Segmentation of Speech for Speaker and Language Recognition

André G. Adami[1], Hynek Hermansky[1,2]

[1]OGI School of Science and Engineering, Oregon Health and Science University, Portland, USA
[2]International Computer Science Institute, Berkeley, California, USA
{adami, hynek}@asp.ogi.edu

Abstract

Current Automatic Speech Recognition systems convert the speech signal into a sequence of discrete units, such as phonemes, and then apply statistical methods on the units to produce the linguistic message. Similar methodology has also been applied to recognize speaker and language, except that the output of the system can be the speaker or language information. Therefore, we propose the use of temporal trajectories of fundamental frequency and short-term energy to segment and label the speech signal into a small set of discrete units that can be used to characterize speaker and/or language. The proposed approach is evaluated using the NIST Extended Data Speaker Detection task and the NIST Language Identification task.

1. Introduction

Many sources of information besides the linguistic message are imprinted on the speech signal. Automatic speech recognition (ASR) aims at getting the linguistic message from speech by describing the signal in terms of discrete units such as words. However, one may also attempt to extract other information from speech, like who is speaking or which language is being spoken. For such applications, the underlying concept of words formed by phonemes (implied in most Large-Vocabulary Connected Speech Recognition (LVCSR) systems) may not be necessary – it is sufficient to consistently convert the continuous acoustic speech signal into a string of discrete labeled units. Along these lines, new approaches for speaker and language recognition, based on simple speaker-specific and/or language-specific models [1-5], started to emerge. Doddington in [1] uses the sequence of words extracted from the speech signal to built statistical models for speaker recognition. Andrews in [4] uses the sequence of phones to capture a speaker's pronunciation. Torres-Carrasquillo in [5] uses a sequence of tokens obtained from a Gaussian mixture model to model language information.

Our research contributes to this emerging direction of research. We use information in prosodic cues (temporal trajectories of a short-term energy and fundamental frequency – f0), as well as coarse phonetic information (broad-phonetic categories - BFC), to segment and label the speech signal into a relatively small number of classes (i.e. significantly less that the context-dependent phonemes of the current LVCSR). We also demonstrate that such strings of labeled sub-word units can be used for building statistical models that can contribute for characterizing speakers and/or languages.

This paper is organized as follows: Section 2 describes techniques for segmentation of the speech signal. In Section 3 and 4, we describe the NIST Language Identification task and the NIST Extended-data Speaker Recognition task. Then, we describe applied systems and demonstrate the performance of the proposed approach in speaker and language identification.

2. Speech Segmentation

Different speakers and different languages may be characterized by different intonation or rhythm patterns produced by the changes in pitch and in sub-glottal pressure, as well as by different sounds of language. Therefore, the combination of pitch, sub-glottal pressure, and duration that characterizes particular prosodic "gestures", together with some additional coarse description of used speech sounds, should be useful in extracting speaker [2, 6] and language information [7, 8]. Thus, converting the continuous speech signal into a sequence of discrete units that describe the signal in terms of dynamics of the f0 temporal trajectory (as a proxy for pitch), the dynamics of short-term energy temporal trajectory (as a proxy for subglottal pressure), and possibly also the produced speech sounds, could be used in for building models that that may characterize given speaker and/or language.

The speech segmentation is divided into 5 steps: 1) compute the f0 and energy temporal trajectories, 2) compute the rate of change 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 starts or ends, and 5) convert the segments into a sequence of symbols by using the rate of change of both trajectory within each segment. Such segmentation is performed over an utterance, which is a period of time when one speaker is speaking.

The rate-of-change of f0 and energy temporal trajectories is estimated using their time derivatives. The time derivatives are estimated by fitting a straight line to several consecutive analysis frames (the method often used for estimation of so called "delta features" in ASR).

The utterance is segmented at inflection points of the temporal trajectories or at the start or end of voicing. First, we detect the inflection points for each trajectory at the zero-crossings of the derivative, as shown by the filled circles in Figure 1. Second, we segment the utterance using the inflection points from both time contours and the start and end of voicing. Finally, each segment is converted into a set of classes that describes the joint-dynamics of both temporal trajectories. Since there are no f0 values on unvoiced regions, the unvoiced segments constitute one class. Table 1 lists the 5 possible classes used to describe the speech segments.

