MODELING PROSODIC DYNAMICS FOR SPEAKER RECOGNITION

Andre G. Adami[1], *Radu Mihaescu*[2], *Douglas A. Reynolds*[3], *John J. Godfrey*[4]

[1]OGI School of Science and Engineering, Oregon Health and Science University, [2]Princeton University, [3]MIT Lincoln Laboratory, [4]U.S. DoD

adami@ece.ogi.edu, mihaescu@princeton.edu, dar@ll.mit.edu, godfrey@afterlife.ncsc.mil

ABSTRACT

Most current state-of-the-art automatic speaker recognition systems extract speaker-dependent features by looking at short-term spectral information. This approach ignores long-term information that can convey supra-segmental information, such as prosodics and speaking style. We propose two approaches that use the fundamental frequency and energy trajectories to capture long-term information. The first approach uses bigram models to model the dynamics of the fundamental frequency and energy trajectories for each speaker. The second approach uses the fundamental frequency trajectories of a pre-defined set of words as the speaker templates and then, using dynamic time warping, computes the distance between the templates and the words from the test message. The results presented in this work are on Switchboard I using the NIST Extended Data evaluation design. We show that these approaches can achieve an equal error rate of 3.7%, which is a 77% relative improvement over a system based on short-term pitch and energy features alone.

1. INTRODUCTION

Current speaker recognition systems are based primarily on modeling the distributions of short-term spectral features [1]. While these systems produce very good performance, they ignore many other aspects of the speech signal that convey speaker information, such as prosodic information from pitch and energy contours. However, it is clear from results in several published studies (e.g., [2, 3, 4, 5] and their references) that prosodic information can be used to effectively improve performance of and add robustness to speaker recognition systems.

Prosodic information has been applied in two main ways. In the first approach, global statistics of some prosodic-based feature are estimated and compared between two utterances. The most common example is comparing the mean and standard deviation of the fundamental frequency between enrollment and test utterances [3]. Alternatively, the prosodic feature may be appended to standard spectral-based features and used in traditional distribution modeling systems. One potential problem with this global statistics approach is that it does not adequately capture the temporal dynamic information of the prosodic feature sequence. This has been addressed in part by using statistics of feature time derivatives and dynamic features derived from segments [2]. The second approach is aimed at explicitly representing and comparing the temporal trajectory of the prosodic contours. The classic example of this approach is applying dynamic time warping (DTW) to compare the pitch contours between two utterances of the same text [6]. This approach has the advantage of potentially being able to capture idiosyncratic speaker-specific temporal dynamic events, but generally requires comparison of the same spoken text to be effective. Due to the lack of control of the spoken text, text-independent applications have generally been limited to using global statistical approaches.

In this paper, we present two new approaches that demonstrate effective ways to model and apply prosodic contours for text-independent speaker verification tasks. The first approach uses the relation between dynamics of the fundamental frequency (f0) and energy trajectories to characterize the speaker's identity. The motivation is that the dynamics of both trajectories can jointly represent certain prosodic gestures that are characteristic of a particular speaker. In addition, the dynamics can also capture the speaking style (for example, excited or monotone) of the speaker. The second approach capitalizes on the increasing accuracy of speech recognition systems on conversational speech to allow explicit template matching of the f0 contours of a predefined set of words and phrases. The motivation is to capture speaker characteristic accent and intonation information from a known set of frequently and naturally occurring words found in conversational speech. For the rest of the paper, we are going to refer to both approaches as prosodic systems.

This paper is organized as follows. In Section 2, we describe the NIST Extended Data Task and the prosodic feature database used in this paper. We then describe systems and performance for a baseline system using simple f0 and energy distributions followed by descriptions of the new approaches using f0 and energy contour dynamics and the text-constrained f0 contour matching. In Section 6, we present some fusion results that demonstrate that these new approaches are producing complementary and beneficial information to the speaker recognition task.

