Joint Prosodic and Spectral Modeling for Robust Speaker Verification

Yuan-Fu Liao[1], Wen-Chieh Chang[2], Zong-You Xie[1], Ding-Yun Zeng[1] and Yau-Tarng Juang[2]

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

Abstract

In this paper, a joint prosodic and spectral modeling framework is proposed instead of traditional score-domain fusion approaches to alleviate the problem of mismatch channel/handset/ambient noise. The basic idea is to embed the concept of hierarchical structure of speech prosody into an ergodic HMM (EHMM), and model the prosodic status transitions and prosodic/spectral features by EHMM's states, state transition probabilities and state-dependent observation distributions, respectively. Experimental results evaluated on the standard single-speaker detection task of NIST 2001 speaker recognition evaluation (NIST-SRE 2001) showed that the proposed approach not only outperformed the spectral feature-based baseline (8.04% vs. 8.64% in equal error rate, EER) but also worked a little bit better than score-domain fusion (8.44%) approach.

1. Introduction

One of the most important issues for speaker verification is the channel/handset/ambient noise mismatch problem. To address this problem, higher level information such as the prosodic cues of a speaker, which may be less sensitive to those mismatch, are attractive recently. For example, several works [1-3] have shown there is a significant benefit to combining prosodic and spectral features.

General speaking, there are three different methods to combine the prosodic and spectral cues including (1) feature-, (2) model- and (3) score-domain fusions. Among the three approaches, feature-domain fusion is the most intuitive way. It usually explores the relationship between the per-frame mel-frequency cepstral coefficients (MFCCs) and pitch/energy contours by directly concatenating them into a single vector stream to build a single model. Score-domain fusion [1-2] is the most popular and successful strategy. But it often ignores the dependency between prosodic and spectral cues and independently establishes one system for one information source. On the other hand, studies on model-domain fusion [3], especially joint modeling of the prosodic and spectral information, are still rare (at least for speaker recognition).

However, model-domain fusion may be a promising way to explore the interaction between prosodic and spectral cues, because an utterance usually can be tokenized into many smaller constituents in different prosodic levels, such as prosodic phrases, words and syllables. According to the recent studies in [4], these units can be hierarchically organized as shown in Fig. 1 where the lower level nodes are subjacent units subject to higher level constraints. In other words, the lower level prosodic and spectral characteristics (prosodic phrase, word or syllable) are susceptible to the higher level organization (prosodic utterance, phrase or word, respectively). For example, a typical trajectory of F0 contours of an utterance is shown in Fig. 2.

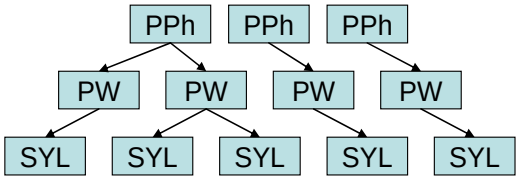

Figure 1: *A schematic representation of the hierarchical structure of speech prosody.*

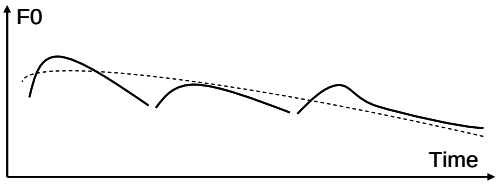

Figure 2: *A schematic illustration of the trajectory of perceived F0 contours of an utterance.*

Therefore, in this paper, a joint modeling framework is proposed that utilizes the concept of hierarchical structure of speech prosody to capture the relationship between spectral and prosodic cues to alleviate the problem of mismatch channel/handset/ambient noise. The basic idea is to embed the hierarchical structure of speech prosody into an EHMM (i.e., prosodic state-embedded EHMM, PEHMM). In this framework, an utterance is tokenized into many small segments and three levels of prosodic features are extracted and modeled including (1) supra-segment- (prosodic phrase/word), (2) segment- (prosodic word/syllable) and (3) frame-level features.

