Proceedings of ICASSP, 2004 (Montreal, Canada)

SPEAKER IDENTIFICATION USING SUPRA-SEGMENTAL PITCH PATTERN DYNAMICS

Farhad Farahani, Panayiotis G. Georgiou, Shrikanth S. Narayanan

Speech Analysis and Interpretation Laboratory, http://sail.usc.edu
Department of Electrical Engineering and Integrated Media Systems Center
University of Southern California, Los Angeles, CA 90089, USA
ffarahan@usc.edu, [georgiou,shri]@sipi.usc.edu

ABSTRACT

Most conventional speaker identification systems rely on short-time spectral envelope features. Recent efforts have yielded significant progress by capturing and modeling speaker-specific aspects of long-term information in the spoken language signal such as prosodic, syntactic and other conversational features. Although significant results have been reported, substantial improvements can be made by using detailed models better describing the specific behavior of each feature. In this paper we focus on modeling pitch pattern dynamics at different prosodic scale levels. Trends in pitch variation are believed to appear at different time-scales – such as microprosody, accent, phrase and discourse levels – making wavelet analysis of the $f0$ contour a suitable choice for investigating the corresponding pitch patterns. We then introduce a transform of the $f0$ contour wavelet coefficients that both results in a compact representation and better reveals the spatio-temporal details in the coefficient sequences representation. In turn, the dynamics of the transformed sequence are modeled by a first order Markov chain, at each scale level. Classification is carried out at each level and the scores of the classifiers operating at the different supra-segmental levels are fused together. The proposed method achieves an EER of 4.8% on the NIST 2001 Speaker ID Extended Data task using a 16-conversation subset, based solely on $f0$-based information.

1. INTRODUCTION

Traditional speaker recognition systems are limited to the use of frame-based spectral features that basically model the different vocal tract shapes of the speaker via GMMs [1]. Hence long-term supra-segmental features conveying information about prosody and speaking style, which are mostly due to psychological and habitual attributes of the speaker rather than physiological characteristics of the vocal tract, are ignored. Exploiting this speaker-specific information in the identification task enriches the feature set by introducing additional informative dimensions to the feature-space. This can result in improved system accuracy and allows for better separation between a large number of speakers due to the reduction of class overlap in the new higher-dimensional space.

Recent efforts [2,3] have shown impressive results by incorporating different levels of linguistic information including conversational [4], lexical [5,6], phonemic [7,8,9] and prosodic [4,10,11] features. This paper focuses on novel ways of modeling fine details at the prosodic level and in turn apply it toward speaker identification.

Previous approaches to pitch and energy pattern modeling for speaker identification fall into the two categories of static and dynamic models. The former approach uses either global statistics [12] or more low level (word, pause, etc) statistics [4], but can not address the dynamic pattern variations in the feature trajectories. The latter approach, either uses Dynamic Time Warping to compare trajectories of frequent words [10], which is ASR-dependent, or uses an $f0$ contour *stylization*, a piece-wise linear approximation, and subsequently employs either statistics [11] or bigrams [10] for modeling the slope and duration of the stylized segments. This work was an important first step towards a comprehensive dynamic model, and has resulted in an EER of 14.1% with the slope-duration system. However, length of the stylized segments are in the range of 50-300 ms and although they are supra-segmental, a first order Markov dependency (bigram) is not adequate to capture variational patterns occurring over longer analysis durations, say in the order of seconds. In this paper we introduce a new modeling scheme to overcome this limitation.

Studies on $f0$ generation models [13], have shown additive trends in feature behavior present along different linguistic levels, i.e. segmental, accent and phrase levels. Each level has the potential of conveying distinct speaker-specific information. For instance, it is believed that the decline in the pitch contour at the phrase level is related to the respiratory cycle of the speaker [14]. One motivation for the present work in deriving signal representations at multiple wavelet scales arises from drawing parallels with observations regarding linguistic information residing at multiple scales in the speech signal.

In this work, summarized in Fig. 1, we have utilized a discrete wavelet transform with a triangular scaling function that leads to a piecewise linear approximation of $f0$ along different time-scales. To efficiently model the dynamic patterns of the coefficients we also propose a transformation that results in a compact discrete representation of the coefficient sequences. The core of our classifier is a bigram in this domain, and therefore we adapt a target model bigram from the background model for each time-scale level. The overall classifier fuses the log-likelihood scores of the level-based classifiers by linear weighting to obtain the overall likelihood.

