PROSODY MODELING AND EIGEN-PROSODY ANALYSIS FOR ROBUST SPEAKER RECOGNITION

Zi-He Chen[1], Yuan-Fu Liao[2] and Yau-Tarng Juang[1]

[1]Department of Electrical Engineering, National Central University, Chung-Li, Taoyuan, 32054, Taiwan
[2]Department of Electronic Engineering, National Taipei University of Technology, Taipei 106, Taiwan
yfliao@ntut.edu.tw, http://www.ntut.edu.tw/~yfliao

ABSTRACT

Unseen handset mismatch and limited training/test data are the major source of performance degradation for speaker identification in telecommunication environment. In this paper, a vector quantization (VQ)-based prosody modeling and an eigen-prosody analysis (EPA) is integrated to transform the close-set speaker identification problem into a full text document retrieval-similar task. The prosody modeling labels the prosodic feature contours of a speaker's speech into sequences of prosody states. EPA then constructs a compact eigen-prosody space to represent the constellation of speakers. Furthermore, EPA is fused with a lower-level *a priori* knowledge interpolation (AKI) handset distortion compensator to complement each other. Experimental results on the HTIMIT database had shown that about 41.0% and 32.8% relative error rate reduction for seen and unseen handsets, respectively, was achieved comparing with the maximum *a priori*-adapted Gaussian mixture model/cepstral mean subtraction (MAP-GMM/CMS) baseline.

1. INTRODUCTION

A speaker identification system in telecommunication network environment needs to be robust against distortion of mismatch handsets. However, some mismatch handsets may not be seen in advance, i.e., unseen handsets, and will cause significant performance degradation. To address this problem, prosodic features, which are known to be less sensitive to handset mismatch, are attractive recently. Several successful techniques have been proposed including the distribution [1], the N-gram [2] and the discrete hidden Markov model (DHMM) [2] approaches.

In the distribution approach, the per-frame pitch and energy values are extracted and modeled using traditional distribution models, such as the conventional Gaussian mixture models (GMMs). In the N-gram approach, the dynamics of the pitch and energy trajectories are described by sequences of symbols and modeled by the n-gram statistics. In the DHMM method, the sequences of prosody symbols are further modeled by the state observation and transition probabilities. However, the distribution approach may not adequately capture the temporal dynamic information of the prosodic feature contours and the N-gram and DHMM methods usually require large amount of training/test data to reach a reasonable performance.

In this paper, a VQ-based prosody modeling and an eigen-prosody analysis approach are integrated together to add robustness to conventional cepstral features-based GMMs close-set speaker identification system under the situation of mismatch unseen handsets and limited training/test data.

The idea of EPA previously proposed in [3] is modified to first build a VQ-based prosodic modeling to label the prosodic feature contours of a speaker's enrollment speech into sequences of prosody states. EPA then treats the sequences of prosody states as a text document which records the detail prosody/speaking style of the specific speaker. The speaker's evaluation speech is also labeled and uses as the query keywords to recall the most related prosody document (speaker). By this way, the speaker identification problem is transformed into a full-text document retrieval-similar task and the latent semantic analysis (LSA) [4] is applied to build an eigen-prosody space to represent the constellation of speakers.

Furthermore, since EPA utilizes prosody-level information, it could be further fused with acoustic-level information to complement each other. In this paper, the *a priori* knowledge interpolation (AKI) [5] approach is chosen. It utilizes the maximum likelihood linear regression (MLLR) to compensate handset mismatch and is capable to deal with unseen handsets.

This paper is organized as follows. Section 2 describes the VQ-based prosody modeling. Section 3 briefly reviews the AKI approach which uses the MLLR model transformation. Section 4 describes the EPA and given some analysis examples. Section 5 reports the experimental results evaluated on the well-known HTIMIT database [6]. Some conclusions are given in the last section.

