Comparing Prosodic Models for Speaker Recognition

Cheung-Chi Leung, Marc Ferràs, Claude Barras, Jean-Luc Gauvain

Comparing Prosodic Models for Speaker Recognition

Cheung-Chi Leung[1], *Marc Ferràs*[1], *Claude Barras*[1,2] *and Jean-Luc Gauvain*[1]

[1]LIMSI-CNRS, BP 133, 91403, Orsay, France
[2]Univ Paris-Sud, F-91405, Orsay, France
{ccleung,ferras,barras,gauvain}@limsi.fr

Abstract

Recently, speaker verification systems using different kinds of prosodic features have been proposed. Although it has been shown that most of these speaker verification systems can improve system performance using score-level fusion with state-of-the-art cepstral-based systems, a systematic comparison of the prosodic modelling algorithms used in these prosodic systems has not yet been performed. This motivated us to review the proposed prosodic modelling algorithms and compare them using a common experimental condition.

These experiments explored different approaches in the sampling/segmentation of prosodic contours and the selection of prosodic features. They show that simple prosodic systems with features extracted from fixed-size contour segments, without knowledge of phone/pseudo-syllable level information, still provide significant performance improvement when fused with a state-of-the-art cepstral-based system. Moreover, some prosodic systems are shown to be complementary to each other. Fusion of these systems with the cepstral-based system can provide further performance improvement on the speaker verification task.

Index Terms: Speaker recognition, prosodic features

1. Introduction

Cepstral features, such as MFCC, and speaker modelling techniques, such as Gaussian Mixture Models (GMM) and Support Vector Machines (SVM), have become the predominant approaches in speaker verification. The performance of such systems is however relatively sensitive to the recording conditions. It is believed that prosodic features are less vulnerable to the channel distortion than cepstral features. Although prosodic features alone cannot perform as well as cepstral features, the fusion of these two types of features has been proposed to further improve the performance of conventional cepstral-based speaker verification systems [1, 2, 3, 4, 5, 6].

Prosody is used to describe many speech characteristics, such as speaking rate, loudness and pitch. Pitch and energy are commonly used in prosodic systems and these features are the main focus of this paper.

Many approaches have been proposed in prosodic systems. For instance, in prosodic contour sampling/segmentation, Carey *et al* [1] and Xie *et al* [2] extracted pitch features from fixed-size contour segments. However, Mary *et al* [3], Shriberg *et al* [4] and Dehak *et al* [5] proposed to segment an utterance into syllable or pseudo-syllable units and extract pitch feature per syllable/pseudo-syllable. In prosodic feature selection, Xie *et al* [2] and Mary *et al* [3] used pitch statistics, such as the mean, minimum and maximum values of pitch, as features. Dehak *et al* [5] used Legendre polynomials to approximate pitch contours. Moreover, Adami *et al* [6] suggested to capture temporal dynamic prosodic information with delta-pitch and delta-energy, and n-gram modelling.

Although it has been shown that prosodic systems can provide performance gains using score-level fusion with cepstral-based systems, a comparison of the prosodic modelling algorithms used in these systems has, to the best of our knowledge, not yet been performed. This motivates us to review these proposed prosodic modelling algorithms and compare them through a common experimental evaluation. In the experiments reported in this paper, we explore different approaches to the sampling/segmentation of prosodic contours and the selection of prosodic features. We also study whether these different approaches can complement each other and if their fusion can provide further performance improvements on the speaker verification task.

The remainder of this paper is organized as follows: Section 2 summarizes the algorithms that we adopt and evaluate. Section 3 describes the experimental conditions and results, followed by conclusions in Section 4.

2. Prosodic models

A prosodic system typically involves four major components: prosodic contour extraction; prosodic contour sampling/segmentation: prosodic feature selection: and speaker probabilistic modelling of the prosodic features.

Prosodic contours on log scale are extracted, being sampled every 10ms with a 30ms analysis window using the Praat toolkit [7]. Pitch estimation is based on the local maxima of the short-term autocorrelation function of the utterance [8]. In the estimation, the pitch floor, the pitch ceiling and the maximum number of pitch candidates are set to 50Hz, 500Hz and 5 respectively. The log energy is normalized by substracting the maximum value in the utterance. The duration feature is extracted from the prosodic contour segmentation.

