A corpus-based study of elderly and young speakers of European Portuguese: acoustic correlates and their impact on speech recognition performance

Thomas Pellegrini[1,*], *Annika Hämäläinen*[2,+], *Philippe Boula de Mareüil*[3], *Michael Tjalve*[4], *Isabel Trancoso*[1,5], *Sara Candeias*[6], *Miguel Sales Dias*[2], *Daniela Braga*[2]

[1]INESC-ID, Lisbon, Portugal
[2]Microsoft Language Development Center, Portugal / Microsoft / ADETTI-ISCTE, IUL, Portugal
[3]LIMSI-CNRS, France - [4]Microsoft / University of Washington
[5]Instituto Superior Técnico, Portugal - [6]Instituto de Tecnologia, Portugal
[*]thomas.pellegrini@inesc-id.pt, [+]t-anhama@microsoft.com

Abstract

This paper presents a study of European Portuguese elderly speech, in which the acoustic characteristics of two groups of elderly speakers (aged 60-75 and over 75) are compared with those of young adult speakers (aged 19-30). The correlation between age and a set of 14 acoustic features was investigated, and decision trees were used to establish the relative importance of the features. A greater use of pauses characterized speakers aged 60 and over. For female speakers, speech rate also appeared to correlate with age. For male speakers, jitter distinguished between speakers aged 60-75 and older. The correlation between the features and speech recognition performance was also investigated. Word error rate correlated mostly with the use of pauses, speech rate, and the ratio of long phone realizations. Finally, by comparing the phone sequences used by the recognizer on the most frequent words, we observed that the young adult speakers reduced schwas more than the elderly speakers. This result seems to confirm the common idea that young speakers reduce articulation more than older speakers. Further investigation is needed to confirm this result by determining whether this is due to ageing or to the generation gap.
Index Terms: acoustic analysis, acoustic correlates, automatic speech recognition, elderly speech, speaker age

1. Introduction

Most elderly people would like to live in their own homes as long as possible. To ensure their autonomy, there is an increased interest in technologies adapted to their needs. As they often have difficulties using computers [1], the naturalness and ease of speech interfaces would be very useful for them. However, speech recognizers do not yet work well with elderly speech for at least two reasons. First, several parameters of the speech signal (e.g. pitch, jitter, shimmer) change with age [2, 3] but the acoustic models needed to recognize speech are typically trained using speech from younger adults, with elderly speakers being underrepresented in the training data. Second, the elderly sometimes interact with computers using everyday language and their own commands, even when a specific syntax is required [4]. As compared with younger adult speech, significantly higher word error rates (WERs) have been reported for elderly speech [5, 6, 7]. While improvements were achieved by using acoustic models adapted to the elderly [7, 8], a better understanding of age-related changes in speech may prove useful for further improvements in recognition performance.

Much research has already been conducted on the acoustic correlates of speaker age [9, 10]. However, to the best of our knowledge, this has not yet been done for European Portuguese (EP). This article reports on our attempts to identify the most salient differences between elderly and younger adult speech in terms of acoustic features. Most of the studied features are language-independent but we also explore features specific to EP, such as vowel formants and an estimate of vowel reduction. We then study their correlation with recognition performance by carrying out detailed analyses of recognition errors made when recognizing elderly speech using acoustic models trained using younger adult speech.

2. Speech material

We used speech from two large corpora of European Portuguese: BD-PUBLICO [11] and EASR [12, 13]. BD-PUBLICO consists of newspaper sentences read by 19–30-year-old speakers from the Lisbon area. EASR comprises a large variety of prompts from isolated digits to phonetically rich sentences read aloud by speakers aged 60 and over. The speakers in this corpus come from different areas of Portugal but none of them have strong regional accents that are very different from the standard Lisbon accent.

BD-PUBLICO contains phonetically rich sentences only. To keep the younger adult speech and elderly speech as comparable as possible, we only used the phonetically rich sentences from EASR. For our analyses, we divided the EASR speakers into two groups: speakers aged 60-75 (hereafter '*S*' for *Senior*) and aged 75-90 ('*E*' for *Elderly*). The young adult speakers from BD-PUBLICO are referred to as *Y*.

The main statistics of the speech data used in this study are presented in Table 1. The young adult subset, which corresponds to the training subset of BD-PUBLICO, comprises a total of 20 hours of 50 male and 50 female speakers. The two groups of older speaker, *S* and *E*, which correspond to the training subset of EASR used in [8] comprise 8.7 and 6.7 hours of speech, respectively. These two subsets include more female speakers than male speakers in the same proportions as in the entire EASR corpus.

