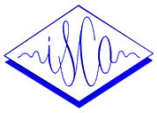

International Symposium on Chinese Spoken Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

UNSEEN HANDSET MISMATCH COMPENSATION BASED ON FEATURE/MODEL-SPACE *A PRIORI* KNOWLEDGE INTERPOLATION FOR ROBUST SPEAKER RECOGNITION

Jyh-Her Yang and Yuan-Fu Liao

Department of Electronic Engineering & Institute of Computer and Communication,
National Taipei University of Technology, Taipei 106
yfliao@ntut.edu.tw, http://www.ntut.edu.tw/~yfliao

ABSTRACT

Unseen but mismatch handset is the major source of performance degradation for speaker recognition in telecommunication environment. In this paper, an unseen handset characteristics estimation method based on *a priori* knowledge interpolation (AKI) is proposed. AKI could be applied in both the feature and model space to interpolate the feature and model transformation functions measured using stochastic matching (SM) and maximum likelihood linear regression (MLLR), respectively. Cross-validation experimental results on HTIMIT database showed that the average speaker recognition rate could be improved from 59.6%/57.8% to 73.8%/66.8% for seen/unseen handsets. It is therefore a promising method for robust speaker recognition.

1. INTRODUCTION

A speaker recognition system in the public telephone switched network (PTSN) needs to be robust against distortion of different handsets. However, some mismatch handsets may not be seen in advance, i.e., unseen handsets, and will cause significant performance degradation. Moreover, the characteristics of the handset and speaker are usually tightly mixed or coupled together. To separate them is essentially a difficult one-to-many mapping problem, unless some knowledge about the unseen handset is available in advance.

Several successful compensation techniques for handset mismatch have been applied. They include cepstral mean subtraction (CMS) [1], RelAtive SpecTrA (RASTA) [2], feature transformation (FT) [3] and speaker model synthesis (SMS) [4]. CMS and RASTA blindly compensate the handset distortion by filtering or subtracting the long-term average of handset-distorted cepstral features vectors. On the other hand, FT and SMS rely on a handset detector to discriminate the types of handsets, such as carbon button or electret, to remove the distortion or to adapt/synthesize speaker models using the pre-trained handset-dependent characteristics.

However, the CMS and RASTA methods may remove not only the characteristics of handset but also the speaker's as well. Moreover, knowledge about the distortion of those unseen handset is usually not available beforehand. Therefore, the handset detector-based approaches may have problem to deal with test speech from unseen handsets. They may select the most likely handset from a set of seen handsets, simply reject them as out-of-handset (OOH) [3] or have to fall back to CMS-based system.

To alleviate the unseen but unavoidable handsets, the *a priori* knowledge interpolation (AKI) approach is proposed in this paper. The concept of AKI is to first collect a set of characteristics of seen handset as the *a priori* knowledge to construct a space of handsets. During evaluation, the characteristic of a test handset is estimated and compensated by interpolating the set of the *a priori* knowledge.

In detail, the estimate $\hat{h}$ of the characteristic of a test handset is defined as:

$$\hat{h} = \sum_{n=1}^{N} \alpha_n h_n \qquad (1)$$

where $H = \{h_n, n = 1 \sim N\}$ is the set of *a priori* knowledge collected from *N* seen handsets, and α_n are the interpolation weights. Moreover, h_n could be the feature-space or the model-space transformation functions between an enrollment handset and the *n*-th seen handsets, measured by SM [5] or MLLR [6], respectively.

This paper is organized as follows: Section 2 briefly reviews the SM and MLLR methods for collecting the *a priori* knowledge. The proposed AKI for unseen handset compensation is described in Section 3. In Section 4, cross-validation experimental results are reported on the well-known HTIMIT database [7]. Some conclusions and future works are given in the last section.

2. *A PRIORI* KNOWLEDGE COLLECTION

SM and MLLR are two mature methods for speaker adaptation.

In this paper, SM and MLLR are applied in the feature and model space for collecting feature and model transformation functions between all available seen handsets and the enrollment handset, respectively.

2.1. SM feature-space transformation

In the feature space, the relationship between speech signal y_n from the *n*-th seen handset and x from the enrollment handset could be represented as follows:

$$y_n = x + b_n \tag{2}$$

where b_n is a bias vectors and could be measured using SM. Therefore, the feature-space *a priori* knowledge is the collection of *N* biases, i.e., $H = \{b_n, n = 1 \sim N\}$.

