Modeling Prosodic Differences for Speaker and Language Recognition

André Gustavo Adami
B. Sc. Computer Science, University of Caxias do Sul, 1994
M. Sc. Computer Science, Federal University of Rio Grande do Sul, 1997

A dissertation submitted to the faculty of the
OGI School of Science & Engineering at
Oregon Health & Science University
in partial fulfillment of the
requirements for the degree
Doctor of Philosophy
in
Electrical Engineering

April 2004

© Copyright 2004 by André Gustavo Adami
All Rights Reserved

The dissertation "Modeling Prosodic Differences for Speaker and Language Recognition" by André Gustavo Adami has been examined and approved by the following Examination Committee:

Hynek Hermansky
Professor
Thesis Research Advisor

Jan P. H. van Santen
Director and Professor

Misha Pavel
Professor

Douglas A. Reynolds
Senior Member of Technical Staff
MIT Lincoln Laboratory

John J. Godfrey
R61, Information Discovery Sciences
National Security Agency

Dedication

To my wife and my parents

Acknowledgments

Many people have contributed to the success of this work. I especially thank my advisor, Hynek Hermansky, for his support and advice, and Jan van Santen, whose collaborative efforts have assisted with many aspects of this work. I thank Douglas Reynolds, Misha Pavel, and Jack Godfrey for their assistance and direction as members of my doctoral committee.

I want to thank the people who helped me decide to come to OGI: Prof. Ron Cole, Prof. Dante Barone, and Andrew Cronk. They played a key role in presenting the opportunities at this school and introducing Hynek Hermansky. Of course, they also told me about the weather in Portland, but still it was really a great choice coming here.

During my doctoral work, I was fortunate to learn about several topics from different professors – Hynek Hermansky, Michael Macon, Eric Wan, Robert Jaffe, Todd Leen, Paul Hosom. In particular, I would like to thank Paul Hosom and Robert Jaffe for their unconditional support and encouragement throughout my doctoral studies and work.

I was fortunate to work with great people in my group. I would like to thank my officemates Sunil, Pratibha, and Sachin for all the discussions and support throughout my research. I want to thank Pavel and Alena for their help (and patience) in computing and administrative issues. I also thank Sachin for reviewing the thesis draft. I cannot forget all the Czech guys (Lucas, Petra I, Franta, Pavel, and Petra II) who for a short period worked with me. In particular, I would like to thank Pavel for his invaluable work in developing the phone recognizer used in this work.

I would like to thank Prof. Hynek Hermansky for encouraging me to attend the summer workshop on Language Engineering at Johns Hopkins University. In the workshop, I had the opportunity of working with knowledgeable people, Joe Campbell,

Jiri Navratil, Douglas Reynolds, Barbara Peskin, Walter Andrews, George Doddington, and Jack Godfrey. The contribution of all these people was invaluable in defining the topic of my thesis.

Several people contributed to this work by providing software and data. I would like to thank Joe, Walter, Doug, Jiri, Andreas Stolcke, Kemal Sonmez, and all the people from the Lincoln labs at MIT and from the Speech Technology and Research Laboratory at SRI International, who provided part of the software and data used in this work.

I would like to acknowledge the unconditional support and love from my parents Oscar and Olga Adami, my brothers Marcos and Ricardo, my wife Adriana, and her family. I am very grateful for having a friend (and business partner) in Brazil like André Segato, who has been always there for me. I thank my colleagues at University of Caxias do Sul, especially Ribeiro, Dorneles, and Notari, for their support and friendship.

I would like to thank the financial support I received from CAPES, without which my research would not have been possible. I also would like to acknowledge Prof. Hynek Hermansky, the National Science Foundation, John Hopkins University, and OGI School of Science and Engineering for providing additional financial support throughout this work.

Contents

List of Tables

List of Figures

Abstract

Modeling Prosodic Differences for
Speaker and Language Recognition
André Gustavo Adami, B.Sc., M.Sc.

Ph.D., OGI School of Science & Engineering
at Oregon Health & Science University

April 2004

Thesis Advisor: Dr. Hynek Hermansky

Prosody plays an important role in discriminating between languages and speakers. Due to the complexity of estimating relevant prosodic information, most recognition systems rely on the notion that the statistics of the fundamental frequency (as a proxy for pitch) and speech energy (as a proxy for loudness/stress) distributions can be used to capture prosodic differences between speakers and languages. However, this simplistic notion disregards the temporal aspects and the relationship between prosodic features that determine certain phenomena, such as intonation and stress.

We propose alternative approaches that exploit the dynamics between the fundamental frequency and speech energy to capture prosodic differences. The aim is to characterize different intonation, stress, or rhythm patterns produced by the variation in the fundamental frequency and speech energy contours. In these approaches, the continuous speech signal is converted into a sequence of discrete units that describe the

signal in terms of dynamics of the fundamental frequency and speech energy contours. Using simple statistical models, we show that the statistical dependency between such discrete units can capture language- and speaker-specific information. On the extended-data task of the 2001 and 2002 NIST Speaker Recognition Evaluation, such approach achieves a relative improvement of at least 17% over a system based on the distribution statistics of fundamental frequency, speech energy and their deltas. We also show that they are robust to communication channel effects when compared to the state-of-the-art speaker recognition system.

Segmental information is incorporated to capture dependencies between segmental and prosodic information. In this approach, a new set of segment classes is estimated from the time-alignment between a sequence of phonemes or phones (i.e., segmental information) and the new prosodic information representation. We show that this approach can characterize speaker-dependent information.

Since conventional recognition systems do not fully incorporate different levels of information, we show that the performance of conventional systems is improved when the proposed approaches are incorporated by fusing the systems. In the 2003 NIST Language Recognition Evaluation, the fusion of the prosodic speech representation and a conventional system yields a relative improvement in performance of 14%. The fusion with the state-of-the-art speaker recognition system achieves a relative improvement of about 28% and 12% for the extended-data task of the 2001 and 2002 NIST Speaker Recognition Evaluation, respectively.

Chapter 1

Introduction

Speech conveys linguistic (e.g., message and language), speaker (e.g., emotional, regional, and physiological characteristics of the vocal apparatus), and environmental (e.g., where the speech was produced and transmitted) information. Even though such information is encoded in a complex form, humans can relatively decode most of it. This human ability has inspired several researchers to study the processes of speech production and perception to develop systems that would emulate such ability. Speech technology has several applications ranging from human-computer interfaces (e.g., dictation and voice commands) to information retrieval (e.g., search for a particular speaker in a radio broadcast). The type of application determines which information in the speech signal is relevant. For example, the linguistic message is relevant if the goal is to recognize the sequence of words that the speaker is producing. The presence of irrelevant information (like speaker or environment information in the previous example) may actually degrade the system accuracy. In this thesis, we deal with automatic systems that recognize who is speaking and the language that is being spoken.

Automatic speaker recognition is the process of recognizing a person's identity from his or her voice [1-3]. Speaker recognition technology makes it possible to use a person's voice to control the access to restricted services (e.g., automatic banking), information (e.g., telephone access to financial transactions), or areas (government or research facilities). In speech-based applications such as information retrieval, recognition of a suspect on a telephone tap, and detection of a speaker in a multi-party

dialogue, speaker recognition can be used to detect a person of interest in a speech recording.

Automatic language recognition is the process by which the language of a spoken utterance is recognized [4, 5]. This technology allows multi-lingual access to automated services and information. For example, telephone customer services can quickly identify the language of foreign callers and route their calls to operators who can speak the language. It also enables multi-lingual speech applications, such as machine translation and speech recognition.

1.1 Motivation

Human listeners often have the ability of recognizing speakers and languages from the speech signal, under varying conditions and contexts, drawing upon several sources of information [5, 6]. For a given speaker, the accent or some stereotype greeting (e.g., speaker can always answers the telephone by saying "greetings and salutations") can provide additional information about his or her identity. Extra information can also be required when some discriminatory information is not available or corrupted. For example, a language can still be identified by its characteristic melody (rhythm and intonation) without any knowledge about the vocabulary or syntax of the language [7]. Human performance seems to be a result of a robust and adaptive method of exploiting several sources of information [8-13].

Despite the richness of information conveyed in the speech signal, most recognition systems rely on the notion that people or languages can be recognized from only one source of information. For example, the most successful automatic language identification systems use the phonotactic content (the rules governing the sequence of allowable sound units of a language) from the speech signal [14-17]. Similarly, automatic speaker recognition systems traditionally use acoustic features extracted from short segments of speech to capture speaker-specific information [18-21]. While such systems have shown reasonably good performance, they ignore the complexities of the speech signal by disregarding other sources of information. Moreover, the performance of such systems is prone to degradation due to varying conditions (e.g., background noise and channel variability) and contexts (e.g., emotional speech) [22-24].

The human speech processing model shows that the reliability and accuracy of conventional systems can be improved by exploiting other sources of information in the speech signal. Not only additional information can improve the system accuracy by providing extra levels of discriminative information, but also it can increase the robustness by providing information that is less susceptible to degradation under varying condition and contexts. In fact, recent work in language identification [12, 25, 26] and speaker recognition systems [27-29] has shown significant gains in accuracy and robustness through the inclusion of different levels of information available in the speech signal.

One source of information that has long been acknowledged to discriminate languages and speakers is prosody [11, 30-34]. Prosodic information is conveyed through stress, intonation, and rhythm phenomena. Differences in the manifestation of these phenomena are perceived by the listener as changes in pitch, loudness, and length (e.g., 'short' versus 'long' phoneme). The manner in which these phenomena are realized can vary across different languages and different speakers. For example, in English, Italian and Dutch, pitch movements (rise or fall) are mainly limited to stressed syllable, whereas, in French, pitch movements occur in relation to word boundaries [35]. Even though certain linguistic effects are bounded by the native language [36], there is a great deal of variability in the methods used by humans to produce a given linguistic phenomenon (e.g., intonation, rhythm, and stress) [37, 38]. For example, changes in the respiratory system (e.g., variation in subglottal air pressure) and laryngeal muscles are important factors in the control of the fundamental frequency (acoustic correlate of pitch) [39-41]. The methods used to produce certain changes in pitch may be consistent for a speaker, but different across speakers [32].

Although perceptual changes in pitch and loudness can encode several prosodic phenomena, such prosodic features have not yet been fully exploited by speaker or language recognition systems. Most of such systems use distribution statistics [10, 42] or the contours [30, 43] of fundamental frequency (or F0, acoustic correlate of pitch) and intensity (acoustic correlate of loudness) to model prosodic information. One problem is that the temporal aspects of F0 or intensity that characterize a prosodic phenomenon are lost when using the distribution statistics because the typical independence assumption in

statistical modeling [44, 45]. In addition, the complex relationship between prosodic features is concealed by approaches the model separately such features. Besides, approaches that use the contours of such prosodic features limit the range of applications by requiring that the speaker must say a pre-defined sentence [2].

The aim of this thesis is to develop methods to quantify the relationship and the temporal variations of F0 and intensity to characterize prosodic differences among speakers and languages. The focus of this thesis is mostly on the speaker-specific information. However, we also present results on language identification to demonstrate the broad impact of our findings.

