SYLLABLE BASED PROSODIC FEATURES FOR SPEAKER IDENTIFICATION

Marcel Kockmann
Postgraduate Program, FIT BUT
E-mail: kockmann@fit.vutbr.cz

Supervised by: Lukas Burget, Jan Cernocky
E-mail: {burget},{cernocky}@fit.vutbr.cz

ABSTRACT

The use of acoustic features for speaker identification has established itself for several years. In recent years the research on automatic speaker identification has been extended to the use of features from a higher level of speech, like the phonetic-, prosodic- or linguistic-layer. This paper deals with the use of prosodic features, namely duration, pitch and energy, as input for a statistically modeling approach. The features are extracted and continuously modeled over a phonetically segment, a so called pseudo-syllable. These feature vectors serve as input for a Gaussian-mixture-model (GMM), as it is used in a standard acoustic system. Results are presented for the NIST SRE 2006 speaker identification task.

1 INTRODUCTION

Automatic speaker identification deals with the task of recognizing a previously trained speaker from a recorded utterance. One has to distinguish between text dependent and text-independent speaker identification. The text-dependent task is mainly used to verify the identity of a given speaker, so the system knows the content of the utterance as well as the speaker. This paper deals with the text-independent identification of a speaker, in which case the system has no prior information about the speech and the speaker itself. In both cases, the speaker has to be enrolled by a certain amount of speech before performing recognition. The most interesting application for text-independent speaker identification is likely to be in the field of forensics.

This work concentrates on building-up a system that makes use of prosodic features, instead of acoustic ones. Acoustic features somehow represent the “acoustic fingerprint” of a speaker and perform very well, but a mayor drawback is the sensibility against variability in recording conditions, e.g. channel, background noise, etc. Higher-level features has shown to be more robust against channel variability and consist of complementary information [1]. Beside developing a pure prosodic system, the aim of this work will be the fusion of acoustic and higher-level systems to gain a better overall performance.

The organization of the paper is as follows: section 2 introduces the main structure of an GMM system, as it is used for the acoustic as well as the prosodic system; section 3 deals with the features and methods to create the final prosodic features; experiments can be found in section 4; conclusions and future work is in section 5.

2 GMM-SYSTEM

The GMM serves as the classificator for the system. GMMs are one of the standard methods to perform speaker identification [2]. The distribution of the input features from each speaker is modeled by a number of Gaussian distributions, where each Gaussian is represented by weight, mean vector and covariance matrix. Beside a GMM-model for each trained speaker, there is also a so called universal-background-model (UBM). The UBM is trained on speech from many different speakers and represents the average target population.
While testing, the likelihoods for a speech segment given each model (including UBM) are computed. The final classification is realized by a hypothesis test, which is a likelihood ratio test given by the probabilities of the two hypothesis H_1: *Segment was spoken by the speaker to be tested* and H_2: *Segment was spoken by someone from the average population.*

3 PITCH AND ENERGY CONTOUR FEATURES

Three kinds of information from the prosodic level are used to form the final prosodic features. The first feature is the duration of the phonetic segment that is modeled. Each segment is a pseudo-syllable that may consist of a number of consonants followed by a vowel. For each segment we take the pitch and energy to model their temporal trajectory.

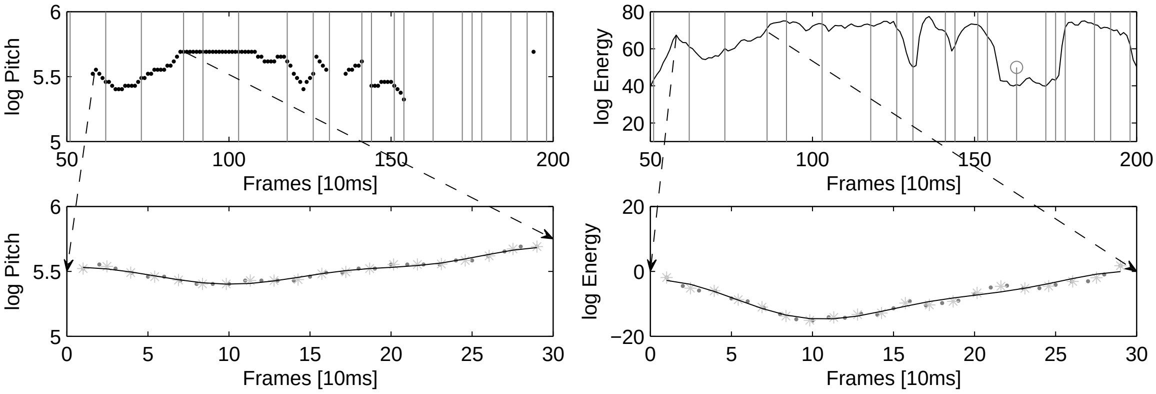

Figure 1: *1st row: Original pitch and energy values with phoneme boundaries. 2nd row: Original and interpolated values (dotted lines) and approximated curve (solid line)*

3.1 EXTRACTION OF DURATION, PITCH AND ENERGY

In order to get the duration and the boundaries of the syllables, phonemes are extracted by a phoneme-recognizer with long temporal context [3]. Pitch is computed with the Average Magnitude Difference Function (AMDF) from the Snack Sound Toolkit [4]. Energy is also obtained by Snack and is normalized by its mean value.

