PROBABILISTIC LATENT PROSODY ANALYSIS FOR ROBUST SPEAKER VERIFICATION

Zi-He Chen[1], Zhi-Ren Zeng[2], 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

In this investigation, two probabilistic latent semantic analyses (PLSA)-based approaches are proposed for use in speaker verification systems to reduce the number of parameters required by prosodic speaker models to (1) estimate reliably speakers' bi-gram models and to (2) reduce the amount of required training and test data. The basic concept is to (1) adopt PLSA to smooth the underlying n-gram-based prosodic speaker models, and to (2) use PLSA to find a compact latent prosody space to represent efficiently the constellation of speakers. The proposed approaches are evaluated on the standard single-speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus, where only one 2minute training enrollment speech and 30s test speech on average are available. Experimental results demonstrated that the proposed approach can reduce the required number of bi-gram parameters from 112 to 88 and 63 per speaker and improve the EERs of MAP-GMM and GMM+T-norm from 12.4% and 9.5% to 10.4% and 8.4%, respectively, and finally to 8.1% after fusing all systems.

1. INTRODUCTION

The most important issue in speaker verification is channel/handset mismatch. Prosodic features, which are known to be weakly sensitive to channel/handset mismatch, have recently been considered to alleviate this problem. Several successful techniques have been developed, including the distribution [1], the n-gram [2] and the discrete hidden Markov model (DHMM) [2] methods.

The dynamic behavior of speech prosody is affected by numerous latent factors, such as the speaker, speaking style, phonetic context or even emotion. Therefore, the variation of the observed prosodic features may be quite large and complex. A large amount of training and test data are usually required to apply reliably prosodic information to speaker verification systems. For example, in the famous NIST 2001 Speaker Recognition Evaluation Extended Data Task [3], eight and one 2minute conversation turns were used for training and testing, respectively.

However, in a real-life situation, only limited training and test data are available. In this work, a latent prosody analysis approach is developed to utilize efficiently prosodic information in the hope of both reliably estimating the parameters of a prosodic speaker model and reduce the required length of training and test speech data to a few minutes and seconds, respectively. The basic idea of efficiently exploiting prosodic information is to apply the concepts of the PLSA [4] to (1) smooth the underlying n-gram-based prosodic speaker models or to (2) find a compact latent prosody space to represent the constellation of speakers. This method differs essentially from previously presented eigen-prosody analysis (EPA) [5] because a different scoring procedure is employed and the evaluation is performed on speaker verification, instead of speaker identification task using a more recent speaker recognition database.

In more detail, after the prosodic contours [2] are stylized by piece-wise curve fitting, the prosodic features of several neighboring segments are concatenated into a prosodic super-vector. A VQ-based prosody model is trained to label automatically the sequences of the prosodic super-vectors into sequences of prosody states. The sequences of prosody states are treated as a text document that records the long-span prosody behaviors of the speaker. Then, two approaches are investigated. They are (1) the use of speaker-specific PLSA to smooth each estimated n-gram model of the sequences of prosody states for each speaker through dimension reduction, and (2) the application of speaker-wide PLSA to analyze jointly the n-grams of all speakers to find a compact latent prosody space to further reduce the number of parameters required to represent each speaker.

Briefly, the proposed approaches differ from the conventional ones in many ways. First, the numbers of parameters of speakers' n-gram models are significantly reduced by PLSA to relax the requirement for a large amount of training and test data. Secondly, long-span prosody behaviors of speakers are captured by PLSA. Finally, the presented methods are evaluated on the standard (instead of the extended data task) one speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus [6] where only 2minute training and 30s test speech (in average) are available.

This article is organized as follows. Section 2 provides information about the 2001 NIST Speaker Recognition Evaluation Corpus and the experimental conditions used throughout this paper. Section 3 describes the proposed approaches in detail and gives some explanatory intermediate simulation results. Section 4 reports the final experimental results. The final section draws some conclusions.

2. NIST 2001 SPEAKER RECOGNITION CORPUS AND EXPERIMENTAL CONDITIONS

All approaches presented herein are evaluated on the one speaker detection task, NIST 2001 Speaker Recognition Evaluation, using only the basic evaluation corpus [6], without extended data [3]. In this task, a total of 174 target speakers, 2,038 target and 20,380 imposter trials are undertaken. Each enrollment and trial lasts approximately 2minutes and 30s on average, respectively.

