INTERSPEECH 2009 BRIGHTON

The Role of Age in Factor Analysis for Speaker Identification

Yun Lei, John H.L. Hansen

CRSS: Center for Robust Speech Systems
University of Texas at Dallas, Richardson, Texas 75083,USA
yxl059200@utdallas.edu, John.Hansen@utdallas.edu

Abstract

The speaker acoustic space described by a factor analysis model is assumed to reflect a majority of the speaker variations using a reduced number of latent factors. In this study, the age factor, as an observable important factor of a speaker's voice, is analyzed and employed in the description of the speaker acoustic space, using a factor analysis approach. An age dependent acoustic space is developed for speakers, and the effect of the age dependent space in eigenvoice is evaluated using the NIST SRE08 corpus. In addition, the data pool with different age distributions are evaluated based on joint factor analysis model to assess age influence from the data pool.

Index Terms: speaker identification, age analysis, factor analysis.

1. Introduction

Automatic speaker identification systems suffer performance loss due to two dominant factors: limited training data and channel mismatch. Gaussian mixture models and support vector machines, are two conventional approaches for text-independent speaker identification, and have been proven to be effective methods. Recently, the factor analysis model has been successfully applied for GMM based speaker identification systems [1, 2, 3, 4]. A series of simplified algorithms for factor analysis have also been suggested in recent studies [5, 6, 7]. In channel compensation, Eigenchannel modeling, as an application of factor analysis, in the model domain [8] or feature domain [9] can be employed for summarizing distortions of the session/utterance by a small number of parameters in a lower dimensional subspace, which are called the channel factors. In speaker model adaptation, there are also many successful approaches based on factor analysis, such as the eigenvoice model [1], and inter-speaker variability model [2]. The eigenvoice model, based on the assumption that a majority of speaker variation can be modeled with a small number of variables, greatly reduces the number of parameters (e.g., speaker factors) to be estimated for the new speaker. Also, the inter-speaker variability model is a combination of the eigenvoice model and classical MAP model, which attempts to leverage the benefits from both models. In eigenvoice, the speaker acoustic space is described by a rectangular matrix; and in the inter-speaker variability model, the speaker acoustic space is described by a combination of rectangular and square matrices. The square matrice reflects speaker specific information which can be considered as an efficient complement to the rectangular matrix that reflects common speakers traits, so as to describe more accurately the entire speaker acoustic space.

This project was funded by AFRL under a subcontract to RADC Inc. under FA8750-05-C-0029.

Although the speaker acoustic space for the eigenvoice model is composed of latent factors, many observable factors are available, such as gender, age, and dialect. Previous attempts to automatically estimate age based on speaker voice have shown that age is an observable and important factor of the speaker traits [10]. As such, it is believed that the age factor plays an important role in the speaker acoustic space, although it is difficult to judge exactly the age of an individual using present day speech algorithms.

In this study, the role of age in the speaker acoustic space of both the eigenvoice and inter-speaker variability models are examined. The primary contribution is a formal understanding of the effect of observable age for the acoustic space on performance of speaker identification systems.

2. Review of Factor Analysis

The Gaussian Mixture Model has been a standard in the fields of speaker verification and identification. Factor analysis, leading to an adaptation model, has been successfully applied for GMM-based systems. In speaker recognition, the supervector obtained by concatenating all mean vectors in the GMM corresponding to a given speaker is applied in the representation of the speaker. Since the supervector m is speaker-dependent, it can be decomposed into a sum of two supervectors, a speaker-dependent supervector s and a speaker-independent (UBM) supervector m_0:

$$m = m_0 + s, \tag{1}$$

where m_0 and s are statistically independent. Both are assumed to be of the same CF dimensions, where C is the number of components in the GMM, and F is the dimension of the acoustic feature vectors. The speaker-independent supervectors m_0 can be obtained using UBM and the distribution of the speaker-dependent supervectors s is assumed to have a hidden variable description of the form:

$$s = v \cdot y + d \cdot z, \tag{2}$$

where v is a rectangular matrixes of low rank and y is a normally distributed random vector with less dimension; d is a $CF \times CF$ diagonal matrix and z is a normally distributed random vector with CF dimension. The model is defined as the inter-speaker variability model in joint factor analysis, where the term vy can be considered to be the eigenvoice MAP and the term dz can be considered as classical MAP. To train the inter-speaker variability model, joint estimation [11] or decoupled estimation [2] can be employed for estimation of the matrices v and d. If only rectangular matrix v and random vector y are considered, the speaker-dependent supervector s can be described using the eigenvoice model:

$$s = v \cdot y. \tag{3}$$

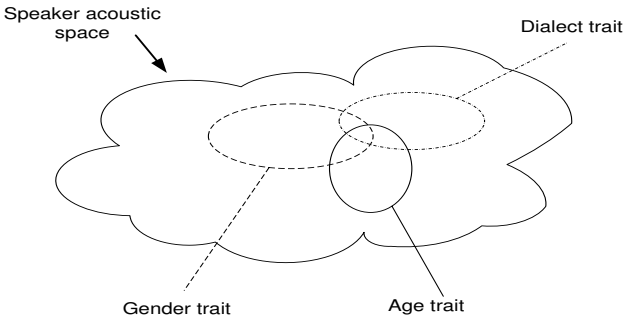

Figure 1: The relationship of gender, age, and dialect factors in the speaker acoustic space.

For training, Maximum likelihood re-estimation can be employed in the processing of each matrix estimation. The likelihood function which serves as the estimation criterion is summarized below:

$$\amalg_s max P_{GMM}(\chi(s)|M_0 + s, \Sigma), \tag{4}$$

where $\chi(s)$ denote the training data for speaker s, and Σ denotes the diagonal matrix. The data pool used for training includes a large amount of speakers with multi-sessions in order to suppress the influence from channel and session variability. The number of speakers is assumed to be balanced and sufficiently large to cover all types of speaker traits.

3. Age Role Analysis

3.1. Age Role in Eigenvoice Model

In the eigenvoice model, the speaker is described by eigenspace v and vector y. Here, eigenspace v describes a speaker acoustic space covering the latent major speaker traits and vector y provides the coordinates of the speaker in the speaker acoustic space. It is noted that the vector y can be considered as a coordinate instead of a Gaussian distribution when the variance of the Gaussian distribution is sufficiently small. However, some details of the speaker traits are observable, such as age, gender, and dialect. It is known that a speaker's voice can be employed to identify speaker gender, dialect, and age by training dedicated classifiers, which means these factors play significant roles in the speaker acoustic space. Fig.1 illustrates the relationship of these factors. Note: the overlap among these factors in the figure reflects a dependency of the factors. This concept will be employed in the proposed method shown in the flow diagram in Fig.2. In this study, the age factor is considered as a means to improve the description of the speaker acoustic space since age knowledge can be obtained in many speaker ID corpora. It is believed that some part of the eigenspace v represents the speaker's age knowledge. Although the speaker's age is not a direct feature of the speaker's voice, the speaker's age can be usually judged on the basis of voice to within a reasonable range [12]. In addition, many data corpora used for training are not balanced for age, such as Switchboard II Phases 2 and 3; Switchboard Cellular Part 1 and 2, and Fisher corpora. Most speakers in these corpora are from 20 to 40 years old, resulting in a shortage of coverage for older speakers' traits. In order to investigate the age factor of the eigenvoice model, the

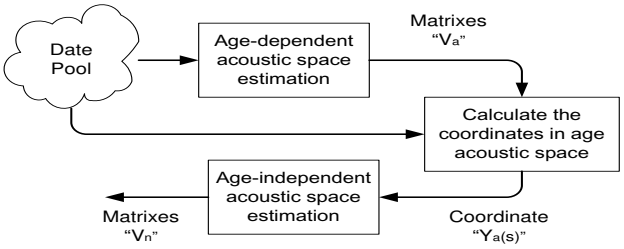

Figure 2: Estimation of acoustic space v_a and v_n.

eigenspace v is split into two parts: age-dependent part and age-independent part. As such, the eigenvoice model is re-written as:

$$s = v_a \cdot y_a + v_n \cdot y_n, \tag{5}$$

where the term $v_a y_a$ represents the speaker age trait and the term $v_n y_n$ represents other traits except age. Note, the splitting here is based on an assumption that these two parts of the model are independent. As such, although the two terms are named age-dependent and age-independent parts, the first term is present in the shared age-related knowledge for all speakers, and the second term is present for other speaker traits independent of the first term but probably still dependent at some level to age.

