INTERSPEECH 2009 BRIGHTON

Age Verification Using a Hybrid Speech Processing Approach

Ron M Hecht[1], Omer Hezroni[1], Amit Manna[1], Ruth Aloni-Lavi[1], Gil Dobry[1], Amir Alfandary[1], Yaniv Zigel[2]

[1] PuddingMedia, Kfar-Saba, Israel
[2] Bio-medical Engineering Dept., Ben-Gurion University, Beer-Sheva, Israel

hadasron@gmail.com, omer.hezroni@puddingmedia.com, amit.manna@puddingmedia.com, ralonilavi@gmail.com, gil.dobry@gmail.com, amir.alfandary@gmail.com, yaniv@bgu.ac.il

10.21437/Interspeech.2009-71

Abstract

The human speech production system is a multi-level system. On the upper level, it starts with information that one wants to transmit. It ends on the lower level with the materialization of the information into a speech signal. Most of the recent work conducted in age estimation is focused on the lower-acoustic level. In this research the upper lexical level information is utilized for age-group verification and it is shown that one's vocabulary reflects one's age. Several age-group verification systems that are based on automatic transcripts are proposed. In addition, a hybrid approach is introduced, an approach that combines the word-based system and an acoustic-based system. Experiments were conducted on a four age-groups verification task using the Fisher corpora, where an average equal error rate (EER) of 28.7% was achieved using the lexical-based approach and 28.0% using an acoustic approach. By merging these two approaches the verification error was reduced to 24.1%.

Index Terms: age verification, age estimation, speech processing, word-based approach, hybrid approach

1. Introduction

In their book, Rabiner and Juang [1] describe the human speech production mechanism as a four-stage process. The speech production process begins with the message formulation stage. During this stage the information that one wants to transmit is materialized into a sequence of words. This stage is followed by the language code stage, which adds the prosody and converts the sequence of words to a sequence of phonemes. The last two stages are the neuro-muscular control stage and the vocal tract system stage. The output of these stages is the speech signal.

The speaker's age affects the speech in several ways, resulting in varied output for each of these four stages. An interesting example of the difference in the message formulation stage among age groups is the variation in each group's vocabulary. For example, in the US, young people are much more likely to use the words "actually" and "like" compared to elderly people. During the language code stage, the phoneme sequence is altered as well. While elderly people pronounce a word using one sequence of phonemes, young people pronounce it using a different set of phonemes [2]. A similar process occurs at the last stages as well. Different phonemes are pronounced differently across the age span, due to anatomic changes in the vocal tract structure and the vocal cords.

Both word-based [2][3][4][5] and acoustic approaches are applied successfully in speaker recognition. However, until now the majority of the published work [6][7][8] conducted on age estimation was focused on the lower level stages of speech production, i.e. the neuro-muscular control stage and the vocal tract system stage. In our current work, an improvement in age-group verification performance was achieved by combining a lower level system: acoustic Gaussian mixture model – support vector machine (GMM-SVM) system with a higher level word-based recognition system.

This work was supported in part by the Ministry of Industry and Trade, grant number 40183

2. Acoustic Age Estimation System

The acoustic age estimation system used is a GMM-SVM system [9]. The proposed acoustic age estimation system is composed of five main phases: four training phases and a single test phase. Each of these phases is a set of several consecutive stages. The block diagram of this system is shown in figure 1.

Training phase *A* is the most basic procedure in the system; it provides the universal background model (UBM) [10]. Training phase *B* then produce a base in the supervector space known as the anchor space [11]. Training phase *C* creates an SVM model for each age group within the anchor space [12]. Lastly, in training phase *D*, a log likelihood ratio (LLR) model is generated within the SVM score space.

Throughout the training and testing the feature extraction output is a feature set that includes 13 mel-frequency cepstral coefficients (MFCC) and their first derivatives such that the total dimension of the feature-space is 26.

In phase *A*, feature vectors – extracted from speech conversations database – are used to estimate a UBM model, a 1024-order GMM, according to the maximum-likelihood (ML) criterion. The GMM UBM estimation procedure is similar to procedure commonly used in speaker verification and language recognition tasks.