2. NIST EXTENDED DATA TASK

The work presented in this paper was developed as part of the SuperSID project [7] in the 2002 JHU Summer Workshop. For this project, the development focus was on the Extended Data Task from the 2001 NIST Speaker Recognition Evaluation[i]. This task was introduced to allow exploration and development of techniques that can exploit significantly more training data than traditionally used in NIST evaluations. For this task, speaker models were trained using 1,2,4,8, and 16 complete conversation halves (where a conversation half is nominally 2.5 minutes long, as opposed to only 2 minutes of training speech. A complete conversation half was used for testing. The 2001 Extended Data Task used the entire Switchboard I conversational, telephone speech corpus in a cross-validation procedure to obtain a large

[i] The 2001 NIST Speaker Recognition Evaluation website:http://www.nist.gov/speech/tests/spk/2001

0-7803-7663-3/03/$17.00 ©2003 IEEE

number target and nontarget trials for the different training conditions.

One reason for focusing on this task was the availability of the SRI prosody database [8]. The SRI database provides time-aligned word and phone transcripts in addition to a wealth of standard and unique prosodic features (f0, pauses, duration, etc.) for the Switchboard I corpus. The f0 and energy features used in this paper were obtained from the SRI prosody database.

The performance measure used to evaluate the described systems is the equal error rate (EER). It represents the system performance when the false acceptance rate (accepting an impostor) is equal to the missed detection rate (rejecting a true speaker). The system results are compared using the target models with 8 conversation halves.

3. BASELINE F0 AND ENERGY DISTRIBUTIONS

A baseline system was developed that used global distributions of energy and f0 features. For each voiced frame, a four-dimensional feature vector consisting of log f0, log energy, and their first-order derivatives estimated over a 5-frame context was created. The first two frames and the last two frames of a voiced segment are discarded to avoid discontinuities in the derivative computations. These features were used to train a likelihood ratio detector consisting of a speaker-independent universal background model (UBM) and a speaker-dependent target speaker model [1]. The UBM is a 512-component Gaussian Mixture Model trained with gender-balanced speech from cross-validation partitions not under test[ii]. The target speaker models are derived by adapting the UBM with the speaker's data. In the testing phase, a likelihood ratio score is obtained as the ratio between the target speaker model and the UBM likelihood scores given a test message. The EER of the baseline for the 8-conversation training condition is 16.3%.

4. MODELING F0 AND ENERGY CONTOUR DYNAMICS

With prosody, as with other aspects of spoken language, speaker information may be found in both static and dynamic forms and may originate from anatomical, physiological, or behavioral differences among individuals. The baseline system experiment described in Section 3 shows that even static and short-term dynamic features, like the statistics of each talker's dynamic range of fundamental frequency and intensity, can be effective to a certain extent. But what we set out to model, if only in a general way, were the dynamics of common local prosodic events called "pitch accents" in English.

Typically, a pitch accent is associated with a lexically stressed syllable, has a time scale in the 100-500 msec range, and is realized as a (usually upward) obtrusion of f0, which then returns toward a global and slowly descending value. Major differences in the shape of these humps in the pitch contour may be associated with, among other things, the amount of emphasis on a word or phrase or its position in an utterance. For example, greater emphasis increases the height and duration of the f0 obtrusion, while a prepausal accent typically has a shorter rise but a longer and deeper fall of f0 and of energy, as well. A full description of pitch accents is beyond our scope here; we simply note that they are likely sources of interspeaker information because they vary greatly in details of execution [9,10,11].

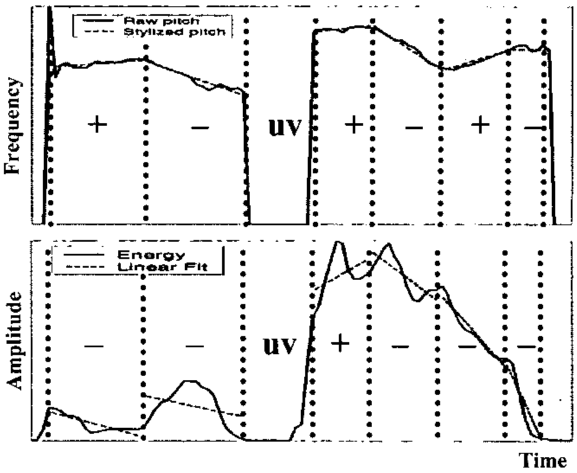

Figure 1 – Example of state symbol sequence estimation for f0 and energy contours.

