Syllable based Feature-Contours for Speaker Recognition

Marcel Kockmann[1 2], *Lukáš Burget*[1]

[1]Speech@FIT, Brno University of Technology, Czech Republic
[2]Siemens AG, Corporate Technology, Munich, Germany
{kockmann|burget}@fit.vutbr.cz

Abstract

The use of acoustic short-time features for speaker identification has established itself for several years. In this work, classical acoustic as well as prosodic features are jointly used in a lexical context. The contours of pitch, energy and cepstral coefficients are continuously modeled over the time span of a syllable to capture the speaking style on phonetic level. Especially the addition of cepstral contours highly improves the performance. Due to the small amount of data, best results are achieved with a factor analysis framework. Results are presented for the NIST SRE 2006 speaker identification task.

1. Introduction

State-of-the-art systems for text independent speaker identification usually make use of acoustic short-time features in a Gaussian Mixture Model (GMM) framework with Universal Background Model (UBM) [1]. As these systems are strongly affected by session variability, new techniques have been successfully developed in the last few years to compensate for these channel effects [2, 3]. Still, most acoustic systems do not make use of information from a higher level of speech, like the phonetic-, prosodic- or lexical-layer. Different studies have shown that adding phonotactic- or prosodic characteristics to an acoustic baseline system can yield to a better overall performance, especially when a large amount of data is available per speaker [4]. Ferrer *et al.* [5] and Dehak *et al.* [6] also reported gain in recognition performance on shorter tasks, where only a few hundred feature vectors are available to train and test each speaker.

The work in this paper is based on the use of classical prosodic features like duration, pitch and energy in a syllable-like temporal context. The trajectories of each feature is continuously modeled over the time span of a syllable and is represented by coefficients from a discrete cosine transformation (DCT). Additionally we also capture the contour of acoustic features in form of Mel-frequency cepstral coefficients (MFCC) and form a single feature vector out of duration and pitch, energy and the MFCC contours. All these features are jointly modeled using a GMM. As this mixed feature vector will also be affected by variations in the channel, established techniques for the compensation of session variability are applied. Since each feature vector represents one syllable in the utterance, there are only a few hundred features per recording, which makes it hard to reliably estimate the channel factors that determine how far a model is shifted in the channel subspace. We will investigate if channel compensation in the model or in the feature domain is more appropriate for this small amount of feature vectors as well as the expansion of this factor analysis model to include speaker factors to give a joint factor analysis model [6].

The performance of the proposed system is presented in terms of equal error rate for the text-independent NIST SRE 2006 speaker identification task.

The organization of the paper is as follows: section 2 describes the extraction of the syllable based features, including the basic features itself, the way the utterance is segmented into syllable-like units and based on this, the actual modeling of the temporal trajectory of the basic features. Section 3 briefly describes the algorithms used to perform the channel compensation. Section 4 presents the experiments and results obtained with the system and conclusions are given in section 5.

2. Heterogeneous syllable based features

This section describes how a feature vector for each syllable is obtained by continuously modeling the temporal trajectory of various frame based features.

2.1. Basic features

Different basic features are extracted at 10-ms intervals. Pitch frequencies are computed with the Average Magnitude Difference Function from the Snack Sound Toolkit [8]. Snack is also used to obtain windowed log power values. All these features are extracted with Snacks default settings. Furthermore 12 Mel-frequency cepstral coefficients (20ms window, 23 bands in Mel filter bank) are used.

2.2. Syllable segmentation

The segmentation into syllable-like units is based on the phonetically alignment from a phoneme recognizer with long temporal context [9]. We use a Hungarian recognizer, whose tokens are mapped to classes pause, consonant and vowel. Then each speech segment between two pauses is equally divided based on the number of vowels in this segment. Each vowel is considered as the nucleus of a syllable. In a second step, the estimated syllable boundary between two vowels can be shifted with regard to the measured pitch at the potential boundary candidates. This is done in order to preserve consecutive pitch contours that proceed for example from a vowel to a voiced consonant.

2.3. Contour modeling

2.3.1. Pre-processing

All basic features are pre-processed before actually modeling the temporal contour of them. Feature warping [10] (blind warping into normal distribution) is applied to all MFCCs and the logarithm is computed for the pitch frequencies. Finally, mean subtraction is applied to all features. Note that the mean was computed over the voiced parts of the whole utterance only (obtained by valid pitch). Small gaps (1 frame) in the pitch contour are smoothed by a median filter.

2.3.2. Temporal trajectory

The temporal contour of each feature can be approximated by a curve fitting tool. We use the first n DCT bases to model the trajectory, which correspond to characteristics of the curve, like mean, slope and finer details. The contour is represented by its DCT coefficients in the feature vector. The advantage of using discrete cosine transformation instead of a simple polynomial curve fitting is, that mapping the contour segment to a fixed length is not necessary and that the coefficients are already decorrelated.