A 1024-mixture universal background model (UBM) [7] is established from the enrollment speech of all 174 speakers to construct a speaker verification baseline system. Then, for each speaker, a maximum *a priori*-adapted Gaussian (MAP-GMM) is established using the UBM and the speaker's own enrollment speech. Thirty-eight mel-frequency cepstral coefficients (MFCCs) including 12 MFCCs, 12 Δ-MFCCs, 12 Δ^2-MFCCs, Δ-log-energy and Δ^2- log-energy were computed with a window size of 30ms and a frame shift of 10ms. Feature domain cepstrum mean subtraction (CMS) and score domain T-norm [8] were also applied to reduce partially the channel/handset distortion.

The pitch and energy contours of all utterances in the corpus were extracted using the popular Wavesurfer/Snack sound toolkit [9] and stylized using the piece-wise curve fitting approach [2], to examine the benefits of the prosodic information.

Five prosodic features are extracted for each found segment following piece-wise stylization. They include (1) the pitch slope, (2) the energy slope and (3) the duration of the segment and (4) the pitch and (5) the mean energy jump between two segments. The prosodic feature vectors were normalized by their global mean and the variance of segments (except pauses). Finally, vectors of *N* neighboring segments are concatenated into a super-vector (of *N**5 dimensions) to normalize partially the variation in speech prosody. In all of the following experiments, the reported speaker detection performances are calculated and plotted using the NIST DET-Curve Plotting software version 2.1 [10].

3. PROBABILISTIC LATENT PROSODY ANALYSIS

In this section, eight- and three-codeword VQs are built and used to label automatically the input sequences of prosodic super-vectors into sequences of prosodic states. Bi-gram speaker models of the prosodic states are then built for the UBM and for each speaker. Then, the PLSA technique is adopted to smooth the bi-gram model and to represent the constellations of speakers in latent prosody space.

3.1. Automatic prosody state labeling and bi-gram speaker models

Clustering the extracted prosodic super-vectors of all registered speakers enabled eight- and three-codeword VQs (see Table 1) for voiced and unvoiced segments, respectively, to be learned and used to model the prosodic characteristics of all speakers. In particular, the trained VQs were used to label automatically the input sequences of an input utterance into sequences of prosodic states.

The sequences of prosodic state labels were utilized to train the bi-gram UBM and speaker models. Cross-checking the values of codewords and the state transition frequencies (Table 2) showed that codewords number 10 and 11 are the major and minor breaks and codewords number 6 and 2 and number 1 are the voiced segments at the beginning and end of prosodic phrases, respectively. Accordingly, these two VQs can automatically label the prosodic states of input utterances.

Finally, an 11*11 bi-gram model of prosodic state sequences was established for each speaker. Notably, the small-scale eight- and three-codeword VQs and 11*11 bi-gram models are selected because of the sparse data problem. Moreover, the standard Good-Turing discounting method has to be applied to solve partially issues of sparse data.

3.2. Speaker-specific PLSA bi-gram smoothing

Training and test data are normally limited in real-life, so the estimated prosodic state bi-gram speaker models may not be reliable even for a small-scale bi-gram. For instance, the 2minute training speech of each speaker in the NIST 2001 Speaker Recognition Evaluation Corpus comprises only approximately 1,000 segments after piece-wise stylization. Therefore, the training data may not suffice even for building a small-scale 11*11 bi-gram speaker model. Additionally, in the preliminary experiment described in section 3.1, the speaker bi-gram models had to be smoothed by Good-Turing discounting to yield a reasonable performance.

The PLSA dimension reduction approach is proposed here to prevent the removal of too much of the unique prosodic characteristic of speakers by conventional discounting or backing-off methods. PLSA is applied to decompose the bi-gram speaker models to find and keep only few principle latent prosody factors (in the sense of probabilities) in order to reconstruct smoother bi-gram models. Figure 1 depicts in detail the procedure includes (1) VQ-based prosodic modeling and automatic prosody state labeling, (2) establishing a prosody state n-gram model for each speaker, (3) dimension reduction of the n-gram of each speaker separately and (4) reconstruction of smoother n-grams. Figure 2 shows a typical example of a PLSA-smoothed bi-gram and its counterpart.

3.3. Speaker-wide PLSA latent prosody space analysis

All n-gram models mentioned in the preceding subsection can be treated as estimates of the long-term characteristics of speakers. Hence, the PLSA technique is proposed here to explore further the relationship between different speakers by jointly finding a compact latent prosody space in order to reduce the number of parameters required for speakers' n-gram models.