In the new model, age factor is displayed directly comparing against latent age factor estimation in the eigenvoice model. In the training of the matrixes v_a, age knowledge of all speakers can be used, which means additional knowledge (age) beside speaker labels, are utilized in the estimation of the speaker acoustic space so that performance improvement is expected with the same number of parameters. In the estimation process of matrix v_n, the term $v_a y_a$ are subtracted for each speaker to keep the independence between the term $v_a y_a$ and the term $v_n y_n$. A review of the processing of factor analysis estimation in [2], requires that we first calculate the order Baum-Welch statistics $\widetilde{F}_c(s)$ as follows,

$$\widetilde{F}_c(s) = \sum_t \gamma_t(c)(Y_t - m_c), \tag{6}$$

where $\gamma_t(c)$ is the posterior probability of Gaussian c for the feature vector Y_t; and m_c is the subvector of m_0 corresponding to the mixture component c. In the proposed model, the statistics $\widetilde{F}_c(s)$ are calculated to estimate v_a in the same manner as eigenvoice. However, the statistics $\widetilde{F}_c(s)$ are updated in the estimation of matrix v_n since the age term $v_a y_a$ must be removed to maintain independence between the two terms. As such, the first order Baum-Welch statistics $\widetilde{F}_c^{v_a}(s)$ and $\widetilde{F}_c^{v_n}(s)$ in estimation of v_a and v_n can be obtained below:

$$\widetilde{F}_c^{v_a}(s) = \sum_t \gamma_t(c)(Y_t - m_c), \tag{7}$$

$$\widetilde{F}_c^{v_n}(s) = \sum_t \gamma_t(c)(Y_t - m_c - v_a y_a(s)), \tag{8}$$

where $y_a(s)$ is the coordinate of the speaker in the age acoustic space. The diagram flow of the estimation of matrices v_a and v_n is presented in Fig.2. Similarly, the supervecter of the enrollment speaker can be obtained using matrices v_a and v_n space in the model adaptation phase. To accomplish this , the coordinate $y_a(s)$ of the speaker in the age acoustic space is calculated. Next, the statistics $\widetilde{F}_c^{v_n}(s)$ are updated using the coordinate y_a

and the age acoustics matrix v_a, to calculate the coordinate y_n. Finally, the supervector of the speaker is generated using m_0, the age-dependent term, and the age-independent term.

3.2. Age Role in Inter-Speaker Variability Model

If the vector y is named "common factors", and vector z can be named "specific factors" [11], then the common factors account for most of the variance in the data, and the term dz serves as a residual to compensate. In [2], Kenny suggests the term dz to model the residual inter-speaker variability which is not captured by a large set of eigenvoices, using the decoupled estimation of v and d with two different data pools. Since the dimension of d is fixed, the age knowledge cannot be added into the matrix d using the similar algorithm in the eigenvoice model. However, the data pool used to train d can be designed based on the age distribution. If the data pool used here are not balanced across age levels, then the term dz will still not capture the traits of the speakers at some ages. If the data used for training v includes more younger speakers, then the d trained on the data with more older speakers can be expected to be a better compensation for the term vy that describe the acoustic space of both younger and older speakers. Alternatively, if both data pools used for training v and d includes more younger speakers, then the models can describe the acoustic space of younger speakers very well, but their reliability decreases for older speakers.

4. Experiments

4.1. Evaluation Data Set

The system is evaluated on the telephone-telephone condition of the NIST 2008 speaker recognition evaluation (SRE). Only male trials are evaluated in our experiments since the age distribution of the data pool used in factor analysis is significantly unbalanced. There are 648 enrollment speakers and 895 test speaker, with a total of 12551.