2.2. MLLR model-space transformation

In the model space, the *s*-th speaker model $\tilde{\Lambda}_{n,s} = \{\tilde{u}_{n,s,m}, \tilde{\Sigma}_{n,s,m}, m = 1 \sim M\}$ generated from the *n*-th seen handset and its corresponding model $\Lambda_s = \{u_{s,m}, \Sigma_{s,m}, m = 1 \sim M\}$ built from the enrollment handset could be related using the following linear regression equation:

$$\tilde{u}_{n,s,m} = A_n \cdot u_{s,m} + b_n \tag{3}$$

$$\tilde{\Sigma}_{n,s,m} = C_{n,m}^T T_{s,m} C_{n,m} \tag{4}$$

where $T_{s,m}$ is the variance transformation to be estimated, A_n, b_n, and $C_{n,m}$ are the mixture mean transformation matrix and bias and the inverse function of Choleski factor of the variance matrix $\Sigma_{s,m}^{-1}$ of the m-th mixture component of the *n*-th seen handset. Therefore, the model-space *a priori* knowledge is the collection of *N* tied mixture mean biases and transformation matrices, and variance transformation matrices, i.e., $H = \{A_n, b_n, T_n, n = 1 \sim N\}$.

3. THE PROPOSED AKI APPROACH

The AKI method is essentially based on the *a posteriori* weighted interpolation method. However, to evaluate the reliability of applying the *a priori* knowledge to test speech, a divergence measure is computed. The divergence is further converted into reliability and utilized to adjust the interpolation weights to emphasize the reliable part of the *a priori* knowledge.

The diagram of the AKI approach is shown in Figure 1. It includes several Gaussian mixture model (GMM)-based handset models, *a posteriori* estimator, a divergence-based reliability measure and the interpolation weights generator. The detail procedures of the AKI will be described in the following section.

3.1 Interpolation algorithm

Assume that there are N seen handsets that speakers are likely to use. In the training phase, the handset characteristics $H = \{h_1, h_2, ..., h_N\}$ of these N seen handsets and their corresponding handset GMMs $\Lambda = \{\Lambda_1, \Lambda_2, ..., \Lambda_N\}$ are first computed from the speech observations of each handset $O_n = \{o_{n,1}, o_{n,2}, ..., o_{n,T}\}$. The set of H and Λ are then used to represent the space of handsets.

During the evaluation, the input test speech observations $O' = \{o'_1, o'_2, ..., o'_{T'}\}$ of a speaker from an unknown handset are fed into the handset models to compute N likelihoods $L(O' \mid \Lambda_n)$. These N likelihoods $L(O' \mid \Lambda_n)$ are further transformed into the *a posteriori* probabilities $p(\Lambda_n \mid O')$ using the following equation:

$$p(\Lambda_n \mid O') = \frac{\exp\left(L(O' \mid \Lambda_i)\right)}{\sum_{n=1}^{N} \exp\left(L(O' \mid \Lambda_n)\right)} \tag{5}$$

To evaluate the reliability of applying the *a priori* knowledge to the input test speech, a divergence measure is utilized to compare the distribution of the *a posteriori* vector P

$$P = \left[p(\Lambda_1 \mid O'), p(\Lambda_2 \mid O'), ..., p(\Lambda_N \mid O')\right] \tag{6}$$

with a uniform distribution reference vector U

$$U = \left[\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}\right] \tag{7}$$

The equation of the divergence measure (or called Jensen difference) $J(P,U)$ is defined as follows:

$$J(P,U) = S\left(\frac{P+U}{2}\right) - \frac{1}{2}\left[S(P) + S(U)\right] \tag{8}$$

where $S(\cdot)$, called the Shannon entropy, is given by

$$S(Z) = -\sum_{n=1}^{N} z_n \log z_n \tag{9}$$

and z_n is the *n*-th component of Z.

The divergence measure $J(P,U)$ is further converted into the reliability measure R using a zero-one $\mathrm{sigmod}(\cdot)$ function defined as:

$$R = \frac{1}{1+\exp(-\lambda(-J(P,U)+\beta))} \quad (10)$$

where λ and β are the parameters of the $\mathrm{sigmod}(\cdot)$ function.

If P approximates U, R is close to zero, then the *a priori* knowledge is not reliable. It means the observation speech O' may come from a handset never seen before. On the other hand, if P is far away from U, R is close to one, the *a priori* knowledge is reliable. It means the observation speech O' may come from a handset which is very similar to one of these seen handsets.