2.1. Prosodic contour sampling/segmentation

A prosodic contour may cover information across several syllable or word units. Speaker-specific characteristics may be found in short-term static or/and dynamic features, such as the statistics of each speaker's dynamic range of pitch values [2] and the rising and falling patterns in prosodic contour segments [6]. To ensure that the features extracted from each contour contain such speaker-specific information, we segment the contours based on phone-level boundaries or pseudo-syllable boundaries, as well as dividing the contours into a number of fixed-size segments.

This work has been partially financed by OSEO under the Quaero program.

Starting with the English word transcriptions provided with the evaluation corpus, the LIMSI automatic speech recognition (ASR) system [9] is used to obtain the phone alignment. The phone-level time labels are then chosen as the segment boundaries. The segment duration of each contour segment is also appended in the feature vector, which will be defined in Section 2.2.

Pseudo-syllable segment boundaries can be located based on the valley points of the energy contour [10]. Similar to the phone segmentation method, the segment duration of each contour segment is also appended in the feature vector, which will be defined in Section 2.2.

The prosodic contours are also chunked into a number of equal-size segments, each of which contains a number of frames extracted from the 30ms analysis window in the Praat toolkit, and with a segment shift of 10 ms.

2.2. Prosodic features

Two types of prosodic features are used, including general statistics of pitch and energy values, and Legendre coefficients of pitch contours.

In the first approach, the features used are the mean, minimum, maximum and delta of the pitch values, and delta of energy values in each contour segment [2]. The delta feature is computed as the difference between the mean values in the first half and the second half of the contour segment. In systems using phone or pseudo-syllable contour segmentation, the segment duration of each contour segment is also appended to the feature vector.

Moreover, we use Legendre coefficients to approximate pitch contours. Similar to [5], each pitch contour segment along time t is approximated by a sequence of Legendre polynomials as

$$f(t) = \sum_{i=0}^{M} a_i P_i(t) \tag{1}$$

where $P_i(t)$ is the i-th Legendre polynomial defined as

$$P_i(t) = \frac{1}{2^i i!} \frac{d^i}{dt^i}[(t^2-1)^i] \tag{2}$$

The first M (M = 4, 6, 8, 10 or 12) coefficients of each contour segment are used to form a M-dimensional feature. In the experiments using phone or pseudo-syllable contour segmentation, the contour segment length is appended, forming the M+1-th dimensional feature. This method and the method using general statistics of pitch values share some identical features. These are a_0 and a_1, which represent the pitch mean and the delta pitch of the contour respectively.

2.3. Speaker probabilistic modelling of prosodic features

GMMs are used to model general statistics of pitch values and the Legendre coefficients, while N-gram models are used to model delta-pitch and delta-energy features.

In our experiments, GMMs are trained by MAP adaptation [11] of the Gaussian means of the corresponding gender-dependent UBM using 3 iterations of the EM algorithm. In the GMM system using general statistics of pitch values in fixed-size contour segments, a 4-dimensional feature vector is used. In systems using general statistics of pitch values in phone or pesudo-syllable segments, segment duration is included in the feature vector and thus a 5-dimensional feature vector is used. In systems using Legendre coefficients as features, the coefficients and the segment duration in each segment form a $M+1$ dimensional feature vector.

When an N-gram is used, the delta-pitch and the delta-energy are quantized into N_p and N_e tokens respectively. Speech data is needed to train the quantization boundaries so that the delta features are equally distributed into their quantized tokens. Unvoiced segment are represented by a "UV" token. In the system extracting features in fixed-size segments, there are $N_p \times N_e + 1$ quantized tokens in the contour segment representation. In the systems using the phone or pseudo-syllable segmentation, the segment duration is also quantized into N_d tokens and included in the contour segment representation. Therefore, $N_p \times N_e \times N_d + 1$ quantized tokens are involved. In [6], $(N_p, N_e, N_d) = (2, 2, 3)$ is used and the pitch and energy contours are segmented according to the local minima and maxima of pitch values. In our experiments, different combinations of (N_p, N_e, N_d) are tested, and the pitch and energy contours are segmented according to the methods described in Section 2.1.

