Exploiting Glottal and Prosodic Information for Robust Speaker Verification

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

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

ABSTRACT

In this paper, three different levels of speaker cues including the glottal, prosodic and spectral information are integrated together to build a robust speaker verification system. The major purpose is to resist the distortion of channels and handsets. Especially, the dynamic behavior of normalized amplitude quotient (NAQ) and prosodic feature contours are modeled using Gaussian of mixture models (GMMs) and two latent prosody analyses (LPAs)-based approaches, respectively. The proposed methods are evaluated on the standard one speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus where only one 2-minute training and 30-second trial speech (in average) are available. Experimental results have shown that the proposed approach could improve the equal error rates (EERs) of maximum *a priori*-adapted (MAP)-GMMs and GMMs+T-norm approaches from 12.4% and 9.5% to 10.3% and 8.3% and finally to 7.8%, respectively.

1. INTRODUCTION

The most important issue for speaker verification is the channel/handset mismatch problem. To address this problem, higher level information including prosodic cues [1, 2] and mode of glottal phonation (or voice-quality) [3] of a speaker, which may be less sensitive to channel/handset mismatch are attractive recently.

The prosodic information, such as the dynamic of pitch/energy contour, lengthening and pause duration, are already known to be informative and complemented with the spectral features-based speaker recognition approaches [1, 2]. On the other hand, there are only few papers working on applying the voice-quality, especially the normalized amplitude quotient (NAQ) [4], to speaker recognition task so far. However, NAQ has been proposed as the 4th prosodic dimension of expressive speech [3]. And the procedure of estimating NAQ, especially, the inverse filtering processing to remove vocal tract influences, may be capable to eliminate any non-glottal-specific distortion. It is therefore very interesting to see if NAQ could add robustness to conventional spectral feature-based speaker verification systems or not.

The dynamic behavior of voice quality may be affected by many different latent factors, such as the speaker himself, his corresponding audiences [3] or even his emotion and intension. However, in this first-try study, the dynamic characteristics of NAQ of speakers, i.e., the per-frame NAQ and its first order derivative, are combined with pitch and energy features and modeled using a Gaussian of mixture model (GMM) [5] approach.

Moreover, the latent prosody analysis (LPA) approach previous proposed in [2] is modified to explore the long-span correlation of successive prosody states (status). The basic idea is to automatically label enrollment utterances of speakers into sequences of prosody keywords (states) and to use chunks of prosody keywords to establish bi-gram speaker models or prosody keyword-speaker co-occurrence matrix. The bi-gram model and the co-occurrence matrix are then analyzed using probabilistic latent semantic analysis (PLSA) to reliably estimating the parameters of bi-gram and to find a compact latent prosody space to represent the constellation of speaker, respectively s. Finally, the NAQ-based GMMs (NAQ-GMMs), and LPAs are fused with conventional mel-scale frequency cepstral coefficients (MFCCs)-based GMMs to complement each other.

This paper is organized as follows. Section 2 gives the information about 2001 NIST Speaker Recognition Evaluation Corpus [6] and the experimental conditions used through this paper. Section 3 gives the procedures to model the dynamic of NAQs. Section 4 describes the proposed LPA-based approaches. Section 5 reports the fusion of acoustic, glottal and prosodic information and the final experimental results. Some conclusions are given in the last section.

2. NIST 2001 SPEAKER RECOGNITION EVALUATION CORPUS AND EXPERIMENTAL CONDITIONS

All approaches proposed in this paper are evaluated on the one speaker detection task of NIST 2001 Speaker Recognition Evaluation [6] using only the basic evaluation corpus, i.e., no extended data is used. In this task, there are in total 174 target speakers. Each speaker comes with 2-minute enrollment speech. Beside, there are 2,038 target and 20,380 imposter trials, respectively. Each trial is about 30-second length in average.

To construct a speaker verification baseline system, a 1024-mixture universal background model (UBM) [5] is established from the enrollment speech of all 174 speakers. Then, for each speaker, a MAP-GMM speaker model is built using the UBM and the speaker's own enrollment speech. 38 mel-frequency cepstral coefficients (MFCCs) including 12 MFCCs, 12 Δ-MFCCs, 12 Δ^2-MFCCs, Δ-log-energy and Δ^2- log-energy are computed with window size of 30 ms and frame shift of 10ms. Feature domain cepstrum mean subtraction (CMS) and score domain T-norm are also utilized to partially resist the channel/handset distortion.