The detailed procedure (Fig. 3) has five steps. Step (1) and (2) are similar to the previous approach. The other steps are (3) calculating the co-occurrence statistics of smoothed n-gram counts or frequencies of speakers to form a prosody n-gram-speakers co-occurrence matrix, (4) decomposing the co-occurrence matrix using PLSA to build a compact latent prosody space and (5) reconstructing the speakers' n-gram models from the compact latent prosody space.

Figure 4 shows a typical example of the compact latent prosody space learned from the enrollment speech of NIST 2001 Speaker Recognition Evaluation Corpus. The figure reveals that the speakers with more major/long (state 10) or more minor/short breaks (state 9 and 11), which are slower and faster speakers, respectively, are presented separately on the bottom and top of the figure. Hence, representing all of speakers in the compact latent prosody space enables the numbers of parameters in the speaker's n-gram models to be further reduced.

4. EXPERIMENTS AND SYSTEM FUSION

In this section, the performances of various speaker verification methods are reported and compared. The experimental conditions of all the following experiments have already been given in Section 2.

4.1. MAP-GMM, T-norm and pitch/energy GMMs

The performance of the MAP-GMM-based speaker models and the popular T-norm score normalization approach were tested. In the T-norm approach, a long list of cohort speakers (50 speakers)

was selected with the closest scores. Figure 5 displays the results and Table 3 presents their corresponding EERs. The figure and table demonstrate that the EER of the MAP-GMM is 12.4% and the T-norm dramatically improved the EER to 9.5%. Therefore, T-norm helps in a mismatch channel/handset mismatch environment.

The distributions of the per-frame log-pitch, log-energy and their delta-terms were modeled using 64-mixture GMMs. EER of 32.3% was achieved using the log-pitch/energy GMMs with a single UBM.

4.2. Prosody state bi-gram speaker models

Figure 5 and Table 3 present the performances of the prosody state bi-gram model obtained using three-segment-long super-vectors and the Good-Turing smoothing method. (In a preliminary experiment, the three-segment-long super-vector is better than the one-segment super-vector). Here, the bi-gram UBM was trained using sequences of prosody states of all registered speakers. Notably, the performance, 31.2% EER, may not be satisfactory for practical applications. However, unlike the extended data task of the NIST 2001 Speaker Recognition Evaluation, only 2minutes training and 30s test data (on average) were available to estimate the bi-gram models and to determine the likelihoods. Furthermore, the performance is compatible to that of the log-pitch/energy-based GMMs.

4.3. Speaker-specific PLSA prosody bi-gram smoothing

PLSA was then adopted instead of the Good-Turing approach to smooth each speaker-specific bi-gram model. The number of latent factors of the bi-gram empirically decreased from 11 to eight. The number of parameters for each speaker's bi-gram model was then reduced from 112 to 88. As presented in Fig. 5 and Table 3, the EER was reduced from 31.2% to 26.8%, revealing the capacities of the PLSA-based smoothing approaches to preserve more unique speaker characteristics.

4.4. Speaker-wide PLSA latent prosody space analysis

Figure 5 shows the performance of the latent prosody space analysis approach. The number of latent factors was empirically set to 90 to analyze globally the n-gram-speaker co-occurrence matrix. The mean number of parameters for each speaker-specific bi-gram model was further reduced to 63. An EER of 26.8% was obtained. This result demonstrates that PLSA responds to long-span prosodic cues and finds a latent prosody space to represent the constellation of speakers.

4.5. System fusion

The scores of the prosody information-based models and MAP-GMMs were fused to determine whether they complemented each other in the situation of limited training and test data, using the popular LNKnet [11] pattern classification software from MIT Lincoln Laboratory. A multi-layer perceptron (MLP) with two output neurons was chosen.

Several combinations of systems were tested. Figure 5 and Table 3 present the results. The figure and table demonstrate that the EERs of the MAP-GMM and MAP-GMM+Tnorm were improved from 12.4% and 9.5% to 10.4% and 8.4%, respectively, by fusing with the prosodic information-based models. Moreover, an EER of 8.1% was achieved by fusing all systems. These results indicate that MAP-GMMs- and PLSA-based approaches complement each other.

5. CONCLUSIONS

In this work, two PLSA-based approaches were developed to reduce the number of parameters required in prosodic speaker models for speaker verification task given limited training and test data. Fusing the PLSA-based systems and the traditional cepstral feature-based GMMs improved the EERs of MAP-GMM and MAP-GMM+T-norm from 12.4% and 9.5% to 10.4% and 8.4%, respectively, and finally to 8.1% after fusing all systems. Notably, only 2minute and 30s (on average) training and test speech were used herein. Hence, the proposed approaches have potential and are worthy of further study as real-life speaker verification systems.