As pitch may be undefined over parts of the syllable, one can consider different approaches to model the other features which are always defined within the syllable. In this work, jointly modeling the unvoiced and voiced part and modeling only the voiced part of each syllable is investigated for the other features.

2.4. Final feature vector

The number of voiced/unvoiced frames inside the syllable also serves as a discrete duration feature. The final feature vector for each syllable consists of the duration followed by the representation of the temporal contour for each basic feature like pitch, energy and MFCCs. Syllable segments that contain less frames than the number of DCT coefficients used to model the contour are omitted.

3. Session and Speaker Variability

Prosodic features like pitch and energy shall be used along with acoustic features like MFCCs. Channel compensation has proved to be beneficial for both of these feature types [6]. Challenging is the use of channel compensation with relatively sparse feature vectors as it is the case here. For this purpose, eigenchannel compensation was performed in both, feature and model domain as it was proposed in [11] and [12]. [6] also reported improvement through the use of a joint factor analysis model on a very similar task. This section gives a brief overview to the jointly used eigenchannel subspace and to the principles of the two different compensation techniques, as well as the use of eigenvoices for the speaker models.

3.1. Eigenchannel Subspace

The eigenchannel subspace is a low dimensional representation of how the means of a GMM representing a speaker can be affected by changing channel. This subspace is estimated as described in [11]. Briefly, a corpus with multiple recordings for each speaker under various conditions is needed. After adapting the UBM to each training utterance, mean supervectors are formed by concatenating all mean vectors and dividing them by corresponding standard deviation. The eigenchannels are the eigenvectors of the average within-speaker covariance matrix.

3.2. Eigenchannel Compensation in model and feature domain

Eigenchannel compensation in model domain is only applied to test conversations. During a single MAP-iteration, channel factors are estimated for the UBM as well as for each speaker model in test. These factors determine, how far each model is shifted towards the test-utterance in the directions defining the eigenchannel subspace.

A more simplified approach of channel compensation leads to the possibility of shifting the features itself, rather than the models as proposed in [12]. The channel compensated features can be used to train and test a standard GMM system.

3.3. Eigenvoice modeling

We expand the channel model to a joint factor analysis model, which is a model of speaker and session variability. In addition to channel directions, also speaker directions are identified and represented in a low-dimensional subspace. The theory and implementation is based on Kennys decoupled estimation as described in detail in [7]. For this data, a pure eigenvoice approach is used to model the speakers ($d = 0$). Initial eigenvoices and eigenchannels are computed by principal component analysis as described above. The final hyper-parameters are estimated iteratively in terms of maximum likelihood.

4. Experiments

4.1. Data

Experiments were performed on the core condition of the NIST 2006 speaker recognition evaluation (SRE) [13], which contains English trials only. The 1-side training 1-side test condition is considered, where approximately $2.5min$ of speech is available from a $5min$ telephone conversation to train each speaker and for each test trial. This set originally contains 462 female and 354 male training utterances (where multiple utterances can arise from one speaker) and 51448 test trials. Results are presented in terms of equal error rate (EER). The UBM model is trained on utterances from Switchboard II, Switchboard Cellular and the NIST 2004 and 2005 SRE data sets. The eigenvoices were estimated on all data, while the eigenchannel subspaces were estimated on NIST SRE data only. The same corpus was used to normalize verification scores via zt-norm [14].

4.2. Framework

The GMM framework used for the whole system is the same as used for an acoustic baseline system [11]. The gender-independent UBM is obtained by Expectation-Maximization (EM) Training and the speaker models are derived by MAP-Adaptation with $\tau = 19$. Discrete as well as continuous features are used within one feature vector, so variance flooring is crucial while EM training. Variances are floored to $1/100$ of the global variance. Initial experiments were carried out with 256 Gaussians, no eigenchannel compensation and no zt-norm.

4.3. Prosodic contour features

First experiments were performed with a classical prosodic feature vector, which comprises the duration of the syllable as well as the approximated pitch and energy contours, which are modeled with 6 DCT coefficients (minimal segment length is $60ms$). Results for different assortments of the feature vector are presented in Table 1. As can be seen it is most beneficial to use duration, pitch and energy jointly which also conforms to similar results in [6].

As the feature vector will grow through the augmentation of MFCC features, we want to use the smallest number of coefficients to properly approximate the temporal contour in terms of recognition performance. Table 2 shows that modeling even finer details is not beneficial and that only a slight degradation has to be accepted by reducing the resolution to 4 DCT coefficients.

