

A COMPARATIVE STUDY OF EUROPEAN PORTUGUESE STOP CONSONANTS AND FRICATIVES IN WHISPERED AND NORMAL SPEECH FOR REAL-TIME OPERATION OF VOICE CONVERSION

João P. Silva¹, Clara F. Cardoso¹, Marco A. Oliveira¹, Luís M. T. Jesus², Aníbal J. S. Ferreira¹

¹ Department of Electrical and Computer Engineering, University of Porto, Portugal

² ESSUA and IEETA, University of Aveiro, Portugal

joaomiguelppsilva@gmail.com, clara.f.cardoso@gmail.com, marcoantmoliveira@gmail.com, lmtj@ua.pt, ajf@fe.up.pt

Abstract: Contrary to normal speech, whispered speech is produced without the contribution of the vocal folds. Therefore, its acoustic projection is weak, its intelligibility is easily hampered by concurrent sounds and noise, and the speaker-specific sound signature is essentially lost. This has motivated the development of an assistive technology aiming at reconstructing normal speech from whispered speech, in real-time, by carefully implanting synthetic voicing on the latter. The success of this approach depends on the phonemic durations in both normal and whispered speech realizations by the same speaker. In this paper, we focus on European Portuguese stop consonants and fricatives. A European Portuguese database has been built that contains isolated words and sentences, uttered both in normal and whispered speech, by female and male speakers. A study of the duration of stop and fricative consonants was carried out to assess if there exist statistically significant differences between normal and whispered speech both in isolated word and sentence contexts. Results show that despite a few non-representative exceptions, in most cases of interest, differences are not statistically significant. This confirms that when reconstructing Portuguese voiced sounds from whispered speech the algorithm operation is not required to enforce any special duration compensation strategy.

Keywords: Stop consonants, fricative consonants, closure duration, whispered speech

I. INTRODUCTION

Speech communication is the most important modality of human social and professional interaction [1, 2]. In normal speech, most sounds involve vocal fold vibration, but when a health condition affects the vocal folds, as in certain cases of laryngectomy, then the associated speech is known as whispered speech. Whispered speech is problematic because its acoustic projection is weak, its intelligibility is strongly affected by concurrent sounds and noise and, although short-distance voice communication is still possible, most of

the sound signature of a specific speaker is lost. This causes communication difficulties, which has a negative impact in professional and social life. This motivated the development of an assistive technology (DyNaVoiceR, www.dynavoicer.com) whose objective is to reconstruct natural speech sounds from whispered speech, in real-time, to allow effective and comfortable communication by patients while using their speech production system seamlessly. The assistive technology that we are developing [3,4,5,6] takes the input whispered speech as a baseline signal, identifies those regions in the signal that would be voiced in natural speech, and implants, in these regions, synthetic voicing creating a replacement for the missing vocal folds contribution. This replacement is carefully shaped in frequency and time such as to enhance the linguistic content of the resulting synthetic speech, to improve voice projection, and to convey elements of the sound signature of a given speaker. The success of this approach depends on the phonemic durations in both natural speech and whispered speech realizations by the same speaker, so that a realistic reconstruction of the former can be done by implanting synthetic voicing on the latter. In this paper, we focus on European Portuguese (EP) stop consonants and fricatives.

Duration studies for fricatives in American English, as reported in [7], and based on listening tests with 12 subjects, concluded that the minimum frication duration required for correct identification depends on the particular fricative, ranging from approximately 30 to 50 ms. The author also notes that, unsurprisingly, identification improves as the duration of the frication noise increases. Previous studies [8] have also looked at the relative importance of the transitions and the frication duration on the perception of the voiceless fricatives /f/ (as in <face>), /s/ (as in <soap>), and /ʃ/ (as in <shame>). The authors note that transition phase spectral characteristics dominate over frication noise duration in terms of fricative identification in several of the tested scenarios. Jesus and Jackson [9] examined the phonetic detail of voiced and voiceless fricatives. In that study, duration statistics were derived from the voicing and frication labels to distinguish between