In more detail, let the states of the PEHMM (called prosodic state from now on) represent the position of a lower level unit inside a higher level one, such as the beginning, middle or ending of prosodic phrase/word. And the distributions of observations, such as the per-segment prosodic and per-frame prosodic/spectral features, of lower-level units in different positions are prosodic state-dependent. Moreover, the transition of prosodic status is modeled by the prosodic state transition probability matrix of the PEHMM.

This paper is organized as follows. Section 2 describes the proposed joint prosodic and spectral modeling framework. Section 3 discusses the speaker verification task and the procedures to initialize and optimize the PEHMMs. Section 4 reports the experimental results on NIST-SRE 2001 corpus [5]. Some conclusions are given in the last section.

2. Joint prosodic and spectral modeling

Fig. 3 shows the proposed joint prosodic and spectral modeling framework. There are two main modules including

(1) tokenization front-end and (2) prosodic state-embedded EHMMs (i.e., PEHMMs).

In this framework, an input utterance is tokenized into smaller prosodic units to extract multi-level features including the per-segment prosodic features, per-frame pitch/energy and spectral features. In addition, a set of prosodic states is introduced and embedded into EHMMs to represent the hierarchical structure of speech prosody (as shown in Fig. 1).

The main goal of utilizing the set of newly introduced prosodic states is to link the interaction between spectral and prosodic features.

2.1. Tokenization front-end

The block diagram of the tokenization front-end is shown in Fig. 4. It firstly extracts the raw prosodic contours (pitch and energy) of an input utterance. The pitch and energy contours are then segmented into smaller units (close to a syllable) by a piece-wise stylization algorithm. The stylization output is a sequence of voiced/unvoiced tags and segment boundaries.

Assuming there are N segments, their corresponding stylized contours are used to extract N sets of segment-level prosodic features, $\mathbf{F}=\{F_1,F_2,\ldots,F_N\}$.

For each voiced segment, nine segment-level prosodic features are extracted include the duration of the segment, the slopes of the pitch and energy contours, left- and right-hand-side pitch and energy jumps and pause durations.

Frame-level features include the per-frame pitch/energy, $\boldsymbol{\rho}_n=\{\rho_{n,1},\rho_{n,2},\ldots,\rho_{n,T_n}\}$, and spectral features, $\mathbf{X}_n=\{X_{n,1},X_{n,2},\ldots,X_{n,T_n}\}$.

On the other hand, for each unvoiced segment, only its duration are utilized.

The output of the tokenization front-end is therefore a sequence of multi-level features, $\{\mathbf{X},\boldsymbol{\rho},\mathbf{F}\}=\{(\mathbf{X}_n,\boldsymbol{\rho}_n,\mathbf{F}_n),n=1\sim N\}$, to represent the prosodic and spectral dynamics of input speech in both the frame and segment levels.

2.2. Prosodic state-embedded EHMM

In this study, a PEHMM-based generative model is adopted since it is not easy to directly model the joint distribution, $P(\mathbf{X},\boldsymbol{\rho},\mathbf{F})$, of the multi-level prosodic and spectral features.

Fig. 5 shows the block diagram of the proposed PEHMM where the hierarchical structure of speech prosody is modeled by (1) a set of prosodic states, $\mathbf{q}=\{q_1,q_2,...,q_Q\}$, and their state transition probabilities, $P(\mathbf{q})$, and (2) state-dependent distributions, $P(\mathbf{X},\boldsymbol{\rho},\mathbf{F}\,|\,\mathbf{q})$, of the multi-level features. The likelihood function of a PEHMM is hence defined as follows:

$$P(\mathbf{X},\boldsymbol{\rho},\mathbf{F})=\sum_{\mathbf{q}}P(\mathbf{X},\boldsymbol{\rho},\mathbf{F}\,|\,\mathbf{q})\cdot P(\mathbf{q}) \quad (1)$$