In phase *B*, in order to have a good representation of anchor models, a large and diverse speaker conversation database is used. For each conversation in the database, a GMM is estimated using UBM MAP adaptation and is transformed to a high-dimensional $(1024 \times 26 = 26624)$ supervector, **s**. Each of these supervectors is an anchor model and together span the anchor space. Combining the supervectors together forms the matrix **A** which is the base of the anchor space. After this stage, every conversation, *j*, is represented by its projection on the anchor space $\left(\mathbf{s}_j^T \mathbf{A}\right)$. By using a set of hundreds of conversations for the creation of the anchor space, the number of dimensions of a model can be reduced from tens of thousands to a few hundred.

In phase *C*, a set of age-group labeled conversations is transformed to a set of points in the supervector space and then projected onto the anchor space as described above. A linear SVM model for each age group can thus be estimated, yielding a linear separator $\boldsymbol{\omega}_i$, where *i* is the age group index. Every conversation, *j*, is now represented within the SVM models' score space:

$$\begin{bmatrix} \mathbf{s}_j^T \mathbf{A}\boldsymbol{\omega}_0 \\ \vdots \\ \mathbf{s}_j^T \mathbf{A}\boldsymbol{\omega}_N \end{bmatrix} \quad (1)$$

In phase *D* the goal is to normalize the different SVM scores produced from different SVM classifiers by using LLR scoring. This is done by estimating the parameters ($\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ - the mean vector and the covariance matrix of the i^{th} Gaussian) of a diagonal Gaussian G_i for each age group in the SVM score space.

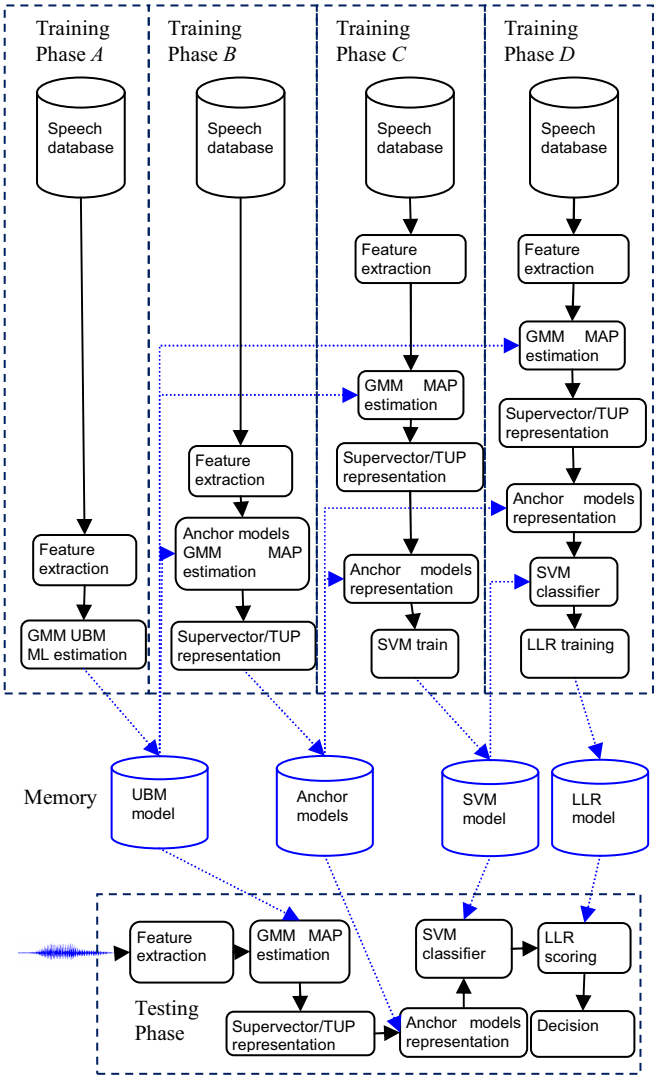

Figure 1: *The acoustic age estimation system.*

The goal of the testing phase in this system is to verify the age-group of an unknown age-group speech utterance. This goal is achieved by the use of the four model types that were estimated in training phases *A* through *D*. Throughout the test phase the utterance changes its representation from raw features and a GMM to a supervector. The supervector is then projected onto the anchor space and used to calculate the SVM scores. These scores form a point in the SVM score space. Ultimately, the verification decision is made according to the likelihood of the LLR models:

$$S_i(\mathbf{s}_j) = G_i\left(\begin{bmatrix} \mathbf{s}_j^T \mathbf{A}\boldsymbol{\omega}_0 \\ \vdots \\ \mathbf{s}_j^T \mathbf{A}\boldsymbol{\omega}_N \end{bmatrix} \middle| \; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i \right) \quad (2)$$

3. Word-based Age Estimation System

The second proposed system is the word-based age-group verification system. It has two training phases and a test phase. The block diagram of the system is shown in figure 2.