The SRI Prosody Database and the NIST 2001 Extended Data task evaluation scheme [15] were used to evaluate our algorithm's performance. The EER of the system is 4.8%, a promising result for just using (ASR-independent) prosodic features[1].

2. LEVEL-BASED FEATURE EXTRACTION

The first step in creating the classifier is representing the input sequence in an appropriate feature space. Fig. 1 summarizes the four steps in our feature extraction process, which are described below.

2.1. Data Cleaning

Although *Lognormal Tied Mixture* (LTM) filtered [16] $f0$ is available in the SRI database, we follow a simpler thresholding approach

[1]Note information at the lexical, syntactic and higher levels is obtained from an automatic speech recognizer in SID systems

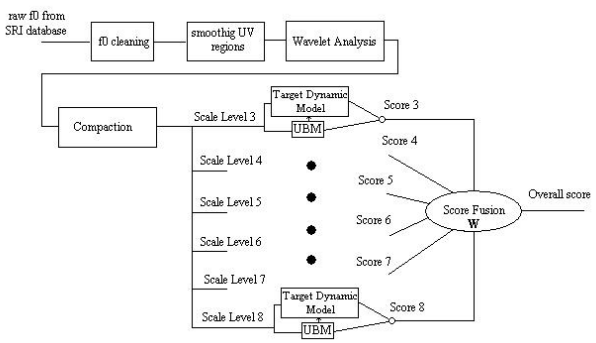

Fig. 1. System Diagram demonstrates the feature extraction process – cleaning the $f0$ contour, smoothing, analyzing using wavelet transformation, and compaction of the feature space – and the dynamic modeling process at multiple levels with classifier score fusion

to fix halving/doubling errors. First the $f0$ median over one utterance serves as the base on which halving/doubling errors are detected and corrected. After the process is iterated 5 times, the sequence is median-filtered to generate the estimate of $f0$.

2.2. Smoothing

Due to $f0$ masking in unvoiced (UV) regions we use a localized cubic polynomial fit to fill in the UV segments.

2.3. Wavelet analysis

Since wavelet transform expands a signal in terms of multiple time-shift and time-scalings of a mother wavelet it works as a suitable framework to present the temporal variations of $f0$ occurring in several time-scales. To identify additive trends in pitch variations we utilize a biorthogonal discrete wavelet transform with a triangular shape decomposition scaling function to capture the piece-wise linear variations along 8 different analysis levels. For our speaker ID task, we found that levels 3 thorough 8 provided reliable information and ignored the first two levels that tend to be noisy. Fig. 2 illustrates the cumulative sums of the composed $f0$ at different time-scales. The various underlying trends in the pitch dynamics manifest themselves in terms of the coefficient sequences of each analysis level, each corresponding to temporal variations in the time-order that microprosody, segmental, phrasal and discourse events occur.

2.4. Compaction

As can be observed from Fig. 2, the resulting wavelet coefficients tend to exhibit specific spatio-temporal behavior viz., cyclical patterns (what we term as "puffs"). For example, if a remove-near-zero global thresholding is performed, 90% of the coefficients can be set to zero with little effect on pitch. This observation suggests the interpretation of these puffs as prosodic events along different levels of linguistic importance; although exploring direct association of this multi-scale signal information, for example, to established prosodic hierarchy such as through ToBI is a topic of future work. We hypothesize that both the timing and shaping of these events can convey speaker-specific information. The key question that remains to be answered relates to issue of representation of the sequences in an efficient manner appropriate for modeling.

To explain the transformation we formulated, consider the simplified case of an unequally-spaced impulse train shown in Fig. 3. The