We can also integrate the duration information in each segment class by adding an extra label with the duration information. Since we are using tokens to built models, the segment classes are further split into "Short" and "Long". For voiced regions, Short is assigned to segments shorter than 8 frames (80 ms). For unvoiced regions, Short is assigned to segments less than 14 frames (140 ms). In this way, we increased the number of segment classes to 10.

Table 1: Set of segment classes used to describe the f0 and energy temporal trajectories

Class	Temporal Trajectory 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

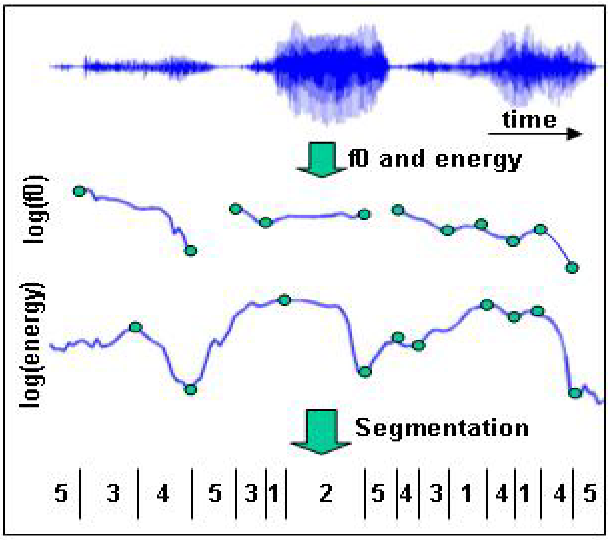

Figure 1: Speech segmentation example extracted from a real conversation. The filled circles show the segmentation boundaries. At the bottom, the vertical bars represent the segmentation boundaries and the number represents the segment classes.

In some experiments, we also investigate the use of broad phonetic category (BFC) segments. First, we segment the speech signal using the broad-phonetic category information. Then, we combine the segments from the f0 and energy temporal trajectories and the broad-phonetic categories to create new segments. Each new segment is labeled using the previous 10 classes plus the broad-phonetic category. Thus, each segment can be labeled using 60 classes (6 broad-phonetic categories times 10 segment classes from f0 and energy trajectories segmentation). The classifier used to obtain the broad-phonetic categories is described in Section 2.1.

Since the broad phonetic classes, f0 and energy trajectories are used to segment and label the speech signal, the identity and duration of each segment can reflect the speaking style of a given speaker [6] or intonation patterns of a given language [8]. Discrete hidden Markov models, binary trees, and n-gram grammars are some of the methods that can be used to build models for generation of these segment sequences. Such models could then describe given speaker and/or language. In this work, we use n-gram models for this purpose, often used in LVCSR. The n-gram model is very similar to the one described by Doddington in [1]. A likelihood ratio detector can be then used to recognize the unknown speaker and/or language, using a speaker- or language-dependent model and a speaker- or language-independent model estimated from a held-out data.

2.1. Broad-phonetic Category Segmentation

TRAP (TempoRAl Pattern) classifier [9] is applied to obtain per frame posterior probability of six broad-phonetic categories (vowels+diphtongs+glides, schwa, stops, nasal, fricatives, and flaps) events. The TRAPS classifier focuses on the temporal characteristics of the speech signal rather than the spectral characteristics. It uses a collection of multilayer perceptrons (MLPs) to estimate class posteriors from features computed from individual critical bands. Such posteriors are subsequently combined (using another "merging" MLP) to produce a global estimate of the posterior probabilities. The input features are one-second temporal trajectories from Bark-scale critical band energies projected into a 15-dimension feature vector using Discrete Cosine Transform. The critical-band MLPs have 100 hidden units (sigmoid activation function) and 7 output units (softmax activation function). The broad-phonetic category label that corresponds to the MLP output with highest posterior is assigned to the current frame. The band-classifier MLPs and the merging MLP are trained, using back-propagation with a cross-entropy error criterion, on the training part of LDC'S NTIMIT Corpus (clean speech passed through a telephone channel).