6. ACKNOWLEDGEMENT

The authors would like to thank the National Science Council of the Republic of China, Taiwan (Contract NSC 94-2213-E-027-003) and the Ministry of Education of the Republic of China, Taiwan (Contract A-94-E-FA06-4-4) for financially supporting this research.

7. 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] NIST - Speaker Recognition Evaluations, http://www.nist.gov/speech/tests/spk/index.htm

[4] Thomas Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis," Machine Learning, 42, 177-196, 2001.

[5] Zi-He Chen, Yuan-Fu Liao and Yau-Tarng Juang, "Prosody Modeling and Eigen-Prosody Analysis for Robust Speaker Recognition", ICASSP'2005

[6] 2001 NIST Speaker Recognition Evaluation Corpus, LDC - Linguistic Data Consortium, http://www.ldc.upenn.edu/

[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.

[8] Roland Auckenthaler, Michael Carey, and Harvey Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," Digital Signal Processing, vol. 10, pp. 42–54, 2000.

[9] The Snack Sound Toolkit, http://www.speech.kth.se/snack/

[10] DET-Curve Plotting software for use with MATLAB, http://www.nist.gov/speech/tools/DETware_v2.1.targz.htm

[11] LNKnet Pattern Classification Software, http://www.ll.mit.edu/IST/lknet/

Table 1. Centroids of the 11-state (8+3) VQ-based prosodic model using (a one-segment-long super-vector : *check)* trained (using OR from) the enrollment speech of 2001 NIST Speaker Recognition Evaluation Corpus.

Codeword	Pitch	Energy	duration	Pitch jump	Energy jump	Pause
1	0.03	-0.12	3.55	-0.10	-0.47	-
2	-0.18	2.63	-0.51	-0.30	1.84	-
3	0.37	-1.21	-0.44	0.45	-1.02	-
4	-1.47	0.22	-0.46	-1.49	-0.02	-
5	0.01	-0.13	1.09	-0.12	-0.39	-
6	0.71	0.90	-0.40	0.71	1.25	-
7	-0.01	-0.36	-0.35	0.18	-0.15	-
8	-0.01	0.39	-0.29	-0.23	0.28	-
9	-	-	-	-	-	10.1
10	-	-	-	-	-	28.3
11	-	-	-	-	-	12.7

Table 2. State transition matrix of the 11-state (8+3) VQ-based prosodic model using a one-segment-long super-vector trained (using OR from) the enrollment speech of 2001 NIST Speaker Recognition Evaluation Corpus.

State	1	2	3	4	5	6	7	8	9	10	11
1	450	43	1029	294	1178	171	1386	443	569	794	417
2	862	2865	1053	1238	2829	2996	3370	4396	1167	750	693
3	151	672	3867	965	654	3455	4073	2055	2397	2100	955
4	303	213	1503	2295	1107	1366	2958	1104	995	835	483
5	1052	113	4109	1084	3740	901	5684	2089	2380	1879	1342
6	1547	5952	1648	3594	5952	6197	5504	12521	2366	1393	1496
7	873	1034	5079	1498	3597	4624	10688	6127	3349	3299	2027
8	1536	470	3028	2194	5316	1466	8532	4225	2716	1659	1818
9	0	1	16	0	0	15896	0	0	-	-	-
10	0	10847	2	0	0	1905	0	0	-	-	-
11	0	8	10	0	0	9185	0	0	-	-	-

Table 3. Comparison of performance (EER in %) of (various OR different) systems in the standard one speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus.

Approach	EER (%)
(1) MAP-GMM	12.4
(2) MAP-GMM+T-norm	9.5
(3) Pitch+Energy	32.3
(4) Bi-gram (Good-Turing)	31.2
(5) Bi-gram (PLSA)	26.8
(6) PLSA	26.8
(7) Fusion: (1)+(5)	10.4
(8) Fusion: (1)+(6)	10.6
(9) Fusion: (2)+(5)	**8.4**
(A) Fusion: (2)+(6)	**8.4**
(B) Fusion: ALL	**8.1**