1.2 Overview of Speaker Recognition Technology

Speaker recognition encompasses two fundamental tasks: speaker identification and speaker verification [1-3, 46]. The goal of speaker identification is to determine who is speaking given a set of known voices. In this task, the system uses only a voice sample to recognize the unknown speaker. There are two modes of operation related to the set of known voices. In closed-set mode, the system assumes that the unknown voice must come from the set of known voices. In open-set mode, the voice sample may belong or not the set of known speakers. In this mode, the system must reject the speakers (also referred to as impostors) who are not from the set of known speakers. Speaker identification can be used for forensic applications to recognize the perpetrator's identity among several known criminals from a voice recording. The goal of the speaker verification task is to determine whether a person is whom he or she claims to be. This task is also known as voice verification or authentication, speaker authentication, talker verification or authentication, and speaker detection [3]. Differently from the speaker identification task, the system requires an identity claim together with the voice sample. It is considered an open-set task because it involves rejecting voices unknown to the system. Speaker verification can be used for security applications, such as, to control telephone access to banking services.

Automatic speaker recognition systems are further classified according to the speech modality: text-dependent or text-independent. In text-dependent mode, the user must speak a phrase, which is fixed or prompted, known to the system. The knowledge of

a spoken phrase can provide better recognition results [2]. In text-independent mode, the system does not require a pre-defined phrase from the user, but it requires enough training data for characterizing speaker's identity.

A typical speaker recognition system involves two phases: enrollment and recognition [3]. In the enrollment phase, the user provides voice samples to the system. Then the system extracts information from the voice samples to build a model of the enrolled speaker. In the recognition phase, a user provides a voice sample that is used by the system to measure the similarity of the user's voice to the model(s) of the previously enrolled user(s) and, subsequently, to make a decision. The speaker associated with the model that is being tested is referred to as the target speaker or claimant [3]. In speaker identification, the system measures the similarity of the voice sample to all pre-computed voice models. In speaker verification, the similarity is measured only to the model for the claimed identity. Finally, the system makes a decision based on the similarity measurement(s). Generally, closed-set speaker identification systems output the identity of the recognized user. Open-set speaker identification systems can also reject the user in case the voice sample does not belong to any of the stored voice models. Speaker verification systems output a rejection or an acceptance decision of the claimed identity.

1.2.1 Basic Structure of Speaker Recognition Systems

Like most pattern recognition problems, a speaker recognition system can be divided into two main parts: feature extraction and classification. The classification part has two components: pattern matching and decision. Figure 1.1 depicts a generic speaker recognition system.

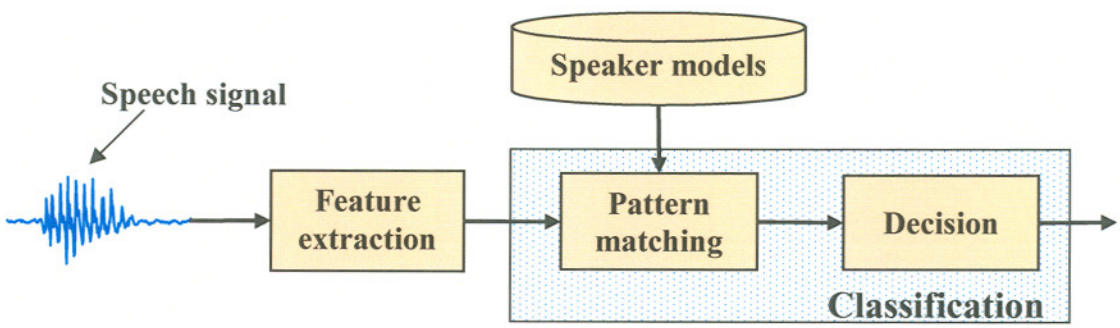

Figure 1.1: Generic speaker recognition system.

Feature Extraction

The feature extraction module extracts a set of features, which represent speaker-specific information, from the speech signal. Even though speech provides a large variety of speaker-specific information, the set of speaker-specific cues should have the following characteristics [47, 48]: occur naturally and frequently in normal speech, be easily measurable, have high variability between speakers, be consistent for each speaker, not change over time or be affected by the speaker's health, not be affected by background noise nor depend on specific transmission characteristics, and show resistance to disguise or mimicry. In practice, not all of these criteria are applied to the features used by the current systems.

Differences between speakers are determined by the physical constraints of the vocal apparatus (e.g., length of vocal tract, rate and acceleration of articulators), and linguistic or semantic choices (e.g., speaker idiosyncrasies, intonation patterns, or dialect) to convey some communicative intent. Such speaker-related differences can be categorized into two levels [49, 50]. Low-level information describes the acoustical characteristics of the speech signal (e.g., nasality, breathiness, and speaking rate). High-level information describes the semantic, linguistic, and prosodic aspects in the speech signal (e.g., dialect, idiosyncrasies, lexical and syntactical patterns, and speaking style).

Most of the speaker recognition systems are based on low-level information [19, 42, 46, 49, 51-53]. Typically, the time-varying acoustic properties of the speech signal are described using a sequence of short-term feature vectors. The underlying assumption is that the properties of the speech signal change relatively slowly in time and therefore short segments (10 to 40ms) have fixed (stationary) acoustic properties [54, 55]. The methods based on this assumption are commonly known as short-term analysis, which are historically inherited from speech coding applications [56]. In the short-term analysis of speech, each feature vector describes the acoustic properties (frequency components of the speech signal – speech spectrum) of a short segment of speech. Many different representations of the speech spectrum have been used for speaker recognition: cepstral-based features [18, 55, 57-59], linear predictive coding (LPC) coefficients [1, 60, 61], and LPC-derived features (e.g., log area ratios [51], reflection coefficients [42], and line spectral pairs [3]). However, the acoustic properties of certain speech events can differ

among different speakers, due to physical constraints of the vocal apparatus or idiosyncrasies in the speaking habits of individuals [21, 62-64]. To exploit the speaker-dependent properties of certain speech events, speaker recognition systems use speech recognition methods to locate such relevant speech events. This approach provides more details about the acoustic properties of the speech signal by having different feature sets for different speech events. In addition, such approach also reduces the problem of the high degree of redundant information in speech by only using certain speech events, instead of the entire speech signal [63]. Some of the speech events include words [65], broad phonetic categories (e.g., vowel, fricatives, nasals, plosives, stops, and liquids) [64, 66, 67] and phonemes [28, 68, 69]. Besides the short-term spectral properties of the speech, additional features have been investigated for speaker recognition, such as vocal tract resonances (formant frequencies and bandwidths) [70, 71], speech intensity [42], and pitch period [47].

Despite the wide variety of high-level information, few speaker recognition systems incorporate this type of information. Given the importance of pitch detection to a variety of speech processing systems [72], pitch was one of the first speech features to be used to describe high-level information. The characteristics of the pitch contour have been used to characterize the variations in prosodic information (e.g., stress and intonation patterns) that can constitute an individual's speaking style [30, 73-76]. Besides prosodic information, several speaker recognition approaches have adopted techniques used in speech recognition systems to incorporate semantic and linguistic information. For example, the sequence of words extracted from the speech signal using an automatic speech recognition system is used to derive the characteristics of the vocabulary, dialect, and idiolect of a speaker. Doddington [27] uses the frequency of word sequences to capture idiolectal differences among speakers. To capture conversational-style characteristics, Weber et al. [28] use the frequency of disfluency classes such as pause-fillers (e.g., *uh, um*), discourse markers (e.g., *you know*), feedback expressions (e.g., *all right, sure, yeah*), speech correction markers (e.g., *I mean*), conjunctions, and sentence fragments. Similar approaches are used by several researchers to exploit the phonetic differences (e.g., differences in the inventory, phonological context, and the frequency of occurrences of phones – unit of sound that contribute to the meaning of a word) among

speakers [16, 29, 77-79]. Since timing information of word or phones is also produced by automatic speech and phone recognizers, duration patterns of such speech units are also used to characterize speaker-specific information [28, 80-82].

Pattern Matching

The goal of the pattern matching module is to compute similarity scores between the features from the voice sample and the model(s) from the set of known speakers. In a verification task, this module outputs a similarity score between the voice sample and the claimed identity. In an identification task, it outputs similarity scores for all pre-computed voice models.

There are many types of pattern matching methods and corresponding models used in speaker recognition [1, 3, 46, 83]:

- **Template matching**: a speaker model (or template) is represented by a sequence of feature vectors (as a function of time) estimated from a fixed phrase. A similarity measure is estimated after time-aligning the voice sample and the speaker template using dynamic time warping (DTW). The goal is to capture the time-varying characteristics of the speech features. This is a typical approach to text-dependent speaker recognition. For example, Doddington [70] and Lummis [71] use the temporal trajectories of several features (pitch, intensity, and the three lowest formant frequencies) estimated from a fixed phrase to represent each speaker; these methods perform a time alignment between the voice sample and speaker template using the second formant frequency (in Doddington's work) and intensity (in Lummis' work) contours.
- **Nearest neighbor:** a speaker model comprises a set of feature vectors estimated from the enrollment data. A similarity measure is estimated using the accumulated distance between each feature vector from the voice sample and its k closest feature vectors (nearest neighbors) in the speaker's feature vectors. For example, Higgins et al. [84] propose a text-independent speaker identification system that uses a modified normalized nearest neighbor distance measure. To reduce significant memory and computational

requirements in testing, the system discards the training frames with distances larger than a pre-defined distance threshold.

- **Vector quantization (VQ)** [61]: a speaker model comprises a small number of representative feature vectors that are generated by clustering the speaker's training feature vectors. A similarity measure is estimated using the accumulated distortion (or distance) between each feature vector from the voice sample and the speaker's representative feature vectors. This approach does not require as much memory as the nearest neighbor approach, and it can be used for text-independent speaker recognition.
- **Artificial neural networks**: a neural network is trained to discriminate between speakers or to estimate the underlying statistics of the speaker's feature space. Some of the neural networks include: auto-associative neural networks [85], multi-layer perceptron [83, 86, 87], and time-delay neural networks [88]. For example, Oglesby and Mason [89] train a modified form of feed-forward neural network based on radial basis functions (RBF) for each speaker in a speaker verification system. Ganchev et al. [90] propose a system that trains a probabilistic neural network for each speaker to capture time correlations between features.
- **Hidden Markov models (HMM)**: a speaker is represented by a stochastic model. The aim of this modeling approach is to capture the temporal and statistical variation of the speaker's speech production. Based on a first-order Markov chain, a HMM is a finite-state machine, where a probability density function is associated with each state. The states are connected by a transition network. Using appropriate HMM topologies, both text-dependent and text-independent systems can be developed. For example, Poritz [91] proposes the use of a five-state ergodic-HMM, where all transitions between states are allowed, for text-independent speaker recognition. Rosenberg et al. [92] use a three-state left-to-right HMM (that is, the system states proceed from left to right) to characterize the speaker's phrase as a sequence of sub-word units.
- **Gaussian mixture model (GMM)**: a speaker is represented by a mixture of Gaussian density functions. The use of GMM for modeling speaker identity is

motivated by the interpretation that the Gaussian components represent some underlying broad phonetic sounds that characterize a speaker's voice [59, 93]. GMM can also be viewed as a single-state HMM with a Gaussian mixture observation density. The wide-spread use of GMMs for speaker recognition is due to its modest computational requirements, ability to model arbitrary densities, and consistent high performance [94]. Reynolds and Rose [93] show that GMM provides a robust speaker representation that outperforms other modeling techniques, such as VQ, RBF, and single Gaussian.