3. Language Identification

The goal of the NIST Language Identification task is the detection of a given language, i.e., to determine whether or not a test segment of speech is from the target language. The data, from Language Data Consortium's (LDC) CallFriend corpus, is a collection of unscripted conversations for 12 languages recorded over digital telephone. The languages are: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. The test segments can last nominally 3 seconds, 10 seconds, and 30 seconds. The performance measure used for this task is the equal error rate (EER). It represents the system performance when the false acceptance probability (detecting the wrong language for a given test segment) is equal to the missed detection probability (rejecting the correct language). The error probabilities are plotted as detection error tradeoff (DET) curves to show the system performances. The results are compared using the 30-second test segments.

We use a trigram model to describe each language. The language-independent trigram model is estimated from the training part of the CallFriend corpus.

The baseline system uses English phones as segment symbols. Using a TRAP-based phoneme recognizer, the speech signal is converted into a sequence of tokens (38 English phones and silence). The recognizer is similar to the one described in Section 2.1., except that the MLPs have 300 hidden units and Viterbi search is applied to the merging MLP output to find the best sequence of phonemes. The EER for a trigram modeling of phones is 24%. Figure 2 shows the performance of the systems using the English phones and the proposed segment classes.

The time derivatives for f0 and energy temporal trajectories are estimated by fitting straight line to 10 analysis frames (100 ms). The f0 and short-term energy values are estimated every 10 ms. The f0 values are estimated from the signal using the normalized cross-correlation function and dynamic programming [10]. The EER for a trigram modeling of the segment classes derived from the f0 and energy trajectories is 35%. Even though this performance is worse than the performance of the phone-based system, it is still an encouraging result, since each language is represented by a sequence of only 5 possible symbols, whereas the phone-based system uses 39 symbols.

The EER for the addition of duration information to the segments represents 30%, which represents a 15% relative improvement over the slope features. The segment duration provides some additional relevant information to the segment classes that can characterize language information.

The EER for the integration of broad phoneme class information to the prosody derived segments is 27%, which represents a 10% relative improvement over the segment classes with duration alone. The same improvement is obtained when this system is merged with the phone-based system – the EER is reduced from 24% to 21.7%. This shows that the segment classes derived from the f0 and energy trajectories and broad-phonetic categories are capturing complementary information with respect to the sequence of phones.

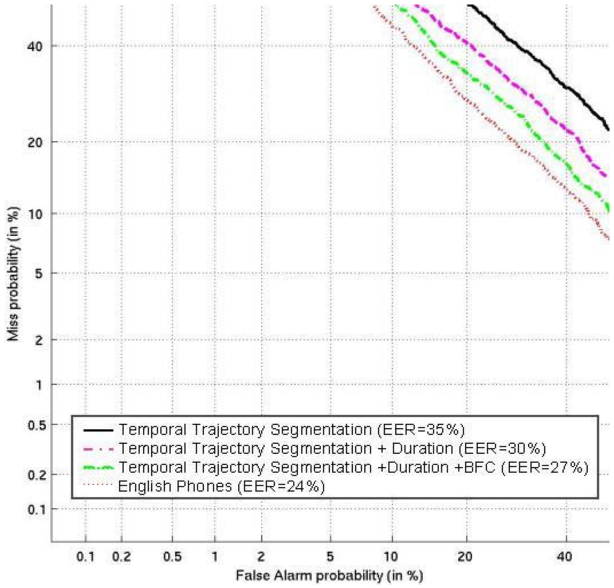

Figure 2: Language identification performance of the phone-based and proposed systems.