2. VQ-BASED PROSODIC MODELING

2.1. Prosodic modeling

In this study, syllables are chosen as the basic processing units. Five types of prosodic features are chosen including the slope of pitch contour and lengthening factor of a vowel segment, average log-energy difference and value of pitch jump between two vowels and pause duration between two syllables. Moreover, the prosody features are normalized according to underline vowel class to remove any non-prosodic effects.

To build the prosodic model, the prosodic features are vector quantized into M codewords using the expectation-maximum (EM) algorithm. For example, in a preliminary experiment, an 8-codeword prosodic modeling was build from the enrollment speech of the HTIMIT database (described in detail in Section 5.1), the centroid of each codeword is shown in Table 1.

After some statistics and cross-examining between the codewords, the distribution of the transition matrix (Table 2)

and the occurrence positions of codewords in utterances, the meaning of the codewords (called states from now on) could be identified. For example, state 6 is the phrase-start; state 3 and 4 are the major-breaks. This model could then be used to label the prosodic status of an input utterance.

Table 1: The centroids of the 8-state prosodic model trained from the enrollment speech of HTIMIT database.

Feature/State	1	2	3	4	5	6	7	8
Pitch slop	-0.1	**0.7**	-0.1	-0.2	0.1	0.3	-0.2	**-2.5**
Energy diff.	-0.4	-0.5	-0.8	**-1.9**	-0.1	0.2	**1.3**	0.1
Pitch jump	-0.2	-0.2	1.3	**1.4**	-0.1	**-0.9**	0.3	-0.6
Lengthening	0.3	**-0.5**	0.3	**1.4**	0.1	-0.1	0.1	0.1
Pause	0.4	**-0.5**	0.5	**2.6**	0.2	-0.3	0.3	0.1

Table 2: The state transition matrix of the 8-state prosodic model trained from the enrollment speech of HTIMIT database.

	1	2	3	4	5	6	7	8
1	3424	1256	854	429	1059	2783	919	304
2	1304	599	255	209	451	1282	344	192
3	347	122	77	55	109	405	109	43
4	20	18	5	3	18	50	10	3
5	1074	510	237	167	348	894	286	91
6	3218	1544	621	364	1005	2804	891	351
7	882	392	255	102	330	829	416	162
8	331	180	100	63	95	349	129	98

2.2. Prosodic labeling

By feeding the prosodic feature contours of a speaker's utterance into the prosodic model, an utterance could be labeled as sequences of prosody state indices. A typical example of the prosody state labeling of an HTIMIT utterance using the 8-state prosodic modeling are shown in Figure 1.

By this way, the prosodic phenomenon of each vowel segment is mapped into a meaningful prosodic state and the sequences of prosodic state indices could be treated as a prosody text document to book the prosodic behavior of the speaker.

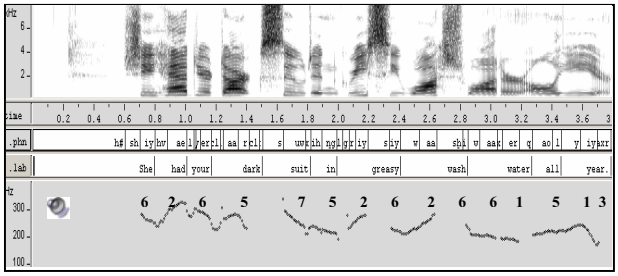

Figure 1: A typical prosodic state labeling of an input utterance using the 8-state prosodic modeling trained from the enrollment speech of HTIMIT database (bottom panel: pitch contour).

3. AKI UNSEEN HANDSET ESTIMATION

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. AKI could be applied in both feature and model spaces. In this paper, model-space AKI using the MLLR model transformation is chosen to compensate the handset mismatch.