- **N-grams**: a speaker model is represented by a set of probabilities that characterizes the frequency of co-occurrences of N discrete units. With the increased research interest in modeling high-level information, the aim of such approach is to capture speech patterns (e.g., pronunciation, idiolect, and intonation) specific to speakers by modeling the statistical dependencies between relevant discrete units, such as words and phonemes. For example, Doddington [27] uses the likelihood of word bigrams (N=2) to capture idiolectal differences between speakers. Andrews et al. [77] use the likelihood of phone trigrams (N=3) to capture differences of pronunciation patterns.

In speaker verification and open-set speaker identification, the estimated features can also be compared to a model that represents the unknown speakers, also known as impostor model or speaker independent model. Typically, the impostor model can be a collection of "cohort" speakers [95] or a single model derived from a large number of speakers [18]. The collection of cohort speakers, also known as likelihood or background sets, can come from other enrolled speakers or from speakers of a different corpus. The approach that uses a large number of speakers is referred to as universal background model (UBM) [59] or speaker-independent model [21, 96].

Decision

In the decision module, the similarity score(s), determined by a statistical or deterministic method in the pattern matching module, is used to make a decision. For closed-set identification applications, a speech utterance from unknown speaker is associated with the speaker identity whose model is the most similar to the voice sample. Therefore, the

decision output is a speaker identity. Since the unknown speaker may not belong to the set of known speakers in open-set applications, a threshold is compared to the similarity score during the decision process. Therefore, open-set speaker identification applications can output an additional decision alternative, "the unknown speaker does not belong to the set of known speakers". In open-set speaker verification, the decision becomes a problem of accepting or rejecting the claimed identity. Given the problem of an unknown speaker being identified as belonging to the set of known speakers in open-set applications, the cost of making a decision error can be incorporated in the decision process. For example, it is more costly for a bank to allow an impostor to withdraw money, than to reject a true bank customer.

The effectiveness of a speaker recognition system is measured differently for different tasks. Since the output of a closed-set speaker identification system is a speaker identity from a set of known speakers, the identification accuracy is used to measure the performance. For open-set systems, there are two types of error: recognizing an impostor as coming from the set of known speakers (false alarm) and false rejection of a known speaker (missed detection). The performance of some systems are reported using the equal error rate (EER), which is the point where the number (or probability) of false alarm errors is the same as the number (or probability) of missed detection errors. The performance measure can also incorporate the cost associated with each error, which depends on the application. For example, in a telephone credit card purchase system, a false acceptance is very costly; in a toll fraud prevention system, false rejection can alienate customers.

1.3 Overview of Language Recognition Technology

With the globalization of our society, there has been an increasing need for technology that enable systems (e.g., speech recognition and information retrieval) or services (e.g., call centers and 911 dispatchers) to deal with more than one spoken language. Language recognition has definitely a very important role in such technology. For example, call centers can provide a better service to foreign customers by automatically identifying the spoken language and routing the call to an operator fluent in the corresponding language. Other applications include: automatic information retrieval (e.g., search for speech

material spoken in some particular language), multi-lingual spoken dialog systems, and automatic translation services.

Similar to speaker recognition, language recognition encompasses two fundamental tasks: identification and detection. In language identification, the goal is to identify the spoken language from a set of known languages. In language detection, the goal is to determine whether a given language is being spoken in a speech segment. The main difference between both tasks is that the identification task requires N comparisons (assuming a set of N languages) to make a decision, whereas the detection task requires only one comparison.

There are several sources of language-specific information that can be used to discriminate between languages [4, 97]:

- Phonology: the sounds of a language can be described in terms of a set of abstract linguistic units called phonemes, which are the smallest contrastive units of sound that contribute to the meaning of word. The differences in the inventory and the frequency of occurrences of phonemes can be used to discriminate languages. For example, some sub-Saharan African languages can be characterized by the "clicks" found in their phonetic inventory. Even though many languages share a common subset of phonemes, the acoustic realization of a phoneme (referred to as phone) can present significant differences. For example, in English, */l/* and */r/* (as in "leaf" and "reef") are two different phonemes, whereas in Japanese they are not. In addition, languages can differ in the phonotactics (the rules governing the sequence of allowable phonemes). For example, phoneme clusters */sr/* and */sp/* are quite common in Tamil and German respectively, but are rare in English.
- Morphology: different combinations of phonemes constitute different words. Thus, languages can differ in the rules of combining phonemes and, consequently, they have different set of words (vocabulary).
- Syntax: when people communicate, the words must be combined in a specific way so that the intent message is understood. All languages have a set of principles (syntactical rules) to combine words, which form phrases, clauses, or sentences.

- Prosody: languages vary in terms of rhythm, intonation, and stress through the variation of pitch, loudness, and duration. For example, tonal languages, such as Mandarin Chinese and Vietnamese languages, use pitch contour patterns on a single syllable to denote different words. Another example is the position of the stress pattern within a word. In Czech and Finish languages, it always occurs in the same position (referred to as fixed-stress languages) within a word, whereas Turkish always has the last syllable stressed.

Given the current speech technology and language resources, only a subset of the sources of information is incorporated in current language recognition systems [4, 5, 97]. For example, the use of morphologic and syntactic rules from a language is limited by the fact that there are systems that produce an accurate orthographic transcription (the sequence of spoken words) for only a small number of languages [4]. In addition, the lack of public-domain speech corpus has restricted the number of languages that can be included in a language recognition system.

1.3.1 Basic Structure of Language Recognition Systems

Given the pattern recognition framework, language recognition systems involve two phases: training and recognition. In the training phase, using language-specific information, one or more models are built for each language. In the recognition phase, a spoken utterance is compared to the model(s) of each language and then a decision is made. Thus, the success of a language recognition system relies on the choice of language-specific information used to discriminate among languages, while being robust to speaker, channel, and context variability.

Driven by the complexity of extracting morphological and syntactical information, most language recognition systems have been using information from prosodic and phonologic sources to discriminate languages. The phonology-based approaches deal with differences of the phonotactics and the acoustic characteristics of the speech signal (i.e., acoustic properties of the phone realizations). The prosody-based approaches deal with differences of the language "melody", result of intonation, stress, and rhythm patterns.

Most approaches to language recognition can be classified into two main categories: short-term and long-term. Short-term methods are based on the information extracted from short-term feature vectors, and long-term methods are based on segments that can span more than one feature vector. Figure 1.2 presents the basic architecture of both approaches to language recognition system that is capable of recognizing *N*-languages.

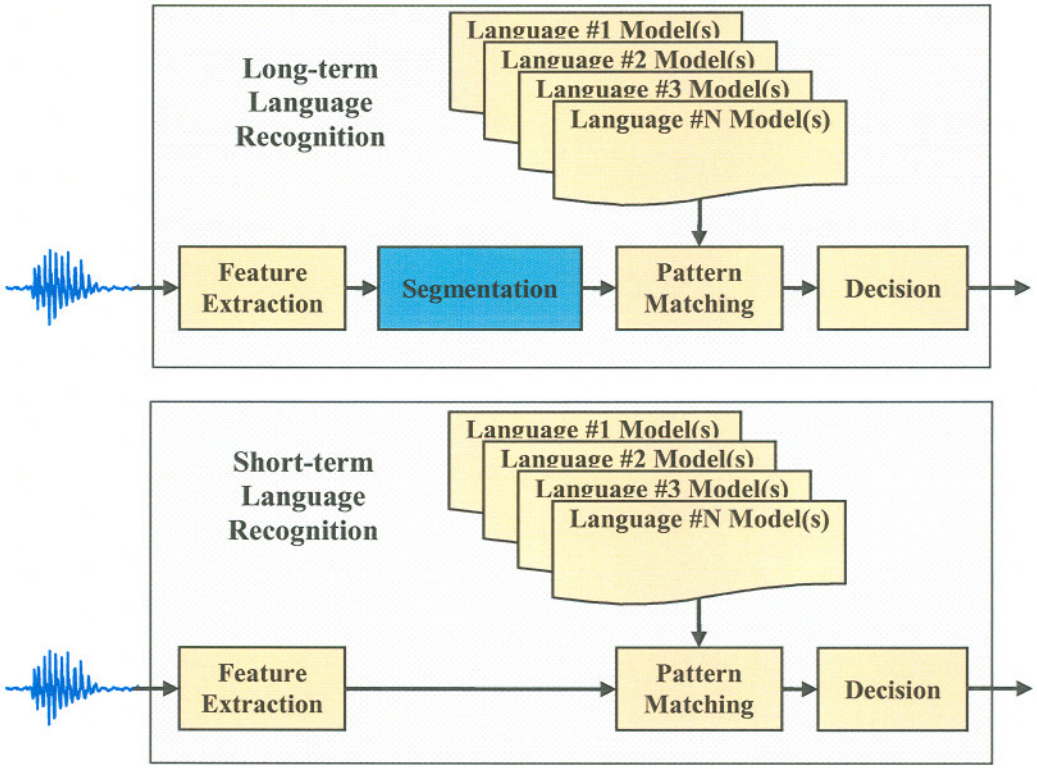

Figure 1.2: Basic architecture of short-term and long-term approaches to language recognition for *N*-languages.

In short-term approaches, each language is modeled using the distribution statistics of the acoustic features (usually, assumed to be independent of each other) extracted from some training data. The goal is to capture acoustic differences between languages, result of the differences in phonetic and prosodic information. Given the similarity to speech recognition and speaker recognition approaches, several systems have been developed based on this approach. Some of the features used for language recognition include linear predictive coding (LPC) coefficients [43, 98], LPC-derived features (e.g., reflection, log area ratios, and prediction error) [99], formant frequencies [100, 101], pitch and energy features (e.g., raw values, and derivative) [100, 102], and

spectral features [4, 103, 104]. Since the language-specific information is represented as a sequence of feature vectors, several methods have been used to classify languages, including hidden Markov models (HMM) [43, 105, 106], Gaussian mixture models (GMM) [103, 104], vector quantization (VQ) [98, 100, 101], polynomial classification [99], quadratic classifiers [100], and artificial neural networks (ANN) [25].

In long-term approaches, the language-specific information is extracted from segments of speech. First, the speech signal is divided into a sequence of segments given certain acoustic patterns. For example, each segment can represent a phone [15], a syllable [10, 107], a broad-phonetic category (e.g., vowels, consonants, fricatives, and sonorants) [108-110], or an acoustic pattern [111]. Then, two techniques can be used to model a set of languages:

1. Similar to short-term language recognition approaches, acoustic features are extracted and used to exploit the acoustic differences between languages conditioned to the segments. For example, Li and Edwards' approach [109] uses LPC coefficients to model different syllables; Parlangeau-Vallès et al. [112] use spectral features to model vowel and consonantal systems; Rouas et al. [107] estimate durational and intonational features (distribution statistics of F0) from pseudo-syllables to discriminate rhythm and intonation properties between languages. Some of the approaches have used GMMs to capture the acoustic properties of the segments [107, 112], or HMMs to not only capture the acoustic properties, but also to model the interactions between the segments [109].
2. The sequence of acoustic pattern categories is used to model the statistical dependencies between the categories (e.g., phonotactic properties of a language). House and Neuburg [108] published the earliest long-term approach based on the different phonotactics of broad-phonetic categories (stop consonant, fricative consonant, nonvocalic consonant, vowel, and silence) using a discrete HMM. The improvement of the speech processing techniques motivated the development of several systems based on the statistical dependency between phones [15, 17], acoustic patterns [111], and

broad-phonetic categories [113, 114]. The method extensively used to model such dependencies is *n*-gram.

