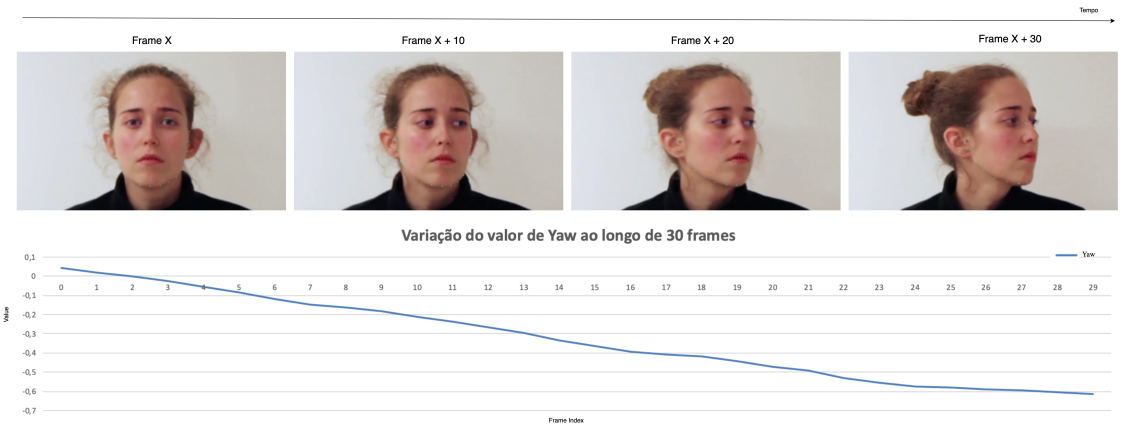


Figura 4.5: Processo de anotação para : *Head Moving Left*.

4.3 Gestos

Para a anotação de gestos executados pelo orador foi desenvolvido um método que, utilizado com *input* os vetores de dados representativos dos pontos de referência corporais, elabora um *output* com a anotação de gestos que se caracterizaram em função do eixo horizontal e em função do eixo vertical.

4.3.1 Gestos Horizontais

Os gestos em função do eixo horizontal foram defendidos em duas vertentes: posicionamento estático e gestos dinâmicos, ou seja, foi feita a diferenciação entre a variação da proximidade entre mãos (para gestos dinâmicos) e posicionamento das mãos relativamente ao nível do tronco (para posicionamento estático).

Em relação ao nível de variação da proximidade das mãos fez-se a distinção em dois grupos: *Approach Hands* e *Separate Hands*. É feita a anotação *Approach Hands* se a distância (euclidiana) entre pulsos em cada dez *frames* consecutivas diminuir. Do mesmo modo, é anotado *Separate Hands* caso a distância aumente.

As Figuras 4.6 e 4.7 ilustram o processo de anotação de gestos para *Approach Hands* e *Separate Hands*, respetivamente. Estas apresentam um conjunto de quatro *frames* e um gráfico representativo do valor de proximidade entre pulsos num intervalo de trinta *frames*. No caso da figura 4.6 é anotado *Approach Hands* uma vez que existe um decréscimo da distância entre pulso. Por outro lado na Figura 4.7 é anotado *Separate Hands* uma vez que

existe um crescimento da distância entre pulsos.

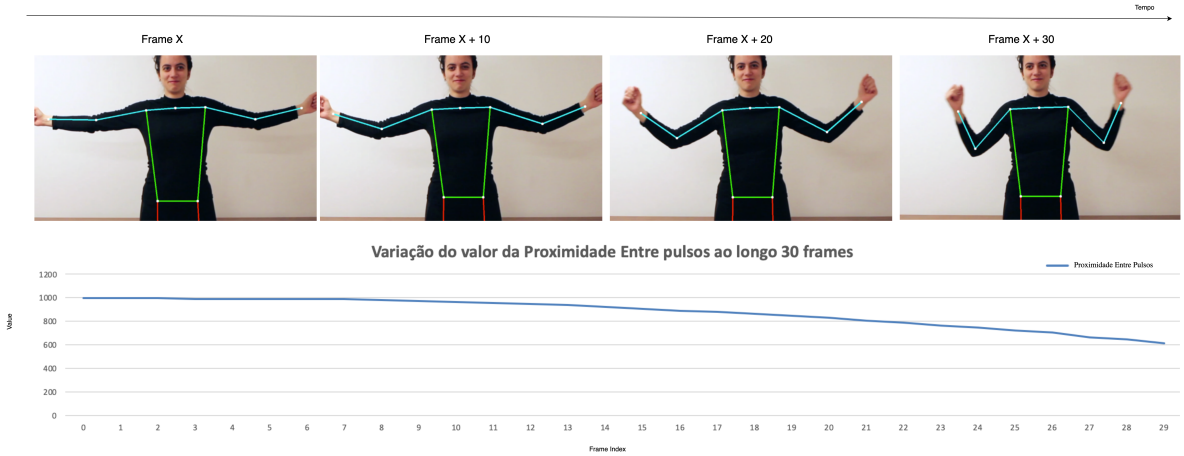


Figura 4.6: Processo de anotação para : *Approach Hands*.

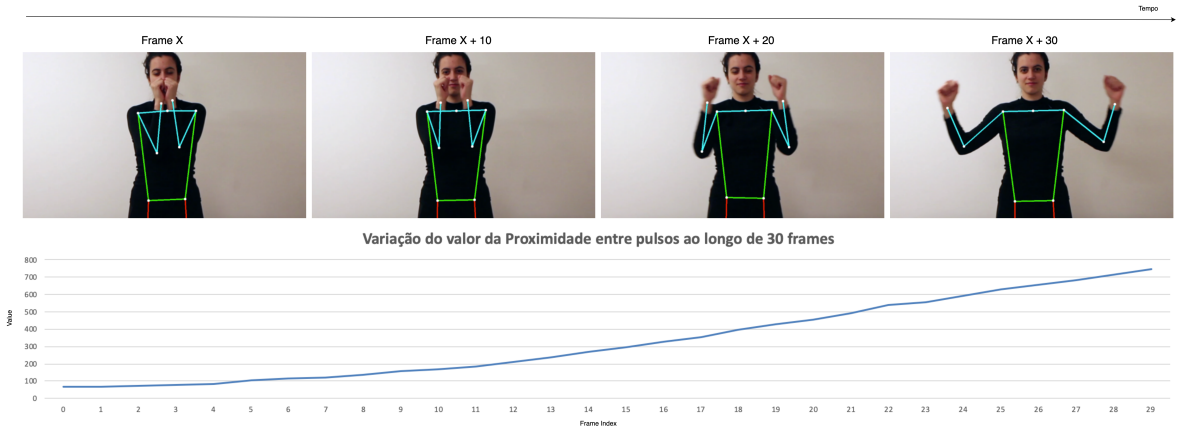


Figura 4.7: Processo de anotação para : *Separate Hands*.

No que diz respeito ao posicionamento estático de cada mãos relativamente ao tronco, fez-se também a distinção em dois grupos: *Inward* e *Outward*. Por exemplo, considerando apenas a mão esquerda, é anotado como *Inward* se o valor de x do *keypoint* do pulso esquerdo for inferior ao valor de x do ombro esquerdo. Por outro lado, se o valor x do pulso esquerdo for superior ao valor de x do ombro esquerdo é anotado *Outward*. As Figuras 4.8 e 4.9 representam as duas anotação consideradas.

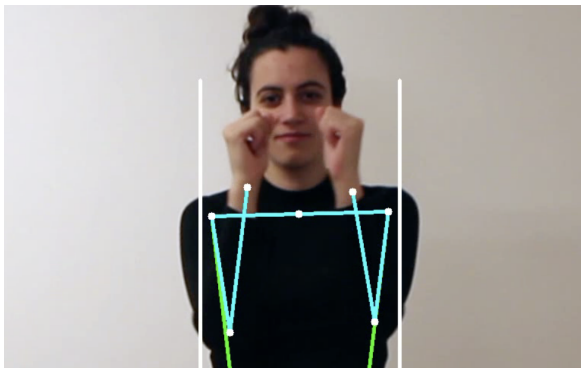


Figura 4.8: Anotação de : *Inward*, para o braço esquerdo e direito. O valor de x do *keypoint* do pulso esquerdo é inferior ao valor de x do ombro esquerdo. Por outro lado, o valor de x do *keypoint* do pulso direito é superior ao valor de x do ombro direito.

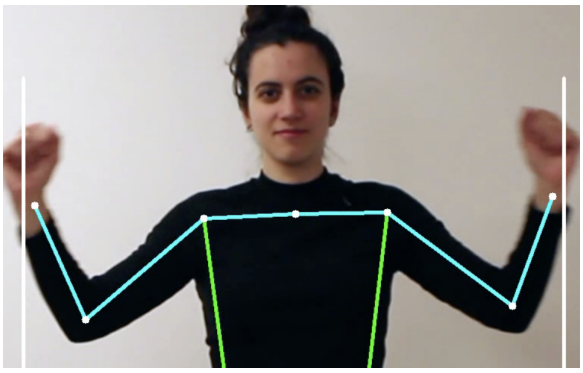


Figura 4.9: Anotação de : *Outward*, para o braço esquerdo e direito. O valor de x do *keypoint* do pulso esquerdo é superior ao valor de x do ombro esquerdo. Por outro lado, o valor de x do *keypoint* do pulso direito é inferiro ao valor de x do ombro direito.

4.3.2 Gestos Verticais

Do mesmo modo para os gestos em função do eixo vertical, foram caracterizados em duas vertentes: posicionamento estático e gestos dinâmicos, ou seja, foi feita a diferenciação entre os movimentos ascendentes e descendentes e o nível de elevação mãos (para posicionamento estático).

Relativamente aos movimentos ascendentes e descendentes, para cada braço, é anotado *Raising Arm* se a coordenada y do *keypoint* do pulso, evoluir para cima, para cada dez *frames* consecutivas. Por outro se essa evolução for para baixo é anotado *Arm Going Down*.