For example, f0 may be raised by increasing vocal fold tension, by increasing subglottal pressure, or by a combination of the two. The pattern of combination, in turn, may be consistent for a speaker, but different across speakers. The exact shape of pitch accents also varies widely, notably the ratio of rise and fall durations, especially when viewed as a function of phonetic context. The role of duration in accented syllables resembles that of pitch and/or energy and can either combine with or complement them. To emphasize a word, one speaker may use lengthening more, another less; one may use the increased duration to carry out a larger rise in f0 and another not.

We, therefore, decided to systematically explore individual variation in the use of duration, f0 and acoustic intensity (as a proxy for subglottal pressure) to accomplish common prosodic gestures, in the hope that it might be consistent enough to aid in recognizing speakers.

For determining the rising/falling state of a contour, we used the piecewise linear model of the f0 contour supplied as part of the SRI prosody database [2]. This stylized f0 track is a series of linear components fit to the f0 trajectory in a voiced region. An example of this stylized f0 track is shown as the thin dashed-line in the upper plot of Figure 1, where we see a voiced region has multiple segments with boundaries defined by each linear segment. The sign of the line's slope is used as the state of the f0 contour over that segment (+ → rising, - → falling). The corresponding energy contour slopes are estimated for each segment found for the f0 contour (vertical thick dotted lines in Figure 1). We calculate the slope by fitting a line to each energy contour segment, as shown by the dotted lines in the lower plot in Figure 1.

For each segment, we can then combine the sign of the f0 and energy slopes into a symbol reflecting their joint state (i.e. ++, +–, –+, —). Unvoiced regions are given a single 'uv' symbol. Thus in Figure 1, the resultant symbol sequence for these contours is: +– — uv ++ — +– —.

The state symbol estimates are performed only for the speech regions as indicated in the ASR transcript. Nonspeech regions, like noise and breaths, are discarded. To avoid the

[ii] Two UBMs were used. A UBM trained on partitions 4-6 was used for testing partitions 1-3 and a UBM from partitions 1-3 was used for testing partitions 4-6.

modeling of slope dynamics across speaker turns, we place <start> and <end> symbols around each turn as indicated in the SRI database.

For modeling, we use a simple bigram model of the symbol sequence, similar to the one used in [12]. A likelihood ratio detector is built using a speaker-dependent target bigram model and a speaker-independent UBM bigram model.

The EER for this system is 19.2%. This result is very promising when compared to the baseline since it only uses five joint contour states ('++', '+–', '–+', '—', and 'uv') with no absolute f0 or energy values.

4.1. Dynamics Duration

Besides the direction of the contour, the duration of the segment can be also integrated in the symbol sequence. The duration can provide a better characterization of the speaking style of the speaker; i.e., for how long the speaker maintains a certain dynamic configuration.

Since we are using n-grams to model the sequence, we quantized the segment durations into 3 levels: Short, Medium, and Long. These quantization levels are set as the 33rd and 67th percentiles of the cumulative distribution of segment durations from held-out data. The quantization boundaries used are 4 and 8 frames. Thus, each segment symbol is now augmented with an additional duration tag (e.g., +–L —L uvM ++M – –M +–M —S).

Using duration tags reduces the EER to 14.1%, a relative improvement of 26% over the system that only uses f0 and energy contour slopes. This shows that the duration of the dynamics is useful for characterizing speaking style.

4.2. Phone Context

Additional information may be added to the contour dynamic by conditioning them on the phone context in which they occur. This is easily done in our system by simply augmenting the f0/energy state symbol with the label of the phone in which they occur (aligned phone labels are available in the SRI database; experiments in this paper used those from truth transcript alignments). State symbols that span more than one phone are broken into 2 or more symbols. To add duration information, phone-dependent duration quantization levels are estimated and applied.

With both phone context and duration used, the EER further decreases to 5.2%. Note that phone sequence modeling by itself is a known technique for speaker recognition [13]. However, the EER when only using the phone information is 10.8%, indicating that the extra f0/energy contour information is, indeed, adding new information.