Moreover, the state transition probabilities in Eq. (1) are further simplified and modeled as a prosodic state bi-gram model:

$$P(\mathbf{q})=P(q_1)\cdot\prod_{n=2}^{N}P(q_n\,|\,q_{n-1}) \quad (2)$$

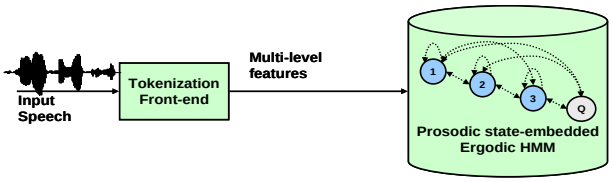

Figure 3: *The proposed joint spectral and prosodic modeling framework for robust speaker verification.*

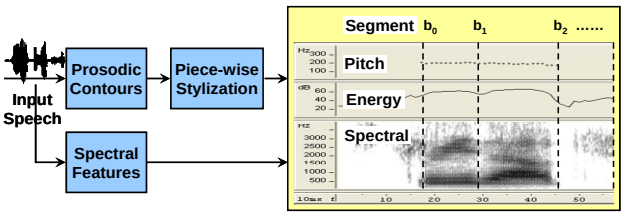

Figure 4: *The block diagram of the tokenization front-end for segmenting an input utterance into smaller units and extracting the corresponding multi-level features*

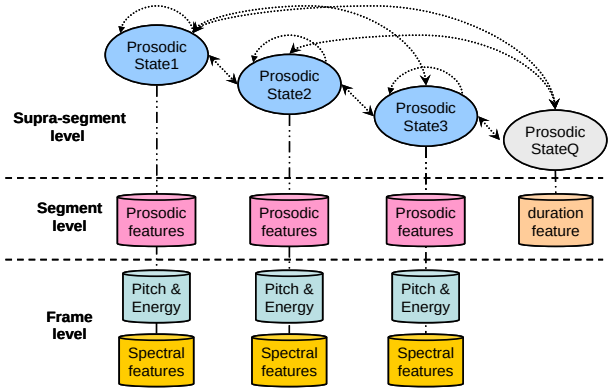

Figure 5: The block diagram of the proposed *prosodic state-embedded EHMM (PEHMM).*

The distributions of $\mathbf{X}_n$, $\boldsymbol{\rho}_n$ and $\mathbf{F}_n$ of the n-th segment in a specific state, q_n, are assumed to be independent of each other, i.e.,

$$P(\mathbf{X}_n,\boldsymbol{\rho}_n,\mathbf{F}_n\,|\,q_n)\approx P(\mathbf{X}_n\,|\,q_n)\cdot P(\boldsymbol{\rho}_n\,|\,q_n)\cdot P(\mathbf{F}_n\,|\,q_n) \quad (3)$$

And the three terms on the right hand side of Eq. (3) are all represented by state-dependent Gaussian mixture models (GMMs) in this study:

$$\begin{cases} P(\mathbf{X}_n\,|\,q_n)=\prod_{t=1}^{T_n}\sum_{m=1}^{M_{q_n}^{X}}w_{q_n,m}^{X}\mathbb{N}\left(X_t\,|\,\mu_{q_n,m}^{X},\sigma_{q_n,m}^{X}\right) \\ P(\boldsymbol{\rho}_n\,|\,q_n)=\prod_{t=1}^{T_n}\sum_{m=1}^{M_{q_n}^{\rho}}w_{q_n,m}^{\rho}\cdot\mathbb{N}\left(\rho_t\,|\,\mu_{q_n,m}^{\rho},\sigma_{q_n,m}^{\rho}\right) \\ P(\mathbf{F}_n\,|\,q_n)=\sum_{m=1}^{M_{q_n}^{F}}w_{q_n,m}^{F}\cdot\mathbb{N}\left(\mathbf{F}_n\,|\,\mu_{q_n,m}^{F},\sigma_{q_n,m}^{F}\right) \end{cases} \quad (4)$$