Finally, the reliability measure R is utilized to adjust the interpolation weights α_n to emphasize the *a priori* knowledge of the dominant handsets using the following equation:

$$\alpha_n = \frac{\exp(R \cdot L(O \mid \Lambda_n))}{\sum_{n=1}^{N} \exp(R \cdot L(O \mid \Lambda_n))} \quad (11)$$

3.2 Score fusion

Moreover, to handle some outliers which are totally uncovered by the pre-collected *a priori* knowledge, the recognition scores S_1 of the AKI and the scores S_2 of the conventional CMS-based speaker recognizer are nonlinearly fused to get the final scores S_f (see Figure 2) using the following equation:

$$S_f = \frac{1}{\gamma}\log\left[\frac{\exp^{\gamma \cdot S_1} + \exp^{\gamma \cdot S_2}}{2}\right] \quad (12)$$

where γ controls the degree of nonlinear fusion. It is worthy to note that by using this fusion function, the final scores will be dominated by AKI/CMS if the AKI/CMS scores are much larger that the other. On the other hand, the final scores will be the average of two recognizers, if their values are close.

4. EXPERIMENTS

4.1. HTIMIT and experiment conditions

To evaluate the effectiveness of the proposed AKI approach, the well-known HTIMIT database [7], which was recorded for studying the handset mismatch problem, was chosen. There were in total 384 speakers, each gave ten utterances using a Sennheizer head-mounted microphone (called senh). The set of 384*10 utterances was then playback and recorded through nine other different handsets include four carbon button (called cb1, cb2, cb3 and cb4), four electret (called el1, el2, el3 and el4) handsets, and one portable cordless phone (called pt1).

However, in this paper, all experiments were performed on 356 speakers including 178 females and 178 males. For training the speaker models, the first 16 seconds speech of each speaker from the senh handset was used as the enrollment speech. The other ten four-second sessions of each speaker from ten handsets were used as the evaluation data, respectively.

To construct the speaker models, a 256-mixture universal background model (UBM) was first built from the enrollment speech of all 356 speakers. Then, for each speaker, a maximum *a posteriori* (MAP)-adapted GMM [8] adapted from the UBM using his own enrollment speech was built. Beisde, 38 mel-frequency cepstral coefficiences (MFCCs) including 12 MFCCs, 12 Δ-MFCCs, 12 Δ^2-MFCCs, Δ-log-energy and Δ^2-log-energy were computed with window size of 30 ms and frame shift of 10ms.

4.2 Cross-validation experiments

First, the speaker recognizer using the conventional CMS method to remove the handset bias was evaluated as the baseline (called MAP-GMM/CMS). The result was shown in Table 1. The average recognition rate of 59.6% was achieved. Compared with the one reported in [7], the baseline results were promising.

Secondly, the leave-one-out cross-validation strategy is used to evaluate the proposed AKI method under the unseen handset mismatch situation. In brief, one of the nine handsets (cb1~4, el1~4 and pt1) was chosen in turn as the unseen handset and removed from the set of the *a priori* knowledge. The remaining nine handsets (including senh) were used as the seen handsets. Therefore, there were in total 9 cross-validation recognition turns (90 experiments).

The proposed AKI compensation method was then tested in the feature space. One feature bias for each handset was measured by SM and used as the *a priori* knowledge. Beside, 256-mixture UBMs were trained for seen handset models. The average speaker recognition rate was significantly improved from 59.6% to 70.4% (see Table 1).

Finally, the proposed AKI compensation method was tested in the model space. The speech was divided into one or three classes (speech or consonant, vowel and silence). For each class, a MLLR mixture mean offset, a transformation matrix and a variance scaling factor were measured for each handset. The average speaker recognition rates were further improved to 73.6% and 73.8% (see Table 1) using one and three classes, respectively.

Moreover, the average speaker recognition rate of the unseen handsets in the nine cross-validation tests were separated and shown in Table 2. It showed that the AKI method could increase the performance from 57.8% (MAP-GMM/CMS) to 64.6%, 66.8% and 66.8% in the feature space and model space using one and three classes, respectively. Therefore, the results in Table 1 and 2 showed that the proposed AKI method could efficiently compensate the mismatch for both seen and unseen handsets.

5. CONCLUSIONS & FUTURE WORKS

In this paper, the AKI method is proposed to alleviate the problem of unseen handset mismatch. Unlike the conventional hard-decision handset detector-based approaches, which may choose incorrect handset type or have to reject OOHs or fall back to CMS-based system, the proposed soft-decision method can seamlessly deal with both seen and unseen handsets. It is therefore a promising method for robust speaker recognition.