4.2. Factor analysis training data

Here, the Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, the Fisher English corpus, and the NIST 2004 and 2005 SRE enrollment data are used to train the gender-dependent UBM with 1024 mixtures. Next, Switchboard II Phases 2 and 3, Switchboard Cellular Part 1 and 2, and the NIST 2004 and 2005 SRE enrollment data (totalling 1090 speakers with 11919 files) are used to train eigenvoice model with 300 factors. Here, only speakers with 5 or more recording sessions are considered. The Fisher English Corpus (totalling 1458 speakers with 4374 files) is used to train the diagonal matrix d. The NIST 2004 and 2005 SRE enrollment data (368 speakers with 2883 files) is used to estimate the channel factor and channel compensation is performed in the feature domain.

To evaluate the role of age, Fig.3 and 4 show the age distribution of the data pools for training the acoustic space v and d. Note, since there are no age labels for the NIST 2004 and 2005 SRE enrollment data, Fig.3 only includes Switchboard II Phases 2 and 3, and Switchboard Cellular Part 1 and 2 (722 speakers with 9036 files).

4.3. Baseline System description

For parametrization, a 39-dimension feature (13 MFCC +Δ + $\Delta\Delta$) with 25 ms analysis windows and 10 ms shift, filtered by feature warping using a 3-s sliding window is employed. In order to set aside silence, the phone recognizer developed from

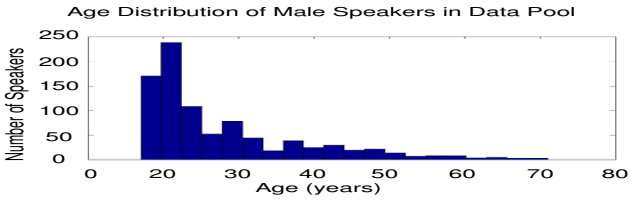

Figure 3: Age distribution of speakers in data pool training acoustic space v.

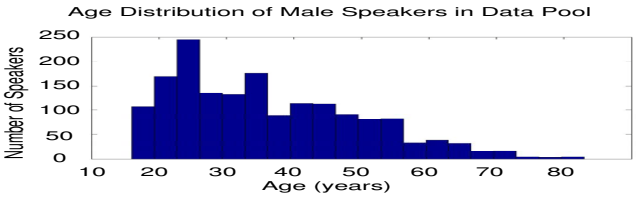

Figure 4: Age distribution of speakers in the Fisher corpus.

BUT is used. Next, gender dependent UBMs consisting of 1024 Gaussians are trained via ML with the number of gender dependent speaker factors set to 300, and the channel factor size of 50 is used to compensate channel mismatch. The matrix d is trained with a fixed v and u. Maximum likelihood re-estimation is used to train all matrices (including v, u, and d), and only mean update is used in the system in order to save computational resources. The posterior probability $\gamma_t(c)$ is calculated from the UBM in all employed models. No normalization steps are used for the baseline system. Finally, a standard 20-best Expected Log Likelihood Ratio is used for scoring. The performance of the baseline system is included in Table 1.

4.4. Age role of Eigenvoice model

In order to investigate the influence of age factor, the total count of factors describing the speakers is fixed in order to ensure that the size of the speaker acoustic space is consistent. In the estimation of the age acoustic space, although exact age knowledge can be obtained form these corpora, age segmentation is employed in our process, since many factors, such as dialect, pitch, and speech rate, can influence the perceived age, which requires that a wide segmentation be used here. As such, there are only two classes (such as younger and older) used to distinguish age knowledge in this study. In order to establish a reasonable age segmentation principle, the distribution of age in the corpora used for training v are considered. Based on Fig.3, two 10-year periods from the age domain are used to represent younger and

Table 1: Basic performances of the system. ("CC" means channel compensation in feature domain)

Algorithms	EER (%)
MAP	12.69
MAP with CC	8.23
300 speaker factors, d = 0 with CC	10.06
300 speaker factors, d ≠ 0 with CC	9.07

Table 2: Performance of speaker identification over eigenvoice models with age factors.

Eigenvoice model	EER (%)
300 speaker factors	10.06
2 gender and 298 speaker factors	9.90

Table 3: Three subsets of Fisher corpus used evaluate the age role in the training of d.