Since different sources of information have been exploited by different approaches, several researchers proposed the fusion of such methods to benefit from the variety of language-specific information. Hazen and Zue [115] use a probabilistic framework to combine phonetic (trigram models from language independent phones), acoustic (GMMs of cepstral features), and prosodic (GMMs of F0 and duration) information for language identification. Some other combinations include phonetic and acoustic models [97] and phonetic and prosodic [116].

1.4 Overview of the Proposed Approach

In this work, we investigate the use of pitch and intensity for producing prosodic phenomena. We hypothesize that different speakers and different languages may be characterized by different intonation, stress, or rhythm patterns produced by the changes in pitch and in intensity. Therefore, the combination of pitch, intensity, and duration that characterizes particular prosodic "gestures" are useful in extracting speaker-specific [76, 117] and language-specific information [12, 97].

We propose alternative approaches to convert the continuous speech signal into a sequence of discrete units that describe the signal in terms of dynamics of the fundamental frequency (as a proxy for pitch) and short-term energy (as a proxy for intensity) contours. We also propose a method to integrate such discrete units to segmental information. Using simple statistical models, we show that the statistical dependency between such discrete units can be used to capture language- and speaker-specific information. We also show that they are robust to communication channel effects, and that they provide complementary information to the conventional systems.

The contributions of this work are following:

- Novel methods to represent the dynamics of fundamental frequency and short-term energy for prosody modeling. We show that this new representation can provide another level of information besides the distribution statistics of the fundamental frequency and energy features.

- Framework to segment speech signal using the dynamics of two or more temporal trajectories. The proposed method can also be used to segment speech using different temporal trajectories (e.g., formant frequencies).
- Complementary information to conventional speaker and language recognition systems. We show that the proposed approaches can improve the performance of the state-of-the art of speaker and language detection.

1.5 Thesis Structure

The remainder of the thesis is organized as follows:

- **Chapter 2** introduces some of the prosodic bases for speaker and language recognition. We describe prosody and its effects with respect to language- and speaker-specific information. A brief review of previous prosody-based approaches to speaker and language recognition is also presented.
- **Chapter 3** describes the speaker recognition paradigm for evaluating the proposed approaches. Since we are proposing a method to complement the current systems, we also describe the state-of-the-art speaker detection system and the method used to fuse systems. A prosodic baseline, which uses information from F0 and short-term energy contours, is also described.
- **Chapter 4** describes several approaches to segment the speech signal using the dynamics of F0 and short-term energy. We show that the statistical modeling of the segment classes can be used to capture speaker-specific information.
- **Chapter 5** describes the method to combine the proposed prosodic modeling approach and segmental information. We show that the combination of the proposed prosodic classes and segmental (or phonetic) information can be used to characterize speaker-specific information.
- **Chapter 6** shows that the proposed prosodic modeling approach can also be used to capture language information. Using the evaluation paradigm defined in the 2003 NIST Language Identification Evaluation, we show that the prosodic modeling approach also provides complementary information to conventional language recognition systems.

Chapter 2

Speaker and Language Recognition using Prosody

The goal of this chapter is to review the speech production system and prosody, and to describe some language- and speaker-specific prosodic differences conveyed by fundamental frequency and intensity contours. Section 2.1 describes the human speech production system. Section 2.2 defines prosody and its manifestation levels. Section 2.3 presents the prosodic differences between speakers and an overview of speaker recognition systems based on prosodic features. Section 2.4 presents the prosodic differences between languages and an overview of language recognition systems based on prosodic features.

2.1 Speech Production

The anatomy of the human speech production system is shown in Figure 2.1. The vocal apparatus comprises three cavities: nasal, oral, and pharyngeal. The pharyngeal and oral cavities are usually grouped into one unit referred to as the vocal tract, and the nasal cavity is often called the nasal tract [118]. The vocal tract extends from the opening of the vocal folds, or glottis, through the pharynx and mouth to the lips (shaded area in Figure 2.1). The nasal tract extends from the velum (a trapdoor-like mechanism at the back of the oral cavity) to the nostrils.

The speech process starts when air is expelled from the lungs by muscular force providing the source of energy (excitation signal). Then the airflow is modulated in

various ways to produce different speech sounds. The modulation is mainly performed in the vocal tract (the main resonant structure), through movements of several articulators, such as the velum, teeth, lips, and tongue. The movements of the articulators modify the shape of the vocal tract, which creates different resonant frequencies and, consequently, different speech sounds. The resonant frequencies of the vocal tract are known as formants, and conventionally they are numbered from the low- to the high-frequency: F_1, F_2, F_3, and so on. The resonant frequencies can also be influenced when the nasal tract is coupled to the vocal tract by lowering the velum. The coupling of both vocal and nasal tracts produce the "nasal" sounds of speech, like /n/ sound of the word "nine".

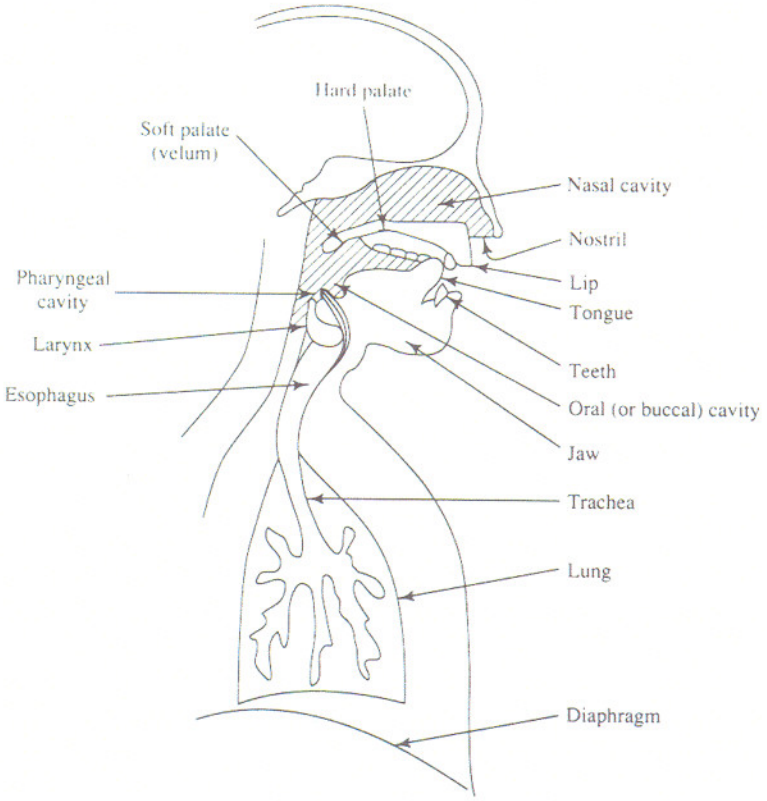

Figure 2.1: The human speech production system [119].

The airflow from the lungs can produce three different types of sound source to excite the acoustic resonant system:

- For voiced sounds, such as vowels, air is forced from the lungs through trachea and into the larynx, where it must pass between two small muscular folds, the vocal folds. The tension of the vocal folds is adjusted so that they vibrate in oscillatory fashion. This vibration periodically interrupts the airflow

creating a stream of quasi-periodic pulses of air that excites the vocal tract. The modulation of the airflow by the vibrating vocal folds is known as phonation. The frequency of vocal fold oscillation, also referred to as fundamental frequency (F0), is determined by the mass and tension of the vocal folds, but is also affected by the air pressure from the lungs. An example of voiced sounds is the vowels.

- For unvoiced sounds, the air from the lungs is forced through some constriction in the vocal tract, thereby producing turbulence. This turbulence creates a noise-like source to excite the vocal tract. An example is the /s/ sound in the word "six".
- For plosive sounds, pressure is built up behind a complete closure at some point in the vocal tract (usually toward the front of the vocal tract). The subsequent abrupt release of this pressure produces a brief excitation of the vocal tract. An example is the /t/ sound in the word "put".

Note that these sound sources can be mixed together to create another particular speech sound. For example, the voiced and turbulent excitation occurs simultaneously for sounds like */v/* (from the word "victory") and */z/* (from the word "zebra").

2.2 Prosody

The literature provides several definitions of the term prosody. In the most classical study in prosody, Lehiste [39] defines prosody as "features whose arrangement in contrastive patterns in the time dimension is not restricted to single segments defined by their phonetic quality". That is, prosody is the speech features (e.g., stress, intonation, and rhythm) that must be examined relative to their context to be understood and that they can extend for more than one phonetic segment (e.g., syllable, word, and paragraph). For example, voicing can extend over a sequence of phonemes but it is not a prosodic feature because it does not require the existence of adjacent unvoiced regions to be determined. Stress is a prosodic feature because it extends over a syllable or more, and requires an adjacent unstressed syllable to be perceived. As the prosodic features typically extend more than a phonetic segment, prosody is often referred to as the suprasegmental property of speech.

Prosody is used to structure the speech flow to convey lexical, syntactic, semantic, and emotional information. The means for obtaining this structure include speech rhythm, intonation, stress, tone, and some aspects of voice quality. Such phenomena are perceived by the listener as variations in pitch, loudness, and length (or quantity). These variations are a result of modulation of the physically measurable dimensions of fundamental frequency, intensity, and duration [48]. Some of the prosodic features include:

- Pitch: is the perceptual correlate of the fundamental frequency of the speech signal. Changes in pitch are also related to changes in the intensity [39].
- Loudness: is the measure of the perceived intensity of the sound. Note that increasing the frequency of vocal folds vibration can also affect the perception of loudness.
- Duration: is the length of a segmental unit such as a phoneme. In many languages, stress is one of the factors that conditions the duration of a sound or a sequence of sounds. For example, in English, a stressed syllable has typically longer vowels than an unstressed syllable.
- Intonation: is the pattern of pitch changes (or melody) over an utterance. There are several functions associated with intonation, such as to differentiate structural units (e.g., phrases), to regulate turn-taking in conversations, and to convey informational structure by highlighting certain words in the utterance.
- Tone: is a perceptual change in the contour or the direction of movement of F0 on syllables. Tone languages (like Mandarin Chinese) use different tones to distinguish certain words from others.
- Stress: is the term given to any form of prominence of syllables. Sentence stress is the most prominent word in a sentence, and word stress (also known as lexical stress) is the stress on the syllables within a word. The acoustic realization of stress comprises two or even three acoustic parameters of prosody (F0, intensity, and duration). For example, prominence of a syllable can be obtained either by syllable lengthening or increasing the intensity of the speech.

- Tempo and rhythm: related to the duration patterns and distribution of events in speech. For example, variations in the use of stress cause different languages to have different rhythms [120]. While tempo primarily indicates the rate of speech, rhythm is a measure of the pattern and manner of speech (e.g., staccato, legato, rhythmic).

2.2.1 Prosodic Phenomena Manifestation

The prosodic phenomena can manifest in four levels [33]: linguistic, articulatory, acoustic, and perceptual. Table 2.1 presents some links between the manifestation levels of prosodic phenomena.

Table 2.1: Links between the manifestation levels of prosodic phenomena.