3. Methodology

To study the differences between young adult and elderly speech, a set of 14 acoustic features commonly used to characterize ageing voices were investigated: source-related fea-

Table 1: *Speech material*

Age group	Gender	# Spk	Duration (h)	# Word Types	# Word Tokens
Y (Young)	female	50	10.8	48.9k	90.6k
	male	50	10.9	49.5k	91.9k
S (Senior)	female	76	5.7	3.7k	41k
	male	38	3.0	2.8k	20k
E (Elderly)	female	70	5.2	3.4k	33k
	male	24	1.5	2.0k	10k

tures (pitch, jitter, shimmer, and harmonics-to-noise ratio), vowel-related features (formants, schwa realization), speech rate (numbers of words and phones per second, phone duration, percentages of short and long phone realizations, and percentage of silent pauses). Strong vowel reduction characterizes spoken EP: especially the schwa (often elided), [u], and [ɐ]. It distinguishes EP from Brazilian Portuguese in particular [14]. For this reason, schwa realization and the first formants of the three vowels mentioned above were used as features. Since some of the features (e.g. pitch and formant frequencies) are gender-dependent, their values were measured separately for men and women.

To compute the feature values, forced alignments between the audio signal and the word and phone sequences were automatically produced with our in-house speech recognition engine Audimus, briefly described in Section 5. Context-independent acoustic models (39 monophones) were used since they have been found to be more suitable for linguistically-motivated research than context-dependent models (e.g. [15]). The pronunciation lexicon comprised 100k lexical entries. 14k of them had multiple pronunciations, allowing in particular the deletion of reduction-prone vowels ([ə], [u], and [ɐ]). Source-related features and vowel formants were calculated with Praat [16].

Each speaker was represented by a single vector comprising the means of the studied features. To understand which features correlate the most with age, decision trees were built using these data. The purpose was not to build a state-of-the-art age classifier but to identify the most discriminant features.

In a second phase of this study (reported in Section 5), the speech data were automatically transcribed with the context-dependent acoustic models of Audimus. As expected, WERs were higher for elderly speech than for young adult speech. To better understand the impact of age on Automatic Speech Recognition (ASR) performance, the correlation between the aforementioned features and the WER obtained for each speaker was investigated.

4. Feature analysis and ranking

Table 2 summarizes, by age and gender, the average values of the 14 features analyzed in this section.

4.1. Source-related features

Contradictory findings are reported in the literature about the impact of speaker age on fundamental frequency (F0). A large number of studies have found that the F0 of women remains fairly constant until menopause, when a drop of about 10-15 Hz usually occurs (e.g. [17]). An increase of about 35 Hz is reported by [17] in the case of men aged 65 and over. However, some studies fail to find any correlation between F0 and age in the case of men, or report a decrease in it (e.g. [18]).

No age-related changes were observed in the F0 values of the men in our data. The average F0 values were 135 Hz, 133 Hz, and 130 Hz for *Y*, *S*, and *E*, respectively. However, in the case of female speakers, a significant drop of 20 Hz was observed for the older speakers: the average F0 values were 212 Hz for *Y*, 190 Hz for *S*, and 191 Hz for *E*.

The other source-related features that we investigated included jitter, shimmer, and harmonics-to-noise ratio (*hnr*). Jitter and shimmer, linked to the perturbations of vocal fold vibration, are cycle-to-cycle variations of pitch and amplitude, respectively. They have commonly been used for the description of pathological voice quality. According to the literature, these features either increase or remain stable with age in both men and women [9]. We computed the three features using Praat. As they are usually computed on long vowel realizations, we only considered vowels with a duration greater than 100 ms. In general, the three features increased with age. However, in the case of men, shimmer decreased with age. Between *S* and *E*, the differences were not significant according to two-sample *t*-tests, contrary to all other pair-wise comparisons ($p<0.001$).

