PROSODIC MODELING FOR SPEAKER RECOGNITION BASED ON SUB-BAND ENERGY TEMPORAL TRAJECTORIES

André G. Adami

University of Caxias do Sul - Department of Informatics
Rua Francisco Getulio Vargas, 1130 - Caxias do Sul, RS - 95001-970 - Brazil
agadami@ucs.br

ABSTRACT

Recent work has proposed the use of a discrete representation of the dynamics of the fundamental frequency and short-term energy temporal trajectories to characterize speaker and/or language information. Since the short-term energy trajectory is affected by several factors, like speaker, phone, and channel information, we propose the use of the temporal trajectories from frequency bands instead of the short-term energy in the speaker modeling. This approach allows us to use only the relevant information (i.e., speaker and phone) and discard the irrelevant information (i.e., channel). The proposed approach is evaluated on the 2001 and 2003 NIST Extended-data speaker detection tasks. We show that the proposed approach can achieve 12% relative improvement in performance over the approach using the short-term energy trajectory.

1. INTRODUCTION

In previous work [1-3], we characterized speaker/language information by modeling the dynamics of temporal trajectories of fundamental frequency (F0) and short-term energy. The idea of this approach is to capture intonation/rhythm patterns produced by a given speaker or language. However, the short-term energy temporal trajectory can be affected by several factors (e.g., environmental and channel) that do not contribute to the speaker characterization. In fact, several authors [4, 5] have shown that different frequency bands are affected by variations due to phoneme, environment, and channel information. For example, the region around 5-6 barks (approximately 500-600 Hz) contains the highest phone variability (voiced phones have high energy in this region and the unvoiced phones have low energy [6]), which is useful for speaker recognition [7]. Besacier et al. [5] show that the low-frequency bands (less than 600 Hz) and the high-frequency bands (more than 3000 Hz) contain more speaker-specific information than the remained bands. Therefore, we investigate the use of temporal trajectories from frequency bands with the F0 trajectory to characterize speaker information. The use of frequency bands for speaker modeling allow us: 1) to use only the bands that carry most of the relevant information (phone and speaker) and discard the irrelevant information (channel), and 2) to deal with noise conditions (environment and channel) that affect only part of the speech spectrum [8]. We describe a system that models the dynamics of the F0 and each frequency band temporal trajectories to characterize speaker information. The proposed approach is evaluated on the 2001 and 2003 NIST Speaker Recognition Evaluation – extended-data one-speaker detection.

This paper is organized as follows: Section 2 describes the 2001 and 2003 NIST Extended-data speaker detection task. In Section 3, we present the baseline system. In Section 4, we describe the proposed method and demonstrate its performance.

2. THE NIST EXTENDED-DATA SPEAKER DETECTION TASK

In 2001, NIST introduced a new speaker detection task that provides large amounts of training data: extended-data one-speaker detection task [9]. The purpose of this task is to support the exploration and development of higher-level and more complex characteristics for speaker recognition. The goal of one-speaker detection task is to determine whether a specified speaker is speaking during a speech segment. It is assumed that the speech segment has only speech from one speaker. The decision must be made based upon a test segment and a target-speaker model. In this task, the target speaker models were trained using 1, 2, 4, 8, or 16 conversation sides (approximately 2.5 minutes of speech per side). A complete conversation side was used for testing.

The data for this task comprises of conversational, telephone speech from LDC's Switchboard corpora in a cross-validation procedure to obtain a large number of trials. The extended-data one-speaker detection task in the 2001 NIST SRE [9] uses data from the Switchboard I corpus, and the 2003 NIST SRE [10] uses data from the Switchboard II corpus (phases 2 and 3). The task in the 2001 NIST SRE consists of 483 speakers with 4,105 target-speaker models and 57,470 trials for the testing

phase. In the 2003 NIST SRE, the task consists of 10,932 target-speaker models and 156,184 trials for the testing phase.

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). In this work we report only the results for 8-conversation training condition. The binomial test for differences in proportion is used to check whether the difference between the EER of the systems is statistically significant [11]. Unless specified, the level of significance is set to $\alpha = 0.05$.

3. F0 AND SHORT-TERM ENERGY BASELINE

The baseline system [2] characterizes speaker information by modeling a sequence of discrete symbols that describe the signal in terms of the dynamics (rate of change) of the F0 and short-term energy temporal trajectories. The sequence of discrete units are estimated from the speech signal as follows: 1) compute the F0 and energy temporal trajectories, 2) compute the rate of change (time derivative) for each trajectory, 3) detect the inflection points (points at the zero-crossings of the rate of change) for each trajectory, 4) segment the speech signal at the detected inflection points and at the voicing boundaries, and 5) convert the segments into a sequence of symbols by using the rate of change of both trajectories within each segment. Since there are no F0 values on unvoiced segments, such segments constitute one class. Table 1 lists the 5 possible classes used to describe the speech segments.