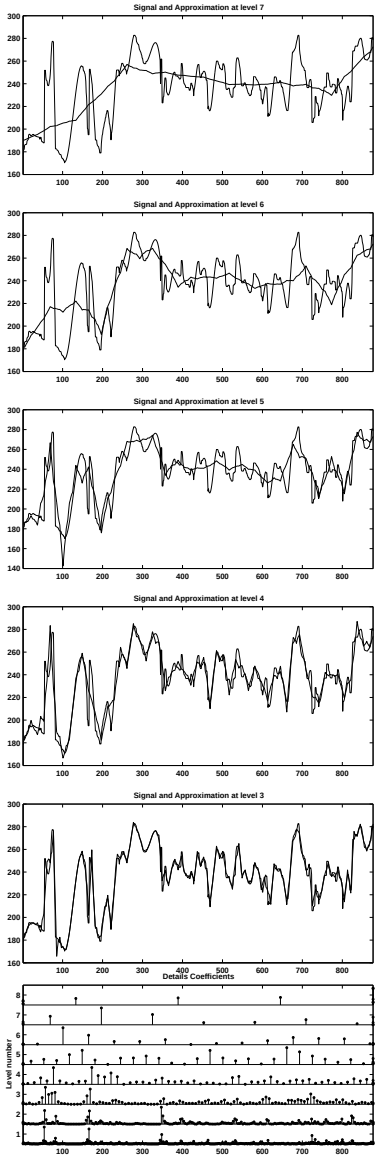

Fig. 2. The scale parameter can be viewed as the scale of a map – the higher the scale number, the lower the detail level. High scales are likely to represent utterance level trends and low scales, microprosodic details. As can be seen from this example scales 1 & 2 are noisy and likely to convey no useful information – compared to the details in scale 3, for example – and are thus omitted.

following transformation generates an impulse sequence in which the impulse magnitudes will correspond to time-intervals:

$$d(t) = \frac{d}{dt}\left[\text{inv}\left(\int |h(\tau)|d\tau\right)\right]$$

where inv stands for inverse of the function.

Although the puffs are not as impulsive, this smooth transform is also suitable for capturing their shape. A discrete version of the above is:

$$d(n) = \text{hist}\left(\sum_{m=-\infty}^{n} |h(m)|\right)$$

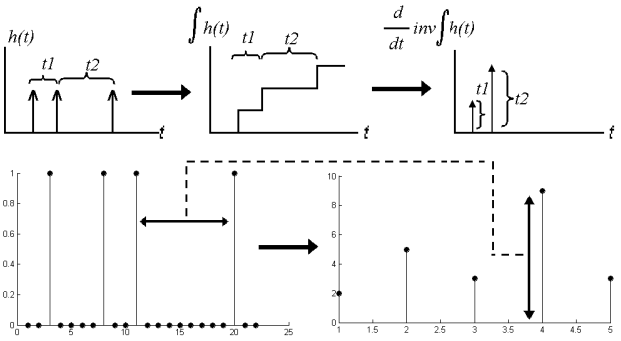

Fig. 3. Compaction procedure illustrated for the simple conceptual case of a continuous and discrete impulse train.

where hist is the binned histogram. Fig. 4 illustrates this transform on the level-3 wavelet coefficients of a sample $f0$ contour.

To determine the number of histogram bins we choose a fixed quantization step q. So:

$$\text{Number of bins} = \left\lceil \frac{1}{q} \sum_{m=-\infty}^{\infty} |h(m)| \right\rceil$$

There is a trade-off for choosing q. Smaller quantization steps lead to less temporal compaction, but more spatial preservation while larger steps have the opposite effect. Attempting to achieve a balance, we set q by trial end error to 20, 50, 80, 100, 110, 130 for levels 3-8 respectively.

3. DYNAMIC MODELING

The output of the compaction procedure is a sequence, with non-negative integer magnitudes, which are typically in the range of 0-20 at the third level and shrink to a narrower range for the higher levels. A bigram can model the probability mass density and the first-order Markov dependence of the compacted sequences. The "dictionary" size is at most 20 and thus, there is ample data to train the bigram for the universal background models (UBMs) which represent a pool of several speakers. Taking 0-20 as our *words*, 16 training conversation-halves provide a corpus comprising around twenty thousand words per speaker. The background speaker corpus will therefore be in the range of millions of words, which can help robustly estimate the model parameters. For target models, we train a bigram using back-off and discounting and then adapt the *Universal Background Model* (UBM) to it with an interpolation weight of $\lambda = 0.2$.

This procedure generates target speaker models and UBM bigram for each level and allows capturing dynamic prosodic variation patterns along the 6 different feature scales.

4. SCORE COMPUTATION

Since we have six Bayes classifiers per target speaker, we compute the log-likelihood scores of the test speaker's pitch-based parameters, with respect to each of the 6 target and background models. Because the length of the feature sequence for each level is different, each log-likelihood is normalized to the length of the respective level sequence.