Table 2: *Feature values for the three age groups*

	Female speakers			Male speakers		
	Y	*S*	*E*	*Y*	*S*	*E*
F0	212	190	191	134	133	130
jitter (%)	1.25	1.46	1.68	1.64	1.75	1.90
shimmer (%)	7.46	7.89	8.83	10.68	9.39	8.87
hnr (%)	7.84	12.62	15.05	12.29	15.12	15.73
F1_[ə]	393	454	452	350	415	418
F1_[ɐ]	505	564	557	437	498	500
F1_[u]	391	456	448	352	408	402
word_sr	2.5	1.9	1.8	2.5	2.0	2.0
phone_sr	13.9	10.7	10.5	14.7	11.3	11.0
%pauses	3.6	9.7	10.3	3.9	8.7	8.8
dur (ms)	69	95	96	65	89	91
%phones≤30	19.9	15.4	15.9	22.1	16.3	16.0
%phones≥200	0.9	6.8	7.9	0.8	4.7	5.4
%schwas	10.4	11.5	12.3	12.0	12.5	12.7
WER(%)	13.3	28.1	38.9	13.8	26.4	29.4

4.2. Vowel-related features

The vowels investigated in this study included the oral vowels [ə], [a], [ɐ], [e], [ɛ], [i], [o], [ɔ], and [u]. Table 2 reports the F1 values of the three vowels most prone to reduction in EP: *F1_[ə]*, *F1_[ɐ]*, and *F1_[u]*. Figure 1 shows the F1/F2 charts for female speakers, with one triangle for each age group. The chart for male speakers is not shown since the age-related tendencies were similar (with smaller triangles).

The *S* and *E* triangles are very similar, whereas the *Y* triangle is more reduced, with lower F1 values. The reduced [ɐ] vowel is proportionally closer to the schwa (noted [ə]) than to the [a] in the case of younger speakers than in the case of elderly speakers. The schwa has a particularly low F1 value, especially in the *Y* triangle where it is located on the [i] – [u] line. This acoustic correlate of close articulation speaks in favor of the [ɨ] symbol sometimes used for EP (e.g. [19]), whereas the central position occupied by the reduced [ɐ] suggests that the [ə] symbol is more appropriate. Our formant measurements suggest a

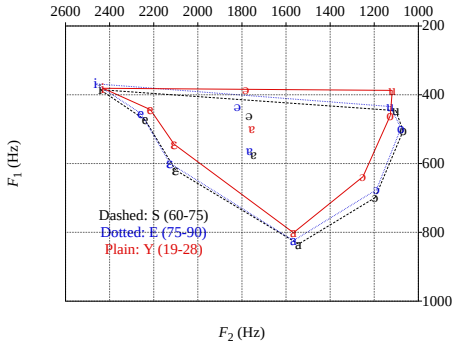

Figure 1: *F1/F2 chart for female speakers.*

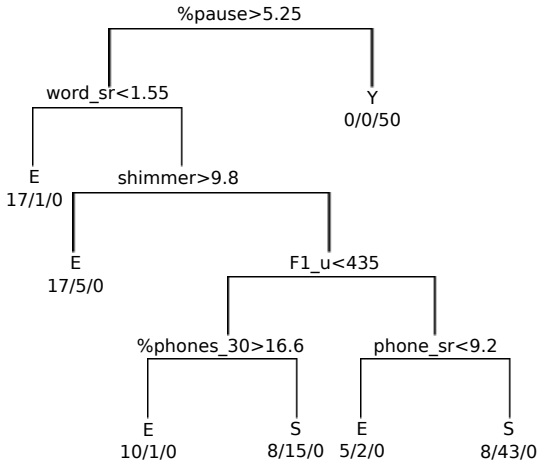

Figure 2: *Decision tree for female speakers. Left-handed branches and leaves correspond to the path when the test condition is true.*

sound change in progress: vowel reduction is more prevalent among young speakers.

4.3. Speech rate related features

Speech rate at word- and phone-levels, noted *word_sr* and *phone_sr*, were computed as the number of words and phones uttered per second. They were expected to decrease with age. As shown in Table 2, the two features did indeed decrease for both genders as compared with *Y*. The differences between *S* and *E* were not significant.

The ratio between the number of silent pauses (longer than 50 ms) and the total number of phones uttered by a speaker, noted *%pauses*, showed a significant increase when comparing *Y* with *E*/*S* ($p<0.001$). Similarly, the average phone duration (*dur*) showed a significant increase with age for both genders: 69 ms (*Y*), 95 ms (*S*), and 96 ms (*E*) for female speakers, 65 ms (*Y*), 89 ms (*S*), and 91 ms (*E*) for male speakers. As expected, the percentages of phones longer than 200 ms (*%phones*$\geq$*200*, measured to capture hesitation phenomena) and shorter than 30 ms (*%phones*$\leq$*30*, measured to capture reduction phenomena) increased and decreased with age, respectively. An exception to the overall patterns was found for the latter feature, which remained stable for the two groups of older male speakers: 16.3% and 16.0% for *S* and *E*, respectively. Interestingly, most short phones corresponded to [ɐ], [ə], and [u], which might be considered as not pronounced.