Standard maximum likelihood estimation and back-off are used for each n-gram model representing a speaker. Bi-gram and tri-gram models are used. To deal with the data sparseness, an interpolation in n-gram probabilities is calculated as

$$p'_m = (1-\alpha)p_m + \alpha p_{ubm} \tag{3}$$

where p'_m is the re-estimated probability, p_m and p_{ubm} are the probabilities from the speaker specific training data and the universal background data respectively, and α is an adaptation weight between 0 and 1. Given a test utterance, a weighted log-likelihood ratio between the target speaker model and the background model is computed.

3. Experiments

3.1. Task and evaluation data

All the systems via speaker verification experiments conducted on conversational telephone speech. The data is that used in the one-conversation two-channel condition task of the NIST SRE'05 evaluation [1].

Given a 5-minute long test conversation and a target speaker, the goal is to decide whether this segment was spoken by the target speaker or not. For each target speaker (274 male and 372 female), a 5-minute long conversation is available for model training. Overall, 2429 test segments (1074 male and 1355 female) need to be scored against roughly 10 gender-matching impostors and against the true speaker. The gender of each target speaker is known. Only the English subset of the evaluation data is considered in our experiments.

The primary performance measure for the NIST speaker verification task is the Detection Cost Function (DCF) defined as a weighted sum of missed detections and false alarms, the normalized cost taking the following form $C_{Norm} = P_{Miss} + 9.9 \times P_{FalseAlarm}$. For all results, we report the Minimal DCF (MDC) value obtained a posteriori for the best possible detection threshold. However, this operating point favors false alarms, so the Equal Error Rate (EER) is also provided as an alternative performance measure.

[1] The NIST year 2005 speaker recognition evaluation, http://www.nist.gov/speech/tests/sre/2005/index.html.

Table 1: Configuration of each prosodic system in the experiments († Duration feature is included in systems with features extracted from phone or pseudo-syllable segments)

Systems		S1	S2	S3	D1	D2	D3	L3
Prosodic contour segment	Fixed-size segment	√			√			
	Phone segment		√			√		
	Pseudo-syllable segment			√			√	√
Prosodic features	Mean, min, max and delta of pitch †	√	√	√				
	Delta-pitch & delta-energy †				√	√	√	
	Legendre coefficients †							√
Speaker model	GMM	√	√	√				√
	N-gram				√	√	√	

3.2. Prosodic systems and MFCC-GMM system

Seven prosodic systems were evaluated in our experiments. The configuration of each system is summarized in Table 1. In the prosodic systems, the training data of each gender-dependent UBM was chosen from 1309 target speakers (770 female and 539 male) in the 1-conv and 8-conv trial conditions in the NIST SRE'04 evaluation. This data was also used in the detection of quantization boundaries in the prosodic *n*-gram systems. 128-mixture GMMs were used in prosodic GMM systems. In the prosodic *n*-gram systems, we used the adaptation weight α = 0.5, which was found to be optimal on the evaluation data.

The MFCC-GMM system was implemented in the same way as in [14]. Each of the gender-dependent UBMs was a 1536-mixture GMM. The training data was chosen from the target speakers in NIST SRE '97-'01 and '03 evaluations and the test speakers in NIST SRE'03 evaluation (for a total of 9041 speech excerpts).

Score normalization was performed using T-norm [13] in the prosodic and MFCC-GMM systems. In the prosodic systems, T-norm model training and the UBM training shared the same data. In the MFCC-system, T-norm model training was chosen from 500 speech excerpts (250 male and 250 female) from the Fisher corpus [2].

Linear logistic regression score-level fusion [15] was used, and a three-fold cross-validation scheme was adopted for the performance evaluation.

3.3. Results

Different parameters in the prosodic systems were tested and their fusion with the MFCC-GMM system was evaluated in the experiments.

First, the effect of the segment size in the fixed-size prosodic contour segments was investigated. Segment sizes ranging from 100ms to 140ms performed well, and the best performing system had a segment size of 120ms.

The effect of selecting different statistics of the pitch values was investigated with system S3. The experiment showed that each of these features contributed to the system performance. The standard deviation of pitch values was also tested in the feature set, but it did not contribute to the system performance.