Uma exemplificação do processo de anotação de *Raising Arm* e *Arm Going Down* encontra-se ilustrado nas Figuras 4.10 e 4.11 respetivamente uma vez que, em cada dez *frames*, na Figura 4.10 existe um decréscimo do valor da coordenada y dos pulsos e na Figura 4.11 existe um acréscimo. Estas apresentam um conjunto de quatro *frames* e um gráfico representativo do valor da coordenada y de cada um dos pulsos (esquerdo e direito) num intervalo de trinta *frames*.

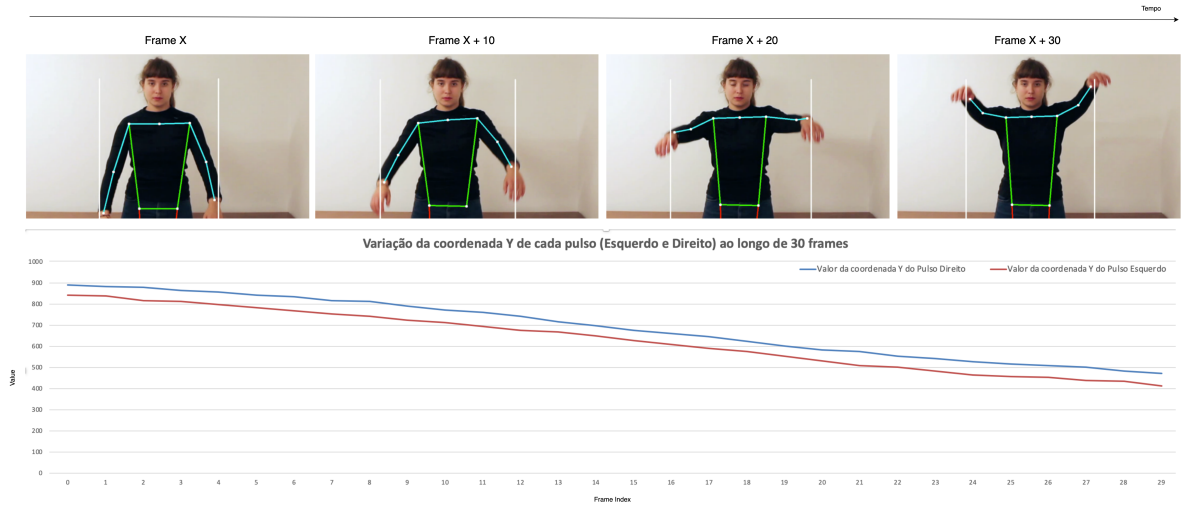


Figura 4.10: Processo de anotação para : *Raising Arm*.

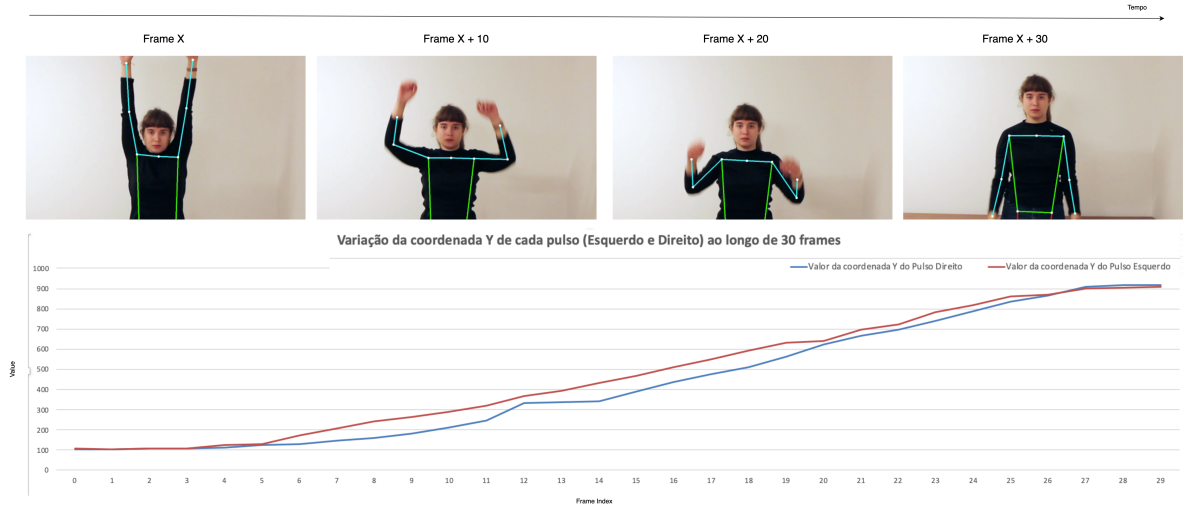


Figura 4.11: Processo de anotação para : *Arm Going Down*.

Para a posicionamento estático dos braços e mãos, foram também definidos 3 grupos: *Arm Up*, *Hand Up* e *Arm and Hand Down*. É anotado *Arm Up* se o valor da coordenada y do *keypoint* cotovelo for inferior valor da coordenada y do *keypoint* ombro. Se o valor da coordenada y do *keypoint* cotovelo for superior ao valor da coordenada y do *keypoint* pulso é denotado *Hand Up*. Uma vez que nenhuma destas condições se verifique é anotado *Arm and*

Hand Down. As figuras 4.12, 4.13 e 4.14 representam as três anotações, *Arm Up*, *Hand Up* e *Arm and Hand Down*, respetivamente.

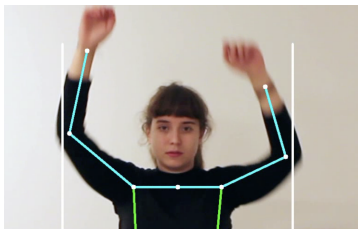


Figura 4.12: Anotação de : *Arm up*. O valor de y do *keypoint*, de cada cotovelo, é inferior ao de cada ombro.

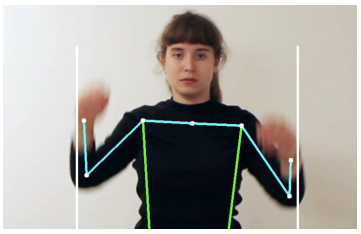


Figura 4.13: Anotação de : *Hand up*. O valor de y do *keypoint*, de cada pulso, é inferior ao de cada cotovelo.

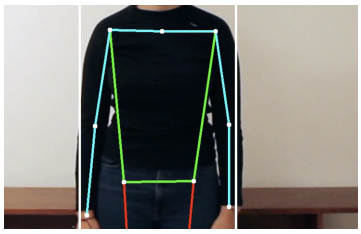


Figura 4.14: Anotação de : *Arm and Hand down*.

4.4 Presença/Ausência da Fala

O algoritmo *Voice Activity Detection* (VAD) é definido como uma maquina de estados finitos com pelo menos dois estados: Presença de fala e Ausência de fala [41].

Nesse sentido, e com a finalidade realizar anotações relativamente aos momentos de pausa ou momentos de fala no discurso por um orador, foi desenvolvido um método que utilizando o algoritmo VAD, baseado em Long *Short-Term Memory Recurrent Neural Networks* (LSTM-RNN), disponível na biblioteca do OpenSmile, procede à extração do valor de probabilidade de existir a presença de fala, que varia entre -1 e 1.

Feita uma análise aos valores de probabilidade extraídos é possível inferir que existe a presença de fala quando o valor da probabilidade é superior a zero e ausência da mesma nos restantes casos. Esta inferência encontra-se descrita na Equação 4.1.

$$\textit{Voice/Non-Voice (VAD)} = \begin{cases} \textit{Voice Detected} & \text{if VAD} > 0 \\ \textit{Silence Detected} & \text{if VAD} \leq 0 \end{cases} \tag{4.1}$$

Depois de obtido o vetor relativo aos momentos de pausa ou fala são ainda calculados o número de ocorrências e respetiva duração, percentagem de ocorrências em relação ao número de *frames* que o vídeo dispõe e ainda o valor médio da duração de cada segmento (pausa e

vozeado).

4.5 Variação da Intensidade da Fala

Uma vez obtidos os vetores relativos a anotação da presença da fala e os vetores de intensidade da voz extraídos usando o método descrito na Secção 3.6 foi desenvolvido um método que, realizado um processamento sobre estes devolve um vetor de anotações relativamente à variação da intensidade da fala.

Este método inicialmente remove todos elementos do vetor de intensidade onde existe a ausência da fala, através da comparação do conjunto de anotações de presença/ausência de fala (utilizando o método da secção anterior), para que não exista a influencia destes na intensidade média. Em seguida, é calculado o valor médio de intensidade que servirá de valor de comparação para anotação das variações da intensidade da voz, isto é, é feita a anotação *Voice Intensity Increase* para todos os elementos que possuam um valor superior ao valor médio e feita a anotação de *Voice Intensity Decrease* para todos os elementos que possuam um valor inferior. Caso a intensidade seja igual a intensidade média é anotado *Voice Intensity Normal*. Estas condições encontram-se expressas na Equação 4.2.

$$\textit{Voice Intensity } (I,\bar{I}) = \begin{cases} \textit{Voice Intensity Increase} & \text{if } I > \bar{I} \\ \textit{Voice Intensity Decrease} & \text{if } I < \bar{I} \\ \textit{Voice Intensity Normal} & \text{if } I = \bar{I} \end{cases} \tag{4.2}$$

4.6 Conclusões

Ao longo deste capítulo são apresentados um conjunto de métodos que permitem a transformação dos dados extraídos para caracterização dos diferentes canais de comunicação em informações, e que passam pela anotação de emoções, orientação da cabeça, gestos, momentos de pausa/vozeados no discurso, e variações de intensidade da voz. Estas anotações são particularmente importantes dado que permitem uma maior legibilidade dos dados e uma melhor identificação das atividades com relevância presentes nos três canais de comunicação considerados.

Capítulo 5

Aplicação dos Métodos Propostos ao Dataset DEP-UA TEDx

Este capítulo ilustra, de forma sucinta, os resultados obtidos por aplicação dos métodos desenvolvidos – nos Capítulos 3 e 4 – ao DEP-UA TEDx Talks Dataset. Estes resultados dizem respeito a dedução de emoções, posicionamento da cabeça e gestos utilizados pelos oradores.