voiceless and voiced fricatives in British English and EP. They concluded that, in normal speech, clusters for voiceless and voiced fricatives are centered at 115 ms and 50 ms, respectively. In a cross-linguistic (Portuguese, Italian and German) devoicing study, Pape and Jesus [10] included stops and fricatives in four vowel contexts and two-word positions and computed the devoicing of the time-varying patterns throughout the stop and fricative duration. They showed that consonant durations are very similar across languages and that considerably longer durations are prevalent for the voiceless consonants when compared to their voiced counterparts. For EP, durations of approximately 100 ms were identified for voiced stop consonants, while the voiceless group presented durations of approximately 150 ms. The above studies, as well as other research results in the literature, consider voiced speech only (*i.e.*, normal speech). In this paper, we focus on a comparison of durational patterns for stops and fricatives between normal and whispered speech. Therefore, an EP database has been created for the DyNaVoiceR project that contains isolated words and sentences uttered in both modes: normal and whispered speech. A study of the stop and fricative consonants was carried out to analyze their duration and to assess whether or not there exist statistically significant differences between normal and whispered speech, both in isolated word and sentence contexts, for female and male speakers.

II. METHODS

Thirty volunteer speakers (15 females and 15 males) were recruited using convenience sampling in the districts of Aveiro and Coimbra, in Portugal, and a database containing whispered and normal speech material was recorded for the DyNaVoiceR project. Recording and manual phonetic annotation tasks for the entire database were performed at the University of Aveiro. The recordings took place in a sound booth with 45 dB sound reduction and using a Sennheiser Ear Set 1 microphone. The sampling frequency was 48 kHz and the sample resolution 16 bits. The database includes 28 isolated words and 6 sentences, among other tasks. Each task was repeated 3 times both in normal speech, and whispered speech modes, by each speaker.

In this paper, we use an underscore W to identify the whispered version of each task (*e.g.*, <nuca_W> represents the whispered version of the Portuguese word <nuca>). The analyses conducted in this study were performed using MATLAB R2016b 64-bit.

In our study, we include both voiceless /f, s, S/ and voiced /v, z, Z/ fricatives, and voiceless /p, t, k/ and voiced /b, d, g/ stops.

All stop consonants and fricatives produced in isolated words and sentences contexts have been analyzed regardless of their syllable and sentence position. However, for closure duration analysis, only voiceless stop consonants in intervocalic contexts were considered. It should be noted that the number of samples (*i.e.*, the number of instances in the database) per consonant is not always the same because in the manual annotation process it was detected that the participants did not always produce the correct stop and fricative consonants. Only correct and clearly identifiable stop and fricative consonants were used in the study. For this same reason, the sample number may also differ between female and male speakers.

Durational patterns of fricative and stop consonants of normal and whispered speech were carefully analyzed. In particular, a detailed statistical analysis of the results was performed focusing on 95% confidence intervals around the means, and on the statistically significant differences between those means.

III. RESULTS

This section presents an analysis of the closure and total duration of stop consonants, and the total duration of fricatives via box-plots, as well as statistical inference results regarding normal and whispered speech. The intervocalic stop consonants' duration labelling was performed manually using the waveform and corresponding spectrogram concerning the second repetition of each word in our database. We carried out a statistical hypothesis Wilcoxon signed-rank test with 5% significance level in order to draw statistical inferences from normal/whispered speech recordings.

Figure 1 shows the particular closure duration distribution for the words containing stop consonants, both in normal and whispered speech modes, and based on the recordings of 15 male participants. Each box-plot reflects 15 data points, one for each of the participants. The symbol '+' represents outliers, the symbol 'x' represents the average value, and the horizontal line corresponds to the median value.

An analysis of the overall closure duration results, for both male and female speakers, shows that, except for the words <nuca> and <ripa> produced by female speakers, the average whispered speech closure duration is slightly longer than the normal speech average closure duration. However, none of the p -values are below the level of significance of 5% ($p > 0.2185$ and $p > 0.2747$ in the case of female and male speakers, respectively), which indicates that there are no statistically significant closure duration differences between normal and whispered speech for the 6 words analyzed in this study. This is expected, as

the mean closure duration differences are rather small between normal and whispered speech realizations.