The effect of Legendre polynomial order used in system L3 was investigated. Since a 6th order polynomial performs the best at most operating points, this setting is used in following experiments.

In the three prosodic *n*-gram systems, the effect of quantization-level of features was investigated. In system D1, (N_p, N_e) = (5,3) performed the best. In system D2, (N_p, N_e, N_d) = (3, 2, 3) performed the best. In system D3, (N_p, N_e, N_d) = (4, 2, 2) performed the best. The effect of the size of the *n*-gram was also investigated. In all three systems, bi-gram models performed better than tri-gram models.

[2]Fisher Corpus, LDC Catalog, http://www.ldc.upenn.edu/Catalog.

Table 2: Performance (in MDC and EER) of various individual prosodic systems

System	MDC	EER (%)
S1	0.908	21.37
S2	0.877	21.04
S3	0.897	20.88
D1	**0.837**	22.62
D2	0.897	25.16
D3	0.953	28.61
L3	0.877	**19.17**

Table 3: MDC and EER of MFCC-GMM system (B) and its fusion with various prosodic systems

System	MDC	Relative improvement in MDC	EER (%)	Relative improvement in EER
B	0.323	—	8.60	—
B + S1	0.309	4.3 %	8.07	6.2 %
B + S2	0.314	2.8 %	**7.99**	**7.1 %**
B + S3	0.323	0 %	8.27	3.8 %
B + D1	**0.300**	**7.1%**	8.15	5.2 %
B + D2	0.318	1.6 %	8.40	2.3 %
B + D3	0.325	-0.6 %	8.52	0.9 %
B + L3	0.324	-0.3 %	8.19	4.8 %

The performance of each individual prosodic system (with the best setting reported previously) is summarized in Table 2. In terms of MDC, system D1 performed the best, whereas in terms of EER, system L3 performed the best.

Each prosodic system was also fused with the MFCC-GMM system. The performance of the MFCC-GMM system and the fusion systems are shown in Table 3. The experiments showed that system D1 provided the best fusion improvement in terms of MDC, where as system S2 provided the best fusion improvement in terms of EER.

We also performed the best-3 score-level fusion test for the prosodic and MFCC-GMM systems. The 3 best performing fusions (in terms of EER) are shown in Table 4. The best fusion

Table 4: MDC and EER of 3 best performing score-level fusion of 3 systems (EER in ascending order)

B	S1	S2	S3	D1	D2	D3	L3	MDC	EER(%)
√		√		√				0.296	7.69
√				√			√	0.309	7.69
√	√			√				0.299	7.73

system provides the EER of 7.69 and the MDC of 0.296 (i.e. relative improvement of around 10% in EER and 8% in MDC). Figure 1 shows that the fusion of two and three systems both provide further performance improvements at most operating points.

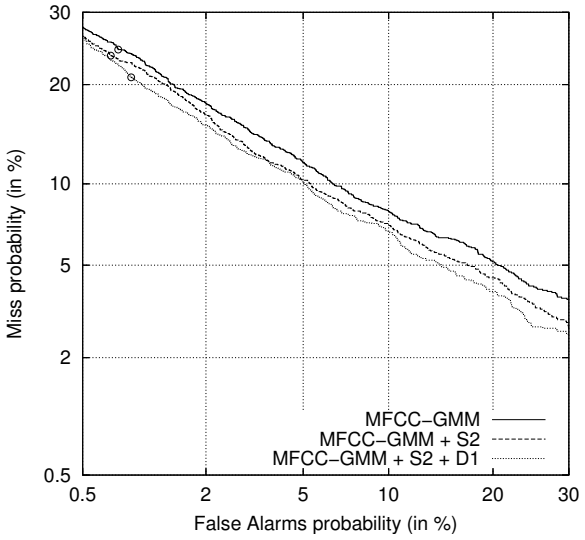

Figure 1: DET curves showing the performance of MFCC-GMM, best-2 and best-3 (in terms of EER) fused systems

System D1 was shown to be one of the most important prosodic systems for the fusion. The addition of the other two prosodic *n*-gram systems, which capture and model similar features as system D1, did not provide any further system fusion improvement. Similarly, since systems S1, S2, S3 and L3 capture and model features in similar ways, it is reasonable that their fusion did not provide any further performance improvement.