To the best of our knowledge, no standard measure of vowel reduction has been defined in the literature. To roughly estimate the level of schwa reduction, we computed the percentage of schwa realizations out of all vowel realizations (*%schwas*). For instance, young female speakers showed lower *%schwas* values: 10.4% for *Y* speakers, as compared with 12.3% for *E* speakers. This result seems to confirm that young speakers reduce vowels more than older speakers. This tendency was confirmed by examining the phone sequences selected by the recognizer (which could select the best pronunciation from forms with or without schwas during forced alignment) for twenty frequent words: *de* ('of'), *se* ('oneself') etc. These words accounted for 27k word tokens in total. The percentage of maintained schwas was 32%, 58%, and 62% for *Y*, *S*, and *E* speakers, respectively.

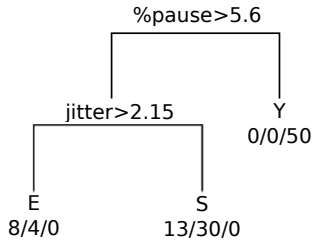

Figure 3: *Decision tree for male speakers.*

4.4. Feature relative importance

Figures 2 and 3 show two decision trees generated with R (`http://www.r-project.org/`) for male and female speakers. They were built using the 14 features discussed above, using a single vector per speaker as input. The trees were used to perform a classification task, with the three age groups *Y*, *S* and *E* as classes. A leaf indicates a class that has been identified, and the three numbers are the proportions of *Y*, *S* and *E* that satisfy all the conditions from the top of the tree to the leaf.

For both genders, the *%pause* ratio was sufficient to distinguish between *Y* and *S*/*E* speakers. Other decision trees were built by removing the most discriminant features. When *%pause* was removed, the percentage of long phones (*%phones*$\geq$*200*) was sufficient to distinguish *Y* speakers. By repeating the removal of the most important feature, the discriminant features between *Y* and *S*/*E* turned out to be, regardless of gender, *%pause*, *%phones*$\geq$*200*, average phone duration (*dur*), phone-level speech rate (*phone_sr*), *F1_[ə]*, and word-level speech rate (*word_sr*).

Since the first test condition identified all the *Y* speakers, the other branches of the trees tried to distinguish between *S* and *E* speakers. For females, word-level speech rate (*word_sr*) appeared to be the most important feature, followed by shimmer, *F1_[u]*, phone-level speech rate, and the percentage of short phones (*%phones*$\leq$*30*). For instance, 17 female *E* speakers were identified by a speech rate slower than 1.55 words/second. For males, the tree was much simpler. Jitter was the only discriminant feature between *S* and *E* speakers. The majority of *S* speakers was identified by low values of jitter (smaller than 2.15%).

5. Analysis of speech recognition results

The speech recognizer used for the study, Audimus, is a hybrid speech recognizer which combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons [20]. The acoustic models for EP were initially trained with 46 hours of manually annotated data collected from television news, and then with 1000 hours of automatically transcribed data collected from television news and selected according to a confidence measure threshold (non-supervised training). The speakers are young and middle-aged adults, thus closer in age to the *Y* speaker group of this study. The language model is a 4-gram with Kneser-Ney modified smoothing.

The bottom line of Table 2 shows the WERs for the three age groups, for male and female speakers. Two-sample *t*-tests confirmed that age-related increases in WER are statistically significant when comparing between the age groups pair-wise ($p<0.001$), except for the *E* and *S* groups of male speakers. This may be due to the lower number of speakers in these groups as compared with the equivalent groups of female speakers.

Several analyses were performed to understand which characteristics of elderly speech are most relevant from an ASR point of view. First, we listened to the elderly speech and compared it with the recognizer output. Second, we analyzed the most frequent substitution, insertion and deletion errors. Third, we computed Pearson correlation coefficients between WERs and the features discussed in Section 4.