The duration information is also integrated in each segment class by adding an extra label representing the duration of the segment. Since the speech representation uses discrete symbols, the segment duration is quantized into "Short" and "Long". For voiced segments (classes from 1 to 4), "Short" is assigned to segments shorter than 80 ms. For unvoiced segments (class 5), "Short" is assigned to segments with duration less than 140 ms. Thus, the number of segment classes is increased to 10.

Table 1: Temporal trajectory segment classes

Class	Dynamics description
1	rising F0 and rising energy
2	rising F0 and falling energy
3	falling F0 and rising energy
4	falling F0 and falling energy
5	unvoiced segment

We use *n*-grams to build target-speaker and speaker-independent models. The speaker detection score is computed using a conventional log-likelihood ratio test [12, 13] between the target-speaker model and the speaker-independent model averaged over all *n*-gram tokens [14]. In all experiments, we used a bigram model for estimating the scores. The EER for the bigram modeling for 8-conversation training is 11.4% on the 2001 NIST SRE and 14.2% on the on the 2003 NIST SRE.

4. SUB-BAND ENERGY MODELING

The sub-band modeling uses the frequency-band energy trajectories instead of the full-band energy (i.e., short-term speech energy). The diagram of the frequency-band based speaker detection system is depicted in Figure 1. We assume that the frequency bands are independent, so that it allows us to score and combine different frequency bands. First, the frequency-localized temporal trajectories are estimated from the speech signal. The temporal trajectories are estimated from non-uniform frequency bands mapped from the speech spectrum to the 15 Bark-scale critical bands (1-Bark spacing between filters). Second, for each band, the sequence of joint-state classes is estimated from the F0 and frequency-band energy temporal trajectories, as described in Section 3. Third, the log-likelihood ratio between the target-speaker model and the speaker independent model is estimated for each frequency-band. Then, the fusion module selects and fuses the likelihood scores from the frequency bands. The likelihood score for

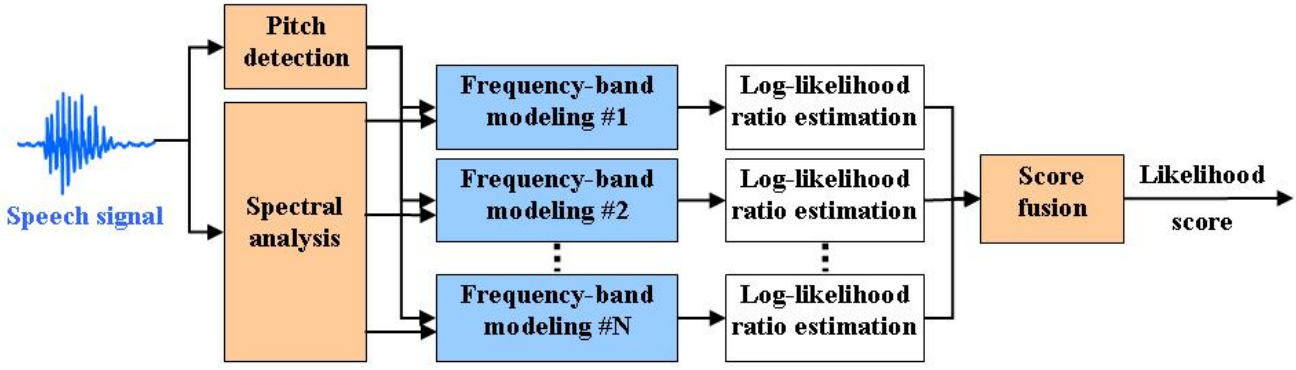

Figure 1: Frequency-band based speaker detection system.

each trial is estimated by averaging the likelihood scores from the selected frequency bands.

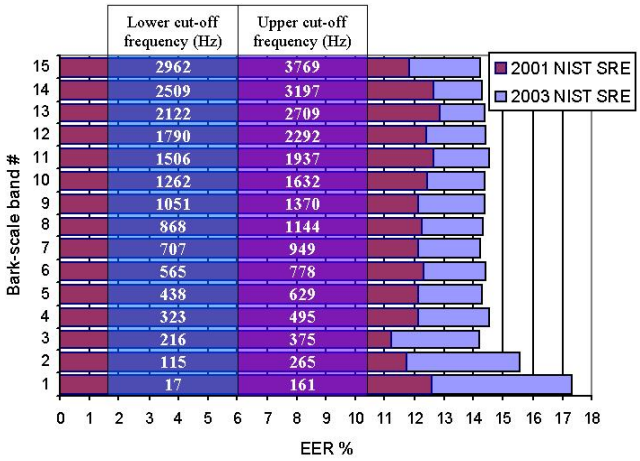

Figure 2: Performance for each bark-scale critical-band on the 2001 and 2003 NIST SREs when used to estimate the joint-state classes plus duration. The lower and upper cut-off frequencies are shown for each critical-band.