Linguistic	Articulatory	Acoustic	Perceptual
Tone, intonation, aspects of stress	Periodic airflow pulses due to glottal closures	Fundamental Frequency (F0)	Pitch
Aspects of stress	Variations in air pressure	Amplitude, energy, intensity	Loudness
Aspects of stress	Amplitude of articulatory movements	Duration	Length (or quantity)

At the linguistic level, the speaker uses prosody to communicate distinction between: types of sentences (e.g., different intonations for questions and statement), elements of a linguistic message (e.g., semantic emphasis of a word or multiple words), chunks of speech (e.g., marking boundaries and defining transition between words, phrases, or sentences), speaking styles (e.g., spontaneous versus read, neutral versus emotional), and discourse functions (e.g., turn-taking, seeking and giving feedback). The differentiated prosodic phenomena are related either to tone, intonation, or stress.

At the articulatory level, prosodic phenomena manifests by modifications of the articulatory behaviors. For example, the articulatory movement involved in the production of a stressed vowel compared to the unstressed variant is the same; the difference is that the articulatory movements for the stressed vowel tend to be larger, and longer in duration and more distinctive from other unstressed vowel movements. Physical observations of prosodic phenomena related to articulatory level typically include variations in the amplitude of articulatory movements, variations in air pressure, or

specific patterns of electric impulses in nerves connected to the articulatory musculature (especially those innervating the larynx).

At the acoustic level, prosodic phenomena are conveyed through significant variations of the fundamental frequency (F0), intensity (amplitude), and duration. For example, stressed syllables can be characterized by their higher F0, greater intensity, and longer duration than comparable unstressed syllables.

At the perceptual level, the listener decodes the information from prosodic phenomena conveyed through the acoustic signal. The prosodic phenomena are perceived through variation in pitch (melody), length, loudness, and speech rate.

2.3 Prosodic Speaker-specific Differences

Speaker differences based on physiological characteristics of the vocal apparatus, social background, and regional characteristics (accents and dialects) contribute to affect the realization of prosody. In this section, we describe some of the speaker related differences that can affect the use of prosody.

Physiological characteristics

The overall complexity of the vocal fold vibration differs for different people, and the shape of the glottal flow waveform varies with vocal effort and other aspects of the voice quality [121]. For example, it is possible to perceive a change in pitch when the fundamental frequency is fixed but signal intensity is slightly varied; an increase in intensity produces a drop in pitch as noted by Lehiste [39]. There are several factors that affect the variation of F0. For example, F0 may be raised by increasing vocal fold tension, by increasing subglottal pressure, or by a combination of the two [39, 40, 122, 123]. The pattern of combination, in turn, may be consistent for a speaker, but different across speakers. In addition, the ability of exploiting these difference regulatory mechanisms makes it possible for speakers to produce voices of considerably different intensity levels [123].

Pitch accent

Typically, a pitch accent is associated with a lexically stressed syllable, has a time scale in the 100 ms to 500 ms 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 pre-pausal accent typically has a shorter rise but a longer and deeper fall of F0 and of energy, as well. 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. A full description of pitch accents is beyond our scope here; we simply note that they are likely sources of inter-speaker information because they vary greatly in details of execution [122, 124, 125].

The location of the pitch accent can also vary across speakers. For example, the peak of a pitch accent may occur earlier or later in a stressed syllable, or even after the syllable with it is phonologically associated [117]. In addition, the phonetic alignment of a F0 event (often a peak, trough, or turning point) during the realization of a phonological target can vary between dialects of a language. Nolan and Farrar [126] show that the peak associated with utterance-initial pitch accents is later in some dialects of English than others.

Stress

Stress and accents manifest by duration, F0 contours, intensity, and degree of articulatory enhancement versus reduction [32]. The relative importance of these components is language specific. 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.

Voice quality

Baken [127] shows that pitch and voice characteristics may be correlated. For example, natural speech is characterized by deviation from strict periodicity (also known as jitter) and other instabilities, such as shimmer (i.e., perturbations on the signal intensity), but the amount of such irregularities is greater when subglottal pressure is falling and fundamental frequency is low, and vice versa.

Intonation

Intonation is considered the main source of speaker variability [32, 117]. The different patterns of F0 contours, combined with other acoustic parameters, are also affected by dialectal variations [128] and social conventions.

Taylor [129] shows that immediate segmental environment affects the local shape of the F0 contour. He lists three types of segmental effects: unvoiced segments, obstruent perturbations, and intrinsic vowel pitch. In unvoiced segments, there is no fundamental frequency at all. In obstruent perturbations, there is a tendency of consonants, especially obstruents, to produce sharp spikes in the contour. The perturbations are normally short in duration (typically less than 30 ms), but can make a F0 excursion into a small pitch accent [130]. In intrinsic vowel pitch, high vowels consistently cause slightly higher F0 values than low vowels [131]. For example, two utterances with the same intonation can have apparently different F0 contours solely due to these utterances having different segment content.

Van Dommelen [31] uses manipulated speech signal to measure the contribution of speech rhythm and pitch to speaker recognition. After synthesizing the manipulated speech samples (two fixed sentences per speaker), two types of tests are performed with a group of eight listeners: speaker identification (identify speaker among the 3 subjects) and speaker discrimination (AX type: whether or not A and X might have been spoken by the same speaker). One of the experiments synthesizes the speech signal from one speaker using the pitch contour from another speaker, while preserving the remaining speaker-specific characteristics. In both tests, the results show that the discrimination between speakers reduces when pitch contour is manipulated.

2.3.1 Prosody-based Approaches

For several decades, researchers have been investigating the use of prosodic features for speaker recognition. In this section, we describe some of the methods used to estimate and model prosodic features for speaker recognition.

Prosodic Features Contour Statistics

In the early 70s, Atal [30] proposed the use of temporal variations of pitch for a text-dependent, closed-set speaker identification system. The method uses linear discriminant

analysis to project the features onto a new space and Euclidean distance to perform the recognition. First, it divides the utterance into 50 ms segments and estimates the average pitch of each segment. Then, the features representing the pitch contour are linearly transformed (projected) onto a new 20-dimensional feature space so that the variance between all speakers is maximized and the variance within a speaker is minimized. The transformation uses the 20 eigenvectors with the highest variability, which is defined by the magnitude of the eigenvalues. The eigenvectors and eigenvalues are estimated over 60 utterances from 10 speakers using the Karhunen-Loève transform, also known as principal component analysis [44]. The classification step uses Euclidean distance to measure the distance between the vectors from a voice sample and the reference vectors from the speaker models. This method obtains a 97% recognition rate on a 10-speaker dataset.

Similar to Atal's work, Markel et al. [42] studied the averaging characteristics of pitch, intensity, and reflection coefficients over various segments lengths of speech for speaker recognition. The experiments are performed on unconstrained conversational speech obtained during interview (15 to 18 minutes duration) of four speakers. The features include average F0, standard deviation of F0, standard deviation of the normalized intensity variation, and spectral features (reflection coefficients extracted from linear prediction analysis). They use the Fisher Discriminant method to investigate the effectiveness of long-term averaging for speaker recognition. Despite the small amount of data, they show that the average F0 carries important information about the speaker and that intensity variation has some discrimination characteristics.

In the effort of capturing local dynamics in intonation to characterize a speaker's speaking style, Sönmez et al. [74] developed a F0 contour stylization method. First, the method eliminates regions of pitch doubling and halving using a lognormal tied-mixture modeling [132] followed by a median filtering. Then a piecewise linear model is fit to the estimated log pitch values in voiced regions. The method generates a sequence of connected linear spline segments for each voiced region. From each segment, they extracted the median F0, slope, and duration, which are independently modeled. In addition to these 3 features, they also extract the duration of the voiced segment and pauses. They refer to these 5 features as prosodic features. The score for verification is

the negative relative entropy between the parametric distributions of the target speaker and test segment (user's voice sample). They report results on the 1998 NIST Speaker Recognition Evaluation data: 500 speakers (250 males and 250 females) in 5000 trials with 2 minutes of telephone, conversational speech for training and test data with nominally 3, 10, and 30 seconds duration. The task is text-independent speaker verification, and the training condition is "one-session" as the 1996 NIST Speaker Recognition Evaluation. They show that prosodic features improve the performance when added to the cepstrum-based GMM system. They also show that the prosodic features require certain amount of data to reliably distinguish speakers.

Peskin et al. [82] examined a variety of prosodic and conversational features using the evaluation data for the Extended data task in the 2001 NIST Speaker Recognition Evaluation [133]. The task is text-independent speaker detection, and the evaluation data comes from the Switchboard I conversational, telephone speech corpus (a more detailed description is in Section 3.2.1). One of the proposed systems employs a nearest-neighbor classifier to compute the distance between the features from a given test conversation and the speaker model. The examined prosodic features include segmental duration (e.g., relative duration of words and phones, and duration of voiced and unvoiced segments), pause related (relative frequency and duration of pauses), and pitch contour related, similar to the Sönmez's work [74] (e.g., maximum, minimum, range, and slope). Given a feature vector estimated from a test conversation, a distance between such feature vector and the speaker model (represented by a set of feature vectors from the speaker's training data) is estimated using a nearest-neighbor classifier. The distance metric used in the nearest-neighbor calculation is the symmetrized Kullback-Leibler. Using the pitch contour related features, they report a 14.8% EER on 8-conversation training condition (approximately 20 minutes) using 3 neighboring points in the nearest-neighbor computation and testing on a conversation side of approximately 2.5 minutes.

Recently, Kajarekar et al. [80] used the statistics of F0 and energy from different regions to capture longer-term patterns that are characteristic of the speaker's speaking style. Two segmentation methods are proposed to define the regions. The pause-based method uses long pauses (greater than 500 ms) to detect a region boundary. The stylization-based method uses the segments detected by the F0 stylization algorithm

proposed by Sönmez [74] as the regions. Several features that describe pitch (e.g., maximum, mean, and slope), energy (e.g., range and slope), pause (e.g., average and maximum duration), and region (e.g., duration of the longest vowel) are modeled using a Gaussian mixture model. The equal error rate of both systems is between 22% and 32% (approximately 20 minutes of training data for each speaker model) on the extended-data one-speaker detection task of the 2003 NIST Speaker Recognition Evaluation [134]. This task uses data from the Switchboard II conversational, telephone speech corpus (a more detailed description is in Section 3.2.1). They also report results of a two-way combination of the prosody-based systems with a spectral-based system that uses a Gaussian mixture model of mel-frequency warped cepstral coefficients to build speaker models. The pause-based segmentation and the stylization-based segmentation systems relatively improve the spectral-based system by 15% and 5%, respectively.

Prosodic Features Contour Matching

Doddington [70] described a text-dependent speaker verification system that performs a pattern matching of the time contour of pitch, intensity, and formant frequencies. The system is evaluated using 8 known speakers and 32 impostors. The feature contours for each known speaker are stored for subsequent recognition. The feature contours are estimated from a pre-defined utterance by sampling the speech signal 100 times per second. After the system extracts the feature contours from a sample utterance, the contours of the sample utterance are time aligned with the stored reference of the claimed identity using a nonlinear time normalization approach. Such approach normalizes the time axes by maximizing the correlation between sample and reference second-formant contour through a piecewise linear continuous transformation of time. He uses the second formant because it has large excursions that are characteristic of the utterance and its shape is relatively consistent across different speakers and repetitions. Then, an accept-reject decision is made based on the dissimilarity between sample and reference functions computed with several heuristically chosen formulas. The equal error rate of the system is on average about 1%.