Table 1: *Different prosodic feature vectors with 6 coefficients per contour.*

Feature Vector	Dim	EER [%]
Pitch Contour	6	29.67
Duration, Pitch Contour	7	29.1
Pitch & Energy Contour	12	28.37
Duration, Pitch & Energy Contour	13	25.73

Table 2: *Pitch & Energy contours modeled by different number of DCT coefficients.*

# of coefficients	EER [%]
4	26.11
5	25.77
6	25.73
7	27.29

The best performing 13-dimensional feature vector was also used to study the treatment of unvoiced parts within a syllable. Either the duration and the energy contour may correspond to the whole syllable or only to the voiced part. As can be seen in Table 3, it is beneficial to use only the voiced part of the syllable. Note also that the mean subtraction of the basic features in the pre-processing step is based only on the voiced parts as well. Using all speech segments as determined by the phoneme recognizer to compute the mean yields to much worse results.

Table 3: *Modeling whole syllable or only voiced part.*

Feature Vector	EER [%]
whole Duration, Pitch & whole Energy Contour	25.73
voiced Duration, Pitch & voiced Energy Contour	24.4

4.4. Expansion of feature vectors

For the following experiments, the number of DCT coefficients was reduced to 4. As the minimal segment length also is reduced to $40ms$, about 10% more feature vectors could be extracted for each utterance. This and additional feature warping of the energy coefficients reduced the EER to 22.3%, which serves as a reference for expanding the feature vector with MFCC contours.

In order to add a simple acoustic information, the prosodic feature vector was augmented with the means of 12 MFCCs over the syllable. This results in a drastic gain in recognition performance to 14.07%. The benefit of adding all coefficients for the MFCC contours can be seen in Table 4. Adding information about the temporal contour of all MFCCs yields to an EER of 9.87%, which is a relative improvement of 55% compared to the purely prosodic system. Even the contours of the higher MFCCs are beneficial and omitting them always results in worse performance (see also Table 4). Also the addition of the cepstral contours does not make the prosodic information negligible, as performance degrades to 10.63% for cepstral contours only.

Table 4: *Augmentation of prosodic feature vector (baseline: duration, pitch & energy contour). Contours are modeled with 4 coefficients, voiced parts only.*

Feature Vector	Dim	EER [%]
Baseline	9	22.3
Baseline + 12 MFCC means	21	14.07
Baseline + 12 MFCC Contours	57	**9.87**
Baseline + 11 MFCC Contours	53	10.14
Baseline + 10 MFCC Contours	49	10.57
Baseline + 9 MFCC Contours	45	11.22
Baseline + 8 MFCC Contours	41	11.27
12 MFCC Contours	48	10.63

4.5. Channel Compensation

The effectiveness of eigenchannel compensation in model and feature domain was investigated for a system trained on a 57-dimensional vector containing duration and the temporal trajectories for pitch, energy and 12 cepstral coefficients. 10 eigenchannels were used in the experiments. Note that only approximately 500 feature vectors are available in this syllable-framework to estimate the channel factors that determine the compensation of each utterance. Table 5 shows the effect of the channel compensation for GMMs with different number of Gaussians. For small models with only 32 Gaussians, the channel factors can be estimated quite well and the compensation in model as well as in feature domain results in 30% relative improvement, while for a model with 512 Gaussians, the gain is only about 5%. Unfortunately the small models perform much worse before applying the channel compensation, and EER is still worse after eigenchannel adaptation. However, for the model with 256 Gaussians the EER could still be reduced by 11% to 8.74%, even with this small amount of data.

Table 5: *Effects of channel compensation for different sized GMMs (10 Eigenchannels) in EER [%].*

# of Gaussians	No CC	Model Domain	Feature Domain
512	9.44	9.06	9.06
256	9.87	8.8	8.74
128	10.89	8.8	8.75
64	12.35	9.3	9.3
32	14.88	10.41	10.42

Eigenchannel compensation in feature domain bears the opportunity to compensate the features on a subspace created on a smaller UBM and do the model training and evaluation with a larger GMM, as it was performed for language identification in [15]. This technique assumes that the properly estimated channel directions and channel factors also fit for the bigger GMM. In our experiments the features were compensated on GMM sizes where the standard compensation showed adequate performance. These compensated features were used to train model sizes that performed best without channel compensation. As can be seen in Table 6, this approach to handle the sparse data results in better performance than the normal eigenchannel adaptation. The relative improvement compared to the standard compensation is 6% and 8% for the GMM sizes 256 and 512, respectively.

Table 6: *Different sized models with features compensated on smaller Eigensubspace (sizes in # of Gaussians).*

Speaker UBM	Subspace UBM	EER [%]
512	128	8.31
512	64	8.36
256	128	8.2
256	64	8.36
128	64	8.9

4.6. Joint factor analysis

A joint factor analysis system with 512 Gaussians, 150 eigenvoices and 35 eigenchannels for the 57-dimensional vectors is used to investigate the effectiveness of this approach for our case. As shown in Table 7, the benefit from the joint use of speaker and channel factors instead of pure eigenchannel compensation (as in Table 6) is quite large. As reported in [6] the joint factor analysis model seems to be very capable to handle sparse amounts of data.

