Impact of age in ASR for the elderly: preliminary experiments in European Portuguese

Thomas Pellegrini[1], Isabel Trancoso[1,2], Annika Hämäläinen[3,4], António Calado[3], Miguel Sales Dias[3,4], and Daniela Braga[3,4]

[1] INESC-ID Lisboa
R. Alves Redol, 9, 1000-029 Lisbon, Portugal
Tel.: +351 213 100 268
thomas.pellegrini@inesc-id.pt
https://www.l2f.inesc-id.pt/wiki/index.php/Thomas_Pellegrini
[2] Instituto Superior Técnico, Lisbon, Portugal
[3] Microsoft Language Development Center, Lisbon, Portugal
[4] ADETTI ISCTE, IUL, Lisbon, Portugal

Abstract. Standard automatic speech recognition (ASR) systems use acoustic models typically trained with speech of young adult speakers. Ageing is known to alter speech production in ways that require ASR systems to be adapted, in particular at the level of acoustic modeling. This paper reports ASR experiments that illustrate the impact of speaker age on speech recognition performance. A large read speech corpus in European Portuguese allowed us to measure statistically significant performance differences among age groups ranging from 60- to 90-year-old speakers. An increase of 41% relative (11.9% absolute) in word error rate was observed between 60-65-year-old and 81-86-year-old speakers. This paper also reports experiments on retraining acoustic models (AMs), further illustrating the impact of ageing on ASR performance. Differentiated gains were observed depending on the age range of the adaptation data use to retrain the acoustic models.

Keywords: ASR, Portuguese, Elderly Speech

1 Introduction

European countries, in particular Western European countries, are about to face a significant social change, brought by an unprecedented demographic change: the ratio of older people is steadily growing, while the ratio of younger people is shrinking. Between 2010 and 2030, the number of people aged 65 and over is expected to rise by nearly 30%-40% relative (according to the statistics of the European Commission from 2010).

Most elderly people would like to live in their own homes as long as possible ("ageing in place"). Thus, research and development of new technologies adapted to older people are becoming strategical, in order to increase their autonomy and

independence. Due to the ageing process and the changes that come with it, this population faces specific difficulties to interact with computers and machines. To overcome this issue, speech appears to be the most natural and effective modality. Thus, speech recognition for the elderly is a key technology in many R&D projects related to the *Ageing Well* problematic.

Due to both cognitive and physiological age-related changes, elderly speech shows specific characteristics that make its processing significantly harder when using models built using speech from younger people. In particular, automatically recognizing the speech of older people is known to be challenging compared with automatically recognizing the speech of younger people, with performance decreases of around 9-12% absolute [1–3]. Various reasons are presented in the literature: ageing causes changes in the speech production mechanism, altering the vocal chords, the vocal cavities and the lungs; it also causes a decline in cognitive and perceptual abilities [4, 5]. Seniors may also interact with machines in a different way than younger speakers do, by using everyday language and their own words to issue commands, even when instructions with a required syntax are given [6].

In the framework of an ongoing national Portuguese project named "AVoz"[5], an in-depth study of ASR for the elderly is conducted in order to improve the global performance in European Portuguese (EP). The goal of this paper is to illustrate the impact of age on ASR performance. Experiments on a large read speech corpus of elderly speech collected by the Microsoft Language Development Center (MLDC) from Lisbon[6] are reported. After an overview of our ASR system for EP, the MLDC elderly speech corpus is briefly described in Section 3. In Section 4, ASR results achieved on this database are reported.

2 Overview of our ASR system

Our automatic speech recognition engine named Audimus [7, 8] is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). The MLPs perform a phoneme classification by estimating the posterior probabilities of the different phonemes for a given input speech frame (and its context). These posterior probabilities are associated to the single state of context independent phoneme HMMs.

Specifically, the system combines three MLP outputs trained with Perceptual Linear Prediction (PLP) features (13 static + first derivative), log-RelAtive SpecTrAl (RASTA) features (13 static + first derivative) and Modulation SpectroGram (MSG) features (28 static) [9]. Each MLP classifier incorporates two fully connected non-linear hidden layers. The number of units of each hidden layer as well as the number of softmax outputs of the MLP networks differs for every language. Usually, the hidden layer size depends on the amount of training data available, while the number of MLP outputs depends on the characteristic

[5] http://avoz.l2f.inesc-id.pt
[6] http://www.microsoft.com/pt-pt/mldc

phonetic set of each language. Finally, the decoder is based on a weighted finite-state transducer (WFST) approach to large vocabulary speech recognition, that maps observation distributions to words.