In another approach similar to Doddington's work, Lummis [71] proposed a speaker verification system that performs time alignment of the sample and reference contours based on the intensity contour. The main differences are: 1) time alignment is

based on the intensity contour instead of the second formant, 2) the reference patterns are constructed differently, 3) different set of dissimilarity measures, and 4) the speech samples are low-pass filtered at 4 kHz. The results are reported using the testing database used by Doddington. Lummis shows that the time alignment of the sample and reference contours using intensity contour yields same performance as using the second formant (about 1%). On a different experiment, the performance of a system that uses only the pitch and intensity contours yields an equal error rate of 0.5% (no statistical significance is reported).

Recently, Adami et al. [76] used a F0 contour template matching approach for text-independent speaker verification system. The system uses the F0 contour of 15 utterances as the speaker template. The utterances are words or phrases that have high occurrence rate and low context or topic dependency, such as, "okay", "absolutely", and "you know". Using an automatic speech recognition system to detect all the utterances from the speech signal, the system uses dynamic time warping (DTW) to perform time alignment and estimate the distance between all the matched utterances [118]. The final score is the average of the DTW distances normalized by a set of cohort speakers. The system yields a 13.3% EER on 8-conversation training condition of extended-data one-speaker detection task of the 2001 NIST Speaker Recognition Evaluation [133]. This task uses data from the entire Switchboard I conversational, telephone speech corpus (Section 3.2 presents a more detailed description of the evaluation task).

Prosodic Features Distribution Statistics

Sönmez et al. [132] proposed a probabilistic model of pitch halving/doubling to characterize speaker-specific information. They show that the pitch has a lognormal distribution and that the halving and doubling of pitch, which is produced by errors in pitch tracking or by the speaker during spontaneous speech, can be also modeled together with the normal pitch using a tied-mixture modeling approach. In this approach, a 3-component lognormal mixture model is used to model the pitch distribution. The means of the mixture components that correspond to halving and doubling pitch are set to be half and twice the mean of normal (or true) pitch, respectively. The variances of the three mixtures are set to be the same. Each speaker is modeled using a lognormal tied-mixture model of the pitch distribution. For each test trial, a lognormal tied-mixture model of the

test utterance is estimated, and the score is estimated using the relative entropy between the pitch distribution of the speaker model and the utterance. The task is open-set, text-independent speaker identification and the system is evaluated on the "one-session" condition (all training data from one phone call) with males for the target and impostor speakers from the March 1996 NIST Speaker Recognition Evaluation data. They show results of the pitch-based systems, a GMM system trained on MFCC, and the fusion between both systems. Despite the worse performance of the pitch statistics compared to the cepstrum-based system, they show that the fusion of the pitch-based system with the cepstra-based system provide a performance gain of more than 20% in false alarm rate, which is a result of the fact that pitch is affected much less by the handset variability than the cepstrum.

Carey et al. [135] showed that the mean and variance of pitch periods can provide complementary information to spectral-based features and that they are less susceptible to handset and channel mismatches. Using the 1995 and 1996 NIST Speaker Recognition Evaluation data (a subset of the conversational, telephone speech corpus Switchboard I), they study the use of prosodic features on the text-independent speaker verification task. First, they show that the mean and variance of pitch and energy variance provide discriminatory information about the speaker when used individually. Second, the fusion of the scores from the pitch (mean and variance features) and spectral envelope (cepstral features) systems provide a relative improvement of 30% on all the 1995 NIST Evaluation tasks. Third, they also compare the effect of handset and channel on the spectral- and pitch-based features. They show that the performance for spectral-based features is more susceptible to handset and channel effects than pitch-based features.

2.4 Prosodic Language-specific Differences

Although prosody exists in all languages, the structure of prosody is not exactly the same for every language. Languages can differ in the manifestation of stress, placement of stress and pauses, the communicative functions of intonations patterns, and the interaction between the parameters F0, intensity, and duration [33].

Some languages may use different prosodic phenomena to convey lexical information. The unrestricted placement of the stress is used in most languages of

European origin, such as English, French, German, and Spanish. For example, in English, the word "increase" can have the stress on the first syllable, "in", signaling a noun, or on the last syllable, "crease", signaling a verb. These languages contrast with the ones that have fixed stress in relation to the word (e.g., Czech, Polish, Hungarian, and Swahili). In accentual languages, such as Swedish and Japanese, the relative pitch difference between words (pitch accent) is used to alter their meaning. In tone languages, such as Chinese, Vietnamese, Thai, and Zulu, a tonal scale is used to differentiate the meaning of words. Besides, differences within each category also exist. For example, Swedish has two contrasting tones, and minimal pairs which differ only when tone exist; the contrast in Japanese is characterized by the presence or absence of a tone [36]. Another example is the number of tone patterns used by the tonal languages: Mandarin Chinese has four tones, plus a "neutral" pitch; Thai has 5 tones; and Cantonese has 10 tones.

Intonation can be different among languages to convey non-lexical information (such as questions versus declarative sentences). In English, declarative intonation is marked by a falling F0 ending, while yes-no question intonation is marked by a rising one. On the other hand, Russian uses strong emphasis on a key word instead of a rising ending. Chinese questions have an expanded pitch range near the end of the sentences.

The manner in which F0, duration, and intensity contribute to the stress of a syllable may also differ among languages. For example, the timing of rises or fall of the F0 contour in relation to the placement of stressed syllables can vary. Some languages use rising F0 at the beginning of a stressed syllable, while others use a rising F0 at the end of a stressed syllable [35].

The variations in the use of stress cause different languages to have different rhythms. Ladefoged [120] classifies the rhythmic differences among languages into those that have variable word stress (such as English and German), those that have fixed word stress (such as Czech, Polish and Swahili), and those that have fixed phrase stress (such as French).

2.4.1 Prosody-based Approaches

Foil [100] used frequency formants and prosodic features extracted from F0 and energy contours to discriminate languages from noisy speech. The motivation for using F0 and

energy contours is that the rhythm and intonation patterns could be described by such contours. The system uses a set of features based on statistics from voicing, pitch and energy contours, and contours derivative. Using three languages to evaluate the system performance, the performance of the prosody-based system is worse than the formant-base system.

Thymé-Gobbel and Hutchins [10] studied the use of several features estimated from pitch and amplitude contour information at the syllable level to discriminate between the following four languages: English, Spanish, Japanese, and Mandarin. In her study, the utterance is segmented into syllables using the pitch and amplitude contours, and then features are extracted from within and across syllables. The statistics of the extracted features are used to examine the discrimination between pairs of languages. She shows that prosodic cues alone can successfully distinguish between some language pairs with results comparable to many non-prosodic systems.

Hazen and Zue [136] described a system that integrates phonotactic, prosodic, and acoustic information of different languages. The prosodic information is modeled through probability densities of pitch information (F0 and its derivative) within a segment (broad-phonetic categories are used to define the segment boundaries). The system is evaluated on 11 languages (Hindi, English, Tamil, Farsi, Mandarin, Japanese, French, German, Spanish, Vietnamese, and Korean) from the OGI Multi-language Speech corpus [137]. They show that, despite the good performance of the system based only on phonotactic constraints alone, the integration of other sources of information (prosodic and acoustic) provides additional gain in performance.

Cummins et al. [25] used the smoothed first difference of F0 and band-limited amplitude envelope to discriminate pair of languages. The amplitude envelope is estimated using a band-pass filter centered at 1000 Hz with a bandwidth of 500 Hz. They claim that the amplitude variation in this frequency range is important for the perceived rhythm of speech. Each pair of languages is modeled using an artificial neural network trained on the first difference of either F0 or the band-limited amplitude envelope. They perform several experiments using data from 5 languages (English, German, Mandarin, Spanish, and Japanese) of the OGI Multi-language telephone speech corpus. They show that F0 is more effective than the amplitude envelope feature. In particular, Mandarin and

Japanese are better discriminated from the European languages, but are highly confused among them.

2.5 Discussion

This chapter presented the prosodic bases for language and speaker recognition. Prosody is used to modify the speech to convey lexical, syntactic, semantic, and emotional information. The structure of prosody differs among languages. Within a language, speakers can differ due to physiological characteristics and the manner in which a speaker employs and produces prosody. However, the complexity of the realization of prosodic phenomena makes difficult to define a set of features that can be used to capture language- and speaker-specific information. Thus, only a few approaches to speaker and language recognition incorporate prosodic features.

Most systems that incorporate prosodic features use the statistics of the distribution of F0 and intensity to model speaker- and language-specific information. This type of approach only captures the variability of prosodic features in a segment but not about the exact sequence of prosodic variations along the time. It disregards the temporal aspects of prosody that determines stress or intonation patterns. In addition, the modeling of prosodic features traditionally assumes statistical independence between them, thereby, disregarding the interaction between such features.

Apart from the exploitation of F0 and intensity statistics, some methods also make use of the information from the contour of F0 or intensity to capture intonational or rhythmic information. Mostly used in text-dependent speaker recognition systems, F0 contours are used to model the intonation characteristics of a word or phrase. Even though this approach does not discard the temporal information, it is limited to text-dependent tasks. Another approach is to segment the speech signal and extract statistics of the contour within the segments (e.g., syllables or fixed-length segments). Despite the use of segmentation, the estimated statistics form the segments do not adequately capture differences in the realization of prosodic features. For example, different contour shapes within a segment can have the same statistics.

Chapter 3

Speaker Detection Framework

This chapter describes the framework for evaluating the proposed speaker detection systems. First, the speaker detection task and the performance measures are described in Section 3.1. Section 3.2 describes the evaluation setup and data, as defined by the NIST Extended-data one-speaker detection task. Sections 3.3 and 3.4 describe the acoustic (the state-of-the-art in the speaker recognition technology) and prosodic baseline systems, respectively. Finally, a fusion system is described in Section 3.5.

3.1 Speaker Detection Task

The goal of 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.

The problem of speaker detection can be formulated as a hypothesis testing of two mutually-exclusive hypotheses:

H_0: target speaker is speaking,

H_1: target speaker is not speaking.

Since there are only two hypotheses, the likelihood ratio test is used to make a decision [44]. The likelihood ratio test is a comparison between the likelihood ratio of two hypotheses and a threshold given by

$$\frac{p(X|H_0)}{p(X|H_1)}\begin{cases}\geq T & \text{accept } H_0\\ < T & \text{reject } H_0\end{cases}$$

where $p(X|\ H_i)$ is the likelihood function for the hypothesis H_i, $i=0,1$, evaluated for the observed speech segment X, and T is the decision threshold (also known as operating point).

There are two types of errors may occur in a detection system. Type I error occurs when the null hypothesis (H_0) is rejected when it is true. The errors of this type are referred to as "misses", meaning that the target speaker is not detected when he or she is speaking. Type II error occurs when the null hypothesis is not rejected when it is false. The errors of this type are referred to as "false alarms", meaning that the target speaker is falsely detected when he or she is not speaking. Furthermore, the type of the application determines the cost for every decision. For example, the cost of a false alarm error has a more damaging effect than a miss error in a telephone credit card purchase system. Therefore, the probability and costs associated with the errors have to be considered when making a decision rule (i.e., selecting the decision threshold). Figure 3.1 shows the main components of a speaker detection system based on likelihood ratio test.