The analysis of the performance for each language shows that segment classes derived from the f0 and energy trajectories perform better than the phone-based approach for some languages. Table 2 shows the performances for English, Mandarin, and German languages and the fusion between the systems. The English language is better described by the phone sequence than by the proposed segment class sequence, which suggests that the proposed segment classes do not provide sufficient information to identify the language. However, Mandarin language, known by its tonal characteristic, is better characterized by the prosody information captured through the segment classes extracted from the f0 and energy trajectories. The performances for the German language show that both approaches can be used to model the language. Besides, the fusion of both systems yields a better performance.

Table 2: Performance (in % EER) comparison between the phone-based and the segment classes with duration systems. Third columns show the performance of the systems fusion.

Language	Segment classes with duration (10 classes)	Phone	Fusion
English	27.5%	17.5%	15.0%
Mandarin	23.8%	26.3%	22.5%
German	21.3%	20.0%	17.5%

4. Speaker Recognition

The goal of the NIST Extended-data Speaker Recognition [11] is, given a conversation side, to determine if the test speaker is the same as the model speaker. The data for this task comes from LDC's Switchboard I conversational, telephone speech corpus in a cross-validation procedure to obtain a large number target and non-target trials for the different training conditions. The target speaker models can be trained using 1, 2, 4, 8, or 16 conversation sides (approximately 2.5 minutes of speech per side). The performance measure used for this task is also the EER. In this task, the false acceptance is the probability of accepting an impostor, and the missed detection is the probability of rejecting a true speaker. The system results are compared using the target models with 8 conversation sides.

Since this task provides less data than the language identification task, we use a bigram model to describe each speaker. The speaker-independent bigram model is estimated from Switchboard I corpus.

The baseline system uses a Gaussian Mixture Model to model the distribution of the four features streams (f0, energy, and their time derivatives). The target models are adapted from a 512-component Universal Background Model trained on a held-out data. The EER of the baseline is 16.3%. Figure 3 shows the DET curves for the baseline and the proposed segment classes described in Section 2.

The time derivatives are estimated over 5 analysis frames (50 ms). Informal experiments indicated that a longer time interval for timer derivative estimation (such as the 10-frame interval applied in language identification system) is not beneficial in this case. The EER for the segment classes derived from the f0 and energy trajectories is 14.2%, a 13% relative improvement over the baseline. This approach uses only the information from the delta parameters, comparing to the baseline, which uses f0 and energy values beside the deltas. There is no significant difference between the performances of both systems for models trained with 1 conversation side, which is probably caused by the sparsity of the available data.

When we add the duration information for each segment, the EER of the bigram modeling of such symbols is 11.4%, a 30% relative improvement over the baseline, and 20% relative improvement over the segment classes. This shows that the segment duration conveys speaker information.

In a cheating experiment, the broad-phonetic categories are obtained by a canonical mapping from the phone labels available in the SRI prosody database [12]. The phone sequences in that database were obtained by force-aligning the signal to a canonical pronunciation of the spoken word. The EER for the system that uses the broad-phonetic categories and segmentation from the f0 and energy temporal trajectories is 8.3%, a 49% relative improvement over the baseline and 27% relative improvement over the system that does not uses the BFC.

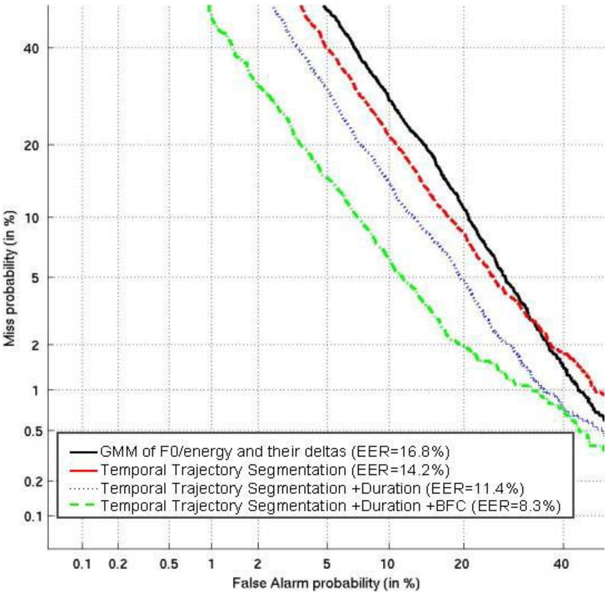

Figure 3: Speaker recognition performance of the baseline and proposed systems.