When carrying out the auditory analysis of elderly speech, we focused on elderly speakers who showed the highest and the lowest WERs. A higher rate of disfluencies was obvious (filled and silent pauses, false starts, repetitions, hesitations, etc.) amongst speakers exibiting the highest WERs. In general, their speech came across as slower and less fluent than that of speakers exibiting the lowest WERs. A higher rate of disfluencies and speech errors, a slower speech rate and more pauses in the case of elderly people have indeed been reported in the literature [21].

No specific phonetic patterns were found for words yielding most frequent recognition errors. As expected, most of them were short, monosyllabic function words, which are usually unstressed and often heavily reduced. We did, however, identify a set of content words producing frequent recognition errors: verbs in the imperative form (e.g. *ligue-me*, 'call me') and verbs in the first person singular (e.g. *fui*, 'I was'). As the language model was built using broadcast news transcripts, a context in which these verb forms are uncommon: problematic words were either out-of-vocabulary or low-probability words.

Pearson correlation coefficients between WERs and acoustic features showed interesting correlations and differences between male and female speakers. There was a high positive correlation between the WERs and the percentage of pauses (*%pauses*), as it appeared in the decision trees, in the case of both female speakers (0.76) and male speakers (0.70). Although this correlation might not first look intuitive, pauses contribute to lowering speech rate, which may increase WER [22, 23]. All the other correlations were either moderate or low, and differed between female and male speakers. In the case of female speakers, there was a moderate negative correlation (-0.53) between the WERs and speech rate (words/second, *word_sr*). There was also a moderate positive correlation (0.50) between the WERs and the percentage of long phones (*%phones≥200*): a higher proportion of very long phones was associated with a higher WER. Jitter, shimmer and *hnr* positively correlated with WER but only slightly (0.34). In the case of male speakers, the second most correlated feature was *%phones≥200* (0.70), followed by the ratio of schwas (0.49). This could be related to the fact that the ASR acoustic models were trained with speech from young adults who were shown to reduce schwas more than older speakers (See Section 4.2 and 4.3). Moderately correlated features also included the average phone duration (*dur*, 0.49), *hnr* (0.45), *F1_[ə]* (0.43), the phone-level speech rate (-0.41), and jitter (0.32).

6. Conclusions

In this article, we presented a study of European Portuguese elderly speech, in which a set of 14 acoustic features was used to compare two groups of elderly speakers (aged 60–75 and over 75) with a group of young adult speakers (aged 19–30). The main feature distinguishing speakers aged 60 and over from the young speakers was a higher frequency of pauses. For females, a slower speech rate and increased shimmer also characterized the two groups of elderly speakers. In the case of males, jitter helped to distinguish between speakers aged 60-75 and older.

The correlation between the 14 acoustic features and speech recognition performance was also investigated. Our analyses suggest that, from the ASR point of view, the most important characteristics of elderly speech are related to pausing, speech rate, average phone duration, and percentage of long phones, regardless of gender. For males, the percentage of schwas and the harmonics-to-noise ratio were also moderately correlated with WER. Jitter and shimmer showed a low-moderate correlation for both genders. Therefore, when modeling elderly speech, it might be particularly important to pay special attention to the modeling of pauses, and to ensure that the acoustic models are adapted to a slower than average speech rate.

A very interesting outcome of this study, which needs further research, concerned a language-dependent feature, linked to schwa reduction. By comparing the phone sequences selected by the recognizer on the most frequent words, we observed that young adult speakers reduced schwas more than elderly speakers did. This result seems to confirm the common idea that young speakers reduce articulation more than older speakers do.

In future research, we will look into more features, in particular voice quality-related features such as creakiness and breathiness. Further investigation is also needed to confirm the results regarding vowel reduction, by determining whether this phenomenon is due to ageing or to the generation gap (or, in other words, whether we are facing a real-time or only an apparent-time change [24]). Finally, it would be interesting to conduct a perception experiment to measure the correlation between ASR performance and the age of the speakers, as perceived by human subjects, in addition to the chronological age investigated in this study.

7. Acknowledgements

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under project PTDC/EEA-PLP/121111/2010 and under project PEst-OE/EEI/LA0021/2013. Microsoft Language Development Center have carried the work out in the scope of the QREN 7900 LUL - Living Usability Lab project, co-funded by Microsoft, the Portuguese Government, and the European Structural Funds for Portugal (FEDER), through COMPETE and QREN.