The estimate of the characteristic $\hat{h}$ of a test handset is defined in Equation 1.

$$\hat{h} = \sum_{n=1}^{N} \alpha_n h_n \tag{1}$$

where $H = \{h_n = (A_n, b_n, T_n), n = 1 \sim N\}$ is the set of *a priori* knowledge, i.e., the tied MLLR mean and variance transformation matrices, collected from N seen handsets, and α_n are the interpolation weights controlled by handset *posteriori* probabilities [5]. Beside, the MLLR model transformation is defined as:

$$\begin{aligned} \tilde{u} &= \hat{h} \cdot u = \hat{A} \cdot u + \hat{b} \\ \tilde{\Sigma} &= C^T \hat{T} C \end{aligned} \tag{2}$$

Where $\tilde{u}$ and $\tilde{u}$ are the original and adapted mixture mean vectors, respectively, $\tilde{\Sigma}$ is the adapted variance and C is the inverse function of Choleski factor of the original variance matrix Σ^{-1}.

4. EIGEN-PROSODY ANALYSIS

The procedures of the EPA (see Fig.2) includes: (1) VQ-based prosodic modeling and labeling to convert the prosodic feature contours of a speaker into sequences of prosody states, (2) segmenting the sequences to extract important prosody keywords, (3) calculating the occurrences statistics of these prosody keywords for each speaker to form a prosody keyword-speaker occurrence matrix, (4) applying the singular values decomposition (SVD) technique to decompose the prosody keyword-speaker occurrence matrix to build an eigen-prosody space to represent the constellation of speakers, and (5) measuring the speaker distance using the cosine of the angle between two speaker vectors in the eigen-prosody space. The procedures (see Figure.2) are briefly described in the following subsections.

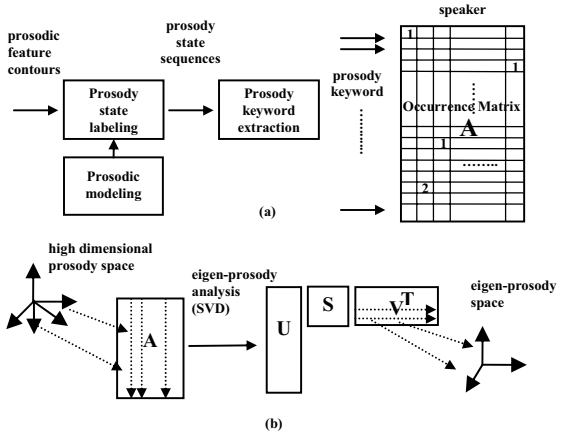

Figure 2: The proposed scheme of the eigen-prosody analysis for robust close-set speaker identification: (a) construction of the prosody keyword-speaker occurrence matrix and (b) eigen-prosody space analysis using SVD.

4.1. Prosody keyword extraction

After the prosodic state labeling, the prosody text documents of all speakers are searched to find important prosody keywords in order to establish a prosody keywords dictionary. Essentially, the compilation of the dictionary can be treated as an unknown-word extraction problem and an N-gram approach for finding high frequency collocations is adopted.

First, all possible combinations of the prosody words, including single words and word pairs (uni-gram and bi-gram), are listed and their frequency statistics are computed. After calculating the histogram of all prosody words, frequency thresholds are set to leave only high frequency ones.

4.2. Prosody keyword-speaker occurrence matrix statistics

The prosody text document of each speaker is then parsed using the generated prosody keywords dictionary by simply giving higher priority to longer words. The occurrence counts of keywords of a speaker are booked in a prosody keyword list vector to represent the long-term prosodic behaviors of the specific speaker. Therefore, the prosody keyword-speaker occurrence matrix A is made up of the collection of all speaker prosody keyword lists vectors. Moreover, to emphasize the uncommon keywords and to deemphasize the very common ones, the inverse document frequency (IDF) [4] weighting method is applied.

4.3. Eigen-prosody analysis

In order to reduce the dimension of the prosody space, the sparse prosody keyword-speaker occurrence matrix A is further analyzed using SVD to find a compact eigen-prosody space. Specifically, given an m by n ($m >> n$) matrix A of rank R, A is decomposed and further approximated using only the largest K singular values as:

$$A \approx U_K \Sigma_K V_K^T \tag{3}$$

where A_K, U_K, V_K, and Σ_K matrices are the rank reduced matrices of the respective matrices.