Figure 2 shows the performance for each frequency band (and their respective lower- and upper-cut-off frequencies) on the 2001 and 2003 NIST SREs. The performance of the first critical-band (12.6% on 2001 NIST SRE and 17.4% on 2003 NIST SRE) is significantly worse than the respective baseline for both tasks. The reduced performance of the first two critical-band performance is expected because we are dealing with narrow-band telephone speech (300-3400 Hz) [5], and because channel variability is higher in lower bands whereas the speaker variability is higher in higher bands [4]. The performance of the second critical-band on the 2003 NIST SRE (15.5% on 2003 NIST SRE) is also worse than the baseline. One reason is that the evaluation data of the 2003 NIST SRE has 50% of the target trials (the hypothesized and target speakers are the same) with matched handset (i.e., target trial has the phone number of the test conversation occurring at least once in the speaker model training data). The evaluation data of the 2001 NIST SRE has about 91% of the target trials with matched handset.

Even though most of the energy concentrates around the low-frequency bands, the performance for high-frequency bands is very similar to the low frequency bands. This result follows the findings that high frequency bands play an important role in speaker recognition [15-18]. Lavner et al. [19] show that the shifting F3 and F4 formant frequencies of vowels affect more the identification rate than shifting F1 and F2 formant frequencies. Lavner's result allows us to speculate that our modeling of the high frequency bands might be capturing some relationship between pitch and the phone formant frequencies.

4.1. Frequency-band Fusion

Since the performances of the individual frequency bands are similar to the performance obtained from the full-band energy, we run several experiments that fuse different combinations of frequency bands. Figure 3 shows the performances of some frequency-band fusions on both NIST SREs.

The first fusion experiment combines the detection scores from all 15 frequency bands. The EER of this fusion is 10.5% on the 2001 NIST SRE and 12.5% on the 2003 NIST SRE. This fusion achieves a significant better performance than the system based on short-term energy. Even though the number of parameters in the fusion system is higher (i.e., number of bands times 10 classes) than the approach that uses the short-term energy contour, the fusion of the frequency-band scores allows the bands that carry more speaker-dependent information to provide sufficient reliable information to the decision process.

The fusion of the upper-half of bark-scale bands (from 8th to 15th bands) performs significantly worse than the lower-half fusion. The main reason is that most of the energy of voiced phones concentrates in the region around 500-600 Hz [6]. Such region has shown to be important for

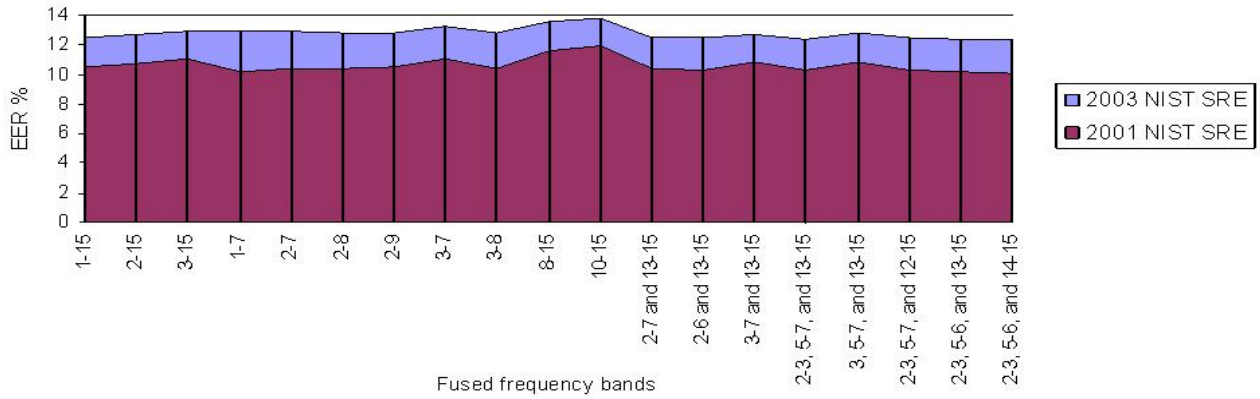

Figure 3: Performance of the frequency-band trajectories fusion on the 2001 and 2003 NIST SREs.

speaker recognition [4].