Definition	# of speakers	# of files
All speakers	1458	4374
Younger Speakers (< 35 years)	667	2002
Older Speakers (> 35 years)	791	2372

older male subjects. If the speaker is older than 20 and younger than 30, he is tagged as 0 (younger); if the speaker is older than 40 and younger than 50, the speaker is tagged as 1 (older). It is believed that two factors are sufficient to describe the age-dependent acoustic space with only two age classes. Here, it is emphasized that the two factors are used to distinguish the age of the speaker, which reflective of an age classifier with an overall soft decision.

In Table 2, we show the results for the eigenvoice model with and without the 2 age factors after channel compensation is performed in the feature domain, where a total of 300 factors are used. It is noted that the eigenvoice model with 2 age factors can produce better results than the eigenvoice model without an age factor with the constraint of the same amount of speaker factors, which means the speaker acoustic space estimated by age dependent matrix and age independent matrix is more accurate than the speaker acoustic space estimated by the eigenvoice model alone without age knowledge. The relative 1.16% EER improvement can be obtained by replacing only 2 speaker factors with 2 age factors from the total 300 speaker factor set which means there are only $2/300$ percentage of the speaker acoustic space used to describe the age dependent acoustic space. Again, the term "age dependent" means shared age knowledge. While this improvement is small, it is important to note it represents an important aspect for researchers to ensure in developing models for speaker ID.

4.5. Age Role for an Inter-Speaker Variability Model

Since the dimension of the diagonal matrix d is fixed, the age dependent factors cannot be considered separately. However, the data pool can be re-designed to balance the age distribution. Three data pool sets are used here to evaluate the age role in the training of d. The definition and size of these data pool sets are shown in Table 3. Since the data pool set used to train eigenvoice includes more younger speakers, The "Older Speakers" data pool is expected to achieve better performance than the other two sets. The performance of these three data pool sets are shown in Table4. It is seen that although the older speaker data pool set includes the least amount data compared with all three data pool sets, the performance is the best versus the other two corpora. This result shows that effective balances of age information can improve overall performance of speaker identification.

Table 4: Performance of speaker identification over different data pool used to train d.

data pool	EER (%)
All speakers	9.07
Younger speakers	9.13
Older speakers	8.81

5. Conclusion

The effect of age factor in the speaker acoustic space for an eigenvoice model has been considered. Using known age knowledge, the role of age was directly described by an age dependent acoustic space, as part of the overall speaker acoustic space in an eigenvoice model. The proposed approach was evaluated on the NIST SRE08 evaluation corps, with an overall improvement in performance versus that from the original eigenvoice model, with an appropriate number of age factors in the size-fixed speaker acoustic space. In addition, different data pool sets were designed with alternative age distribution, in the training of the diagonal matrix d in the joint factor analysis models and matrix d, as a residual of the eigenvoice models, trained on a well-designed data pool. The resulting study shows that while age distribution is generally ignored in speaker ID, careful attention in developing the age sensitive eigenvoice models can result in an improved overall system.

6. References

[1] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. on SAP*, vol. 13, pp. 345–354, 2005.

[2] P. K. P. Oueleet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification," *IEEE Trans. on ASLP*, vol. 16, pp. 980–988, 2008.

[3] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernochy, "Analysis of feature extraction and channel compensation in a gmm speaker recognition system," *IEEE Trans. on ASLP*, vol. 15, pp. 1979–1986, 2007.

[4] R. Vogt, B. Baker, and S. Sridharan, "Factor analysis subspace estimation for speaker verification with short utterances." INTERSPEECH-2008.

[5] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification." INTERSPEECH-2005.

[6] D. Matrouf, N. Scheffer, B. Fauve, and J. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification." INTERSPEECH-2007.

[7] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified." ICASSP-2005.

[8] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification." ICASSP-2004.

[9] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso, and P. Laface, "Channel factors compensation in model and feature domain for speaker recognition." IEEE Odyssey: Speaker and Lang. Rec. Workshop, 2006.

[10] N. Minematsy, M. Sekiguchi, and K. Hirose, "Automatic estimation of perceptual age using speaker modeling techniques." EUROSPEECH-2003.

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. on ASLP*, vol. 15, pp. 1435–1447, May 2007.

[12] E. Ryan and H. Capadano, "Age perceptions and evaluative reactions toward adult speakers," *Journal of Gerontology*, vol. 33, pp. 98–102, 1978.