A typical example of the eigen-prosody space by the analysis of the senh enrollment set is shown in Figure 3. By this way, EPA is capable to give a compact eigen-prosody space to model the long-term prosodic behaviors of the speakers.

4.4. Score measurement

The problem of the speaker identification is now formulated as a pseudo-document testing as in the LSA approach. The test utterances of a test speaker are first labeled and parsed to form the pseudo query document y_Q, and then transformed into the query vector v_Q in the eigen-prosody speaker space by

$$v_Q = y_Q^T U_K \Sigma_K^{-1} \tag{4}$$

The distance between the test speaker and the i-th registered speaker is defined as the cosine of the angle between the query vector v_Q and the i-th speaker vector $v_{K,i}$

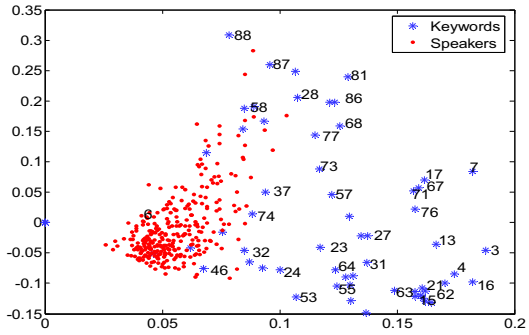

Figure 3: A typical distribution of the prosody keywords and speakers castellation on the two dimensional eigne-prosody space using the 8-state prosodic modeling trained from the enrollment speech of HTIMIT database.

5. SPEAKER IDENTIFICATION EXPERIMENTS

5.1. HTIMIT database and experiment conditions

To evaluate the effectiveness of the proposed EPA approach, the well-known HTIMIT database [6], 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 302 speakers including 151 females and 151 males which have all the ten utterances. For training the speaker models, the first seven utterances of each speaker from the senh handset were used as the enrollment speech. The other ten three-utterance 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 302 speakers. Then, for each speaker, a MAP-adapted GMM (MAP-GMM) [7] adapted from the UBM using his own enrollment speech was built. 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. The pitch contours of all utterances were extracted using the popular Wavesurfer/Snack sound toolkit. Moreover, phone-level segmentations from TIMIT corpus were used to extract THE five prosodic features.

5.2. Cross-validation experiments

First, the MAP-GMM speaker identification using the CMS method to remove the handset bias was evaluated as the baseline (called MAP-GMM/CMS). The average identification rate of 60.5% (shown in Table 3) was achieved. Compared with the one reported in [6], the results was promising.

Secondly, the leave-one-out cross-validation strategy is used to evaluate three fusion approaches including AKI+, EPA+

and EPA+AKI+MAP-GMM/CMS 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 identification turns, 90 experiments.

The proposed AKI+MAP-GMM/CMS fusion system was tested. The speech was divided into three classes (speech or consonant, vowel and silence). For each class, a MLLR mixture mean offset and transformation matrix and a variance scaling factor were measured for each handset. The average speaker identification rate was rising to 72.2% (see Table 3).

The proposed EPA+MAP-GMM/CMS fusion approach was also evaluated. From the senh subset, a 32-state prosodic modeling was trained and 367 prosody keywords were extracted to form a sparse 367*302 dimensional matrix A. While using three dimensional eigen-prosody space, identification rate of 70.5% was achieved. These two results indicate that the both EPA and AKI approaches are promising approaches for robust speaker identification under the mismatch unseen handset condition.

Finally, EPA, AKI and MAP-GMM/CMS approaches were all fused together to form a EPA+AKI+MAP-GMM/CMS system. From Table 3, identification rate of 76.7% was achieved. This indicates that the EPA, AKI and MAP-GMM/CMS methods are complement to each other.

