Speaker age estimation for elderly speech recognition in European Portuguese

Thomas Pellegrini\textsuperscript{1}, Vahid Hedayati\textsuperscript{2}, Isabel Trancoso\textsuperscript{2,3}, Annika Hämäläinen\textsuperscript{4}, Miguel Sales Dias\textsuperscript{4}

\textsuperscript{1}Université de Toulouse; UPS; IRIT; Toulouse, France
\textsuperscript{2}INESC-ID, Lisbon, Portugal
\textsuperscript{3}Instituto Superior Técnico, Lisbon, Portugal
\textsuperscript{4}Microsoft Language Development Center, Lisbon, Portugal / ISCTE - University Institute of Lisbon (ISCTE-IUL), Lisbon, Portugal

pellegri@irit.fr, vahidhdyt@gmail.com, isabel.trancoso@inesc-id.pt, t-anhama@microsoft.com, miguel.dias@microsoft.com

Abstract

Phone-like acoustic models (AMs) used in large-vocabulary automatic speech recognition (ASR) systems are usually trained with speech collected from young adult speakers. Using such models, ASR performance may decrease by about 10\% absolute when transcribing elderly speech. Ageing is known to alter speech production in ways that require ASR systems to be adapted, in particular at the level of acoustic modeling. In this study, we investigated automatic age estimation in order to select age-specific adapted AMs. A large corpus of read speech from European Portuguese speakers aged 60 or over was used. Age estimation (AE) based on i-vectors and support vector regression achieved mean error rates of about 4.2 and 4.5 years for males and females, respectively. Compared with a baseline ASR system with AMs trained using young adult speech and a WER of 13.9\%, the selection of five-year-range adapted AMs, based on the estimated age of the speakers, led to a decrease in WER of about 9.3\% relative (1.3\% absolute). Comparable gains in ASR performance were observed when considering two larger age ranges (60-75 and 76-90) instead of six five-year ranges, suggesting that it would be sufficient to use the two large ranges only.

Index Terms: automatic speech recognition, elderly speech, automatic age estimation, i-vector extraction

1. Introduction

In the context of a Portuguese national project called “AVoz”, we studied European Portuguese (EP) elderly speech with the objective of improving speech recognition for elderly speakers. There is no standard age boundary to define the elderly. However, to give an idea, speakers aged above 75 are often called elderly speakers in the literature. Independent of the language in question, speech recognizers’ performance is significantly worse in the case of elderly speech than in the case of young adult speech ([1, 2, 3]). There are several reasons for this. First, some parameters of the speech signal (e.g. speech rate, F0, jitter, shimmer) change with age ([4, 5, 6]), while the acoustic models of speech recognizers are typically trained using speech from younger adults, with elderly speakers not appearing at all or being under-represented in the training data. Second, the elderly usually interact with computers using everyday language and their own commands, even when a specific syntax is required ([7]). Improvements in ASR performance can be achieved by using acoustic models (AMs) specifically adapted to the elderly ([3, 8]). In order to automate the selection of age-specific adapted AMs, one could estimate the age of the speakers automatically.

A number of studies have explored automatic age estimation. In an early study, Minematsu et al. ([9]) estimated speakers’ age only using acoustic (i.e., no linguistic) information. They used Gaussian Mixture Models (GMMs) to distinguish between two groups of speakers defined using the results of previous listening tests: speakers whose speech sounded very old to the judges (“subjective elderly”) and a control group with the rest of the speakers in their databases (“non-subjective elderly”). A correct automatic identification rate of 91\% was achieved, and the rate further increased to 95\% when using additional prosodic features. More recent studies have continued to use techniques derived from the speaker recognition field, such as GMM supervectors ([10, 11, 12]), and, more recently, i-vectors, based on the so-called total variability model ([13]). These techniques involve estimating vectors that somehow characterize the speaker’s voice. Thus, age is a factor that may also be represented in the vectors. Channel compensation techniques may be applied to the vectors to focus on the speaker characteristics only. In these studies, after gathering the supervectors or the i-vectors, Support Vector Machines (SVM) for classification or regression are used to estimate either the age range or the specific age of the speaker. i-vectors have the advantage of producing low-dimension vectors, typically between 200 and 400. In [13], their use also resulted in the best AE performance, as compared with the GMM-supervector technique that had a mean absolute error (MAE) of 7.6 years.

In this study, after giving a bird’s-eye view of AE techniques in Section 2, we report experiments on a subset of a large corpus of read elderly speech in European Portuguese. The corpus is described in Section 3. We used AE to select age-specific adapted AMs and report the results of our ASR experiments in Section 5. To the best of our knowledge, apart from another study of ours [14], previous studies on AE do not report ASR experiments exploiting the results of the AE.