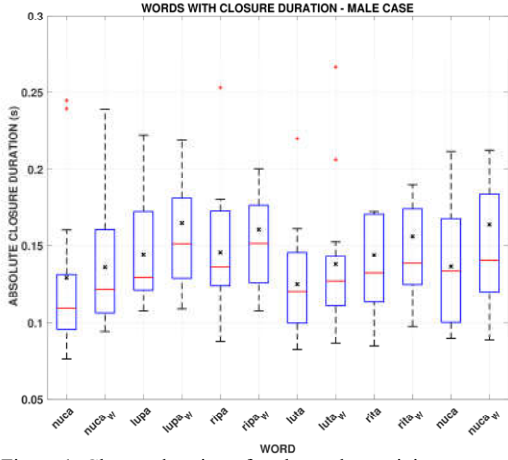


Figure 1: Closure duration of each word containing a stop consonant, produced by male speakers, for both normal and whispered speech modes.

A similar analysis of the stop consonants total duration was also carried out, as illustrated in Figure 2.

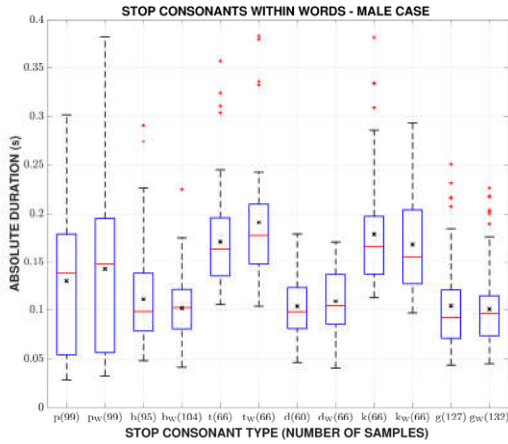


Figure 2: Duration of stop consonants in isolated words, produced by male participants, for both normal and whispered speech scenarios.

It can be observed that, in most cases, the average total duration of whispered stop consonants tends to be slightly longer than the corresponding duration in normal speech. However, in general, the differences are small and not statistically significant, with all $p > 0.07$ in the case of isolated words, similarly to the sentences results with only one significant difference $p = 0.0187$ for the stop consonant pair [g]-[g_w]. In the case of female speakers, the average whispered stop consonants total duration is also slightly longer than that of the voiced counterparts. Albeit the pair [d]-[d_w] in isolated words showing a statistically significant difference ($p = 0.0053$), this is not relevant because none

of the differences in the case of sentences has been found to be statistically significant (all $p > 0.1137$).

A similar duration analysis was carried out for fricative consonants with the similar goal to ascertain if there exist statistically significant differences between normal speech and whispered speech, both in words and sentences contexts, for both female and male speakers.

As an illustrative example, Figure 3 shows the distribution of the duration of each fricative in isolated words regarding male participants. Differences in the mean duration results in the case of words and sentences contexts are minor, and mixed, without a clear trend of a tendency for whispered fricatives to be longer or shorter than normal speech fricatives. With respect to sentences, none of the differences were found to be statistically significant (all $p > 0.66$), however, with respect to words, 5 in 6 cases show p -values less than the level of significance ($p < 0.044$), which means that the average duration of fricatives in isolated words tends to differ significantly. However, isolated words are not as representative of normal speech as sentences are.

In the case of female speakers, similar conclusions were reached concerning the isolated words tests. However, in the case of the sentence tests, 2 out of 6 cases were found to exhibit statistically significant differences ($p < 0.026$) between the average duration of fricatives, specifically in the case of the fricative pairs [S]-[S_w], and [z]-[z_w].