4.3. Training Data Requirement

Speaker recognition systems that use prosodic features are known for requiring large amounts of data for training [2,4,12]. Table 1 shows the performance for the baseline (GMM-Pitch), slope and duration, and slope with phone context systems when training with 1, 2, 4, and 8 conversations. It is clear that the slope-duration system requires more than 1 conversation half (~2 minutes) for training to outperform the baseline system. It also shows that, when the phone context is added to the dynamics, the prosodic system performs better than the baseline no matter the number of training conversations.

Table 1 – Systems performance (EER) per number of training conversations

# Training conversations	Baseline	Slope-Duration	Slope-Phone-Duration
1	20.3%	22.2%	18.9%
2	18.3%	16.9%	10.8%
4	16.8%	15.1%	7.4%
8	16.3%	14.1%	5.2%

5. DYNAMIC TIME WARPING ON WORD F0 CONTOURS

In the second approach for comparing prosodic dynamics, we use the output of a speech recognition system to allow application of classic text-dependent f0 contour template matching to text-independent speech. In text-dependent f0 contour matching, the f0 contour from an enrollment phrase is used as the reference template for a speaker that is matched, using dynamic time warping, to the f0 contour of the same phrase from an unknown speaker. To apply DTW, it is important that the same phrase be used for both the template and the reference contours. For this text-independent application, the time-aligned text transcription from the SRI speech recognition system is used to provide the needed word knowledge for selecting and comparing common words or phrases. A similar approach was used for text-constrained Gaussian mixture modeling [14].

The selection of words to use was driven by two criteria. First, we wanted to select words that occur frequently enough so that they are likely to appear in conversational speech used for training or testing. But, we did not want to have too many occurrences, since this will increase the number of DTW matches required for testing, thus increasing the computational cost. Second, we wanted to select words that were likely to have very low dependency on context or topic and, thus, contain information with low intraspeaker variability. Based on these criteria, we looked for words and phrases among the frequent back-channel words and discourse markers. The following set of 15 words and phrases was selected: {*right, okay, well, uhhuh, true, really, like, sure, yeah, absolutely, I mean, I know, you know, I think, I see*}. In the Switchboard I corpus, these words account for roughly 5% of all word tokens and collectively accounted for roughly 30 words per conversation half.

The speaker model consists of the f0 contours from each occurrence of each word in the list found in the speaker's training data, supplying multiple templates per word. We used the median-filtered f0 values from the SRI database, since they were found to work better than the raw f0 values. A set of 20 male and 20 female models obtained from held-out data are used for cohort background scoring. The DTW algorithm employed used the absolute difference between f0 values of aligned frames, as the distance function, subject to constraints on the number of frames that can be aligned to the same frame. Matches that failed the path constraints were discarded. Unvoiced frames were given a f0 value of 0, which helped bias against voiced to unvoiced matches in the DTW. The final raw score from the DTW is the log of the sum of the frame-wise distances normalized by the number of frames in the reference and test sequences.

In verification, the male or female cohort set is first selected as the set with the lowest average distance to the test words. For each occurrence of a word/phrase in the test utterance, the average of the best 15% DTW match distances to a model's corresponding word templates is computed. The speaker's score is then normalized by subtracting the mean and dividing by the standard deviation of the cohort models' scores. The final test score is computed as the average, weighted by the number of occurrences, of the normalized scores over all test utterance words.

On the Extended Data Task for the 8-conversation condition, this system produces an EER of 13.3%. This result is quite encouraging given the small number of words used. With further tuning of system parameters, we believe the EER should reduce even further.

6. SYSTEM FUSION

Since the baseline system is modeling the absolute f0 and energy values while the slope system is modeling the relative f0 and energy contour dynamics, it is expected that a fusion of these systems should produce better performance than the individual systems. In Table 2, we show the results of fusing the various systems using a single layer perceptron.

Table 2 – Performance of the fused systems

Fused Systems (individual EER)		EER
1	Baseline GMM f0/energy (16.3%) + F0 Contour-DTW (13.3%)	11.4%
2	Baseline GMM f0/energy (16.3%) + Slope-Duration (14.1%)	9.2%
3	Baseline GMM f0/energy (16.3%) + Slope-Phone-Duration (5.2%)	4.1%
4	Baseline GMM f0/energy (16.3%) + F0 Contour-DTW (13.3%) + Slope-Phone-Duration (5.2%)	3.7%

Rows 1, 2 and 3 in Table 2 show that fusion of the global distribution baseline system with the dynamic modeling systems provides improvements. The template matching system appears to have less complementary information than the slope dynamics since it too uses the static f0 values in contour matching. The fusion between all systems (row 4) further decreases the EER to 3.7%.