5.1 Visualização dos dados extraídos

Os métodos apresentados nos capítulos anteriores comportam um grande conjunto de dados e informações consideradas relevantes para a caracterização dos diferentes canais de comunicação. Nesse sentido, foi considerado importante, numa primeira fase, propor uma ferramenta de visualização que, permitisse não só avaliar e validar os métodos desenvolvidos, mas também verificar o comportamento das ferramentas computacionais utilizadas no desenvolvimento deste trabalho. Nessa lógica, foi desenvolvida uma ferramenta que realiza o *overlay* dos dados e conjuntos de informação, sobre o vídeo, em tempo real. A Figura 5.1 ilustra a ferramenta de visualização desenvolvida, onde é possível verificar o *overlay* dos dados e conjuntos de informação, como por exemplo, intensidade das *Action Units* e emoção expressa pela face.

Como já referido, uma das motivações para a proposta desta ferramenta de visualização foi a de permitir avaliar e validar os métodos desenvolvidos e as ferramentas utilizadas, para

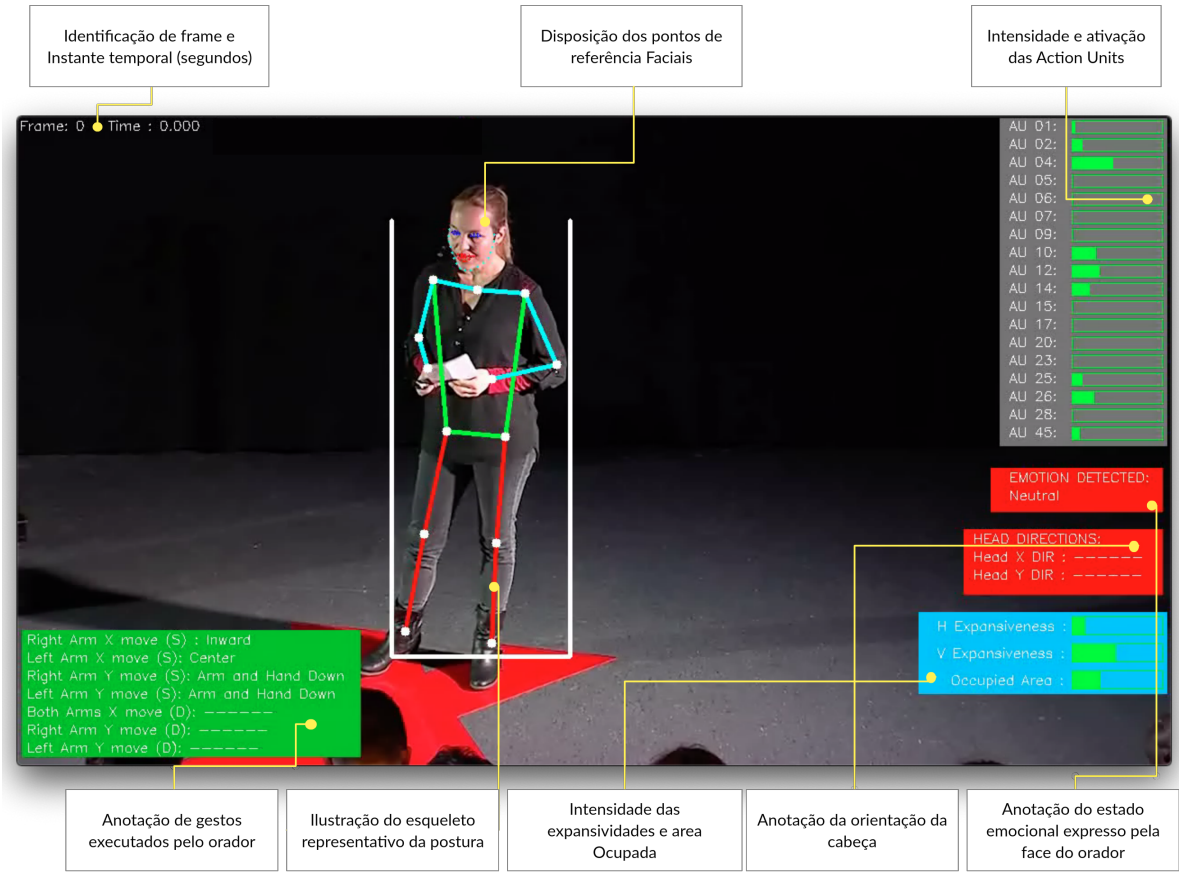


Figura 5.1: Ilustração da ferramenta de visualização e identificação da disposição dos dados e conjuntos de informação extraídos.

que pudessem ser detetados e resolvidos problemas como falta da deteção do esqueleto ou da face humana. A Figura 5.2 ilustra um comportamento errático da ferramenta OpenPose, onde existe a falha da deteção do esqueleto humano. Ainda que alguns dos erros de computação pudessem ser detetados por análise dos valores numéricos obtidos, a visualização do que eles significam, sobre o vídeo, permitiu perceber, por exemplo, as possíveis causas (e.g., mudanças de plano).

Por outro lado, a visualização dos dados extraídos e anotações, em contexto, permitiu aferir da sua validade e suportou o processo de desenvolvimento e teste dos métodos propostos.

Finalmente, esta ferramenta pode suportar, no futuro, uma análise crítica, em contexto, por parte de investigadores, e.g., da área da Educação, sobre os dados e anotações propostos.

Figura 5.2: Ilustração da ferramenta de visualização: Comportamento errático da ferramenta OpenPose na deteção do esqueleto.

5.2 Deteção de Emoções

De forma a ilustrar as anotações de atividades relevantes relativamente os estado emocional do orador, tendo em consideração as expressões faciais, foram aplicados, ao DEP-UA TEDx Talks Dataset, os métodos de extração das intensidade das *Action Units* e subsequente dedução das emoções. As Figuras 5.3 e 5.4 apresentam um gráfico relativo à variação das intensidades das sete emoções consideradas, acompanhadas com três *frames* ilustrativas da face do utilizador, num determinado momento, para um conjunto de *frames* consecutivas para dois vídeos do dataset.

Na Figura 5.3 é exemplificado de que forma é atribuído o estado emocional de *"Happy"* ao orador para trinta *frames*. Este facto é concretizado devido ao valor de intensidade desta emoção ser superior, ao longo de todas as *frames*, a todas as outras emoções. É de realçar ainda que, na fase final deste conjunto, existe um decréscimo no valor da intensidade da emoção *"Happy"* que é refletido pela *frame* representativa da face do orador mais à direita, na figura.

Figura 5.3: Ilustração da variação das intensidades da emoções para um conjunto de trinta *frames*.

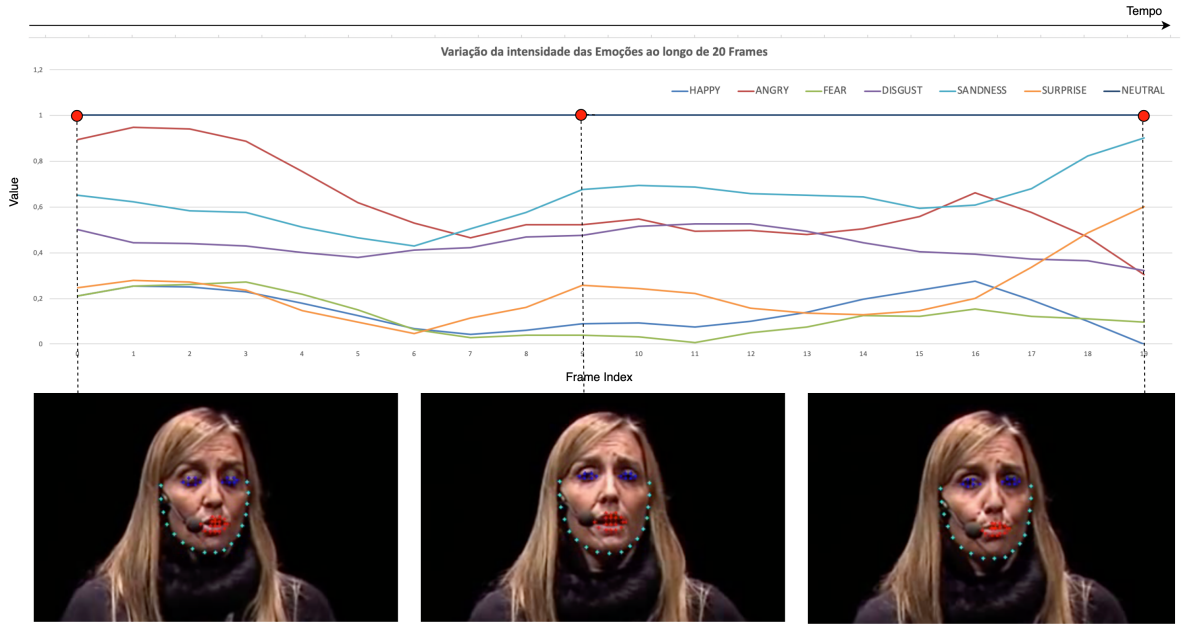


Figura 5.4: Ilustração da variação das intensidades da emoções para um conjunto de vinte frames *frames*. No final do gráfico existe o crescimento da intensidade da emoção *"Sadness"* que é o reflexo da *frame* mais a direita da figura.

Por outro lado, na Figura 5.4 é exemplificado como é atribuída a anotação dos estado emocional de "*Neutral*", para um conjunto de vinte *frames*, dado que as intensidades das diferentes emoções não apresentam valores suficientemente intensos para a atribuição de uma emoção. É ainda possível evidenciar que, no final do conjunto de *frames*, existe o crescimento da intensidade da emoção *"Sadness"* que é o reflexo da expressão facial exibida na *frame* mais à direita da figura.

5.3 Estimativa da Posição da Cabeça

De maneira a representar a anotação da orientação da cabeça do orador em torno do eixo vertical e horizontal, foram aplicados os métodos relativos a extração dos valores de *Pitch* e *Yaw* e transformação em movimentos da cabeça.