The baseline ASR system used in this work is exactly the ASR system for EP described in [10]. The acoustic models were initially trained with 46 hours of manually annotated broadcast news (BN) data collected from the public Portuguese TV, and in a second time with 1000 hours of data from news shows of several EP TV channels automatically transcribed and selected according to a confidence measure threshold (non-supervised training). The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a very significant part of all the transition units present in the training data. Details on phone transition modeling with hybrid ANN/HMM can be found in [11].

The Language Model (LM) is a statistical 4-gram model that was estimated from the interpolation of several specific LMs: in particular a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005, and a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts. The final language model is a 4-gram LM, with Kneser-Ney modified smoothing, 100k words (or 1-gram), 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram. The multiple-pronunciation EP lexicon includes about 114k entries.

These models, both AMs and the LM, were specifically trained to transcribe BN data. The Word Error Rate (WER) of our current ASR system is under 20% for BN speech in average: 18.4% obtained in one of our BN evaluation test sets (RTP07), composed by six one hour long news shows from 2007 [10].

Table 1. *Number of speakers and speech durations according to the age ranges in the all corpus (after removing speakers with less than 2min of speech).*

Age	# Speakers	Duration (h)
60-65	371	64.1
66-70	183	31.9
71-75	155	28.3
76-80	87	15.4
81-85	55	10.2
86-90	27	5.0
91-95	2	0.3
96-100	1	0.2

3 Elderly speech corpus

The speech corpus is comprised of about 150 hours of read speech (including silences) that was collected by MLDC. A total of 1038 speakers between 60 and 100 years of age read up to 160 prompts among a broad variety of prompts, from isolated digits to phonetically rich sentences. On average, this corresponds to 12 minutes of speech per speaker. For this work, speakers with less than 2 minutes of speech were removed from our datasets, so that the total speaker number was 881. Speaker age information is reported using 5-year ranges: 60-65, 66-70 and so on. Many more female than male speakers were recorded: 641 and 240 respectively. The number of speakers and the duration of the recordings according to the age ranges are presented in Table 1. Speakers in the 60-65 age range were the most numerous ones with a total of 64 hours of recordings, whereas only 5 hours were collected from speakers in the 86-90 age range. The corpus also provides speech from younger speakers, but with no precise information about their age (indication of a 0-59 age range), hence this data was not used in this work.

A test set comprised of about 10% of the corpus, totaling 15h of speech, was randomly selected. Speakers from this subset do not appear in the rest of the corpus. The proportions of the age range and gender in the full corpus were respected. Speech from the last two age ranges (91-95 and 96-100) was not considered since the corresponding durations were much shorter than for the other age ranges. Table 2 summarizes the characteristics of the subset.

Table 2. *Test subset. Number of speakers and Speech durations according to the age ranges.*

Age	# Speakers	Duration (h)
60-65	35	6h22
66-70	18	3h04
71-75	17	2h49
76-80	10	1h34
81-85	6	1h05

4 Results

In this section, performance results are reported, first gathered with our baseline system, second with the same system but with several sets of acoustic models that were adapted to each age range. The Out-Of-Vocabulary (OOV) rate with the 100K word vocabulary was 0.65% and the perplexity estimated with the 4-gram LM was 150 for the test set.

4.1 Age impact on the baseline system performance

Table 3 presents the WERs obtained with our baseline system. For the entire test set, the WER was 35.3%. As stated earlier, the same system achieved a 18.4% WER with BN speech that, generally speaking, is much more difficult to transcribe than read speech. The much higher WER observed with the present corpus may be explained by the inappropriate LM that is suited for BN data and not for this corpus, which is comprised of a diversity of prompts. Another reason may be the discrepancy of the AMs due to the age mismatch between the speech used to train the baseline MLPs and the elderly speech.

The difference in WER between male and female speakers, 33.5% and 36.0% respectively, was not found to be statistically significant by a one-sided t-test that gave a *p-value* of 0.5539. The greater diversity of female speakers may explain this difference.

Finally, the bottom part of the table reports the WERs according to the subsets of the test data distinguished by the age range of the speakers. A clear increase in WER can be observed with increasing speaker age. One-sided t-tests were performed to assess statistical significance of the WER differences. The alternate hypothesis was: 'the true difference in means is less than 0' between the WERs of the speakers of the first age range (60-65) and the WERs of the speakers of each of the larger age ranges. A p-value of 0.6252 indicated no significant difference with the closest 66-70 age range, but much slower p-values were obtained with the larger age-range (71 and above), with values about 0.03, validating the alternate hypothesis.