2. Automatic age estimation

Until recently, the GMM supervector paradigm was the state-of-the-art technique in the field of speaker recognition. In a nutshell, this approach consists of training a Universal Background
tances can be projected onto this sub-space to be represented by a sub-space called the total variability sub-space. Speech utterances are extracted ([15]). Instead of using GMM supervectors as such, with standard dimensions greater than 10k, this new approach proposes to represent channel variability and speaker characteristics simultaneously with a low dimensional i-vectors, which have a low dimension that is typically between 200 and 400.

Both the GMM-supervector and the i-vector approaches have already been used for AE in various studies, such as [10, 11, 13]. In [13], the authors used the GMM-supervector approach. Five age classes were used to estimate the age of children between five and ten years of age. The authors used GMM-supervectors and an SVM classifier. Overall precision and recall of 83% and 60% were achieved. The authors also reported slightly worse results when using Support Vector Regression (SVR). However, SVR produced more balanced results among the different age ranges. This approach is similar to the one used in [10], except that the authors compare the effect of both the GMM-supervector and the i-vector approaches on AE performance, on large sets of telephone speech data. The i-vector approach outperformed the GMM-supervector approach, with MAE rates of 7.6, and 7.9 years, respectively.

Seeing that the total variability approach is state-of-the-art both in speaker recognition and in AE, we adopted the i-vector/SVR approach also for our study.

Table 1: Main statistics of the speech material

<table>
<thead>
<tr>
<th>Set</th>
<th>Gender</th>
<th># Spk</th>
<th>Duration</th>
<th># Word Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>female</td>
<td>528</td>
<td>6h30</td>
<td>4.4k</td>
<td>32.3k</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>179</td>
<td>2h17</td>
<td>2.5k</td>
<td>11.6k</td>
</tr>
<tr>
<td>Test</td>
<td>female</td>
<td>225</td>
<td>2h43</td>
<td>2.7k</td>
<td>12.2k</td>
</tr>
<tr>
<td></td>
<td>male</td>
<td>75</td>
<td>1h01</td>
<td>1.4k</td>
<td>5.3k</td>
</tr>
</tbody>
</table>

Figure 1: Histograms of the number of utterances in the training and test sets containing male speakers.

Model (UBM) with a large set of speakers and then performing UBM adaptation (usually using Maximum A Posteriori (MAP) adaptation) to gather speaker-dependent high-dimensional supervectors, composed of the concatenated means of the adapted GMMs.

Significant improvements have been obtained using a new technique, referred to as the total variability approach, in which i-vectors are extracted ([15]). Instead of using GMM supervectors as such, with standard dimensions greater than 10k, this new approach proposes to represent channel variability and speaker characteristics simultaneously with a low dimensional sub-space called the total variability sub-space. Speech utterances can be projected onto this sub-space to be represented by i-vectors, which have a low dimension that is typically between 200 and 400.

Both the GMM-supervector and the i-vector approaches have already been used for AE in various studies, such as [10, 11, 13]. In [13], the authors used the GMM-supervector approach. Five age classes were used to estimate the age of children between five and ten years of age. The authors used GMM-supervectors and an SVM classifier. Overall precision and recall of 83% and 60% were achieved. The authors also reported slightly worse results when using Support Vector Regression (SVR). However, SVR produced more balanced results among the different age ranges. This approach is similar to the one used in [10], except that the authors compare the effect of both the GMM-supervector and the i-vector approaches on AE performance, on large sets of telephone speech data. The i-vector approach outperformed the GMM-supervector approach, with MAE rates of 7.6, and 7.9 years, respectively.

Seeing that the total variability approach is state-of-the-art both in speaker recognition and in AE, we adopted the i-vector/SVR approach also for our study.

Figures 1 and 2 show histograms of the number of utterances used for training and testing in the case of male and female speakers. Only six five-year ranges were considered: 60-65, 66-70 and so on. 72% of the speakers in the age corpus are female.

The number of speakers and the duration of the data used in this study are presented in Table 1. For this work, we only used 10 utterances per speaker, corresponding to about one minute of speech per speaker. We limited the amount of speech per speaker to have an experimental setup similar to that used in NIST speaker identification evaluations, in which the amount of speech is limited to 20-160 seconds per speaker ([17]). The training and test sets contained about 70% and 30% of the utterances in the whole corpus, corresponding to almost 9h and 3h45 of speech, respectively. None of the speakers in the training set appear in the test set.

Figure 2: Histograms of the number of utterances in the training and test sets containing female speakers.

3. Speech material

As in our previous work, we used the EASR Corpus of European Portuguese Elderly Speech ([16]). The corpus contains about 190 hours of read speech, including silences. A total of about 1000 speakers aged 60 or over read out 160 prompts representing 14 different prompt types ranging from isolated digits to phonetically rich sentences. The exact age of the speakers is not known; the age of the speakers is reported using five-year ranges: 60-65, 66-70 and so on. 72% of the speakers in the corpus are female.