8. References

[1] V. Teixeira, C. Pires, F. Pinto, J. Freitas, M. Sales Dias, and E. Mendes Rodrigues, "Towards elderly social integration using a multimodal human-computer interface," in *Proc. AAL*, Vilamoura, 2012.

[2] S. Xue and G. Hao, "Changes in the human vocal tract due to aging and the acoustic correlates of speech production: A pilot study," *Journal of Speech, Language, and Hearing Research*, vol. 46, pp. 689–701, 2003.

[3] C. Reynolds and J. Czaja, S.and Sharit, "Age and perceptions of usability on telephone menu systems," in *Proc. of the Human Factors and Ergonomics Society Annual Meeting*, vol. 46, no. 2, 2002, pp. 175–179.

[4] S. Anderson, N. Liberman, E. Bernstein, S. Foster, E. Cate, B. Levin, and R. Hudson, "Recognition of elderly speech and voice-driven document retrieval," in *Proc. ICASSP*, Phoenix, 1999, pp. 145–148.

[5] J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, Atlanta, 1996, pp. 349–352.

[6] A. Baba, S. Yoshizawa, M. Yamada, A. Lee, and K. Shikano, "Acoustic models of the elderly for large-vocabulary continuous speech recognition," *Electronics and Communications in Japan*, vol. 87:7, pp. 49–57, 2004.

[7] R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proc. Interspeech*, Brisbane, 2008, pp. 2550–2553.

[8] T. Pellegrini, I. Trancoso, A. Hämäläinen, A. Calado, M. Dias, and D. Braga, "Impact of age in ASR for the elderly: preliminary experiments in European Portuguese," in *Proc. IberSPEECH*, Madrid, 2012.

[9] S. Schötz, "Acoustic analysis of adult speaker age," *Speaker Classification I*, pp. 88–107, 2007.

[10] C. T. Ferrand, "Harmonics-to-noise ratio: an index of vocal aging," *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002.

[11] J. Neto, C. Martins, H. Meinedo, and L. Almeida, "The Design of a Large Vocabulary Speech Corpus for Portuguese," in *Proc. European Conference on Speech Technology*, Rhodes, 1997.

[12] A. Hämäläinen, F. Pinto, M. Dias, A. Júdice, J. Freitas, C. Pires, V. Teixeira, A. Calado, and D. Braga, "The first European Portuguese elderly speech corpus," in *Proc. IberSPEECH*, Madrid, 2012.

[13] F. Pinto, A. Júdice, C. Galinho Pires, V. Duarte Teixeira, J. Freitas, D. Braga, A. Calado, and M. Sales Dias, "European Portuguese elderly speech corpus: data collection methodology and results," in *Proc. IberSPEECH*, Madrid, 2012.

[14] M. Cruz-Ferreira, "European Portuguese," *Journal of International Phonetic Association*, vol. 25:02, pp. 90–94, 2009.

[15] B. Vieru, P. Boula de Mareüil, and M. Adda-Decker, "Characterisation and identification of non-native French accents," *Speech Communication*, vol. 53, no. 3, pp. 292–310, 2011.

[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[17] W. Brown, R. Morris, H. Hollien, and E. Howell, "Speaking fundamental frequency characteristics as a function of age and professional singing," *Journal of Voice*, vol. 5, pp. 310–315, 1991.

[18] L. Ramig and R. Ringel, "Effects of physiological aging on selected acoustic characteristics of voice," *Journal of Speech and Hearing Research*, vol. 26, pp. 22–30, 1983.

[19] J. Veloso, "Schwa in European Portuguese: The phonological status of [ɨ]," 5^{es} *Journées dÉtudes Linguistiques*, pp. 55–60, 2007.

[20] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast News Subtitling System in Portuguese," in *Proc. ICASSP 2008*, Las Vegas, 2008, pp. 1561–1564.

[21] K. Georgila, M. Wolters, V. Karaiskos, M. Kronenthal, R. Logie, N. Mayo, J. Moore, and M. Watson, "A fully annotated corpus for studying the effect of cognitive ageing on users interactions with spoken dialogue systems," in *Proc. of LREC*, vol. 65, 2008.

[22] M. Siegler and R. Stern, "On the effects of speech rate in large vocabulary speech recognition systems," in *Proc. ICASSP*, Detroit, 1995, pp. 612–615.

[23] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," in *Proc. ASRU*, Madonna di Campiglio, 2011.

[24] W. Labov, *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press, 1972.