3.2 PITCH AND ENERGY CONTOUR

The contour of pitch and energy over each syllable is approximated by some kind of curve fitting in order to get coefficients that represent the trajectory in a more compact and decorrelated way than the values itself. Two approaches are used, polynomials and discrete cosine transformation

(DCT). For the polynomial curve fitting, the coefficients of the polynomial $c_n * x(t)^{n-1} + c_{n-1} * x(t)^{n-2} + ... + c_2 * x(t) + c_1$ that fits the data best in a least-squares sense are computed. For the DCT features, the first n coefficients of the transformed values are taken. In both cases, the coefficients represent characteristics of the curve, like mean, slope, curvature and fine details. Before approximating the curves, the segment is time normalized by linear interpolation, so that contours from different sized segments are comparable and can serve as input vectors for the GMM system. These steps are visualized in Figure 1. The first row shows the extracted pitch and energy values, respectively. The vertical lines are the phoneme boundaries obtained by the phoneme-recognizer. In the second row you can see how a pseudo-syllable over three phonemes is created from the consecutive pitch and energy values.

4 EXPERIMENTS

Several different types of pitch and energy contour features are generated and evaluated on the core condition of the NIST 2006 speaker identification task [5]. That comprises approximately 2.5 min of speech for training and testing. GMMs with 512 mixtures and diagonal covariances are used. Speaker models are derived by standard MAP-Adaptation from the UBM model [2]. The UBM is trained on data from a previous NIST evaluation (2004). The results are presented in terms of Equal Error Rate (EER), which is the point denoting equal false-rejection- and false-alarm-rate.

Table 1 shows results for best performing configuration: feature vectors consisting of 6 coefficients for pitch and 6 for the energy contour. Experiments has been performed with and without the duration feature. As can be seen, adding the duration feature always decreased the EER. Overall, the DCT coefficients seem to perform better than the polynomial coefficients. The best achieved result is an EER of 26.2% which is comparable with results that have been published on similar work [7].

Table 1: *EER [%] for different prosodic feature vectors*

	Polynomial	DCT
Pitch,Energy	28.7	27.4
Duration,Pitch,Energy	27.9	26.2

Subsequently, methods that are working well to enhance an acoustic system have been applied to the baseline prosodic system. These include approaches in the feature- as well as the model-domain. After augmenting the feature vectors with their time derivatives (so called delta-features) a feature transformation is applied (HLDA) to decorrelate and reduce the feature vectors. In the model domain the so called eigenchannel-adaptation is performed while testing the files, in order to compensate for possible channel mismatch between the training and the test utterances. All these methods are described and evaluated for the same task on an acoustic system in [6].

Results for the enhanced prosodic system are shown in Table 2. It can be seen that the EER could be reduced successively down to 23.6% by applying these methods. Especially for the Eigenchannel Adaptation the gain in performance is not as good as for the acoustic system, but the results in Table 2 show that these techniques also do work for prosodic features.

Table 2: *Results for the enhanced prosodic systems*

	EER [%]
Baseline	26.2
Baseline + Deltas	25.3
Baseline + Deltas + HLDA	24.3
Baseline + Deltas + HLDA + Eigenchannel	23.6

5 CONCLUSION

Encouraging first results for this stand-alone prosodic system could be achieved. Even if the EER is worse than for a acoustic system, it has been shown in a similar work with a more sophisticated modeling approach [7], that those prosodic features increased the overall performance of a fused acoustic-prosodic system relatively by 12%. The focus for future work should be the further enhancement of the prosodic features and the fusion with an acoustic GMM System. Feasible approaches could be the use of more complex feature vectors [8], clustering the features to acoustic classes (e.g. determined by the vowel in the syllable) and the use of a more appropriate training algorithm [7].

REFERENCES

[1] 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 on

[2] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn: Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing 10, 19-41 (2000)

[3] Schwarz Petr, Matejka Pavel, Cernocky Jan: Hierarchical structures of neural networks for phoneme recognition, In: Proceedings of ICASSP 2006, Toulouse, FR, 2006, p. 325-328

[4] http://www.speech.kth.se/snack/

[5] The NIST Year 2006 Speaker Recognition Evaluation Plan, On: http://www.nist.gov/speech/tests/spk/2006/

[6] Burget, L. Matejka, P. Schwarz, P. Glembek, O. Cernocky, J.: Analysis of Feature Extraction and Channel Compensation in a GMM Speaker Recognition System, Audio, Speech, and Language Processing, IEEE Transactions on

[7] Dehak, N. Dumouchel, P. Kenny, P.: Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification, In: Audio, Speech, and Language Processing, IEEE Transactions on

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