Since the proposed segmentation captures the variation in temporal trajectories and the baseline models the characteristic f0 and energy of the speaker, it is expected that both systems could yield complementary information. Using a single hidden layer perceptron [2] with 2 neurons to fuse the scores, the EER of the fusion between the scores of the baseline system and the bigram model of the symbol sequence derived from f0 and energy temporal trajectories is 5.6%. This shows that both systems have complementary information.

5. Conclusions

We presented a method for converting the speech signal into a sequence of symbols that can be used for characterizing speaker and/or language information. We demonstrated that these symbols could capture speaker and/or language information and that they can provide complementary information to the conventional systems.

We show that the inflection points estimated from f0 and energy temporal trajectories provide a better characterization of the speaker information. In the previous work [2], the inflection points were derived only from temporal trajectory of the f0 and the EER was 14.1%. The performance of the proposed segmentation is 11.4%, a 19% relative improvement over the previous work.

Even though the performance in the speaker detection task is better than the performance of the baseline system, the performance does not improve at the same rate as the number of training conversation decreases. It appears that the proposed segmentation system requires more training data to reliably characterize the speaker identity.

In language identification, although the overall performance is worse than the phone-based system, we show that the prosody dynamics can capture language-dependent information. The performance of the individual languages shows that the prosody dynamics may characterize some languages better than the phone-based approach. This could happen when recognizing languages that are known by their particular prosody (intonation and rhythm), such as Mandarin Chinese.

For future work, we plan to develop a different method to obtain the duration quantization that is application independent and investigate different methods to model the segment classes. We also want to incorporate the modeling of the unvoiced regions dynamics.

6. References

[1] G. Doddington, "Speaker Recognition based on Idiolectal Differences between Speakers," In Proc. of Eurospeech, Aalborg, Denmark, pp. 2521-2524, 2001.

[2] A. Adami, R. Mihaescu, D. A. Reynolds, and J. Godfrey, "Modeling Prosodic Dynamics for Speaker Recognition," In Proc. of ICASSP, Hong Kong, 2003.

[3] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, and J. Abramson, "Combining Cross-Stream and Time Dimensions in Phonetic Speaker Recognition," In Proc. of ICASSP, Hong Kong, 2003.

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

[5] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. D. Jr., "Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features," In Proc. of ICSLP, Denver, CO, 2002.

[6] F. Nolan, "Intonation in Speaker Identification: An Experiment on Pitch Alignment Features," *Forensic Linguistics*, vol. 9 (1), pp. 1-21, 2002.

[7] J. Navratil, "Spoken Language Recognition - A Step Toward Multilinguality in Speech Processing," *IEEE Transaction on Speech and Audio Processing*, vol. 9 (6), pp. 678-685, September 2001.

[8] F. Ramus and J. Mehler, "Language Identification with Suprasegmental Cues: A Study Based on Speech Resynthesis," *J. Acoust. Soc. Am.*, vol. 105 (1), pp. 512-521, 1999.

[9] S. Sharma, D. Ellis, S. Kajarekar, P. Jain, and H. Hermansky, "Feature Extraction Using Non-linear Transformation for Robust Speech Recognition on the AURORA Data-base," In Proc. of ICASSP, Istanbul, Turkey, pp. 1117-1120, 2000.

[10] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)," in *Speech Coding and Sysnthesis*, W. B. Kleijn and K. K. Paliwal, Eds. New York: Elsevier, 1995.

[11] A. Martin, "NIST 2001 Speaker Recognition Evaluation Plan," http://www.nist.gov/speech/tests/spk/2001/doc, 2001.

[12] E. Shriberg, A. Stolcke, D. Hakkani-Tur, and G. Tur, "Prosody-Based Automatic Segmentation of Speech into Sentences and Topics," *Speech Communication*, vol. 32 (1-2), pp. 127-154, 2000.