The best performance achieved on this database is 0.7% EER using a GMM/UBM approach on mel-cepstrum coefficients [7]. The performance of fusing the prosodic approaches with the GMM/UBM over mel-cepstra is 0.3% EER (55% relative improvement over the GMM/UBM approach). This result additionally indicates that the prosodic approaches have complementary information to standard spectral information.

7. CONCLUSIONS

The proposed approaches are shown to capture speaker characteristics through the modeling of the f0 and energy contours dynamics. Moreover, the modeling of such dynamics works better than their global static distributions. Another advantage of these prosodic features is that, because of the dynamics quantization, these features are expected to be more robust to errors in the f0 and energy estimation. However, the sparsity of these prosodic features requires a considerable amount of training data. In our experiments, we observed that the f0/energy slope system requires more than 1 conversation half to outperform the baseline system.

We also showed that the prosodic approaches improve the performance of systems that use short-term information. The fusion of the prosodic systems with a system based on short-term f0 and energy features improved the performance by 43% (relative). The addition of context information to the prosodic features improves the performance by 75% (relative). In addition, when all systems (prosodic and short-term) are fused with a system based on mel-cepstrum coefficients prosodic, the performance of the mel-cepstrum system is improved by 55% (relative).

For future work, we plan to look into different streams, like formants, that can carry more speaker characteristics. We also intend to find and study a larger set of words and phrases that carry speaker information.

8. REFERENCES

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, "Speaker Verification Using Adapted Mixture Models", Digital Signal Processing, v. 10, pp 181-202, 2000.

[2] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling Dynamic Prosodic Variation for Speaker Verification," In Proc. of ICSLP, Vol. 7, pp. 3189-3192, 1998.

[3] M. Carey, E. Parris, H. Lloyd-Thomsa, and S. Bennett, "Robust Prosodic Features for Speaker Identification," In Proc. of ICSLP, Vol. 3, pp. 1800-1803, 1996.

[4] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, "Using Prosodic and Lexical Information for Speaker Identification," In Proc. of ICASSP, Vol. 1, pp. 141-144, 2002.

[5] K. Bartkova, D. Le-Gac, D. Charlet, and D. Jouvet, "Prosodic Parameter for Speaker Identification," In Proc. of ICSLP, pp. 1197-1200, 2002.

[6] B. Atal, "Automatic Speaker Recognition Based on Pitch Contours," JASA, Vol. 52, pp. 1687-1697, 1972.

[7] SuperSID webpage:
http://www.clsp.jhu.edu/ws2002/groups/supersid.

[8] E. Shriberg, A. Stolcke, D Hakkani-Tur, and G. Tur, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," Speech Communication, Vol. 32, No. 1-2, pp. 127-154, 2000.

[9] J. J. Godfrey and J. N. Brodsky, "Acoustic Correlates of Emphasis," JASA 80, S49(A), 1986.

[10] R. Silipo and S. Greenberg, "Automatic Transcription of Prosodic Stress for Spontaneous English Discourse," In Proc. of ICPhS, San Francisco, August 1999, pp. 3:2351-2354.

[11] R. Silipo and S. Greenberg, "Prosodic stress revisited: Reassessing the role of fundamental frequency," NIST Speech Transcription Workshop, College Park, MD, May 16-19, 2000.

[12] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," In Proc. of EUROSPEECH, September, Vol. 4, p. 2521-2524, 2001.

[13] W. Andrews, M. Kohler, J. Campbell, et al., "Gender-Dependent Phonetic Refraction for Speaker Recognition," In Proc. of ICASSP, Vol. 1, p. 149-152, 2002.

[14] D.E. Sturim, D.A. Reynolds, R.B. Dunn, and T.F. Quatieri, "Speaker Verification using Text-Constrained Gaussian Mixture Models," In Proc. of ICASSP, May 2002, Vol. 1, p. 677-680.