It is worth noting that the simple prosodic systems extracted features from fixed-size contour segments, without the knowledge of phone/pseudo-syllable level information, still provide satisfactory results. The system fusion of the MFCC-GMM system and two simple prosodic systems (system S1 and D1) provides a relative improvement of 10% in EER in the English subset of the NIST SRE'05 evaluation data. This result is comparable to the results reported in other prosodic systems. The prosodic system (similar to system L3) in [5] includes the energy contour in the Legendre polynomial approximation and factor analysis is used to compensate for the inter-session variability in the Legendre coefficients. Its fusion with a MFCC-GMM system provides a relative improvement of 12% in EER on the English subset of the NIST SRE'06 evaluation data. The prosodic system in [4] uses the SNERF approach with a SVM classifier and a speech recognition system is needed in this approach. Its fusion with a MFCC-GMM system provides a relative improvement of 14% in EER in the English subset of the NIST SRE'06 evaluation data.

4. Conclusions

This paper has investigated various methods used in prosodic contour sampling/segmentation and prosodic feature selection in some proposed SRE systems. Our experiments show that the simple prosodic systems with features extracted from fixed-size contour segments, without the knowledge of higher level information, still provide comparable performance gain in their fusion with a state-of-the-art cepstral-based system. Moreover, some prosodic systems are shown to be complementary to each other and their system fusion with the cepstral-based system can provide further performance improvement on a speaker verification task.

5. References

[1] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennet, "Robust Prosodic Features for Speaker Identification," *Proceedings of ICSLP*, pp. 1800-1803, 1996.

[2] Y. L. Xie, X. Zhou, Z. Q. Yao, J. X. Chen, and M. H. Liu, "University of Science and Technology of China SSIP Laboratory NIST SRE 2005 System," *NIST SRE Workshop*, June 2005.

[3] L. Mary and B. Yegnanarayan, "Prosodic features for speaker verification," *Proceedings of ICSLP*, pp. 917-920, 2006.

[4] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling Prosodic Feature Sequences for Speaker Recognition," *Speech Communication*, pp. 455-472, 2005.

[5] N. Dehak, P. Kenny, and P. Dumouchel, "Continuous Prosodic Features and Formant Modeling with Joint Factor Analysis for Speaker Verification," *Proc. of Interspeech*, pp. 1234-1237, 2007.

[6] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," *Proceedings of ICASSP*, pp. IV-788-91, 2003.

[7] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," *http://www.praat.org*.

[8] P. Boersma, "Accurate Short-term Analysis of the Fundamental Frequency and the Harmonics-to-noise Ratio of a Sampled Sound," *IFA Proceedings 17, University of Amsterdam*, pp. 97-110, 1993.

[9] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, 37(1-2):89-108, 2002

[10] C.Y. Lin and H.C. Wang, "Language Identification Using Pitch Contour Information," *Proc. of ICASSP*, pp. I-601-604, 2005.

[11] C. H. Lee and J.-L. Gauvain, "Speaker Adaptation Based on MAP Estimation of HMM Parameters," *Proceedings of ICASSP*, pp. 558-561, 1993.

[12] B. Baker, R.Vogt, M. Mason, and S. Sridharan, "Improved Phonetic and Lexical Speaker Recognition through MAP Adaptation," *Odyssey: The Speaker and Language Recognition Workshop*, pp. 94-99, 2004.

[13] C. Barras and J.-L. Gauvain, "Feature and Score Normalization for Speaker Verification of Cellular Data," *Proceedings of ICASSP*, pp. II-49-52, 2003.

[14] M. Ferras, C.C. Leung, C. Barras, and J.-L. Gauvain. " Constrained MLLR for Speaker Recognition," *Proceedings of ICASSP*, p.p. 53-56, 2007.

[15] S. Pigeon, P. Druyts, and P. Verlinde, "Applying Logistic Regression to the Fusion of the NIST'99 1-Speaker Submissions," *Digital Signal Processing*, 2000.