Na Figura 5.5 é apresentado um gráfico correspondente a variação do valor de *Yaw* ao longo de parte de um vídeo (trinta *frames*), com a exposição de quatro *frames* representativas dos valores de *Yaw* identificados no gráfico. No processo de variação da orientação da cabeça, é para este período de tempo feita a anotação de *"Head Moving Right"* dado que existe uma variação positiva dos valores de *Yaw*. É ainda possível evidenciar através da análise da figura que existe uma grande variação entre as *frames* 0 e 10 e uma pequena variação entre as *frames* 19 e 29, que é justificável mediante a observação das duas imagens mais à esquerda e mais à direita, respetivamente.

Figura 5.5: Ilustração da variação do valor de *Yaw* com imagens associadas aos valores identificados no gráfico para a variação da cabeça da esquerda para a direita

Em relação à orientação da cabeça em torno do eixo vertical a figura 5.6 é o reflexo da anotação de *"Head Moving Down"* e *"Head Moving Up"* para um conjunto de quarenta *frames*. Dado que nas vinte primeiras *frames* existe uma variação positiva do valor de *pitch* é feita a anotação de *"Head Moving Down"*. Por outro lado, nas ultimas vinte *frames* o valor do *pitch* tem uma variação negativa que se reflete na anotação de *"Head Moving Up"*, para o restante conjunto de *frames* .

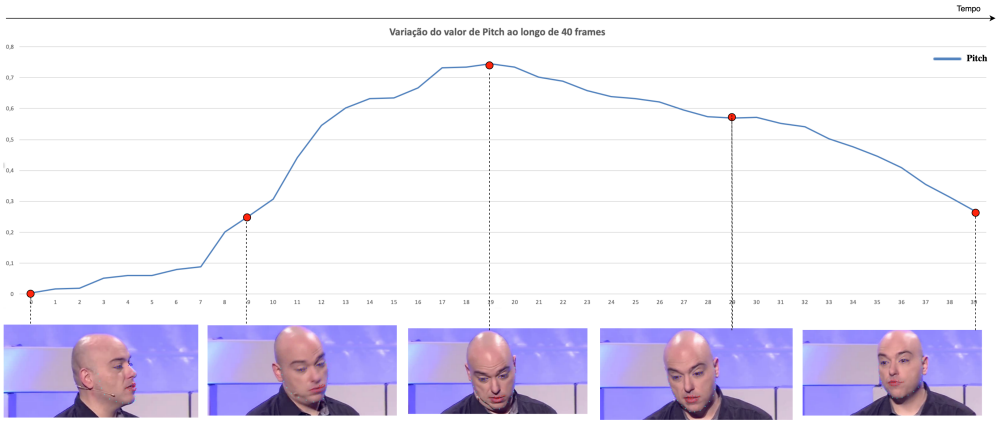


Figura 5.6: Ilustração da variação do valor de *pitch* com imagens associadas aos valores identificados no gráfico para a movimentação da cabeça de cima para baixo e de baixo para cima respetivamente

5.4 Gestos

De forma a ilustrar a anotação de gestos horizontais e verticais, mas também a variação das expansividades e área ocupada foram também aplicados os métodos correspondestes a extração dos pontos de referência corporais, obtenção dos valores de expansividades e área ocupada e obtenção de gestos.

A Figura 5.7 ilustra um conjunto de *frames* para as quais foi feita a anotação de gestos em função do eixo vertical e horizontal. Usando, como exemplo de análise, o primeiro conjunto de dez *frames* é possível observar que existe um decréscimo do valor da distancia entre pulsos, que se reflete na anotação de *"Approach Hands"*. Consequentemente existe também, um decréscimo relativamente a expansividade horizontal. É também possível evidenciar uma pequena, mas crescente, variação no nível de elevação da pulso direito que se traduzirá numa anotação de *"Raising Arm"*. Estas duas anotações podem ser verificadas tendo em conta

a diferença da posição dos braços do orador observáveis através das duas imagens mais a esquerda, da figura.

Por outro lado, nos últimos dois conjuntos de dez *frames* é possível constatar que existe um crescimento do valor da distancia entre pulsos que se traduz numa anotação de *"Separate Hands"*. Este facto é possível de verificar através da observação das diferenças entre as três imagens mais a direita na figura. É ainda possível observar o crescimento da expansividade horizontal e a área ocupada devido ao crescimento da distância entre pulsos.

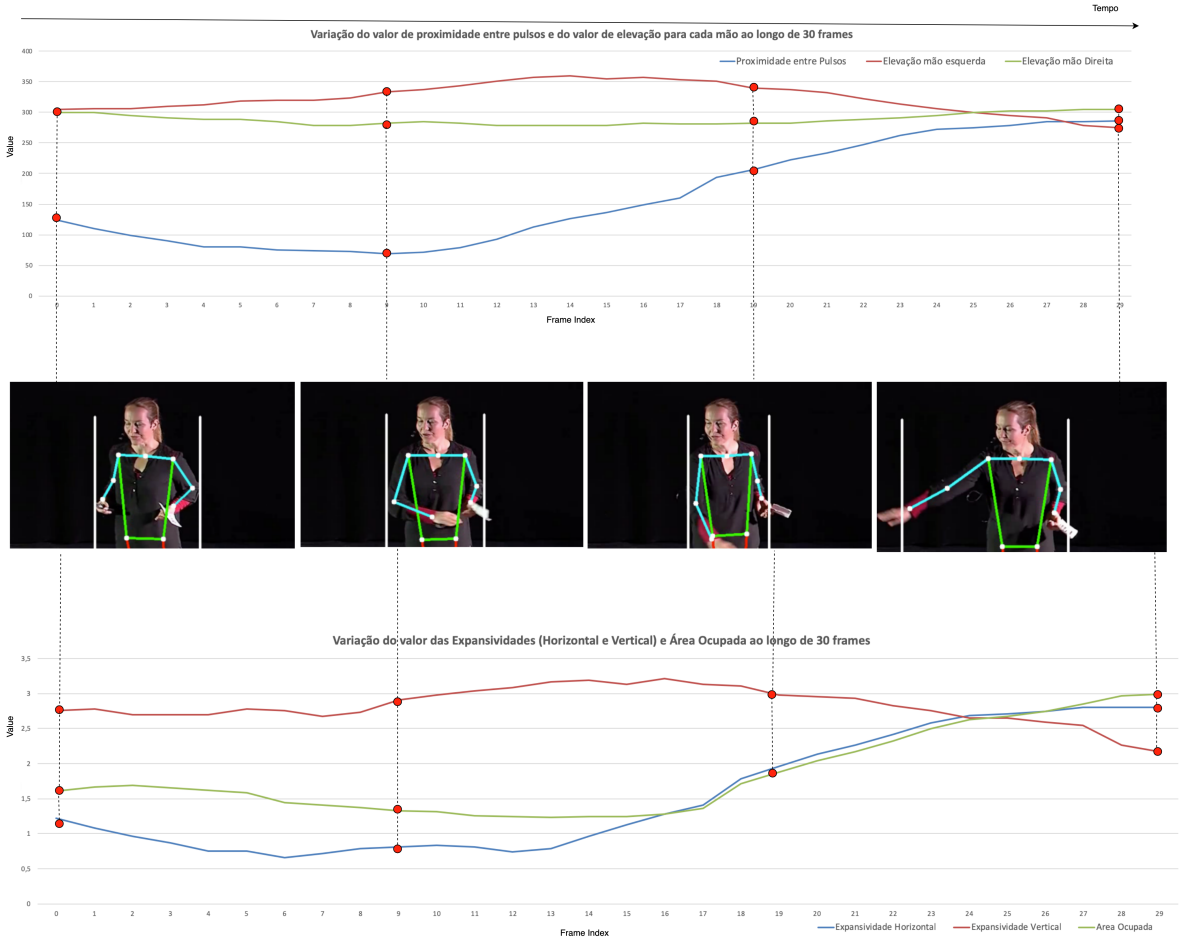


Figura 5.7: Ilustração da variação das expansividades, área ocupada (gráfico superior) e variação do nível de proximidade entre pulsos e nível de elevação para cada pulso com a utilização representativa de quatros *frames* que identificam os valores nos dois gráfico

5.5 Disponibilização dos Contributos a Terceiros

A disponibilização de um dataset enriquecido com informações relativas aos três canais de comunicação é essencial neste cenário do DEP-UA TEDx Talks, onde a comunicação em público é a temática central. Nesse sentido, foi desenvolvido um método que, uma vez extraídas as características descritoras dos canais de comunicação e anotações de atividades relevantes são, para cada vídeo, gerados um conjunto de três ficheiros *JavaScript Object Notation* (JSON) que contém as informações relativas aos três canais de comunicação: Face, Postura e Voz. Desta forma, os dados e informação resultantes deste trabalho podem ser facilmente disponibilizados a terceiros, num formato que é legível e facilmente importado para ser usado no âmbito de outros trabalhos de análise computacional.

5.6 Conclusões

Neste capítulo, foram apresentados alguns exemplos ilustrativos dos resultados obtidos por aplicação dos métodos desenvolvidos a um dataset de comunicação em público para extração de características e atividades relevantes. Foi, ainda, apresentada uma ferramenta de visualização para validação dos métodos e ferramentas computacionais considerados e que poderá contribuir para suportar, futuramente, uma análise crítica dos contributos deste trabalho por investigadores de outras áreas, e.g., Educação.

Capítulo 6

Exemplos de Aplicação

O trabalho desenvolvido, descrito nos capítulos anteriores, cria uma base quantitativa (objetiva) sobre a qual se pode trabalhar para aumentar a nossa compreensão sobre diferentes aspetos, não só no que se refere à comunicação em público, mas também sobre a forma como a assistência olha e classifica os oradores. Tendo isto em mente, neste capítulo serão apresentados alguns exemplos de possíveis caminhos que podem ser seguidos fazendo uso dos conjuntos de dados extraídos para cada canal de comunicação considerando o DEP-UA TEDx Talks Dataset. Estes exemplos não pretendem ser exaustivos, mas apenas deixar indicações sobre o potencial dos contributos deste trabalho e lançar questões que motivem novas direções dos trabalhos futuros.