The number of speakers and the duration of the data used in this study are presented in Table 1. For this work, we only used 10 utterances per speaker, corresponding to about one minute of speech per speaker. We limited the amount of speech per speaker to have an experimental setup similar to that used in NIST speaker identification evaluations, in which the amount of speech is limited to 20-160 seconds per speaker ([17]). The training and test sets contained about 70% and 30% of the utterances in the whole corpus, corresponding to almost 9h and 3h45 of speech, respectively. None of the speakers in the training set appear in the test set.

Figures 1 and 2 show histograms of the number of utterances used for training and testing in the case of male and female speakers, respectively. Only six five-year ranges were considered: 60-65, 66-70, 71-75, 76-80, 81-85, and 86-90. Older speakers were not considered due to lack of data. The age range and gender distributions in the full corpus were respected when creating the training and test sets. As can be seen in the figures, speakers in the 60-65 age range are the most numerous. In the training sets, there are 706 and 2,070 utterances from the oldest male and female speakers, respectively, as compared with 35 and 99 utterances from the oldest male and female speakers (in the 86-90 age range).

4. Age estimation results

Gender-dependent UBMs of 1024 Gaussian mixtures were trained on the male and female training sets. The acoustic features consisted of 13 Mel-Frequency Cepstral Coefficients (MFCCs), including energy, with their first order derivatives, resulting in 26-d feature vectors extracted every 10 ms with 20 ms Hamming window frames. Energy-based speech activity detection was used, followed by mean-variance feature normalization. 200-d i-vectors were extracted for each utterance, both for the training and the test sets. The ALIZE toolbox was used to...
perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.

As shown in Section 3, there are progressively less data from speakers in each five-year age range from 60-65 to 86-90. Therefore, a second experiment was carried out by down-sampling the number of i-vectors from the youngest speakers before training the SVR model. In this experiment, each five-year range had the same number of i-vectors as the training set. In total, six sets of AMs were derived from the baseline MLP, corresponding to the six age-specific data from the training set. In Table 2, the original unbalanced training set was used to adapt the baseline MLP to perform the UBM training and the i-vector extraction ([18]). Since there was only one session per speaker, no channel compensation was applied. The same recording setup was used for all the speakers, so the total variability matrix is expected to model the speaker characteristics, including age. To perform the regression, we used the WEKA machine learning toolbox [19].

Table 2 shows the age estimation results in terms of MAEs. With the original unbalanced training set and six age ranges, the experiments yielded global MAE values of 5.45 and 5.68 years for male and female speakers, respectively. These values are slightly larger than the five-year maximum precision that we could expect. The age of the speakers in the EASR Corpus is reported using five-year ranges, from 60-65 to 86-90, so the maximum precision is 5 years. We also considered two age ranges only: from 60 to 75 and from 75 to 90. As expected, the MAE is smaller in this case: 4.18 and 4.53.
In this study, we investigated automatic age estimation as a pre-processing step to selecting age-specific AMs for elderly speech recognition. Our objective was to verify that AE based on i-vectors is efficient enough for a model selection frontend in ASR. A positive result, which we obtained, suggests that it might not be necessary for elderly users to spend time adapting ASR systems to their voices. We used a large corpus of read speech from European Portuguese speakers aged 60 or over. The exact age of the speakers in the corpus is not known; the age of the speakers is reported using five-year age ranges. We performed AE using i-vectors and support vector regression. When using six five-year age ranges from 60-65 to 86-90, we obtained mean error rates of about 5.4 and 5.7 years for male and female speakers, respectively. When only using two larger age ranges, 60-75 and 75-90, the error rates decreased to 4.2 and 4.5. The selection of five-year-range adapted AMs, based on the automatically estimated age ranges of the test speakers, led to a decrease in WER of 9.3% relative (1.3% absolute) over the WER of 13.9% obtained using a baseline ASR system without AM adaptation to elderly speech. Only small differences were observed when only using the two larger age ranges. Thus, it only seems necessary to use two sets of adapted AMs to recognize speakers aged 60 and over.

In future work, we will compare these results with speaker adaptation in the context of a hybrid HMM/MLP system. Furthermore, in collaboration with the Microsoft Language Development Center in Lisbon, we are currently carrying out AE experiments with speakers representing a wide range of ages from three years old until old age [14].

7. Acknowledgments

This work was partially supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under projects PTDC/EEA-PLP/121111/2010, PTDC/EIA-CCO/122542/2010, and PEst-OE/EEI/LA0021/2011, and also by the QREN 5329 Fala Global project, which is co-funded by Microsoft and the European Structural Funds for Portugal (FEDER) through POR Lisboa (Regional Operational Programme of Lisbon), as part of the National Strategic Reference Framework (QREN), the national program of incentives for Portuguese businesses and industry.

8. References