For utilized prosodic information, the pitch and energy contours of all utterances in corpus are extracted using the popular Wavesurfer/Snack sound toolkit and are stylized using the piece-wise curve fitting approach.

Five prosodic features are extracted for each found segment after the piece-wise stylization. They include the (1) pitch slope, (2) energy slope and (3) duration of the segment and (4) pitch and (5) energy mean jumping between two segments. The prosodic feature vectors were normalized by their global mean and variance of segments (except pauses). Finally vectors of N neighboring segments are concatenated into a super-vector (N*5 dimensions) to partially normalize the variation of speech prosody.

Furthermore, in all the following experiment, the reported speaker detection performances are calculated and plotted using the NIST DET-Curve Plotting software version 2.1.

3. NORMALIZED AMPLITUDE QUOTIENT

In this paper, NAQs are treated as speaker-specific features and the procedures to extract and model NAQ contours will be described in detail:

3.1. Automatic NAQ contour extraction

NAQ is usually measured by first estimating the glottal speech waveform derivative of a short and stable input speech signal through an inverse filtering processing using time-varying optimized formants [4]. It then picks one period of stable signal and calculates the ratio of the largest peak-to-peak amplitude and the largest amplitude of the cycle-to-cycle minimum derivative and finally normalizes the amplitude quotient ratio based on the underlying fundamental frequency (F0).

However, in this study, it is desired to extract the NAQ contour of the whole enrollment utterances of a speaker in order to learn the dynamic phonation behavior of the speaker. For this purpose, the HUT Aparat toolkit [7] is modified to extract the values of per-frame NAQ under the guiding of corresponding pitch contours. The per-frame NAQs are further smoothed to become a NAQ contour and to suppress some estimation noise. A typical example of the extracted NAQ contour from the enrollment speech of the speaker no. 5007 of NIST 2001 Speaker Recognition Evaluation Corpus is shown in Fig. 1 and the generated NAQ contour is shown in Fig. 2.

3.2. GMM-based NAQ dynamic modeling

A Typical example of histogram of the smoothed per-frame NAQs is shown in Fig. 3. From the figure, it is worthy noting that per-frame NAQs already have certain discriminative capacities. However, Fig. 2 also suggests that it is necessary to model the intrinsic dynamic nature of NAQ contours. Therefore, in this study, the per-frame values of the NAQ contour and their first-order derivative are combined into a two-dimensional feature vector to represent the dynamic of NAQ contours. GMMs are then trained from the enrollment speech of each speaker to learn the phonating behavior of each speaker.

Furthermore, it is found in a preliminary experiment that although that NAQs are already normalized by the underlying F0 to remove any correlation between NAQs and F0s. Some weak connections between NAQ and F0 are still observed. The reason is not yet clear but may be due to the estimation error of pitch or NAQ. Therefore the pre-frame log-pitch and log-energy are further combined with NAQs to form a six-dimensional feature vectors to train the GMM-based speaker models.

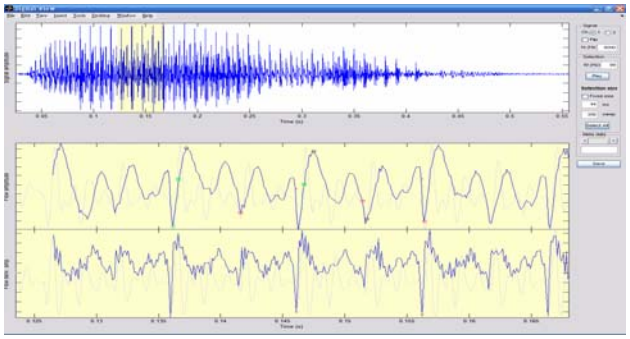

Figure 1. A typical example of the NAQ contour extraction using HUT Aparat toolkits on the enrollment speech of the speaker no. 1830 of NIST 2001 Speaker Recognition Evaluation Corpus.