Table 3. *Word error rates (WER) of the baseline system on the test set. Detailed WERs on age-range subsets are given in the bottom part of the table. M: Male, F: Female speakers.*

Gender	WER(%)
all	35.3
M	33.5
F	36.0
Age range	WER(%)
60-65	29.1
66-70	28.1
71-75	36.1
76-80	45.1
81-85	41.0
86-90	54.9

Table 4. *WERs of the baseline and the six adapted systems on the test set. (AM for Acoustic Models)*

System	WER(%)
Baseline	35.3
AM-60-65	31.5
AM-66-70	31.4
AM-71-75	31.1
AM-76-80	30.0
AM-81-85	30.0
AM-86-90	33.4

Table 5. *P-values achieved with the MP test performed between the adapted systems.*

	AM-66-70	AM-71-75	AM-76-80	AM-81-85
AM-60-65	.582	.054	.001	.001
AM-66-70		.142	.001	.001
AM-71-75			.001	.001
AM-76-80				.741

4.2 Impact of specific age MLP retraining

In order to further investigate the impact of age on ASR performance, basic adaptation of the acoustic models was tested by simply retraining the baseline MLPs with age-specific data from the train set. All the adapted MLPs shared the same MLP structure as the baseline MLP: 2 hidden layers with 2000 units each and an output layer with 500 units. All the remaining components were identical (the LM, the pronunciation lexicon and the decoding parameters).

Many prompts appear in both the adaptation ("train") and test sets. These prompts were removed from the train set used to adapt the AMs. Furthermore, the 86-90 age range was the one with the least data available: 2 hours (5 hours minus the common sentences with the test set). Experiments not reported here showed that this amount of data to retrain the MLPs led to limited improvements (the MLPs have about 5.7 million weights to re-estimate and the 500 output units need some representation in the adaptation corpus). Hence, we limited the adaptation data amount to 6 hours that was the amount of training data available for the 80-85 age range. Five sets of MLPs were adapted with 6 hours of data for the five age ranges from 60-65 to 80-85. The last one, 86-90, was adapted with the only 2h available. Each set is comprised of three MLPs for the three different feature streams (PLP, RASTA, and MSG), exactly as the baseline system.

Table 4 reports the WER of the baseline and the WERs of the six adapted systems achieved on the test set. 'AM-60-65' for example corresponds to the system where the AMs were adapted with data from 60-65 years old speakers. All the adapted MLPs showed improvement over the baseline, ranging from

10.7% to 15.0% relative. The smaller improvement observed for AM-86-90 may be explained by the smaller amount of adaptation data available for this age range (almost one-third less data).

Since all the systems were tested on the same test data, statistical significance can be assessed directly on the word outputs by a Matched Pairs Sentence-Segment Word Error (MAPSSWE or MP) test with the help of the NIST sc_stats tool. Each of the six adapted system's outputs was tested again the baseline output. All the one-to-one tests showed to be significant at the level of a 0.001 p-value. To determine whether the differences between the adapted systems were significant, the same test was applied to each pair of adapted system word outputs. The p-values are given in Table 5. In general, the outputs of two systems adapted with data of close age ranges did not present significant differences, whereas outputs from disjoint age ranges did, with 0.001 values. This seems to confirm that using adaptation material that matches the speaker age of the test data lead to improvement.

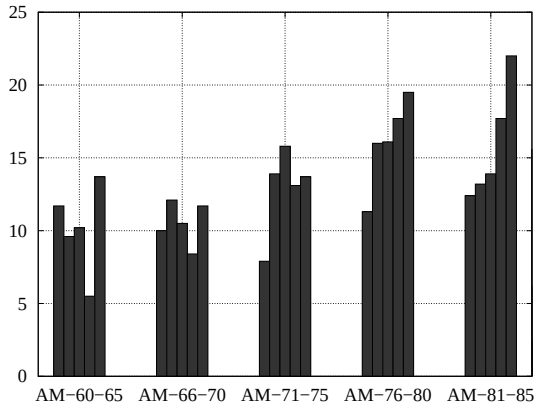

Fig. 1. *Relative differences in WER between the baseline and each of the five systems with age-adapted AMs, for the five age-specific test subsets.*

Results are further illustrated in figure 1 where the Y-axis corresponds to the relative WER differences between the baseline and the WERs obtained with the adapted AMs. The higher the bar, the better the improvement. For each of the five age-specific adapted MLPs on the X-axis, five bars were plotted to give the detail of the improvements according to the five age-specific test subsets. The results of the 86-90 range are not shown since the improvements are smaller due to less adaptation data. For each group of bars, the first one on the left corresponds to the 60-65 test subset, the first neighbor one to 66-70, etc, until the most right-handed bar that corresponds to the 81-85 test subset. As it can be observed, using adaptation data from older speakers gave better results on the test subsets with larger age ranges. For instance, AM-60-65 and AM-81-85

respectively showed 13.7% and 22.0% relative improvements over the baseline for the 81-85 test subset (5.6% and 9.0% absolute respectively). Figure 2 shows the WER points of one of the best adapted system, AM-81-85, with the baseline ones as a function of the age specific test subsets. The adapted curve globally follows the baseline one, with the largest relative gains obtained for the 66-70 and 81-85 age ranges.