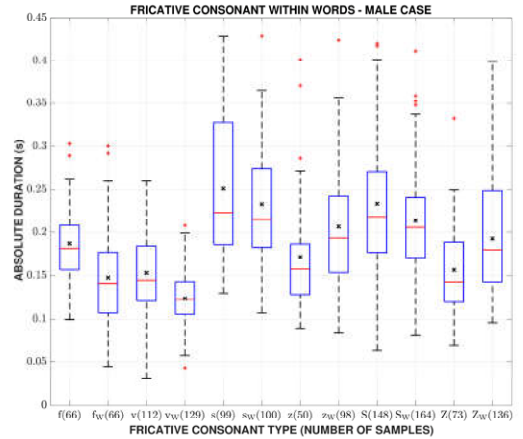


Figure 3: Duration of fricative consonants in isolated words, produced by male participants for both normal and whispered speech scenarios.

IV. DISCUSSION

The paper discusses two sets of results: those regarding intervocalic stop consonants (including closure duration and total duration) and those regarding voiced and voiceless fricatives. Results show that in

both stops and fricatives, while in some non-representative cases statistically significant differences can be found, in most cases of interest, especially regarding sentence contexts, differences were not found to be statistically significant. This important outcome confirms that when reconstructing “on the fly” Portuguese voiced sounds from whispered speech, in real-time, in addition to a careful phoneme-oriented segmentation, the algorithm operation does not need to adopt any special compensation strategy in the whispered speech to normal speech conversion process and regarding the duration of fricatives or stop consonants.

V. CONCLUSION

While most stop and fricative duration studies available in the literature consider voiced speech only (*i.e.*, normal speech), in this paper we focused on a comparison of durational patterns between normal and whispered speech, using male and female recordings.

The first conclusion that can be drawn from our work is that the voiceless stop consonants average closure duration tends to be slightly longer in whispered speech than in normal speech. Despite a few non-representative exceptions, a statistical analysis comparing whispered speech and normal speech shows that there are no consistent statistically significant differences between the two speech modes.

Regarding stop consonants total duration, in the case of male speakers, no statistically significant differences were found between whispered and normal speech realization in word contexts, and only one statistically significant difference was found in sentence contexts. Regarding female speakers, the opposite was verified.

Therefore, in general, it can be concluded that regarding the average closure duration and the average total duration of the stop consonants analyzed in this paper, there is a tendency for the whispered speech realizations to be slightly longer than in normal speech, however, differences are rather small and negligible.

Regarding fricatives, considering the results of both male and female speakers, it was observed that the average duration of fricatives in isolated words tend to differ significantly although not in a consistent manner. In sentence contexts, which are more representative of normal speech, fricative duration differences are not statistically significant in the case of male speakers, and, in the case of female speakers, in only 2 (out of 6) cases differences were found to be significant.

As a summary, representative and systematic statistically significant stop and fricative duration differences between whisper and normal speech realizations have not been found. This important outcome confirms the real-time operation feasibility of the DyNaVoiceR assistive technology converting

Portuguese whispered speech into naturally sounding synthetic speech. This is because in its “on the fly” operation, the algorithm does not need to implement any stop/fricative consonants duration compensation. The linguistic implications of this decision will be fully assessed as the DyNaVoiceR algorithm approaches the final stages of development, in the near future.

ACKNOWLEDGMENTS

This work was financially supported by Project PTDC/EMD-EMD/29308/2017 - POCI-01-0145-FEDER-029308 - funded by FEDER funds through COMPETE2020 - POCI and by national funds (PIDDAC) through FCT/MCTES. Support was also received from National Funds through the FCT - Foundation for Science and Technology, in the context of the project UIDB/00127/2020.