In the future, the AKI method will be improved using eigen-vector analysis to product a compact and orthogonal handset space when the number of seen handsets is enough. Beside, the *a priori* knowledge interpolation weights α_n will be optimized using the maximum likelihood algorithm.

6. REFERENCES

[1] R. Mammone, X. Zhang, R. Ramachandran "Robust speaker recognition", *IEEE Signal Processing Magazine*, p.58-71, September 1996.

[2] H. Hermansky and N. Morgan, "RASTA processing of speech", *in IEEE Transactions on Speech and Audio Processing*, vol. 2, num. 4, pp. 578-589, Oct, 1994.

[3] S. Y. Kung, M. W. Mak, and C. L. Tsang:"Stochastic feature transformation with divergence-based out-of-handset rejection for robust speaker verification", *EURASIP J. on Applied Signal Processing*, 2003.

[4] R.Teunen, B.Shahshahani and L.P.Heck, "A modelbased transformational approach to robust speaker recognition", *Proc. ICSLP*, 2000.

[5] Ananth Sankar and Chin-Hui Lee, "A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 4, no. 3, pp.190-202, May 1996.

[6] Leggetter, C.J. and Woodland, P.C., "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech Lang.*, Vol. 9, pp. 171-185, 1995.

[7] D. A. Reynolds: "HTIMIT and LLHDB: Speech corpora for the study of handset transducer effects", *in Proc. ICASSP'97*, Vol. II, pp. 1535-1538, 1997.

[8] D. Reyolds, T. Quatieri and R.Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, Vol. 10, pp. 19-41, January 2000.

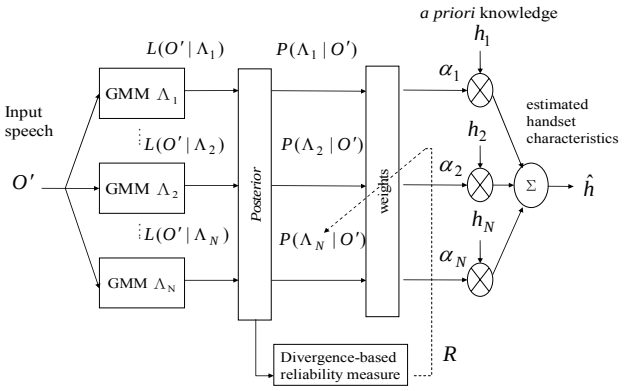

Figure 1: The proposed *a priori* knowledge interpolation (AKI) scheme for handset characteristics estimation.

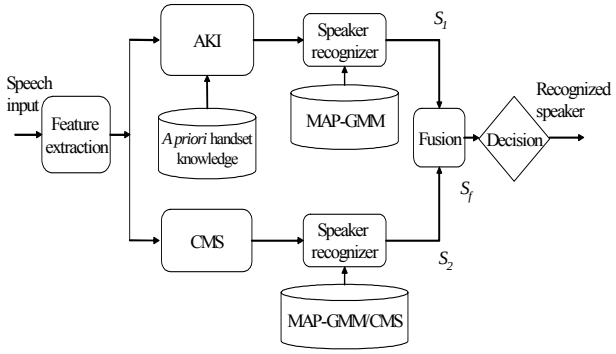

Figure 2: The proposed AKI and MAP-GMM/CMS fusion scheme for handset characteristics estimation.

Table 1: The average speaker recognition rates (%) of the nine cross-validation evaluations on the HTIMIT database achieved by the MAP-GMM/CMS and AKI+fusion methods in the feature and model space, respectively.

	#. of classes	Average
MAP-GMM/CMS	-	59.6
Feature-space AKI+Fusion	1	70.4
Model-space AKI+Fusion	1	73.6
	3	73.8

Table 2: The speaker recognition rates (%) of the unseen handsets in the nine cross-validation evaluations on the HTIMIT database achieved by the MAP-GMM/CMS and AKI+fusion approaches in the feature and model space, respectively.

	#. of classes	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average
MAP-GMM/CMS	-	69.4	71.9	28.1	37.6	74.7	62.4	59.6	63.2	53.1	57.8
Feature-space AKI+Fusion	1	79.8	78.1	29.2	45.5	84.0	66.0	68.5	71.9	58.7	64.6
Model-space AKI+Fusion	1	80.1	78.9	35.7	54.5	83.7	66.0	68.3	73.0	60.7	66.8
	3	78.7	79.5	38.8	56.7	83.7	67.1	65.7	71.1	60.1	66.8