All the phases have the same first two stages: word recognition and word level feature extraction. In the first stage, the audio is transformed to a sequence of words using a large-vocabulary continuous speech recognition (LVCSR) system. In the next stage, the word sequence is transformed to word level features. While there are a variety of word level feature types [3][4][5], we decided to extract *N*-gram features [2]. Our experiments were conducted on unigrams $p(w_i)$, where w_i is the i^{th} word, unigrams of pairs of words $p(w_i, w_{i-1})$ and bigrams $p(w_i|w_{i-1})$. For each of these features only the *M* most common *N*-grams were used. The values of *M* varied from a small vocabulary of 50 words that consists mainly of stop words to a large vocabulary consisting of thousands of words that covers a significant part of the spoken words. All the *N*-grams that were estimated are unsmoothed.

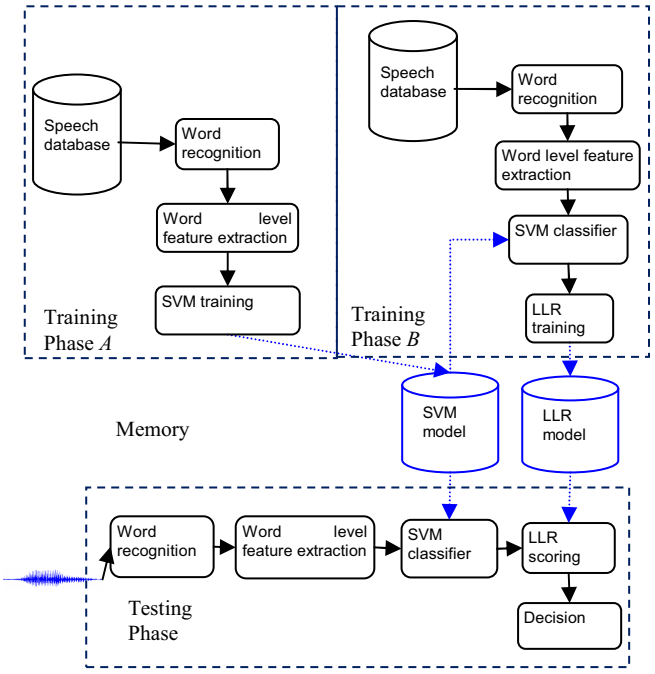

Figure 2: *The word-based age estimation system.*

In training phase *A*, SVM models with a radial base function (RBF) [13] kernel are trained. For each age group a single model was trained to discriminate it from the other groups. In training phase *B* the goal is to normalize the different SVM scores by LLR scoring. This is done by estimating a diagonal Gaussian for each age group in the SVM score space (as described in the previous section). The outputs of the training phases are LLR models and SVM models.

During the Testing phase, the word-level features are first scored by the SVM models. The results of the different SVM models are then rescored by the LLR Gaussian model and verification decision is made.

4. Merging Technique

The two proposed systems, acoustic and word-based, are of different nature; however both contain valuable information for speaker age verification. While the word-based system processes the first stage of the speech production mechanism, where information is transformed into words, the acoustic system utilizes the last stage, where the words are communicated through the vocal tract. Therefore, we assume that the systems' errors should be uncorrelated. Intuitively, the hybrid approach will generate better results because although a speaker may have a voice or use vocabulary that is not typical for his age, the likelihood that the speaker has both traits is significantly lower.

Since the two systems produce uncorrelated log-likelihood scores, our first merging technique was to sum the scores of the two systems. As a second alternative, a linear SVM combiner [3][4][5] was trained since it can be viewed as a generalization of the former technique.

5. Experiments and Results

This section describes a set of experiments that were conducted on LDC Fisher English part I and II speech corpora [14]. The task that was chosen was a four age-groups verification task where the age group of a given conversation is verified. The ranges for the age groups were: below 25 years old, between 25 and 40 years old, between 41 and 55 years old, and above 55 years old. Each side of a conversation was treated as a different call.

5.1. Corpora Description

Currently the largest and most diverse from the speaker point of view corpora are the Fisher corpora. Due to those two unique characteristics they were chosen and all of the experiments were conducted on them. Fisher has Thousands of calls for each age group and given each call is about ten minutes long, we were provided with tens of thousands of minutes of relevant audio. These calls were recorded from thousands of distinct speakers.