REFERENCES

- [1] Gunnar Fant, “Acoustic theory of speech production”, The Hague, Netherlands. Mouton, 1970
- [2] Douglas O’Shaughnessy, “Speech Communication: Human and Machine”, *Wiley-IEEE Press*, 1999
- [3] Aníbal Ferreira, “Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information”, *ISIVC2016* (invited paper), Tunisia, 2016
- [4] J. P. Silva, M. A. Oliveira, C. F. Cardoso, A. J. Ferreira, “Manipulation of the Fundamental Frequency Micro-Variations Using a Fully Parametric and Computationally Efficient Speech Model”, *IEEE IWSPS*, Portugal, 2020
- [5] A. J. Ferreira, J. Silva, F. Brito, D. Sinha, “Impact of a Shift-Invariant Harmonic Phase Model in Fully Parametric Harmonic Voice Representation and Time/Frequency Synthesis”, *ICASSP2020*, Spain, 2020
- [6] J. Silva, M. Oliveira, A. Ferreira, “Flexible parametric implantation of voicing in whispered speech under scarce training data”, *EUSIPCO 2020*
- [7] A. Jongman, “Duration of frication noise required for identification of English fricatives”, *JASA* 85 (4), 1989, pp. 1718–1725.
- [8] K. Nataraj, P. Pandey, H. Dasgupta, “Effect of frication duration and formant transitions on the perception of fricatives in VCV utterances”, *ICASSP 2020*, pp. 6259- 6263
- [9] L. Jesus, P. Jackson, “Frication and voicing classification” In A. Teixeira, V. Lima, L. Oliveira, and P. Quaresma (Eds.), *Computational Processing of the Portuguese Language*, 2008, pp. 11-20. Berlin: Springer-Verlag.
- [10] D. Pape, L. Jesus, “Stop and fricative devoicing in European Portuguese, Italian and German”, *Language and Speech* 58(2), 2015, pp. 224–246.

**MODELS AND ANALYSIS OF VOCAL
EMISSIONS FOR BIOMEDICAL
APPLICATIONS**

12TH INTERNATIONAL WORKSHOP

**December 14-16, 2021
Firenze, Italy**

**Edited by
Claudia Manfredi**

Firenze University Press
2021

Models and Analysis of Vocal Emissions for Biomedical Applications : 12th International Workshop, December, 14-16, 2021 / edited by Claudia Manfredi. – Firenze : Firenze University Press, 2021.
(Proceedings e report ; 131)

<https://www.fupress.com/isbn/9788855184496>

ISSN 2704-601X (print)

ISSN 2704-5846 (online)

ISBN 978-88-5518-448-9 (Print)

ISBN 978-88-5518-449-6 (PDF)

ISBN 978-88-5518-450-2 (XML)

DOI 10.36253/978-88-5518-449-6


Cover: designed by CdC, Firenze, Italy.

FUP Best Practice in Scholarly Publishing (DOI https://doi.org/10.36253/fup_best_practice)

All publications are submitted to an external refereeing process under the responsibility of the FUP Editorial Board and the Scientific Boards of the series. The works published are evaluated and approved by the Editorial Board of the publishing house, and must be compliant with the Peer review policy, the Open Access, Copyright and Licensing policy and the Publication Ethics and Complaint policy.

Firenze University Press Editorial Board

M. Garzaniti (Editor-in-Chief), M.E. Alberti, F. Vittorio Arrigoni, E. Castellani, F. Ciampi, D. D'Andrea, A. Dolfi, R. Ferrise, A. Lambertini, R. Lanfredini, D. Lippi, G. Mari, A. Mariani, P.M. Mariano, S. Marinai, R. Minuti, P. Nanni, A. Orlandi, I. Palchetti, A. Perulli, G. Pratesi, S. Scaramuzzi, I. Stolzi.

 The online digital edition is published in Open Access on www.fupress.com.

Content license: except where otherwise noted, the present work is released under Creative Commons Attribution 4.0 International license (CC BY 4.0: <http://creativecommons.org/licenses/by/4.0/legalcode>). This license allows you to share any part of the work by any means and format, modify it for any purpose, including commercial, as long as appropriate credit is given to the author, any changes made to the work are indicated and a URL link is provided to the license.

Metadata license: all the metadata are released under the Public Domain Dedication license (CC0 1.0 Universal: <https://creativecommons.org/publicdomain/zero/1.0/legalcode>).

© 2021 Author(s)

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

*This book is printed on acid-free paper
Printed in Italy*