6.1 Estudo Exploratório das Características de Comunicação

Uma vez obtido um conjunto de dados que descrevem a atividade em cada canal de comunicação, pode ser interessante, para o estudo exploratório dos canais de comunicação, a adoção de métodos computacionais.

6.1.1 Posturas Constritas/Expansivas como Indicadores de Confiança e Assertividade

Segundo a literatura, como referido no capítulo 2, as pessoas mais expansivas e abertas projetam sinais de confiança e assertividade. Dessa forma, usando os dados da área ocupada, recorrendo a medidas estatísticas (média, desvio padrão, máximo e mínimo), foram aplicados

métodos de *Machine Learning* não supervisionados com o intuito de distribuir os oradores do DEP-UA TEDx Talks Dataset, na sua prestação geral, para a avaliação relativamente a estes julgamentos sociais presentes na comunicação. Assim, obteve-se a distribuição dos oradores, relativamente à área ocupada, utilizando o algoritmo *Agglomerative Clustering* [42] com distância *Ward*, ilustrada na Figura 6.1 em que se verifica a distribuição dos mesmos em três grandes grupos.

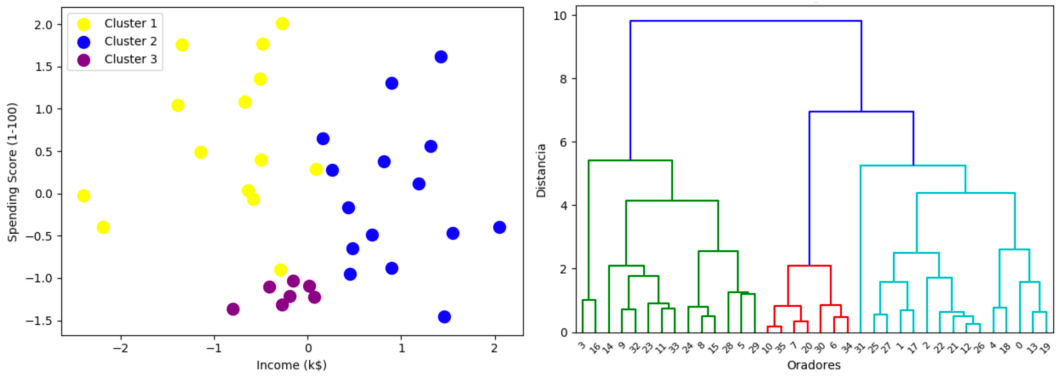


Figura 6.1: Ilustração do *Cluster* e dendrograma para a distribuição dos oradores pela *feature*: Área Ocupada.

Uma vez obtida a distribuição dos oradores pelos diferentes grupos, foi calculado o valor médio, máximo, mínimo e desvio padrão de cada um dos grupos para o valor da área ocupada, que se traduziu nos valores apresentados na Tabela 6.1. É possível verificar que existe efetivamente um agrupamento dos oradores por esta *feature*, onde os mais expansivos, em média, encontram-se no *Cluster* 3 que é assinalado pela cor azul mais claro no dendrograma.

	CLUSTERS		
Área Ocupada	Cluster 1	Cluster 2	Cluster 3
Média	27,4%	30,8%	37,2%
Desvio Padrão	5,4%	2,8%	4,2%
Máximo	40,8%	40%	52%
Mínimo	14%	23,2%	26,2%

Tabela 6.1: Valores estatísticos da *feature* área ocupada para os três *Clusters*. *Cluster* 1 identificado pela for verde , *Cluster* 2 identificado pela cor vermelha e *Cluster* 3 identificado pela cor azul claro no dendrograma.

Para provar o conceito que é referido na literatura relativamente às posturas constritas e expansivas, foram obtidos para cada um dos *Clusters* calculadas as pontuações médias dos julgamentos sociais de assertividade e confiança, que se encontram representadas na Tabela 6.2, recorrendo à anotação do dataset.

	CLUSTERS		
Julgamento Social	Cluster 1	Cluster 2	Cluster 3
Assertividade	57	50	48
Confiança	61	51	49

Tabela 6.2: Pontuação média dos valore de assertividade e confiança para os três *Clusters* *Cluster* 1 identificado pela for verde, *Cluster* 2 identificado pela cor vermelha e *Cluster* 3 identificado pela cor azul claro no dendrograma.

.

Tendo sido obtida a pontuação média referente a cada um destes aspetos para os diferentes *Clusters* é possível verificar que em média os oradores do *Cluster* 1, representados pela cor verde no dendrograma, são melhor classificados que os oradores dos *Clusters* 2 e 3. Dessa forma e comparativamente com os valores da área ocupada representados na Tabela 6.1, pode afirmar-se, ainda que o dataset carece de dados, que as pessoas não aparentam interpretar julgamentos sociais como assertividade e confiança apenas com base nas posturas expansivas/constritas. Assim, esta análise apresenta uma pista muito interessante, isto é, a comunicação não se restringe a apenas a um aspeto, mas possivelmente a um conjunto de aspetos presentes na comunicação verbal e não verbal.

6.1.2 Perceção de Dominância

Outra abordagem interessante, pode consistir em realizar o processo inverso da secção anterior, ou seja, realizar o agrupamento dos oradores pelo valor das anotações que o dataset dispõe, usando o valor médio e o desvio padrão da anotação, e verificar o comportamento dos mesmos no que diz respeito aos diferentes canais de comunicação.

Nessa lógica, foi realizando o agrupamento dos oradores pelo julgamento social de dominância que se obteve a distribuição ilustrada na Figura 6.2. Nesta figura constata-se a existência de dois grandes grupos, que complementados com a tabela 6.3 é possível verificar

que os oradores do *Cluster* 1, identificados pela cor verde no dendrograma são, em média, mais dominantes relativamente aos oradores do *Cluster* 2.

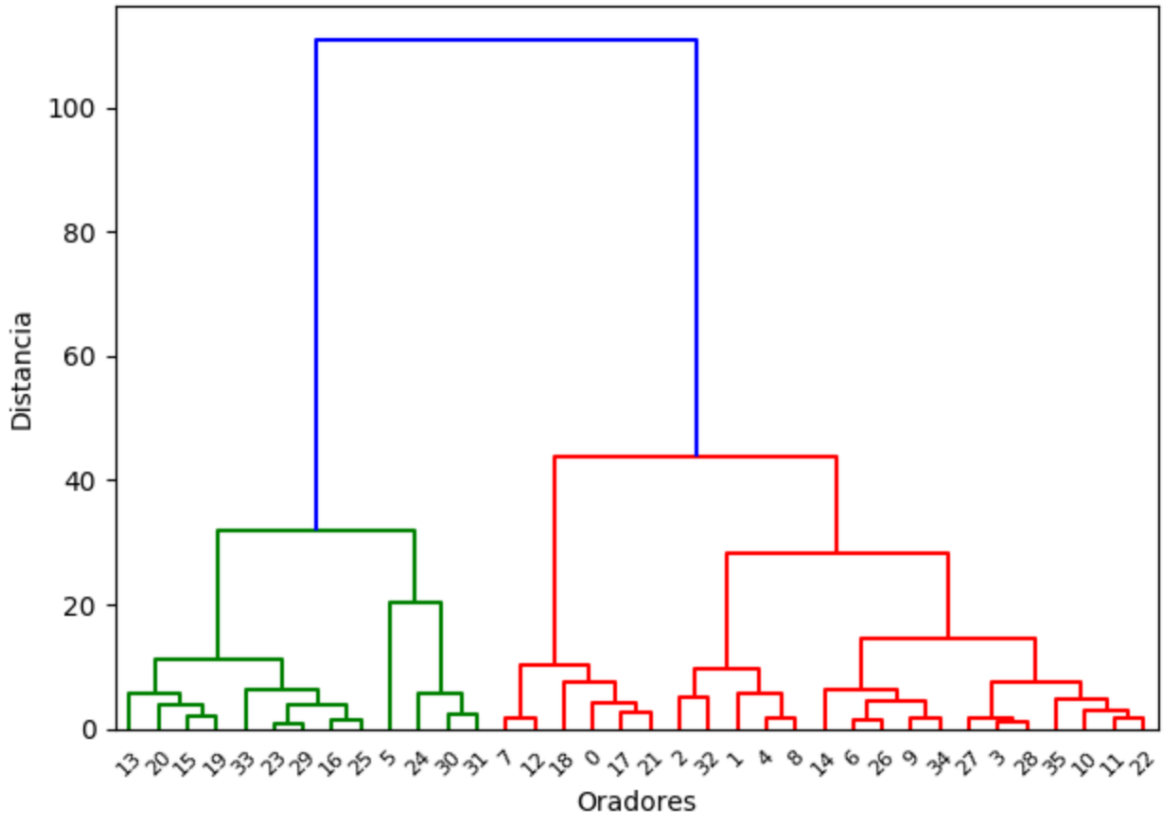


Figura 6.2: Ilustração do dendrograma para a distribuição dos oradores pela anotação: Dominância.

	CLUSTERS	
Dominaçia	Cluster 1	Cluster 2
Valor Médio	62,8	35,7
Desvio Padrão	24,6	24,2

Tabela 6.3: Identificação do valor médio e desvio padrão para o julgamento social de dominância para os dois *Clusters*. *Cluster* 1 identificado pela for verde e Cluster 2 identificado pela cor vermelha no dendrograma.

Uma vez distribuídos os oradores pelos diferentes grupos, no sentido de compreender o que

influencia a perceção de dominância nos seres humanos, foram analisadas algumas características extraídas dos diferentes canais de comunica como: expansividade horizontal, expressões faciais, posição da cabeça e momentos de silêncio. Os resultados para estas diferentes *features* encontram-se descritos na Tabela 6.4.

		CLUSTERS	
Features		Cluster 1	Cluster 2
Expansividade Horizontal	Média	22%	23%
	Desvio Padrão	6%	5%
Expressões Faciais (Action Unit 04)	Máximo	45%	48%
Expressões Faciais (Action Unit 12)	Máximo	36%	34%
Pitch (Face)	Máximo	74 º	67 º
	Mínimo	-14 º	- 26 º
Silêncio / Pausas no Discurso	Ocorrências	9%	12%
	Duração Média	0.4s	0.7s

Tabela 6.4: Análise das *features* para caracterização do julgamento social: Dominância. *Cluster* 1 identificado pela for verde e Cluster 2 identificado pela cor vermelha no dendrograma.