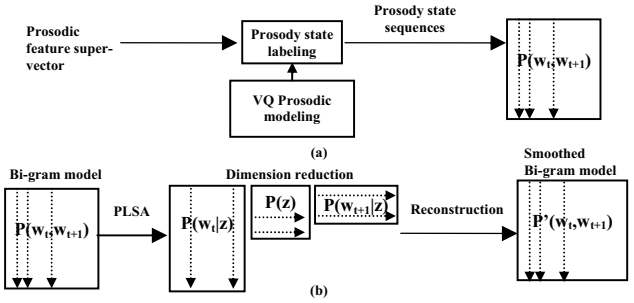

Figure 1. Proposed PLSA-based n-gram speaker model smoothing approach: (a) construction of the n-gram-speaker model, (b) PLSA-based dimension reduction, where w_t, w_{t+1}, d and z are the indices of the n-gram terms and the latent prosody factors, respectively.

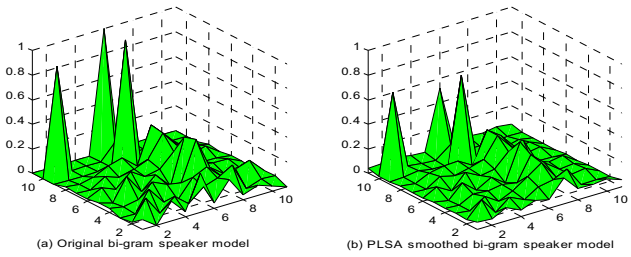

Figure 2. Comparison of the conditional probability functions of (a) the original bi-gram speaker model and (b) its PLSA smoothed version.

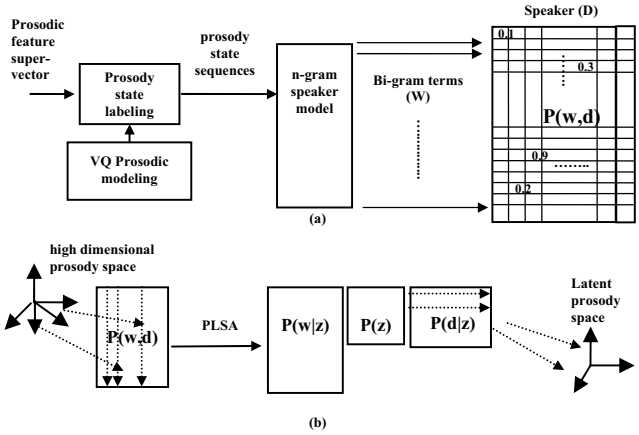

Figure 3. Block diagram of the proposed PLSA-based latent prosody analysis: (a) construction of the n-gram-speaker co-occurrence matrix, (b) PLSA-based dimension reduction, where w, d and z are the indices of n-gram terms, speaker and latent prosody factors, respectively.

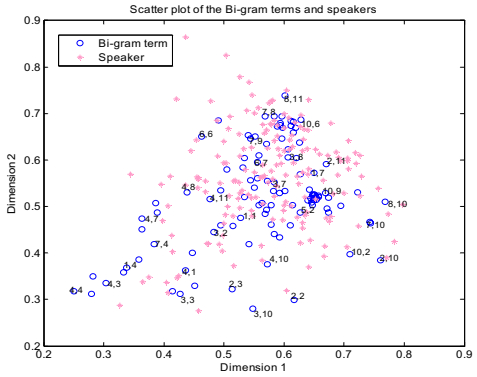

Figure 4. Typical distribution of the bi-gram terms and speakers in the compact latent prosody space, using the 11-state (8+3) prosodic model trained using the enrollment speech of NIST 2001 Speaker Recognition Evaluation Corpus.

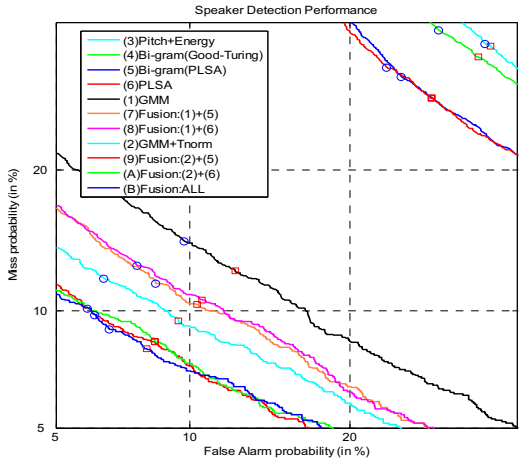

Figure 5. Speaker detection performance evaluation of various speaker verification methods in the standard one speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus (EERs in descending order).