Moreover, the average speaker identification rate of the unseen handsets in the nine cross-validation tests were separated and shown in Table 4. It showed that the AKI+, EPA+ and EPA+AKI+MAP-GMM/CMS methods could improve the performance from 58.9% (MAP-GMM/CMS) to 68.2%, 70.5% and 72.4%, respectively. Therefore, the results in Table 3 and 4 showed that the proposed fusion system could efficiently compensate the mismatch for both seen and unseen handsets.

Table 3: The average close-set speaker identification rates (%) of the nine cross-validation evaluations on the HTIMIT database achieved by the MAP-GMM/CMS, AKI+, EPA+ and EPA+AKI+MAP-GMM/CMS approaches, respectively.

	Average
MAP-GMM/CMS	60.5
+AKI	72.2
+EPA	70.5
+EPA +AKI	**76.7**

6. CONCLUSIONS

This paper presents an EPA+AKI+MAP-GMM/CMS fusion approach. It integrates together a prosody-level EPA, an acoustic-level AKI and a MAP-GMM/CMS baseline for robust close-set speaker identification on the situation of unseen mismatch handset and limited available data.

Unlike conventional fusion approaches, which usually require a lot of speech data to build a reasonable prosodic modeling and may have difficulty to deal with unseen handsets, the proposed method requires only few training/test utterances (in this case, seven/three utterances) and could alleviate the distortion of unseen mismatch handsets. Experimental results on the HTIMIT database have shown that, a remarkable improvement, about 41.0% and 32.8% relative error rate reduction for seen and unseen handsets, respectively, comparing with the conventional MAP-GMM/CMS baseline, could be achieved. It is therefore a promising method for robust speaker identification under mismatch environment and limited available data.

7. ACKNOWLEDGEMENT

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 93-2213-E-027-014 and Ministry of Education under the project with contract A-93-E-FA06-4-4.

8. REFERENCES

[1] Kemal Sonmez, Elizabeth Shriberg, Larry Heck, Mitchel Weintraub, "Modeling Dynamic Prosodic Variation For Speaker Verification," In Proc. of ICSLP, Vol. 7, pp. 3189-3192, 1998.

[2] D. A. Reynolds et. al., "The superSID project: exploiting highlevel information for high-accuracy speaker recognition," Proc. ICASSP'03, vol. IV, pp.784-787, 2003.

[3] Zi-He Chen, Yuan-Fu Liao and Yau-Tarng Juang, "Eigen-Prosody Analysis for Robust Speaker Recognition under Mismatch Handset Environment", To appear in Electronics Letters.

[4] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman, "Indexing by Latent Semantic Analysis," Journal of the American Society of Information Science, 1990.

[5] Jyh-Her Yang and Yuan-Fu Liao, "Unseen Handset Mismatch Compensation Based on *A Priori* Knowledge Interpolation for Robust Speaker Recognition", To appear in ICSLP'2004.

[6] D. A. Reynolds, "HTIMIT and LLHDB: speech corpora for the study of handset transducer effects," in ICASSP'97, vol. 2, pp. 1535-1538, 1997.

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

Table 4: The close-set speaker identification rates (%) of unseen handsets in the nine cross-validation evaluations on the HTIMIT database achieved by the MAP-GMM/CMS, AKI+MAP-GMM/CMS, EPA+MAP-GMM/CMS and EPA+AKI+MAP-GMM/CMS, respectively.

	cb1	cb2	cb3	cb4	el1	el2	el3	el4	pt1	Average
MAP-GMM/CMS	70.2	70.9	30.8	40.1	74.5	61.6	59.6	64.6	56.0	58.9
+AKI	78.8	79.5	41.1	58.9	82.1	65.9	71.5	71.5	64.2	68.2
+EPA	78.1	80.5	42.1	53.3	81.8	73.8	71.2	74.2	66.2	70.5
+EPA+AKI	83.4	85.1	51.0	67.2	86.4	74.2	78.1	78.8	73.5	**72.4**