3. PEHMM-based speaker verification

For speaker verification task, a universal background PEHMM (UBM), $\mathbf{\Phi}_{UBM}$, is first trained using all available training utterances of many speakers (usually hundreds of speakers). The UBM represent the average prosodic and spectral dynamic characteristics of all non-target speakers. In addition, one PEHMM, $\mathbf{\Phi}_s$, is built for each target speaker by adapting the UBM using the maximum *a posteriori* (MAP) [6] algorithm and his corresponding training utterances. The speakers' PEHMMs, $\{\mathbf{\Phi}_s, s=1\sim S\}$, therefore represent the speaker-specific behaviors of the target speakers.

The decision rule for the speaker verification task is based on the log-likelihood ratio test and defined as follows:

$$LLR \underset{<}{\overset{H_0}{\geq}} Threshold, LLR = \log\left(\frac{P(\mathbf{X},\boldsymbol{\rho},\mathbf{F}\mid\Phi_s)}{P(\mathbf{X},\boldsymbol{\rho},\mathbf{F}\mid\Phi_{UBM})}\right) \quad (5)$$

where s is the claimed speaker in a trial, H_0 and H_1 are the null and alternative hypotheses, respectively. In the following subsections, two implementation issues will be discussed including the PEHMM (1) initialization and (2) optimization.

3.1. PEHMM initialization

To build the UBM, $\mathbf{\Phi}_{UBM}$, and speaker PEHMMs, $\mathbf{\Phi}_S$, an automatic prosodic state labeler is necessary to first convert the input segment prosodic feature vectors into a sequence of prosodic states in order to calculate the prosodic state bi-gram and initialize all prosodic state-dependent parameters as defined in Eqs. (2) and (4), respectively.

Usually, to learn the relationship between the prosodic features and prosodic states, a classifier that is trained with supervised training using manually annotated prosodic feature-state pairs is required. For example, tones and break index (ToBI) annotation and artificial neuron networks (ANNs) are often adopted. However, manual annotation is always labor intensive and expensive.

Therefore, an unsupervised vector quantization (VQ) is utilized to automatically learn a codebook from the training data of all speakers. Because the segment prosodic features are used to capture prosodic events, each codeword in the codebook may represent a specific prosodic status. The VQ codebook is then used to automatically label each input utterance of a speaker into a sequence of prosodic states.

3.2. PEHMM optimization

Once the PEHMMs are initialized, the segmental K-mean algorithm (for simplifying the computation) is applied to iteratively fine-tune the UBM and all speaker models. The procedures to optimize a speaker model $\mathbf{\Phi}_S$ (it is similar for UBM) are shown as follows:

Step 1: re-label all training utterances using Viterbi search algorithm to find the best prosodic state sequences:

$$\hat{\mathbf{q}} = \max_{\mathbf{q}}\left\{P(\mathbf{X},\boldsymbol{\rho},\mathbf{F}\mid\mathbf{q},\mathbf{\Phi}_S)\cdot P(\mathbf{q}\mid\mathbf{\Phi}_S)\right\} \quad (6)$$

Step 2: re-estimate the parameters of the speaker PEHMMs according to the found prosodic state label sequences given in Step 1.

Step 3: go to Step 1 until converged.

Table 1: Comparison of performance (equal error rate, EER in %) of various systems evaluated on the standard one speaker detection task of the NIST-SRE 2001. Here PD means "prosodic state-dependent".

System	EER (%)
(1) Spectral GMMs	8.64
(2) Pitch/energy GMMs	28.62
(3) Prosodic GMMs	31.65
(4) Prosodic State bi-gram	35.78
(5) PD spectral GMMs	8.28
(6) PD pitch/energy GMMs	28.61
(7) PD prosodic GMMs	31.89
(8) Scoring fusion (1)+(2)+(3)+(4)	8.44
(9) PEHMM (4)+(5)+(6)+(7)	8.04

4. Experimental results

4.1. Evaluation corpus and experimental settings

All experiment results presented in this paper were evaluated on the standard one speaker detection task of NIST-SRE 2001 using only the basic evaluation corpus, i.e., no extended data were used. In this task, there are in total 174 target speakers. Each speaker comes with two minutes of enrollment speech. Besides, there are 2,038 target and 20,380 imposter trials, respectively. Each trial is about 30 seconds long, on average.