Analisando os valores para a expansividade horizontal, para os diferentes *clusters*, é notório que não existe uma variação significativa destas. Da mesma foram, analisando as expressões faciais, é também notório que não existe uma diferença significativa para o valor máximo de intensidade para as duas *Action Units* analisadas. Relativamente à posição da cabeça em função do eixo vertical, as pessoas menos dominantes voltam a cabeça para baixo de forma mais acentuada que os mais dominantes. Por último, relativamente aos momentos de silêncio os mais dominantes, ainda que a diferença não seja muito significativa, aparentam fazer menos pausas e com uma duração menor do que as pessoas menos dominantes.

Dessa forma, tal como referido na secção anterior, as pessoas não aparentam perceber o fator de dominância baseando-se apenas nas posturas expansivas, mas também tendo em conta outros aspetos presentes na comunicação verbal e não verbal.

6.2 Métodos Computacionais para Previsão

Tendo em conta a importância de uma boa capacidade de comunicação em público, seria interessante a aplicação de métodos computacionais com a finalidade de tentar fornecer algum *feedback* automático para a classificação prestação de um orador.

Utilizando o DEP-UA TEDx Talks Dataset e recorrendo a métodos de *Machine Learning* supervisionados, dado que estes têm sido fortemente explorados pela comunidade na temática da classificação, nesta secção são apresentados alguns exemplos ilustrativos que podem ser aplicados aos conjuntos de informações extraídos dos diferentes canais de comunicação.

6.2.1 Previsão de Aspetos de Comunicação: Postura e Gestos

Um exemplo da aplicação, recorrendo a métodos supervisionados de *Machine Learning*, pode consistir na previsão de aspetos de comunicação, como por exemplo, postura e gestos adotados. Nesse sentido e, beneficiando da anotação do dataset relativamente a postura e gestos foi, para cada um dos oradores, obtida a pontuação referente a cada um destes aspetos, através do cálculo da curva média de variação ditada pelas quarenta pessoas que participaram na experiência (anotação do dataset), resultando num valor entre 0 e 100. De seguida, os oradores foram distribuídos por quatro grupos, relativamente estes dois aspetos. A Tabela 6.5 ilustra a forma como foi efetuada a distribuição dos oradores nos quatro grupos com base na pontuação média de postura e gestos.

Grupo	Pontuação Média dos Julgamentos Sociais
1	Igual ou Superior a 75
2	Igual ou Superior a 50
3	Igual ou Superior a 25
4	Restantes

Tabela 6.5: Distribuição dos oradores pelos quatro grupos com base na pontuação, de postura e gestos, ditada pelos participantes.

Em seguida, para cada um dos trinta e seis oradores, que compõem o dataset, foram obtidas medidas estatísticas (média, desvio padrão e máximo) dos valores referentes às expansividades, velocidade dos gestos e área ocupada. Posteriormente foi aplicado o método de aprendizagem

supervisionada para classificação, utilizando como conjuntos de entrada os descritores da comunicação através do corpo e saídas desejadas a distribuição pelos quatro grupos. A Tabela 6.6 apresenta os resultados obtidos utilizando o algoritmo *Support-vector machine* (SVM) [43] e *Cross-validation* para validação do modelo, para os dois aspetos de comunicação.

Algoritmo	Postura	Gestos
SVM (kernel = RBF)	75 % ACC (+/- 1%)	78 % ACC (+/- 5%)

Tabela 6.6: Resultados Obtidos na classificação para: Postura e Gestos.

Naturalmente, os resultados não devem ser tomados como totalmente representativos, dado que o dataset possui, ainda, poucos dados, contando apenas com 36 amostras, para este efeito, servindo apenas, como um exemplo de aplicação. Para melhores e mais confiáveis resultados seria interessante a existência de volumes dados superior.

6.2.2 Conjuntos de informação como avaliadores de Confiança

Outra anotação que o dataset contém é o nível de confiança percebido pelos participantes relativamente a prestação dos oradores, em tempo real. Nesse sentido, seria também interessante o estudo de métodos computacionais que permitissem a previsão deste julgamento social.

Dessa forma, para cada um dos oradores, foram selecionados conjuntos de dados dos três canais de comunicação, em cada instante (*frame* por *frame*) que se encontram descritos na Tabela 6.7. Este conjunto de dados foram utilizados para dados de treino uma vez que estes aparentemente apresentam um impacto significativo na perceção do nível de confiança.

	Corpo	**Face**	**Voz**
Descritores	- Expansividade Horizontal - Expansividade Vertical - Área Ocupada	- *Action Unit 12* - *Action Unit 04* - *Pitch* - *Yaw*	- *MFCC 1 -12,* - *Pitch* - *Loudness,* - *HNR* - *Intensity ,* - *VAD*

Tabela 6.7: Conjuntos de dados selecionados (dos diferentes canais de comunicação) para treino para previsão do nível de confiança.

Posteriormente, foi utilizada a mesma abordagem descrita na secção anterior para a dis-

tribuição dos oradores pelos quatros grupos, através do cálculo da curva média de confiança ditada pelos participantes da experiência, com o intuito de atribuir uma saída desejada aos conjuntos de dados de treino. Contudo, esta divisão é feita a cada instante, sendo possível um orador variar de grupo para grupo em instantes (*frames*) diferentes.

Todavia, o nível de confiança não é algo que possa ser interpretado e respondido de modo instantâneo. Nesse sentido, ponderou-se qual a janela de variação que melhor se ajustava ao conjunto de dados de treino (descritores dos canais de comunicação) em função do tempo de resposta das pessoas que participaram na experiência.

A figura 6.3 apresenta os resultados para as diferentes janelas de variação utilizado o algoritmo *k-Nearest Neighbors* (KNN) [44] e *Learning Curve* para validação do modelo usando 70% dos dados para treino e 30% para teste.

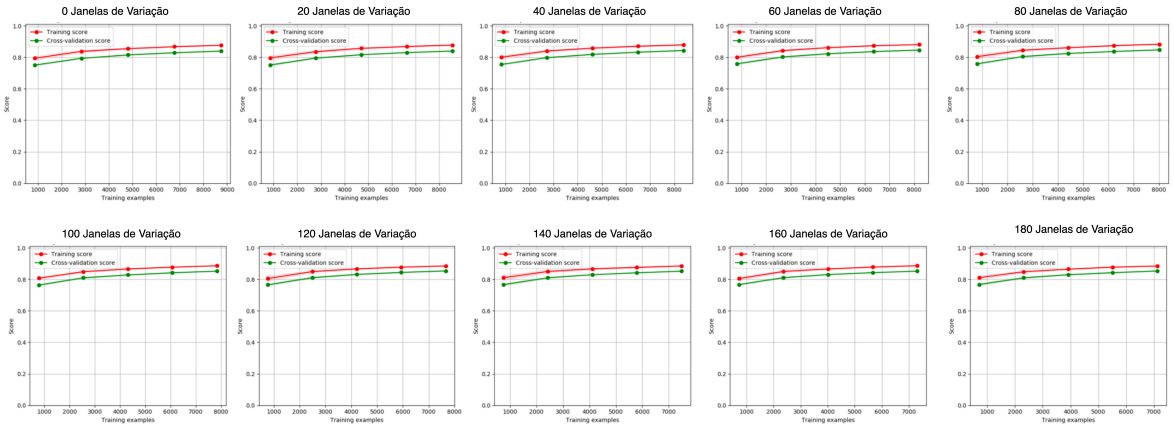


Figura 6.3: Resultados obtidos paras as diferentes janelas de variação.

Contudo, a abordagem seguida e o resultado obtido não revelaram o impacto desejado, dado que era expectável uma variação significativa em diferentes janelas de variação. Uma análise posterior levou a que se colocasse a hipótese de, não sendo a confiança um valor absoluto, e que é interpretado de maneiras diferentes pelos seres humanos, que essa variabilidade pudesse estar a ocultar o que se passa.

Para fazer face a este problema, foi pensada uma outra abordagem para que o valor da confiança, calculada com base nas votações de todos os participantes, para cada vídeo, não fosse a de considerar o seu valor absoluto, mas sim como uma curva de probabilidade de a confiança, subir ou manter-se.

Esta nova abordagem consistiu em avaliar, para cada participante, alterações no nível de confiança. Em cada análise, por participante, é criado um vetor de zeros de dimensão igual ao vetor de votações. Em seguida para cada instante, é analisado alterações do nível de confiança. Caso este aumente ou decresça em relação ao instante anterior é, ao vetor de zeros, modificado o valor para 1 e -1, respetivamente, nas sessenta *frames* que se seguem. Depois de obtidos os vetores de variação do nível de confiança de todos os participantes, é feito o somatório de cada instante, obtendo uma curva resultante que dá uma indicação de quando há ou não maior concordância entre os participantes numa alteração (positiva ou negativa) do valor da confiança.

As Figuras 6.4, 6.5 e 6.6 apresentam uma ilustração da construção destas curvas para um intervalo de 5 *frames* (em vez das 60 usadas).

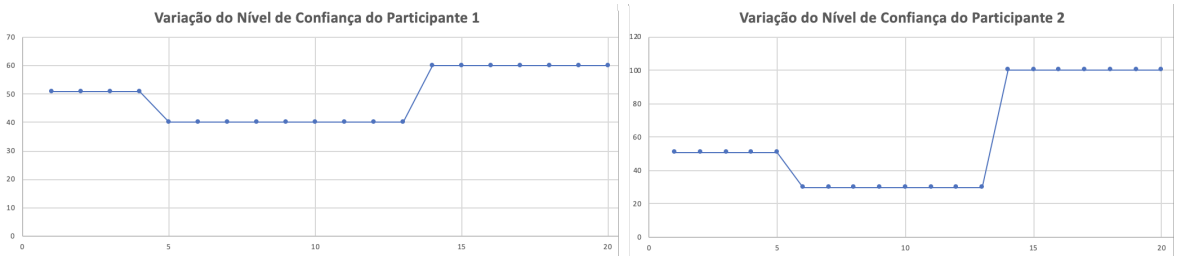


Figura 6.4: Representação gráfica da variação do nível de confiança percebida por 2 participantes.