Approximately 11,000 conversations from the Fisher part II corpus were used for training, while tests were conducted on more than 8,000 conversations from the Fisher part I. Conversations in the original Fisher part I corpus were omitted from the tests if the speaker of a conversation was present in the Fisher part II corpus. More detailed information on the sizes of the training and testing corpora is given in Table 1. The merging models were trained on the Fisher I using a two-fold jackknife approach.

Table 1. *Corpora training and testing sizes (in number of conversations). Each conversation is about 10 minutes long.*

Age-group	Training	Testing
0-25	2595	1894
26-40	4202	3309
41-55	2853	2155
56+	1114	868

5.2. Results

The experiments focused on a word-based approach and on merging techniques for age group verification. Therefore all of the experiments were conducted using the same acoustic system. The performance of this system was 28.0% average EER (average EER of a random classifier is 50.0%). The results of all the experiments are shown in Figure 3.

Figure 3 demonstrates that one's choice of vocabulary indicates one's age. The hybrid approach outperformed all other systems which indicates that the acoustic system and the word-based system are uncorrelated. Three sets of word-based features were used: word unigram, pair unigram and word bigram $P(w_i), P(w_i, w_j), P(w_i|w_j)$. The word unigrams result in lower average EER than the other two sets when M is small, however, when M increases, the performance of the word bigrams and word pairs unigrams sets improves. Throughout the experiments the pair unigrams system was consistently more accurate than the word bigrams'. The difference in performance between the two merging techniques was small and therefore only the SVM approach results are shown in Figure 3.

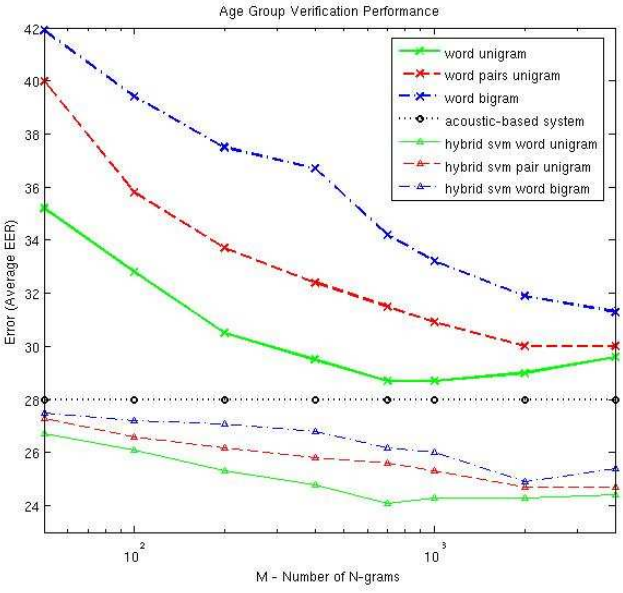

Figure 3: *Results of four age groups verification errors on Fisher part I corpus.*

6. Discussion

A good intuition to the advantages of the hybrid approach can be obtained by looking at the "under 25 years old" verification task aiming in verifying that a speaker's age is less than 25. The results of this task are summarized at figure 4. The X axis represents the LLR score estimated during the test phase of the "under 25 years old" verification task (Figure 4.a - word based LLR scores, Figure 4.b - acoustic LLR scores). The Y axis represents the real age of the speakers.

It is expected that during the test phase a high score would be given for speakers under 25 and low scores for speakers above that age. Each row in the figure is the conditional distribution of the LLR scores for a given true age. The value of z (gray level) at location (x, y) on the graph is the conditional probability for that point $z = P(s = x \mid True\_age = y)$. For example, the probability that a 34 years old speaker will have a score value of 8 at figure 4.a is about 0.2.

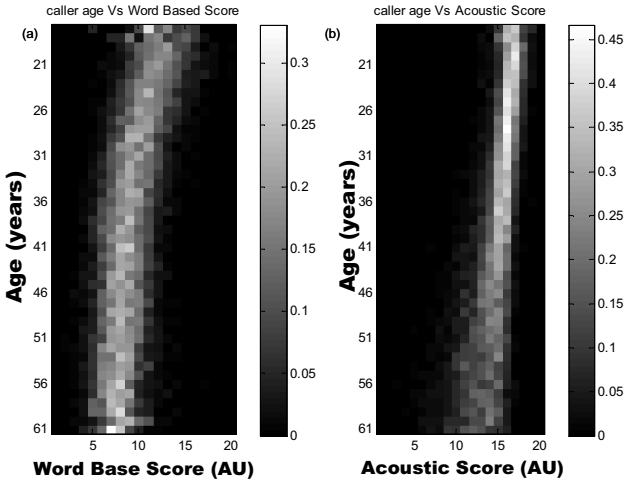

Figure 4: *The conditional distribution of the "under 25 years old" score in arbitrary units (AU) (a) for the word-based system, (b) for the acoustic system.*