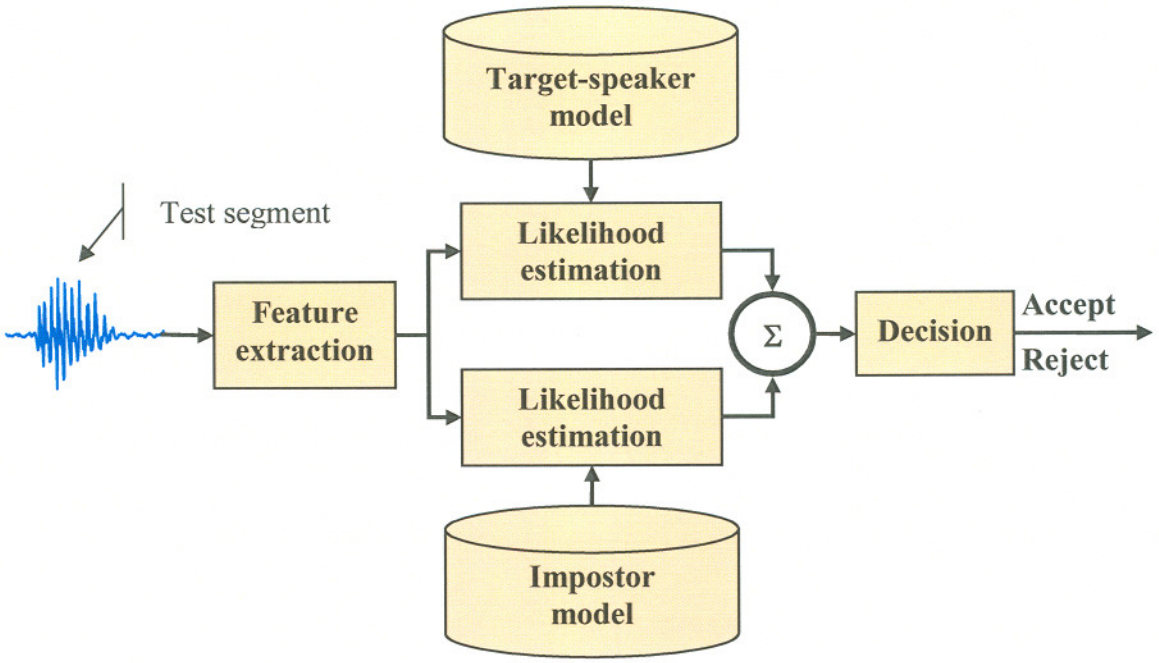

Figure 3.1: Speaker detection system based on likelihood ratio test.

The features extracted from the test segment are used to compute the hypotheses likelihood. The likelihood of the null hypothesis is estimated using the target-speaker model. The likelihood of the alternative hypothesis is estimated using an impostor model that characterizes all the speakers, but the target speaker. The estimation of the likelihoods depends on the distribution used to model the respective feature space. For example, a system can assume that the feature space is represented by a Gaussian distribution, so that the models are the mean and variance parameters.

The target-speaker and impostor models are estimated *a priori*. The target-speaker models are estimated using training data from the respective speaker. The estimation of the impostor model poses a more complex task because it must represent the speaker space that is complementary to the target speaker. The method to define a speaker set that represents such speaker space is still under research [59, 138].

3.1.1 Speaker Detection Performance

The performance of a detection system is evaluated using the probabilities and costs associated with the decision errors (false alarm and miss). One performance measure frequently used in speaker detection tasks is the equal error rate (EER). The EER is the point where the false alarm and miss probabilities are equal, and both errors have the same cost. Another performance measurement is the detection cost function (DCF), which is adopted by the NIST evaluations of speech, speaker, and language recognition systems. The DCF is defined as a linear combination of the miss ($P_{Miss|Target}$) and false alarm ($P_{FalseAlarm|NonTarget}$) probabilities:

$$C_{Det} = C_{Miss} * P_{Miss|Target} * P_{Target} + C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * P_{NonTarget}$$

where C_{Det} is the detection cost, $C_{FalseAlarm}$ and C_{Miss} are the relative costs of detection errors, and P_{Target} is the *a-priori* probability of a particular target speaker (so that, $P_{NonTarget} = 1 - P_{Target}$). Defined by the NIST evaluation paradigm, the DCF parameters are set as following: $C_{FalseAlarm} = 1$, $C_{Miss} = 10$, and $P_{Target} = 0.01$. This means that the false acceptance error is considered about 10 times stronger than a false rejection error. Despite that DCF is a more generic measure, performance comparisons focus on EER because the costs associated with each error can vary across different applications. Besides the EER, we report the minimum DCF value obtained a posteriori for the best

detection threshold. The minimum DCF value is always going to follow the EER in between parenthesis.

The performance of a system can also be described in terms of a graph representing the probabilities involved in the detection task. Since there is a tradeoff between the miss and false alarm errors, the system can work under different operating conditions. A performance curve is estimated by continuously changing the value of the decision threshold. The receiver operating characteristic (ROC) curve has been traditionally used for this purpose. The ROC curve uses the false alarm rate and the correct detection rate as the x- and y-axes, respectively. However, ROC curves are not used in this work because systems that are close to the optimal point (upper left corner of the graph) tend to group together near this corner, which makes difficult the visualization of different systems. Instead, we use the detection error tradeoff (DET) [139] curve to plot the systems performance.

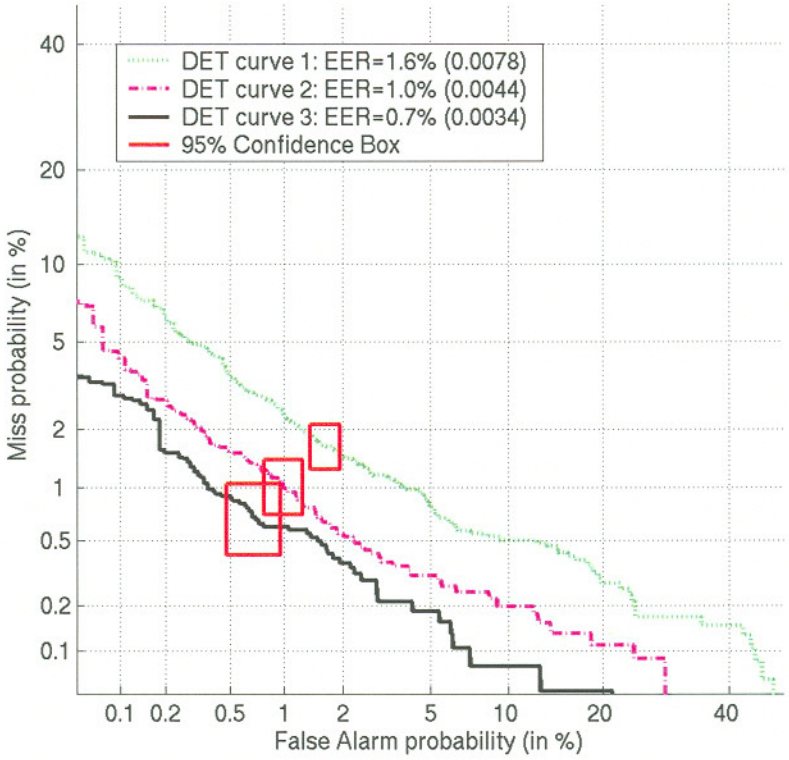

Figure 3.2: Example of detection error tradeoff (DET) curve. The minimum DCF value is in between parenthesis following the EER. The boxes represent the 95% confidence interval.

The DET curve uses the false alarm and the miss probabilities as the x- and y-axes, respectively, in normal-deviate scale, as shown in Figure 3.2. Suppose that the

distribution of the likelihood ratio values are normally distributed for true speakers and impostors, and the variances of the distributions are equal. Plotting the normal deviates that correspond to the false alarm and miss probabilities results in a straight line performance curve (consequence of the normality assumption) with unit slope (consequence of equal variances). Note in Figure 3.2 that the use of the normal deviate scale moves the curves away from the lower left corner when the performance is worse, making visual comparisons easier.

The rectangles in Figure 3.2 represent the 95% confidence interval around the EER point. The confidence interval is computed under the assumption that each detection test is an independent trial, that misses and false alarms are uncorrelated errors, and that the data comes from a binomial distribution. In case of non-overlapped rectangles, like DET curve 1 and 2 in Figure 3.2, the difference between the systems is statistically significant (significance level at 0.05). However, the converse is not true, which always is going to be followed by a significance test. We use the binomial test for differences in proportion [140] to check whether the difference between the EER of systems is statistically significant. Unless specified, the level of significance is set to $\alpha = 0.05$.

3.2 NIST Extended-data One-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 [133, 141]. The purpose of this task is to support the exploration and development of higher-level and more complex characteristics for speaker recognition [27].

3.2.1 Evaluation Data

The data for this task comprises of conversational, telephone speech from LDC's Switchboard corpora. The extended-data task in the 2001 NIST Speaker Recognition Evaluation (SRE) uses data from the Switchboard I corpus, and the 2002 NIST SRE uses data from the Switchboard II corpus (phases 2 and 3).

The Switchboard I corpus consists of approximately 2,400 telephone conversations averaging 6 minutes from all areas of the United States [142]. There are 543 speakers (302 male, 241 female) ranging from 20 to 60 years old. The education

level of most of the participants (about 90%) was at least college. The collection process was controlled by a computer system (also known as "Robotoperator"). Upon receiving an incoming call, the system selected and dialed a second person, and introduced a topic for discussion. Then, the entire conversation was recorded from the two subjects into separate channels. The search for the second person was restricted to the participants who had not spoken with the caller. There were about 70 topics provided for discussion, of which about 50 were used frequently. No one spoke more than once on a given topic.

The Switchboard-II Phase 2 corpus consists of 4,472 5-minute telephone conversations from 6 states (Illinois, Iowa, Minnesota, Michigan, Ohio, and Wisconsin) of the United States. The 679 participants (352 Female, 327 Male) were mainly college students from several mid-western university campuses. The Switchboard-II Phase 3 corpus consists of 2,728 5-minute telephone conversations from southern United States, under a variety of telephone (land line) handsets. The 640 participants (292 Male, 348 Female) were mainly college students. The collection process of both phases is similar to the one used for Switchboard I. Each participant was allowed to initiate and receive at most one call per day. Each participant would receive five calls at a designated number and make five calls from distinct phones lines.

3.2.2 Evaluation Setup

NIST provided a control file listing target and test speakers, along with training and testing conversation list. The training list has all the target-speaker models and their respective conversations for training. A speaker can have several models that vary the conversations or the number of conversations used for training. The target-speaker models are trained using 1, 2, 4, 8, or 16 conversation sides. The testing list has the hypothesized speaker (test segment) and multiple target-speaker models. One conversation side will serve as a test segment. The system must produce a score and decision for each trial, which comprises a target-speaker model and its hypothesized speaker. There are two types of trials: target trials (the hypothesized and target speakers are the same) and impostor trials (the hypothesized and target speakers are not the same).

To supply a large number of target trials, impostor trials, and target-speaker models, the evaluation used a cross-validation processing of the entire corpus. The cross-

validation experiments were defined by NIST [133, 141]. In both NIST SREs, the evaluation data is divided into several partitions. All trials within a partition have only target-speaker models and test segments from the speakers within that partition only; data from other partitions are available for other purposes, such as background model building and normalization. In the 2001 NIST SRE, the evaluation data was divided into 6 partitions of 80 speakers each. The detection task consists of 483 speakers with 4,105 target-speaker models and 57,470 (target and impostor) trials for the testing phase. In the 2002 NIST SRE, the evaluation data was divided into 10 partitions of 106 speakers each. The detection task consists of 10,932 target-speaker models and 156,184 trials for the testing phase.