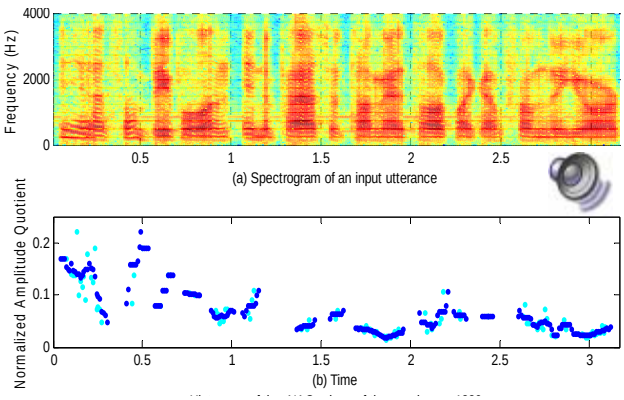

Figure 2. A typical example of the extracted NAQ contour on the enrollment speech of the speaker no. 5007 of NIST 2001 Speaker Recognition Evaluation Corpus. Here those light-blue dots are the raw per-frame NAQs estimated using the modified HUT Aparat toolkits and those dark-blue dots indicate the smoothed NAQ contour.

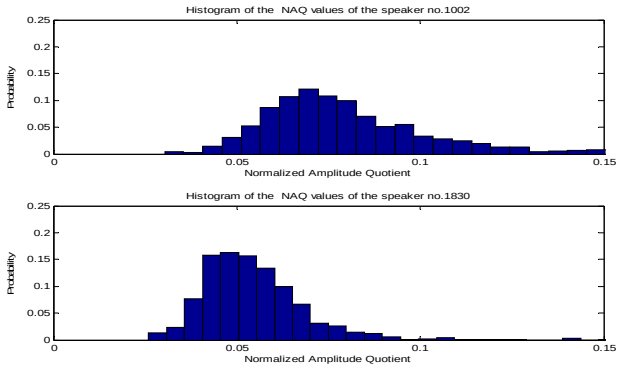

Figure 3. Histograms of NAQ features of the speakers no. 1002 and 1830, respectively, of NIST 2001 Speaker Recognition Evaluation Corpus, respectively. In this example, speaker no. 1002 speaks in a relax way and speaker no.1830 is more tense.

4. LONG-SPAN LATENT PROSODY ANALYSIS

The procedures of applying LPA for bi-gram model smoothing and prosody keyword-speaker co-occurrence matrix

decomposition are shown in Fig. 4 and 5, respectively. Both approaches include the same prosodic contours stylization, VQ-based automatic prosody labeling and prosody keyword parsing modules in the front-end. It is worthy noting that, by using the parser, the length of prosody keyword could be expanded from single- to multi-state words and the long-span temporal information could be well explored.

4.1. PLSA-based bi-gram smoothing

Since the amount of training and trial data are usually limited in real-life applications, the estimated prosody keyword bi-gram speaker models may not be reliable even for a small-scale bi-gram. Usually, this problem is alleviated by conventional discounting or backing-off method. However, those methods do not consider the behavior of speech prosody and may remove the unique prosodic characteristic of speakers. Therefore, in this study, PLSA is utilized to decompose the bi-gram speaker models to find the principle prosody cues and to remove the noisy dimensions with small eigen-values in order to reconstruct smoother bi-gram models. The detail procedures are shown in Fig. 4.

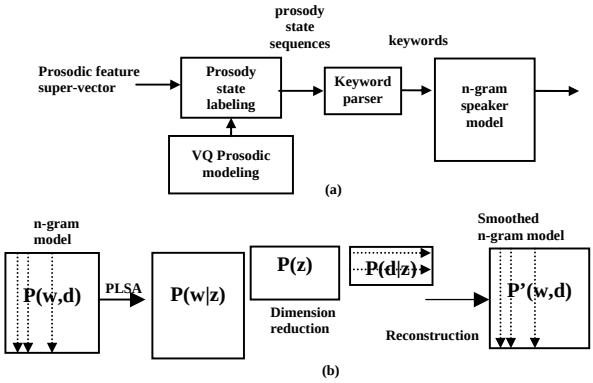

Figure 4. Block diagram of the proposed PLSA-based n-gram speaker model smoothing approach: (a) construction of the n-gram speaker model, (b) PLSA-based n-gram smoothing, where **w**, **d** and **z** are the indices of n-gram terms and latent prosody factors, respectively.

4.2. PLSA-based latent prosody space analysis

The whole n-gram models generated in previous subsection could be treated as the estimates of long-term prosodic characteristics of speakers. Therefore, PLSA technique is proposed here to further explore the relationship between different speakers and keywords. The purpose is to find a compact latent prosody space to further reduce the amount of the required parameters of speakers' n-gram models.