In all experiments, 38 spectral features, including 12 MFCCs, 12 Δ-MFCCs, 12 Δ2-MFCCs, Δ-log-energy and Δ2-log-energy were computed with window size of 30 ms and frame shift of 10 ms. Cepstral mean subtraction and variance normalization were utilized to partially resist the channel/handset distortion. As prosodic features, the raw pitch and energy contours were extracted using the Snack Toolkit (ESPS's get_f0 function). Moreover, all the fusion weights were empirically set.

4.2. Four baseline systems and score-domain fusion

A 1024-mixture spectral feature GMM is first built from the enrollment speech of all 174 speakers. Then, one MAP-adapted GMM (MAP-GMM) was estimated for each speaker using the speaker's own enrollment speech. The per-frame pitch/energy features and segment prosodic features were modeled using 128- and 32-mixtures GMMs, respectively. In addition, 8 voiced and 3 unvoiced codewords were learned in the VQ codebook and an 11-state bi-gram model was trained.

The experimental results, i.e., EERs of the four baselines were shown in Table 1. The EERs of the per-frame spectral, pitch/energy, segment prosodic features GMMs and prosodic state bi-gram are 8.64%, 28.62%, 31.65% and 35.78%, respectively. Table 1 also shows the result of fusing the scores of the four systems. The combined system gave an EER of 8.44%. These results demonstrate that the prosodic and spectral information complement each other.

4.3. Proposed PEHMM joint modeling

In the following subsections, the behavior of the VQ-based automatic prosodic state labeler used to initialize the proposed PEHMM and the prosodic state-dependent GMMs are first built and examined, respectively. The experimental result of the proposed PEHMM is reported in the end.

4.3.1. Analysis on learned prosodic states

4.3.1.1. Justification of the automatic prosodic state labeling

First, the codewords of the VQ-based automatic prosodic state labeler learned from the enrollment speech of all speakers is shown in Fig. 6. By cross-examining the relations of those codewords and the transition matrix of the labeled prosodic state sequences, a meaningful prosodic state transition diagram was created, which is shown in Fig. 7.

These 11 codewords may be interpreted (post-hoc) as 11 different prosodic states that represent, respectively, one major and one minor prosodic phrase-start state (MajPhStart (State 3) and MinPhStart (State 7) with longer left-pause, energy, and pitch jump up), one major- and one minor-prosodic phase-end state (MajPhEnd (State 5) and MinPhEnd (State 2) with longer right-pause and pitch jump up), four transient states (Trans (States 1, 4, 6, and 8)), and short, major and minor breaks (States 9, 10 and 11).

This state diagram may represent the typical prosody behaviors of all seen speakers in the training set. Therefore, any speaker-specific deviation from these typical behaviors may be used as the informative cues for speaker recognition.

4.3.1.2. Effectiveness of the prosodic state-dependent GMMs

Given the prosodic state tags, all training material was divided to build the prosodic state-dependent spectral, pitch/energy and prosodic GMMs.

The EERs of these prosodic state-dependent GMMs are shown in Table 1. From the table, it was found that the EER of the spectral GMMs could be improved from 8.64% to 8.28%. This result indicates that the characteristics of spectral features may be different in different prosodic states. However, it is not the case for per-frame pitch and energy contours and segment prosodic features.

4.3.2. Experimental results

Finally, the experimental result of the proposed PEHMM joint modeling framework is also shown in Table 1 and EER of 8.04% was achieved. This result indicates that the proposed PEHMM approach not only outperformed the spectral GMMs (8.64%) but also worked a little bit better than the score-domain fusion one (8.44%). The reasons for the relatively small improvement may be that Viterbi search was applied (instead of a forward-backward algorithm) to compute the likelihoods. Beside, more iterations or even Baum-Welsh re-estimation may be needed to further improve the performance.