Table 7: *EER [%] for systems without channel compensation, eigenchannel compensation in model and feature domain and factor analysis model (FA).*

No CC	Model Domain	Feature Domain	FA
9.44	9.06	9.06	**5.91**

As we always used a gender-independent framework (in contrast to [6, 7]), a completely gender-dependent system was also investigated. UBMs as well as the hyper-parameters were trained on male and female utterances, respectively. Results are presented in Table 8 for two types of feature vectors. While the purely prosodic system enhances only slightly, the EER for the big system is reduced by nearly 40% relatively.

Table 8: *EER [%] for gender-dependent (GD) and gender-independent (GI) FA-Systems with different number of eigenvoices (EV) and eigenchannels (EC).*

Feature Vector	EV	EC	GI	GD
Dur, F0 & Energy	50	20	15.27	14.62
Dur, F0, Energy & MFCCs	150	35	5.91	**3.63**

5. Conclusions

We have shown that syllable based prosodic feature vectors can be successfully expanded and jointly modeled with acoustic cepstral features by the use of DCT coefficients to represent the temporal contour of each phonetically motivated segment. The addition of cepstral contours achieves over 50% improvement compared to a classical prosodic system with duration, pitch and energy only. Without any compensation for session variability, the performance of such a system is comparable to a frame-based acoustic system and comprises complementary information through different kinds of features like pitch and a different temporal context. As the effect of channel compensation decreases for the proposed system due to the small amount of features in the test utterance, it could be shown that a gender-dependent joint factor analysis system highly improves the performance and gives results similar to short-time eigenchannel systems.

6. Acknowledgments

This work was partly supported by European projects AMIDA (FP6-033812), Caretaker (FP6-027231) and MOBIO (FP7-214324), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 201/2006. Lukas Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383.

7. References

[1] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., “Speaker Verification Using Adapted Gaussian Mixture Models”, Digital Signal Processing 10, 19-41 (2000).

[2] Kenny, P. and Dumouchel, P., “Disentangling speaker and channel effects in speaker verification”, in Proc. ICASSP, 2004, pp. 37–40.

[3] Brümmer, N., “Spescom DataVoice NIST 2004 system description”, in Proc. NIST Speaker Recognition Evaluation 2004, Toledo, Spain, Jun. 2004.

[4] Douglas et al., “The SuperSID Project: Exploiting High-level Information for High-accuracy”, Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP ’03). 2003 IEEE International Conference.

[5] Ferrer, L., Shriberg, E., Kajarekar, S. and Sonrnez, K., “Parameterization of Prosodic Feature Distributions for SVM Modeling”, in Speaker Recognition, Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference

[6] Dehak, N., Dumouchel, P., Kenny, P., “Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification”, in Audio, Speech, and Language Processing, September 2007, Volume 15. pp. 2095–2103.

[7] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P. “A Study of Inter-Speaker Variability in Speaker Verification” , IEEE Transactions on Audio, Speech and Language Processing, July 2008

[8] Sjölander, K., “The Snack Sound Toolkit”, Online on: http://www.speech.kth.se/snack.

[9] Schwarz, P., Matějka, P., Černocký, J., “Hierarchical structures of neural networks for phoneme recognition”, in Proceedings of ICASSP 2006, Toulouse, FR, 2006, p. 325-328.

[10] Pelecanos, J. and Sridharan, S., “Feature Warping for Robust Speaker Verification”, in proc of A Speaker Odyssey, Paper 1038, 2001. pp. 3117–3120.

[11] Burget, L., Matějka, P., Glembek, O., Schwarz, P. and Černocký, J. H., “Analysis of feature extraction and channel compensation in GMM speaker recognition system,” in IEEE Trans. on Audio, Speech and Language Processing, September 2007, Volume 15. pp. 1979–1986.

[12] Castaldo, F., Colibro, D., Dalmasso, E., Laface, P. and Vair, C., “Compensation of Nuisance Factors for Speaker and Language Recognition”, in IEEE Trans. on Audio, Speech and Language Processing, September 2007, Volume 15. pp. 1969–1978.

[13] “The NIST Year 2006 Speaker Recognition Evaluation Plan”, Online on: http://www.nist.gov/speech/tests/spk/2006.

[14] Auckenthaler, R., Carey, M. and Lloyd-Tomas, H., “Score normalization for text-independent speaker verification systems”, in Digital Signal Processing, vol. 10, pp. 42–54, 2000.

[15] Hubeika, V., Burget, L., Matějka, P. and Schwarz, P., “Discriminative Training and Channel Compensation for Acoustic Language Recognition”, in Proc. Interspeech, 2008, submitted.