It can be observed in the two images of figure 4 that the acoustic scores and the word based scores exhibit differences in behavior for different age values. The first difference is in the LLR scores' variance. While the scores' variance of the word base system only slightly changes for different age values, the acoustic scores' variance becomes broader as the age increases.

The second difference is the change is the means of the scores. The word-based system scores deteriorate rapidly at the top of the figure, as the age of the speakers increases until the age of thirty. After the age of thirty, the scores remain relatively constant in the low range of the scores. In the acoustic system a different picture is observed. Until the age of forty there is a constant decline of the score. After the age of forty there is a rapid increase in the score variance. The differences in the behavior of the two systems results in a synergetic hybrid approach.

7. Conclusions

Three sets of age verification experiments were conducted. All the experiments were performed on the same corpora. In the first set of experiments an acoustic system was used while the second set was based on a word-based, lexical-level system. The last set of experiments involved a hybrid system that consisted of the two former systems.

It was demonstrated that on both levels of the speech signal, the acoustic and word levels, the signal was altered due to the speaker's age, proof that these levels contain information that is relevant to the verification of one's age. In addition, it is shown that the two levels are orthogonal in nature and that the combination of these approaches yields improved performance.

Intuitively one can imagine a scenario that a person's speech does not match their age by acoustic means or that a person's vocabulary is not appropriate for their age. However, the likelihood of a third scenario in which a person's speech will not match their age by acoustic means and by vocabulary is significantly lower than the two former scenarios.

8. References

[1] Rabiner, L. and Juang, B. H., Fundamental of Speech Recognition, pp 48-49, Pearson, 1993.

[2] Doddington, G., "Speaker Recognition Based on Idiolect Differences between Speakers", in Proc. of Interspeech, 2001.

[3] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A. and Leek, T. R., "High-Level Speaker Verification with Support Vector Machine ", Proceeding of ICASSP, 2004.

[4] Ferrer, L., Graciarena, M., Zymnis, A. and Shriberg, E., "System Combination Using Auxiliary Information For Speaker Verification", in Proc. of Interspeech, 2008.

[5] Shriberg, E., "High-Level Features in Speaker Recognition", Speaker Classification I, pp 241-259, Springer, 2007.

[6] Metze, F., Ajmera, J., Englert, R., Bub, U., Burkhardt, F., Stegmann, J., Muller, C., Huber, R., Andrassy, B., Bauer, J. G. and Little, B., "Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications", in Proc. of ICASSP, 2007.

[7] Minematsu, N., Sekiguchi, M. and Hirose, K., "Automatic Estimation of One's Age with His/Her Speech Based Upon Acoustic Modeling Techniques of Speakers", in Proc. of ICASSP, 2002.

[8] Bocklet, T., Maier, A., Bauer, J., Bukhardt, F. and Noth, E., "Age and Gender Recognition for Telephone Applications Based on GMM Supervectors and Support Vector Machines" in Proc. of ICASSP, 2008.

[9] Aronowitz, H. and Noor, E., "Efficient Language Identification Using Anchor Models and Support Vector Machines" in Proc. of Odyssey, 2006.

[10] Sturim, D. E., Reynolds, D. A., Dunn, R. B. and Quatieri, T.F., "Speaker Verification Using Text-Constrained Gaussian Mixture Models", in Proc. of ICASSP, 2002.

[11] Collet, M., Mami, Y., Charlet, D. and Bimbot, F., "Probabilistic Anchor Models Approach for Speaker Verification", in Proc. of Interspeech, 2005.

[12] Dehak, R., Dehak, N., Kenny, P. and Dumouchel, P., "Kernel Combination for SVM Speaker Verification" in Proc. of Odyssey, 2008.

[13] Cristianin, N. and Shawe-Taylor, J., An introduction to support vector machines and other kernel-based learning methods, Cambridge University Press, 2000.

[14] Cieri, C. et al., "Fisher English Training" Linguistic Data Consortium, Philadelphia, 2005.