The detail back-end procedures (see Fig. 5) include: (1) calculating the co-occurrences statistics of smoothed n-gram counts (or frequencies) of speakers to form a prosody n-gram terms-speakers co-occurrence matrix, (2) decomposing the co-occurrence matrix using PLSA to build a compact latent prosody space (3) projecting and reconstructing back the speakers' n-gram models from the compact latent prosody space. By projecting all speakers in the compact latent prosody space, the numbers of parameters of speaker's n-gram models could be further reduced.

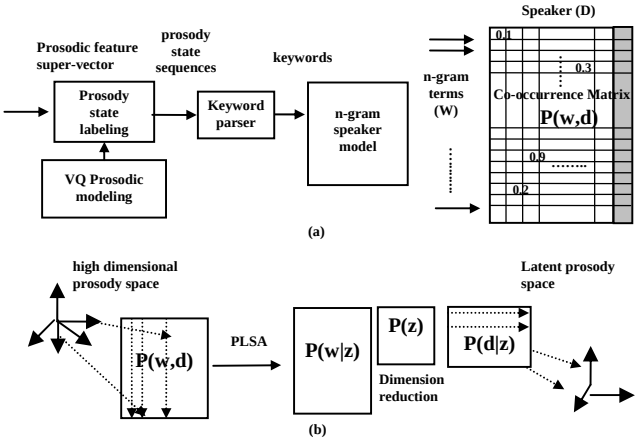

Figure 5. Block diagram of the proposed PLSA-based latent prosody analysis for decomposing the prosody keyword-speaker co-occurrence matrix: (a) construction of the n-gram term-speaker co-occurrence matrix, (b) PLSA-based dimension reduction, where **w**, **d** and **z** are the indices of n-gram terms, speaker and latent prosody factors, respectively.

5. EXPERIMENTS AND SYSTEM FUSION

In this section, the performances of several different speaker verification approaches are compared and fused together under the experimental conditions previously described in Section 2.

5.1. MAP-GMM and prosody bi-gram speaker model

First of all, the performance of the conventional MAP-GMM-based speaker models and the popular T-norm score normalization approach were tested. For T-norm approach, a long list of cohort speakers was selected by pick up 50 most closed speakers (in the sense of recognition scores). The results are shown in Fig. 6 and their corresponding EERs are given in Table 1. Could be seen from the figure and table, the EER of the MAP-GMM are 12.4% and T-norm could dramatically improve the EER to 9.5%. Therefore, T-norm is quite helpful under mismatch channel/handset mismatch environment.

The performance of the prosody keyword bi-gram model with Good-Turing smoothing method is also shown in Fig. 6 and Table 1. It is worth nothing that the performance, 31.2% and 29.6% EERs for single- and multi-state prosody words, respectively, may not be satisfied. However, there are only 2-minute training and 30-second trial data (in average).

5.2. NAQ-GMMs

The NAQ-GMM approaches are then evaluated. A GMM-based speaker model for each registered speaker and one UBM are trained from the enrollment NAQ features and the verification scores are normalized using T-norm algorithm.

The results are shown in Fig. 6 and Table 1. It shows that NAQ-GMM could decrease the EER of pitch+energy system from 32.3% to 30.3%. Moreover, after fusing with the scores of NAQ-GMM, the performance of MAP-GMM+T-norm was improved from 9.5% to 9.4%. This result indicates that NAQ is informative and useful.

5.3. PLSA-based bi-gram smoothing and decomposition

One single- and one multi-state prosody word bi-grams with 11 and 20 states, respectively, were tested. The number of latent factors of the bi-gram speaker model was empirically decreased from 20 to eight and 11 to eight, respectively. The number of parameters for each speaker's bi-gram model was then reduced from 391 to 160 and 118 to 88, respectively. As shown in Fig. 6 and Table 1 the EERs were reduced from 31.2% and 29.6% to 26.8 and 26.5%, respectively.

The performance of the latent prosody space analysis approach is also shown the Fig. 6 and Table 1. Here, the number of latent factors was empirically set to 90 and the numbers of parameters for each speaker's bi-gram model were reduced to 202 and 58 in average, respectively. EERs of 26.8% and 25.9% were achieved, respectively. These results show that PLSA could catch long-span prosodic cues and find a compact latent prosody space to represent the constellation of speakers.