Additionally, in cases when the test utterance is not long enough to have a meaningful discrete wavelet transform, especially on high time-scale levels, we ignore those levels: $\log_2(\text{length}(f0) - 10)$

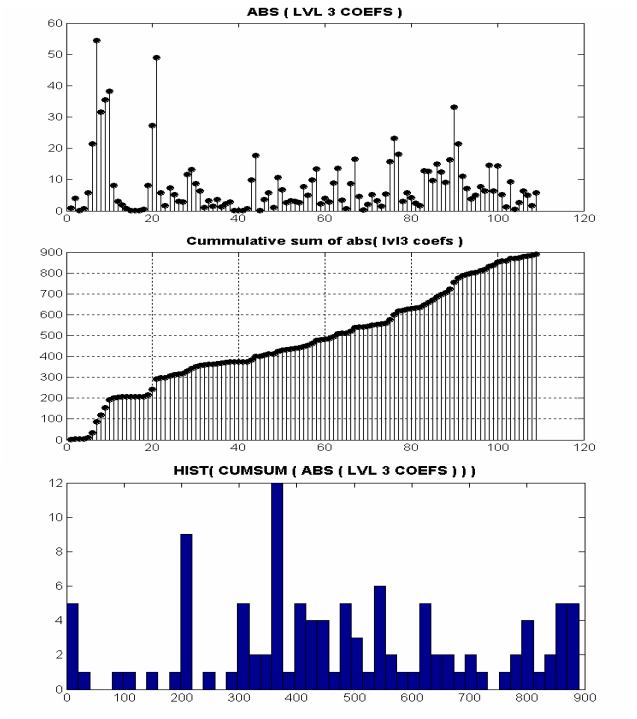

Fig. 4. An example of compaction: (top) level 3 wavelet coefficients, (middle) the cumulative sum, and (bottom) the resulting compaction, discrete both in time and magnitude.

works as a threshold for deciding this, where 10 is the length of the high-pass filter in the wavelet transform.

Once the per-classifier scores are computed and normalized to the background model score, we fuse them by using a linear weighting approach. For this work, the fusion weights were experimentally derived to be $w = [0.35,\ 0.25,\ 0.15,\ 0.1,\ 0.1,\ 0.05]$.

5. RESULTS

We used a subset of the 16-conversational NIST 2001 extended data task scheme, which uses 16 conversation halves for training and then tests on a selection of true and impostor targets. The standard NIST tests recommends the use of a whole conversation-half which is 50 utterances (180 s) in average, for testing. Under these test conditions, our method, which is solely based on the pitch contour for feature extraction, resulted in an EER (Equal Error Rate) of 4.8% as seen on the DET curve shown in Fig. 5.

We also explored the effect of decreasing the test segment duration by segmenting the test conversation-sides into sets of 20, 10, 5, and 2 utterances and re-evaluating the performance of the system. This resulted in equal error rates of 8%, 13%, 18%, and 28% respectively which indicates the expected degradation of the system when the test segment duration is decreased. Nevertheless, this is still promising and demonstrates the robustness of the system to reduction of test segment duration to the order of 10 seconds.

Although EER [17] is a popular performance measure for classifiers, and holds a close relation with recognition rate measure, we would like to point the difference between them. EER deals with pdf's of likelihood scores of true and impostor speakers and gets nor-

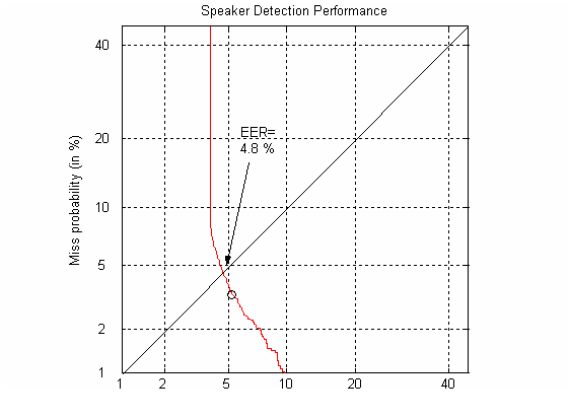

Fig. 5. DET curve and EER of the overall system that uses fused classification information: features from each of the six analysis levels are modeled by a first order Markov chain. The data are from sixteen conversation-halves from the 2001 NIST test set.