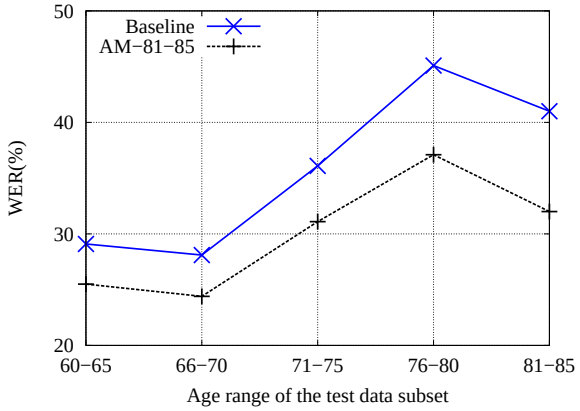

Fig. 2. *Word error rates (WERs) of the baseline and one of the best adapted systems (AM-81-85) as a function of the age-range specific subsets of the test data.*

5 Discussion and future work

In this paper, we presented ASR experiments that illustrate the impact of speaker age on ASR performance. Standard ASR systems use acoustic models typically trained with speech collected from young adult speakers. Hence, ASR performance is expected to decrease when recognizing elderly speech. The impact of aging on speech production and its consequences for ASR have already been well illustrated in the literature but this article reports results achieved on Portuguese, for which no similar study has been published to the best of our knowledge.

A large read speech corpus of European Portuguese elderly speech allowed us to measure statistically significant performance differences among different age groups with 60- to 90-year-old speakers. For instance, an increase of 41% relative (11.9% absolute) in the word error rate was observed between speakers in the 60-65 and 81-86 age groups.

To further illustrate the impact of ageing, preliminary retraining experiments showed that consistent gains in performance can be achieved by simply retraining the baseline MLPs with age-specific data. Differentiated impacts were observed

according to the age range of the adaptation data. However, the limitation of these experiments lies in the fact that the adaptation data was very similar to the test data (similar prompts). Hence, additional experiments that use a completely different test set are needed to draw firmer conclusions on the impact of AM adaptation.

We plan to devise and test other adaptation techniques, for instance the adaptation of the MLP output layer alone may help in case of small amount of adaptation data. To be able to use age-specific ASR systems, one would need to detect the speaker age automatically if no *a-priori* information on it is available. Since chronological age is not a consistent indicator of ageing in speech production, other features (such as jitter and shimmer) will be investigated in order to build a classifier. Linguistic characterization of the errors observed in the ASR experiments will be performed with the objective of better understanding the special needs of elderly speech recognition. Finally, in the long term, we plan to collect elderly speech in a Wizard-of-Oz framework in order to study the interaction of elderly people with dialog systems.

6 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/LA0021/2011.

References

1. J. Wilpon and C. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. ICASSP*, Atlanta, 1996, pp. 349–352.
2. 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.
3. R. Vipperla, S. Renals, and J. Frankel, "Longitudinal study of ASR performance on ageing voices," in *Proc. Interspeech*, Brisbane, 2008, p. 25502553.
4. L. Baeckman, B. Small, and A. Wahlin, "Aging and memory: cognitive and biological perspectives," *Handbook of the psychology of aging*, pp. 349–377, 2001.
5. J. Fozard and S. Gordon-Salant, "Changes in vision and hearing with aging," *Handbook of the psychology of aging*, pp. 241–266, 2001.
6. 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.
7. 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, USA, 2008.
8. H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," Ph.D. dissertation, IST, Lisbon, Portugal, 2008.
9. H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: a broadcast news speech recognition system for the european portuguese language," in *proceedings of PROPOR*, Faro, 2003, pp. 9–17.

10. H. Meinedo, A. Abad, T. Pellegrini, J. Neto, and I. Trancoso, "The L2F Broadcast News Speech Recognition System," in *Proc. Fala*, Vigo, 2010, pp. 93–96.
11. A. Abad and J. Neto, "Incorporating Acoustical Modelling of Phone Transitions in a Hybrid ANN/HMM Speech Recognizer," in *proceedings of INTERSPEECH*, Brisbane, 2008, pp. 2394–2397.