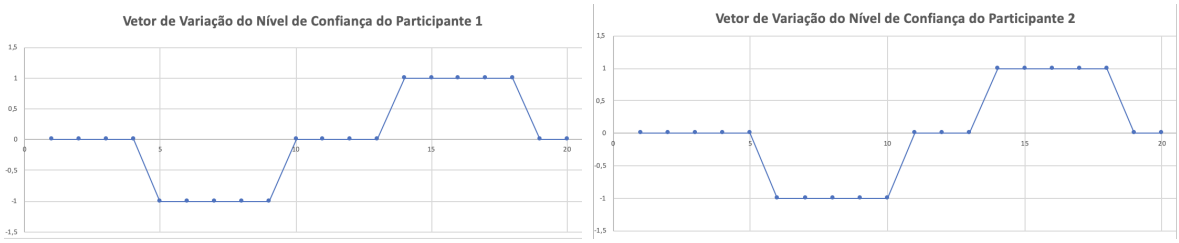


Figura 6.5: Representação gráfica da curva de variação do nível de confiança percebida por 2 participantes.

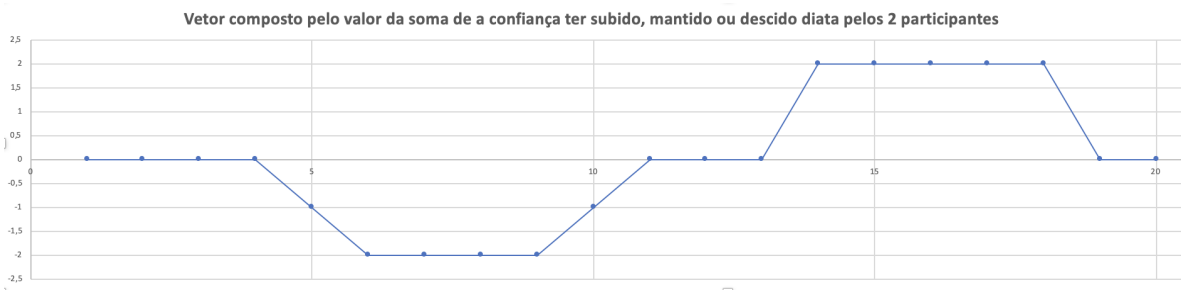


Figura 6.6: Representação gráfica da curva resultante que dá uma indicação de quando há ou não maior concordância entre os participantes numa alteração (positiva ou negativa) do valor da confiança.

De maneira a atribuir uma saída desejada ao conjunto de dados de treino, em cada instante, foi feita a distribuição dos oradores, pelos três grupos, baseada no valor probabilístico, que é obtido através da divisão da curva resultante pelo total de participantes, usando a abordagem descrita na Tabela 6.8.

Gurpo	Valor Probabilístico
1	Igual ou Superior a 0.3
2	Igual ou Inferior a 0.3
3	Restantes

Tabela 6.8: Distribuição dos oradores pelos três grupos com base no valor de probabilidade.

Uma vez feita a distribuição dos oradores por em cada um dos grupos em instantes diferentes e utilizados os conjuntos de dados descritos na Tabela 6.7 foi aplicado o método de *Machine Learning* supervisionado usado na abordagem anterior. Os resultados deste método, para as diferentes janelas de variação, encontram-se descritos na Figura 6.7.

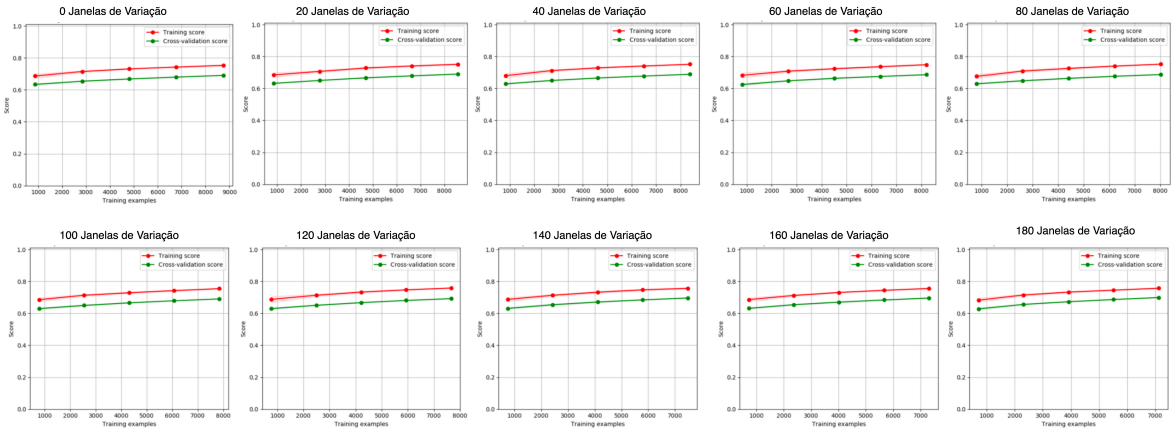


Figura 6.7: Resultados obtidos paras as diferentes janelas de variação.

Do mesmo modo os resultados obtidos, tal como na abordagem anterior, não revelam uma variação significativa entre diferentes janelas de variação.

6.3 Conclusões

Neste capítulo são apresentados um conjunto de métodos que podem ser aplicados aos conjuntos de dados extraídos. Os métodos ilustrados neste capítulo, ainda que sejam abordagens simples, permitiram obter alguma compreensão sobre os dados e lançam ideias e questões que podem orientar a investigação futura.

Do ponto de vista do estudo dos fatores determinantes da qualidade da comunicação em público, os resultados do estudo exploratório, para este dataset, não demonstram uma relação tão clara, como a literatura parece apontar, entre a expansividade da postura e a dominância, confiança e assertividade transmitida. Naturalmente, a quantidade de dados é ainda pequena, mas isto parece apontar para a possibilidade de que a perceção da prestação de um orador é obtida à custa de uma análise multidimensional em que todos os fatores considerados são pesados pela assistência. Ou seja, ainda que a postura seja um fator revelador, o seu impacto é mediado pelo que se passa noutras vertentes. Este aspeto deve motivar um estudo mais aprofundado sobre como a consideração pesada de dados dessas diferentes dimensões (canais) pode ajudar a compreender esse fenómeno. Naturalmente, a proposta de novas medidas para a caracterização do que se passa, em cada um dos canais de comunicação, e das sinergias entre

eles (e.g., voz e gestos) pode ser, também, um passo importante.

Outro dos aspetos que se tornou evidente durante a execução dos exemplos aqui apresentados, é que existe, apesar de tudo, alguma uniformidade de comportamento entre os diferentes oradores, ou seja, não existem casos muito pronunciados dos extremos da capacidade de comunicação na base de dados o que pode, também, tornar mais complexa a tarefa de análise.

Finalmente, e como explicado, os dados anotados referentes à confiança, em tempo real, e aos julgamentos sociais, são difíceis de trabalhar dada a forma relativa como cada participante usa as escalas. Este pode, também, ser um fator que não deixa ter uma visão mais clara sobre a prestação dos oradores. Foram feitas algumas tentativas simples de normalizar, por exemplo, o valor da confiança, mas o cálculo desta, a partir das votações individuais dos participantes, que anotaram os vídeos, deve, do nosso ponto de vista, ser alvo de análise.

Capítulo 7

Conclusões e Trabalho Futuro

Neste capítulo é feita uma discussão do trabalho desenvolvido nesta dissertação, expondo dde que forma os objetivos propostos foram atingidos e fazendo algumas considerações relativamente aos resultados obtidos. Por fim, neste capítulo, são ainda apresentados alguns caminhos possíveis para trabalho futuro.

7.1 Discussão

Os objetivos definidos para o trabalho a desenvolver nesta dissertação foram atingidos de forma satisfatória. Tendo em conta que a comunicação em publico é uma área que ainda é muito pouco explorada, este trabalho conseguiu contribuir, de forma positiva para os avanços da mesma.

Com base na literatura, foi possível selecionar um conjunto de características presentes na comunicação verbal e não verbal que, de grosso modo, apresentam um impacto relevante na comunicação do ser humano em público, dado que é percetível que a comunicação é multimodal, ou seja, não se restringe apenas a um aspeto, mas sim a uma junção de vários.

Através do uso de diferentes ferramentas computacionais foi possível extrair, de forma plausível, um conjunto de elementos para a caracterização de três diferentes canais utilizados na comunicação humano-humano: corpo, face e voz. Através destas, foi ainda possível a implementação de alguns métodos de anotação de atividade considerada relevante que facilitam a descrição e interpretação de um conjunto de ações/conteúdos que ocorrem durante a comunicação. De realçar, os métodos propostos são generalizáveis a outros vídeos, isto é, todos os

métodos enumerados que foram desenvolvidos não servem apenas para o dataset utilizado.

A disponibilização de uma base de dados audiovisual onde a temática da comunicação em publico é primordial, complementada com as anotações agora disponibilizadas poderá servir para os avanços do estudo nesta área, mas também servir de base para a criação de novas e mais refinadas bases de dados.

No que diz respeito à forma como os diferentes conjuntos de dados disponibilizados podem ser usados, foram apresentados alguns exemplos ilustrativos de como podem ser úteis, por exemplo, para estudos exploratórios. Relativamente ao que é referido na literatura, indicando as posturas expansivas como sinal de confiança, dominância e assertividade, recorrendo aos exemplos ilustrativos, ainda que o dataset que foi utilizado apresente poucos dados, é possível referir que não sobressai uma relação direta entre expansividade e a atribuição destes julgamentos sociais. Por outro lado, relativamente aos métodos automatizados para fornecimento de *feedback* estes apresentam resultados que, claramente, poderão ser melhorados, no futuro. As duas abordagens seguidas, para a atribuição de saídas desejadas aos conjuntos de dados de treino podem não ser a melhor abordagem possível dado que a avaliação da confiança não é um valor absoluto e depende do referencial pessoal utilizado por cada ser humano. Contudo, os resultados obtidos e as abordagens propostas para obter as curvas de confiança, a partir dos dados anotados pelos participantes, para cada vídeo do dataset, poder ser tomadas como pistas importantes para motivar a proposta de formas alternativas de calcular os valores de confiança a partir das votações dos diferentes participantes.