The best performance is achieved by the fusion of the 2nd, 3rd, 5th, 6th, 14th, and 15th critical-bands. The performance for both databases yields a 12% relative improvement over the full-band energy based modeling (EER=10% for the 2001 NIST SRE and EER=12.4% for the 2003 NIST SRE). Even though there is no significant improvement over the fusion of all 15 bands, this fusion uses only 6 frequency bands. This result is very similar to the findings in a speaker identification experiment on TIMIT database (clean, telephone speech) done by Besacier et al. [5].

5. CONCLUSIONS

We presented a new multi-band based approach to characterize speaker information. This approach extends the concept of modeling the dynamics of two different streams by replacing the short-term energy by frequency-band energy temporal trajectories. The motivation of this approach is that different frequency bands are affected differently by phone, speaker, and channel information. Besides, the independence between frequency bands provides a more robust approach to channel effects.

The modeling of the dynamics of the temporal trajectories of F0 and a set of some frequency bands provides a 12% relative improvement over the full-band energy based modeling. This result shows that the dynamics of some frequency bands and F0 can characterize speaker information. Note that the best performance is achieved by using only 6 frequency bands (i.e., 4 bands below 600 Hz and 2 bands above 2500 Hz), which have been long acknowledged to carry more speaker-specific information than the remainder frequencies.

As future work, we plan to investigate the relationship between the dynamics of the frequency-band energy and F0 temporal trajectories and prosodic phenomena (e.g., stress and intonation). We also plan to develop a new duration quantization process that is tuned according to the frequency band being modeled.

6. REFERENCES

[1] A. Adami, R. Mihaescu, D. A. Reynolds, and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," presented at ICASSP, Hong Kong, 2003, pp. 788-791.

[2] A. Adami and H. Hermansky, "Segmentation of Speech for Speaker and Language Recognition," presented at EUROSPEECH, Geneva, Switzerland, 2003, pp. 841-844.

[3] A. Adami, "Modeling Prosodic Differences for Speaker and Language Recognition," Ph.D. thesis, ECE, OGI School of Science & Engineering at OHSU, Portland, OR, 2004.

[4] S. Kajarekar, "Analysis of Variability in Speech with Applications to Speech and Speaker Recognition," Ph.D. thesis, ECE, OGI School of Science and Engineering at OHSU, Portland, 2002.

[5] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and Selection of Speaker-specific Information with Statistical Modeling," *Speech Communication*, vol. 31, pp. 89-106, 2000.

[6] H. Hermansky and N. Malayath, "Spectral Basis Functions from Discriminant Analysis," presented at ICSLP, Sydney, Australia, 1998, pp. 1379-1382.

[7] S. Kajarekar and H. Hermansky, "Speaker Verification Based on Broad Phonetic Categories," presented at 2001: A Speaker Odyssey, Crete, Greece, 2001, pp. 201-206.

[8] S. Sharma, "Multi-Stream Approach to Robust Speech Recognition," Ph.D. thesis, ECE, Oregon Graduate Institute of Science and Technology, Portland, USA, 1999.

[9] A. Martin, (03/01/2001), "NIST 2001 Speaker Recognition Evaluation Plan," Available: http://www.nist.gov/speech/tests/spk/2001/doc.

[10] A. Martin, (02/10/2003), "NIST 2003 Speaker Recognition Evaluation Plan," Available: http://www.nist.gov/speech/tests/spk/2003/doc/2003-spkrec-evalplan-v2.2.pdf.

[11] L. Gillick and S. J. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," presented at ICASSP, Glasgow, Scotland, 1989, pp. 532-535.

[12] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, "Gender-dependent Phonetic Refraction for Speaker Recognition," presented at ICASSP, Orlando, FL, 2002, pp. 149-152.

[13] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," presented at Eurospeech, Aalborg, Denmark, 2001, pp. 2521-2524.

[14] F. Jelinek, *Statistical Methods for Speech Recognition*. Cambridge: MIT Press, 1997.

[15] I. Pollack, J. M. Picket, and W. H. Sumby, "On the Identification of Speakers by Voice," *Journal of the Acoustical Society of America*, vol. 26, pp. 403-406, 1954.

[16] A. J. Compton, "Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally," *Journal of the Acoustical Society of America*, vol. 35, pp. 1748-1752, 1963.

[17] S. Hayakawa and F. Itakura, "Text-dependent Speaker Recognition using the Information in the Higher Frequency," presented at ICASSP, Adelaide, Australia, 1994, pp. 137-140.

[18] S. Furui and M. Akagi, "Perception of Voice Individuality and Physical Correlates," *Journal of the Acoustical Society of Japan*, vol. J66-A, pp. 311-318, 1985.

[19] Y. Lavner, I. Gath, and J. Rosenhouse, "The Effects of Acoustic Modifications on the Identification of Familiar Voices Speaking Isolated Vowels," *Speech Communication*, vol. 30, pp. 9-26, 2000.