5.4. System fusion

The scores of the long prosody keyword-based PLSAs with different length, NAQ-based GMMs and the conventional MFCC-based MAP-GMMs were then fused to see if they are complemented to each other on the situation of limited training and trial data. For this purpose, the popular LNKnet pattern classification software from MIT Lincoln Laboratory was applied. A multi-layer perceptron (MLP) with two output neurons was chosen. Several different combinations of systems had been tested. The results are shown in Fig. 6 and Table 1.

It could be seen from the figure and table, the EERs of the MAP-GMM and MAP-GMM+Tnorm were improved from of 12.4% and 9.5% to 10.3% and 8.3%, respectively. Furthermore, best EER of 7.8% was achieved by fusing all systems. These results show that MAP-GMMs, the PLSAs and the NAQ-GMM approaches are complement to each other.

6. CONCLUSIONS

In this paper, NAQ-GMMs and two PLSA-based approaches were proposed to improve the performance of speaker verification system under the situation of limited training and trial data. By fusing together the glottal-, prosodic- and spectral-level information, the EERs of MAP-GMM and MAP-GMM+T-norm were improved from 12.4% and 9.5% to 10.3% and 8.3% and finally to 7.8%, respectively. It is worth noting that only 2-minute and 30-second (in average) training and trial speech are used here, respectively. Therefore, the proposed approaches are promising and worthy further studying for real-life speaker verification systems.

7. ACKNOLEDGEMENT

This work was supported by the National Science Council, Taiwan, under the project with contract NSC 94-2213-E-027-003 and Ministry of Education under the project with contract A-94-E-FA06-4-4.

8. REFERENCES

[1] 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.

[2] Zi-He Chen, Zhi-Ren Zeng, Yuan-Fu Liao, and Yau-Tarng Juang, "Probabilistic Latent Prosody Analysis for robust speaker verification", submitted to ICASSP06.

[3] Campbell, N.; Mokhtari, P., 2003. Voice Quality: the 4th Prosodic Dimension. *15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2417-2420.

[4] Alku, P.; Bäckström, T.; Vilkman, E., 2002. Normalized amplitude quotient for parameterization of the glottal flow. *Journal of the Acoustic Society of America*, 112 (2), 701-710.

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

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

[7] Airas, M., Pulakka, H., Bäckström, T., and Alku, P., "A Toolkit for Voice Inverse Filtering and Parametrisation," in Proceedings of the 9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech), pp. 2145-2148, Lisbon, Portugal, September 4-8, 2005. http://aparat.sourceforge.net/index.php/Main_Page.

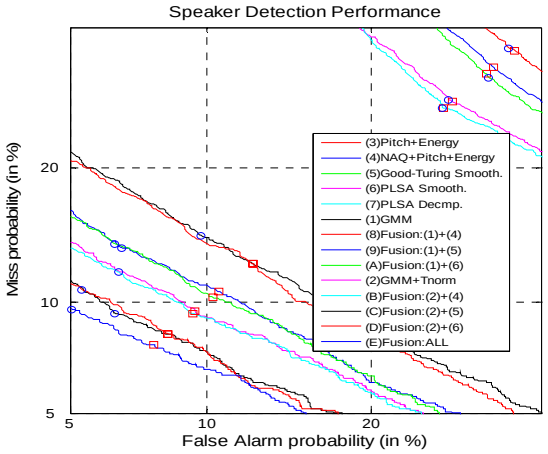

Figure 6. Speaker detection performance evaluation of various speaker verification methods on the standard one speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus (EERs in descent order, multi-state prosody words only).

Table 1. Performance (EER in %) comparison between different systems on the standard one speaker detection task of the 2001 NIST Speaker Recognition Evaluation Corpus.

	EER (%)	
Keyword length	Single-state words	Multi-state words
(1) MAP-GMM	12.4	
(2) MAP-GMM+T-norm	9.5	
(3) Pitch+Energy	32.2	
(4) NAQ+Pitch+Energy	30.3	
(5) Good-Turing Smooth.	31.2	29.6
(6) PLSA Smooth.	26.8	26.5
(7) PLSA Decomp.	26.8	25.9
(8) Fusion: (1)+(4)	12.4	
(9) Fusion: (1)+(5)	10.4	10.6
(A) Fusion: (1)+(6)	10.6	10.3
(B) Fusion: (2)+(4)	9.4	
(C) Fusion: (2)+(5)	8.4	8.3
(D) Fusion: (2)+(6)	8.4	8.3
(E) Fusion: ALL	**8.0**	**7.8**