7.2 Trabalho Futuro

Relativamente a trabalho futuro, embora a dissertação apresente um conjunto de métodos que permitem descrever, de forma objetiva, os conteúdos/ações de diferentes canais de comunicação, estes podem ainda ser trabalhados para a obtenção de descritores e anotações mais sofisticados dos diferentes canais de comunicação. Alguns exemplos de possíveis aspetos a explorar são:

- Refinamento da forma como é tratada a expansividade horizontal considerando aspetos como a simetria.
- Exploração mais aprofundada da comunicação verbal, através dos recursos áudio ex-

traídos, dados que estes, por limitação de tempo, não foram tão explorados quanto o pretendido.

- Considerar métodos supervisionados de *Machine Learning* para a determinação de estados emocionais a partir das expressões faciais.

Outro aspeto que deve ser tido em conta é a natureza da base de dados audiovisual, isto é, é necessário compreender qual o impacto de usar vídeos retirados de ambientes controlados e como a respetiva duração dos mesmos pode afetar diretamente os resultados obtidos nas experiências.

No que diz respeito a anotação do dataset, relativamente aos diferentes julgamentos sociais, por parte dos participantes que os visualizaram, existem fatores que talvez mereçam uma análise. Um deles passará por avaliar qual a melhor forma de realizar as diferentes anotações uma vez que os valores podem ser sujeitos a interpretações ambíguas. Outro, pode ser a aquisição das anotações uma vez que os valores obtidos, por exemplo para a assertividade, são passíveis de ser interpretados de forma diferente por cada participante e assim tornam complexa a análise conjunta dos dados. Nesse contexto, formas alternativas de medir esses julgamentos ou técnicas de normalização desses dados, por participante, poderão ser exploradas.

Bibliografia

[1] Silvia Bonaccio, Jane O'Reilly, Sharon L. O'Sullivan, and François Chiocchio. Nonverbal Behavior and Communication in the Workplace. *Journal of Management*, 42(5):1044–1074, jul 2016.

[2] B. de Gelder, A. W. de Borst, and R Watson. The perception of emotion in body expressions, 2015.

[3] Markus Koppensteiner, Pia Stephan, and Johannes Paul Michael Jäschke. From body motion to cheers: Speakers' body movements as predictors of applause. *Personality and Individual Differences*, 74:182–185, feb 2015.

[4] Markus Koppensteiner, Pia Stephan, and Johannes Paul Michael Jäschke. Moving speeches: Dominance, trustworthiness and competence in body motion. *Personality and Individual Differences*, 94:101–106, may 2016.

[5] Dana R. Carney, Amy J.C. Cuddy, and Andy J. Yap. Power Posing. *Psychological Science*, 21(10):1363–1368, oct 2010.

[6] Rui Sacchetti, Tiago Teixeira, Bruno Barbosa, António J R Neves, Sandra C Soares, and Isabel D Dimas. Human Body Posture Detection in Context : The Case of Teaching and Learning Environments. *SIGNAL 2018: The Third International Conference on Advances in Signal, Image and Video Processing*, (SIGNAL 2018, The Third International Conference on Advances in Signal, Image and Video Processing):79–84, may 2018.

[7] Kazuki Sekine and Sotaro Kita. The listener automatically uses spatial story representations from the speaker's cohesive gestures when processing subsequent sentences without gestures. *Acta Psychologica*, 179:89–95, sep 2017.

[8] Francesco Ianì and Monica Bucciarelli. Mechanisms underlying the beneficial effect of a speaker's gestures on the listener. *Journal of Memory and Language*, 96:110–121, oct 2017.

[9] Anna Zhen, Stephen Van Hedger, Shannon Heald, Susan Goldin-Meadow, and Xing Tian. Manual directional gestures facilitate cross-modal perceptual learning. *Cognition*, 187:178–187, jun 2019.

[10] Yen Liang Lin. Co-occurrence of speech and gestures: A multimodal corpus linguistic approach to intercultural interaction. *Journal of Pragmatics*, 117:155–167, aug 2017.

[11] Hillel Aviezer, Noga Ensenberg, and Ran R Hassin. The inherently contextualized nature of facial emotion perception. *Current Opinion in Psychology*, 17:47–54, oct 2017.

[12] Fabiola Becerra-Riera, Annette Morales-González, and Heydi Méndez-Vázquez. Facial marks for improving face recognition, oct 2016.

[13] Susan W. White, Lynn Abbott, Andrea Trubanova Wieckowski, Nicole N. Capriola-Hall, Sherin Aly, and Amira Youssef. Feasibility of Automated Training for Facial Emotion Expression and Recognition in Autism. *Behavior Therapy*, 49(6):881–888, nov 2018.

[14] Najmeh Sadoughi, Yang Liu, and Carlos Busso. Meaningful head movements driven by emotional synthetic speech. *Speech Communication*, 95:87–99, dec 2017.

[15] Judith Holler, Louise Schubotz, Spencer Kelly, Peter Hagoort, Manuela Schuetze, and Asli Özyürek. Social eye gaze modulates processing of speech and co-speech gesture. *Cognition*, 133(3):692–697, dec 2014.

[16] Jens Kreitewolf, Angela D. Friederici, and Katharina von Kriegstein. Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *NeuroImage*, 102(P2):332–344, nov 2014.

[17] Theodoros Giannakopoulos. pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PLOS ONE*, 10(12):e0144610, dec 2015.

[18] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, feb 2012.

[19] Eldar Insafutdinov, Mykhaylo Andriluka, Leonid Pishchulin, Siyu Tang, Evgeny Levinkov, Bjoern Andres, and Bernt Schiele. ArtTrack: Articulated multi-person tracking in the wild. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, volume 2017-Janua, pages 1293–1301, 2017.

[20] Eldar Insafutdinov, Leonid Pishchulin, Bjoern Andres, Mykhaylo Andriluka, and Bernt Schiele. Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 9910 LNCS, pages 34–50, 2016.

[21] OpenPose - Realtime Multiperson 2D Keypoint Detection from Video | Flintbox.

[22] Joan Alabort-I-Medina, Epameinondas Antonakos, James Booth, Patrick Snape, and Stefanos Zafeiriou. Menpo: A Comprehensive Platform for Parametric Image Alignment and Visual Deformable Models. 2014.

[23] Brais Martinez, Michel F. Valstar, Xavier Binefa, and Maja Pantic. Local evidence aggregation for regression-based facial point detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1149–1163, may 2013.

[24] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 59–66. IEEE, may 2018.

[25] Florian Eyben, Martin Wöllmer, and Björn Schuller. OpenEAR - Introducing the Munich open-source emotion and affect recognition toolkit. In *Proceedings - 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009*, pages 1–6. IEEE, sep 2009.

[26] Turgut Özseven and Muharrem Düğenci. SPeech ACoustic (SPAC): A novel tool for speech feature extraction and classification. *Applied Acoustics*, 136:1–8, jul 2018.

[27] Paul Boersma and David Weenink. Praat, a system for doing phonetics by computer, 2013.

[28] Florian Eyben, Martin Wöllmer, and Björn Schuller. Opensmile. In *Proceedings of the international conference on Multimedia - MM '10*, page 1459, New York, New York, USA, 2010. ACM Press.

[29] Felix Weninger. open-Source Media Interpretation by Large feature-space Extraction. (December), 2015.

[30] Steven R. Livingstone and Frank A. Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLOS ONE*, 13(5):e0196391, may 2018.

[31] Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010*, pages 94–101. IEEE, jun 2010.

[32] Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, dec 2008.

[33] Soroosh Mariooryad and Carlos Busso. Generating Human-Like Behaviors Using Joint, Speech-Driven Models for Conversational Agents. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(8):2329–2340, oct 2012.

[34] Paul Ekman and W V Friesen. *Facial Action Coding Consulting*. Consulting Psychologists Press, Palo Alto California, 1977.

[35] Brais Martinez, Michel F Valstar, Bihan Jiang, and Maja Pantic. Automatic Analysis of Facial Actions: A Survey. pages 1949–3045, 2017.

[36] Daniel Lopes and António Neves. A Study on Face Identification for an Outdoor Identity Verification System. In *Lecture Notes in Computational Vision and Biomechanics*, volume 27, pages 689–699, 2017.

[37] Euclides Arcoverde, Rafael M. Duarte, Rafael M. Barreto, Joao Paulo Magalhaes, Carlos C. M. Bastos, Tsang Ing Ren, and George Cavalcanti. Enhanced real-time head pose estimation system for mobile device. *Integrated Computer Aided Engineering*, 21:281–293, 2014.

[38] Ailbhe Cullen, Andrew Hines, and Naomi Harte. Perception and prediction of speaker appeal – A single speaker study. *Computer Speech and Language*, 52:23–40, nov 2018.

[39] Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, M Christian, Spoken Language, Processing Group, Deutsche Telekom, and A G Laboratories. The INTERSPEECH 2010 Paralinguistic Challenge German Research Center for Artificial Intelligence (DFKI), Saarbr. *Language*, (September):2794–2797, 2010.

[40] Sudha Velusamy, Hariprasad Kannan, Balasubramanian Anand, Anshul Sharma, and Bilva Navathe. A method to infer emotions from facial action units. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 2028–2031. IEEE, may 2011.

[41] Tae Jun Park and Joon Hyuk Chang. Dempster-Shafer theory for enhanced statistical model-based voice activity detection. *Computer Speech and Language*, 47:47–58, jan 2018.

[42] Louise Francis. Unsupervised learning. In *Predictive Modeling Applications in Actuarial Science: Volume I: Predictive Modeling Techniques*, pages 280–312. 2014.

[43] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, sep 1995.

[44] T. M. Cover and P. E. Hart. Nearest Neighbor Pattern Classification. *IEEE Transactions on Information Theory*, 13(1):21–27, jan 1967.