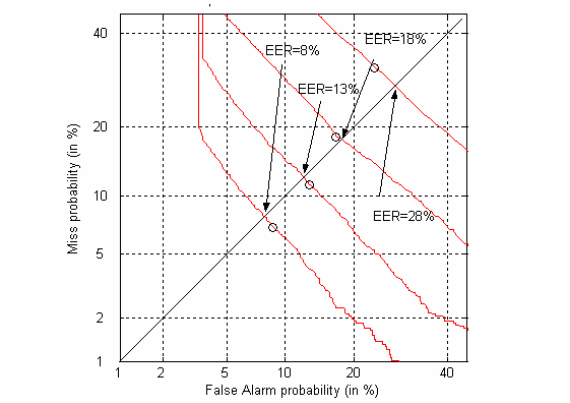

Fig. 6. DET curve and EER of the fused system for 20,10,5,2 utterance test segments.

malized by the number of classes, and hence stands for the efficiency of the classifier structure.

6. CONCLUSIONS AND FUTURE DIRECTIONS

In this paper we demonstrated a framework representing the dynamic variations of pitch, a key component among prosodic features, and utilized it in a speaker identification task. Results show, that this framework can capture the speaker-specific trends in pitch patterns along different temporal scales. Since pitch is generally considered uncorrelated from spectral envelope features, it will enrich traditional speaker recognition systems by adding new dimensions to the classification space. The results may be combined with spectral envelope features, and others, to further improve the speaker identification performance.

In future work we plan to evaluate the system on the entire set of NIST 2001 evaluation plan for 1-16-conversational tasks to determine the effect of both different training and testing conditions on system performance. Furthermore we are interested in incorporating additional features known to carry speaker information such as energy and unvoiced region duration and in optimizing the overall system parameters.

Use of supra-segmental information demands further exploration both in terms of uncovering novel features, and finding more elaborate modeling schemes appropriate for each stream. In addition to speaker identification, such prosody modeling may be employed in other applications such as automatic prosody labeling or user-state assessment, some of which we are currently investigating.

7. ACKNOWLEDGMENTS

The authors are grateful to Elizabeth Shriberg of SRI International for kindly sharing the SRI Prosody Database. The work was supported by in part by NSF and DARPA.

8. REFERENCES

[1] D. Reynolds, "Robust text-independent speaker identification using Gaussian mixture speaker models," *SpeechCommunication*, vol. 17, pp. 91–108, 1996.

[2] D. Reynolds *et al.*, "The SuperSID project: exploiting high-level information for high-accuracy speaker recognition," Proc. ICASSP, 2003.

[3] L. Ferrer, *et al.* "Modeling duration patterns for speaker recognition," *Proc. Eurospeech*, 2003.

[4] B. Peskin *et al.* "Using Prosodic and Conversational Features for High-Performance Speaker Recognition: Report from JHU WS'02," *ICASSP*, 2003.

[5] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," *Proc. Eurospeech, 2001*.

[6] F. F. Weber, L. Manganaro, B. Peskin and E. Shriberg, "Using Prosodic and Lexical Information for Speaker Identification," *Proc. ICASSP*, 2002.

[7] Q. Jin *et al.*, "Combining cross-stream and time dimensions in phonetic speaker recognition," *Proc. ICASSP*, 2003.

[8] J. Navratil, Q. Jin, W. Andrews, and J. Campbell, "Phonetic Speaker Recognition Using Maximum Likelihood Binary Decision Tree Models," *Proc ICASSP*, 2003.

[9] D. Klusacek, J. Navratil, D. Reynolds, and J. Campbell "Conditional Pronunciation Modeling In Speaker Detection," *Proc ICASSP*, 2003.

[10] A. Adami, R. Mihaescu, D. Reynolds, and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," *Proc ICASSP*, 2003.

[11] K. Sonmez, E. Shriberg, L. Heck and M. Weintraub, "Modeling Dynamic Prosodic Variation for Speaker Verification," *Proc. ICSLP*, 1998.

[12] M. Carey, E. Parris, H. Lloyd-Thomas and S. Bennett, "Robust Prosodic Features for Speaker Identification," *Proc. ICSLP*, 1996.

[13] H. Fujisaki, S. Ohno, "The use of a generative model of F0 contours for multilingual speech synthesis", *Proc. ICSP* 1998.

[14] E. Keller *et al*, *Improvements in speech synthesis*, John Wiley and Sons, England , 2002.

[15] `http://www.nist.gov/speech/tests/spk/2001`

[16] K. Sonmez, L. Heck, M. Weintraub and E. Shriberg, "A log-normal tied mixture model of pitch for prosody-based speaker recognition," *Proc. Eurospeech*, 1997.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowski and M. Przybocki, "The DET Curve In Assessment Of Detection Task Performance," *Proc. Eurospeech*, 1997.