5. Conclusions

In this paper, a joint prosodic and spectral modeling framework, PEHMM, is proposed to alleviate the problem of mismatch channel/handset/ambient noise. Experimental results evaluated on the standard one-speaker detection task of NIST-SRE 2001 showed that PEHMM outperformed the spectral GMMs and were comparable with the score domain fusion method. It is therefore a promising approach

6. Acknowledgement

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 96-2221-E-027-100-MY2.

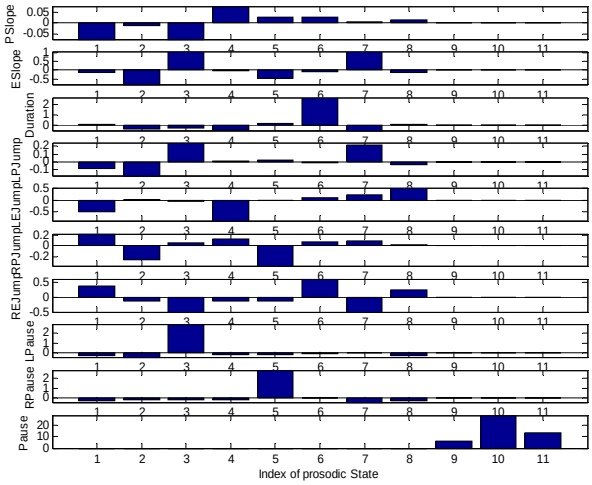

Figure 6: *The centroids of the vector quantization codebook learned from the enrollment speech of all speakers in the training set of NIST-SRE 2001.*

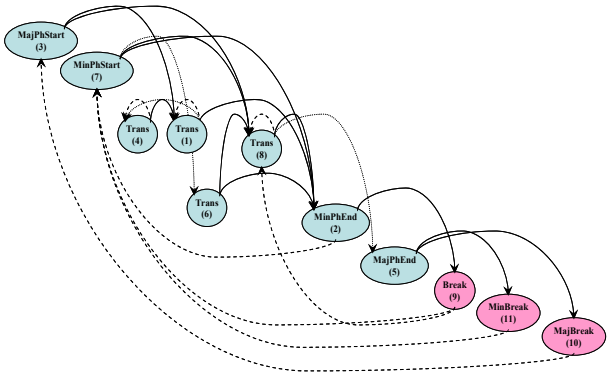

Figure 7: *The major prosodic state transition flow of the VQ-based automatic prosodic state labeler learned from the enrollment speech of all speakers in the training set of NIST-SRE 2001*

7. References

[1] D. A. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The SuperSID project: Exploiting high-level information for high-accuracy speaker recognition," Proc. ICASSP'03, vol. IV, pp. 784-787, 2003.

[2] Zi-He Chen, Zhi-Ren Zeng, Yuan-Fu Liao, and Yau-Tarng Juang, "Probabilistic Latent Prosody Analysis for robust speaker verification," Proc. ICASSP'06, vol. 1, pp. 105-108, 2006.

[3] M. Arcienega, A. Alexander, P. Zimmermann and A. Drygajlo, "A Bayesian network approach combining pitch and spectral envelope features to reduce channel mismatch in speaker verification and forensic speaker recognition," Proc. INTERSPEECH'05, pp. 2009-2012, 2005.

[4] Tseng, C., Pin, S., Lee, Y., Wang, H. And Chen, C., "Fluent speech prosody: Framework and modeling," SPEECH COMMUNICATION, vol. 46:3-4, pp. 284-309, 2005.

[5] 2001 NIST Speaker Recognition Evaluation Corpus, LDC – Linguistic Data Consortium, http://www.ldc.upenn.edu/.

[6] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," Digital Signal Processing, vol. 10, pp. 19-41, 2000.