Due to the limited number of speakers/models, the results for the 16-conversation training condition were found to have high statistical variation so we will generally cite results only up to the 8-conversation training condition.

3.3 Acoustic Baseline System

Although this work proposes a prosody-based system, we want to show that the proposed prosodic approaches can provide complementary information to acoustic-based systems. The most successful approach to text-independent speaker detection is based on a likelihood ratio detector (as depicted in Figure 3.1) that uses Gaussian mixture models (GMMs) for estimating the likelihoods [20, 59]. Since this approach is the state-of-the-art on the NIST speaker recognition evaluation [94], the parameters configuration of this system are the same as the ones used in the evaluations according to Reynolds et al. [59].

3.3.1 Feature Extraction

The feature extraction processing consists of three steps, as depicted in Figure 3.3. First, the speech signal is divided into overlapping segments of 20 ms, with a 10 ms overlap. Then, a speech activity detector is used to discard silence-noise frames. The speech activity detector is a self-normalizing, energy-based detector that tracks the noise floor of the signal and can adapt to changing noise conditions [143]. This procedure basically estimates the instantaneous signal to noise ratio (SNR) as the ratio of the short-time and

the long-time signal energy, and removes signal parts for which the SNR is below a certain threshold.

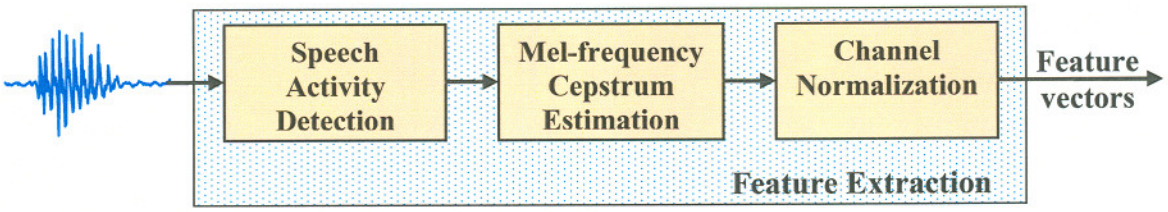

Figure 3.3: Sequence of steps to extract features from the speech signal.

A 19-dimensional mel-scale cepstral (MFCC) feature vector is extracted from the frames not discarded by the speech activity detector. First, each frame is multiplied by Hamming window to reduce the discontinuities at the edges of the frame [119]. Second, the frame is transformed to the frequency domain using short-time Fourier transform. Since the Fourier transform of a real signal is a complex and symmetric signal, the signal is converted to a real signal by computing the square of its magnitude. The resulting signal is referred to as "speech spectrum". Then, motivated by the non-uniform frequency resolution of the human hearing, the speech spectrum is converted to a non-uniform frequency scale by applying a set of simulated triangular-shaped filters (spaced according to the Mel-scale [119]). Such filters compute the average spectrum around each center frequency with increasing bandwidths. Since speech signal was acquired trough a band-limited telephone, only the Mel-filters within the frequency range 300 Hz and 3138 Hz are retained for further processing. A non-linear transformation (logarithmic function) is applied to the resulting Mel-frequency spectrum. Finally, the Mel-frequency spectrum is projected on a discrete cosine basis. The coefficient that corresponds to the spectral energy is discarded from the feature vector.

The mel-frequency cepstrum coefficients are augmented with transitional spectral information often referred to as delta features [51, 144]. These features provide complementary information to the mel-frequency cepstrum and are less affected by communication channel effects. Delta features are computed using a first-order orthogonal polynomial temporal fit of each cepstral trajectory over a finite length window (in time) [57, 145]. Let $c_k(t)$ be the k^{th} mel-cepstrum coefficient from a frame of speech at

time t. The first-order orthogonal polynomial coefficient, or the generalized spectral slope (in time), denoted as $\Delta c_k(t)$, has the following form:

$$\frac{\partial \mathbf{c}_k(t)}{\partial t} \approx \Delta \mathbf{c}_k(t) = \frac{\sum_{i=-I}^{I} i\mathbf{c}_k(t+i)}{\sum_{i=-I}^{I} i^2} \tag{3.1}$$

where $2I + 1$ is the length of the window. The delta features are estimated over a 50 ms time interval (I=2 in Equation 3.1). The addition of the delta features to the mel-frequency cepstrum coefficients produces a 38-dimensional feature vector.

Finally, RASTA (RelAtive SpecTrA) filtering [146] is applied to compensate for the varying acoustic properties of the different telephone handsets or channels (e.g., land-line or cellular) encountered in training and testing speech data. Assume the case where the handset or even the channel that was used for collecting training data for a given speaker is switched in the testing phase. This change can affect the recognition performance because the speech from the new handset or channel is not represented in the training data. However, the characteristic of the sources of these non-speech components is often deterministic, which makes the effects of the speech signal predictable. For example, the frequency characteristic of a communication channel is often fixed or slowly varying in time, and it shows as an additive component in the logarithmic spectrum of speech (convolutional effect). In addition, the rate of change of these components in speech often lies outside the typical rate of change of the vocal tract shape. The RASTA filtering exploit these differences to reduce the effects of changes in the communication channel, by suppressing the spectral components that change more slowly than speech. This filtering technique also suppresses spectral components that change faster than speech, thereby reducing the effects of convolutional noise.

3.3.2 Speaker Modeling and Detection

Given the likelihood ratio detector framework described in Section 3.1, two different models are used to estimate the likelihoods: universal background model (UBM) and target-speaker model. Extensively used for modeling in text-independent speaker recognition applications [21, 46, 93, 94, 96, 147], each model is represented by a Gaussian mixture model (GMM).

Gaussian Mixture Models

A Gaussian mixture model describes the probability distribution of a given data set as a linear combination of several Gaussian densities. In this model, each *d*-dimensional random vector $\mathbf{x}$ is assumed to be drawn independently from a mixture density given by the equation

$$p(\mathbf{x}|\Theta) = \sum_{i=1}^{M} \varpi_i p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \qquad 0 \le \varpi_i \le 1 \text{ and } \sum_{i=1}^{M} \varpi_i = 1$$

where ϖ_i defines the mixing weight of the i^{th} Gaussian component (for all $i = 1, 2, \ldots, M$) given by the relative importance of each component in the density function, $p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ represents the i^{th} *d*-dimensional Gaussian component with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$ given by

$$p(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\},$$

and $\Theta = \{\varpi_1, \ldots, \varpi_M, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_M, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_M\}$ represent the mixture density parameters. The mixture density parameters are estimated using a maximum-likelihood approach.

The expectation maximization (EM) [148] is an iterative algorithm for finding the maximum-likelihood parameter estimates for the case of incomplete data, where in the mixture of Gaussians the probability of assigning a sample $\mathbf{x}_n$ to the i^{th} Gaussian component is unknown. It is assumed that the samples in $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ are independent and identically distributed (i.i.d.). Assuming a generative model of the data, each sample $\mathbf{x}_n \in X$ is generated by only one of the Gaussian components. The goal is to obtain the parameter values $\hat{\Theta}$ which maximize the likelihood of X given the data, as follows

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmax}}\, p(\mathbf{X}|\Theta) = \underset{\Theta}{\operatorname{argmax}} \prod_{n=1}^{N} p(\mathbf{x}_n|\Theta).$$

The maximum-likelihood estimate $\hat{\Theta}$ is the value of Θ that maximizes $p(\mathbf{X}|\Theta)$. Given an initial estimate of the mixture density parameters, the EM algorithm iterates the following two steps:

E-step: estimate the probability that $\mathbf{x}_n$ is assigned to the i^{th} mixture component given the current parameter estimate $\Theta^{(p)}$ and the data X using

$$p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right) = \frac{\varpi_i^{(p)} p\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_i^{(p)}, \boldsymbol{\Sigma}_i^{(p)}\right)}{\sum_{j=1}^{M} \varpi_j^{(p)} p\left(\mathbf{x}_n \middle| \boldsymbol{\mu}_j^{(p)}, \boldsymbol{\Sigma}_j^{(p)}\right)},$$

where $\Theta^{(p)} = \left\{\varpi^{(p)}, \boldsymbol{\mu}^{(p)}, \boldsymbol{\Sigma}^{(p)}\right\}$ represent the mixture density parameters after the p^{th} iteration.

M-step: re-estimate the mixing weights, means, and covariances of the Gaussians densities using the data set weighted by $p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right)$ to maximize the likelihood of the data, according to the following equations:

$$\varpi_i^{(p+1)} = \frac{1}{N} \sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right)$$

$$\boldsymbol{\mu}_i^{(p+1)} = \frac{\sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right) \mathbf{x}_n}{\sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right)}$$

$$\boldsymbol{\Sigma}_i^{(p+1)} = \frac{\sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_i^{(p+1)}\right)\left(\mathbf{x}_n - \boldsymbol{\mu}_i^{(p+1)}\right)'}{\sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \Theta^{(p)}, \mathbf{X}\right)}$$

There are several reasons for using a GMM in speaker recognition. First, individual mixture components may capture the characteristics of some set of acoustic classes. For example, speech can be viewed as a sequence of acoustic classes represented by a set of phonemes. Such phonemes can be realized differently across speakers due to physiological characteristics. A mixture component may represent the acoustic characteristics of a phoneme by encoding the average and the variation of the acoustic characteristics using the mean and covariance parameters, respectively. Second, a linear combination of Gaussian basis functions is capable of forming smooth approximations of arbitrarily shaped densities. However, the strong assumption that the observations are independent conceals the temporal aspects of the speech signal.

Universal Background Model (UBM)

Since the UBM must represent the feature space of all speakers except the target speaker, it is common to use a GMM with a considerable number of mixture components. In this work, the UBM has 2048 mixtures components with diagonal covariance matrix. Despite the large number of mixture components, the number of parameters that needs to be estimated is reduced by the use of diagonal covariance matrix. Moreover, diagonal covariances provide a simpler implementation and require less computational power than full covariance matrices. Besides, it has been shown that the modeling capability of using a set of full-covariance Gaussians can be equally achieved by using a larger set of diagonal-covariance Gaussians [93]. The UBM is estimated using the Switchboard II conversational, telephone speech corpus. The EM algorithm is initialized by selecting 2048 features vectors from the training data as the means of the Gaussians distributions. Then, the EM algorithm is iterated five times to estimate the UBM parameters.

Target-speaker Model

The target-speaker models are estimated by employing a Bayesian adaptation (also known as *maximum-a-posteriori* estimation) from the UBM using the target-speaker's training data [59]. Like the EM algorithm, the adaptation is a two-step estimation process. In the first step, the estimates of the sufficient statistics of the speaker's training data are computed for each mixture component in the UBM. This is identical to the E-step of the EM algorithm. Given the mixture density parameters of the UBM and the speaker's training data $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$, the sufficient statistic estimates of the target-speaker are computed using the posterior probability $p\left(\mathbf{x}_n, i \middle| \boldsymbol{\Theta}^{UBM}, \mathbf{X}\right)$, i.e., the probability of $\mathbf{x}_n$ being generated by the i^{th} mixture component of the UBM:

$$\gamma_i = \sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \boldsymbol{\Theta}^{UBM}, \mathbf{X}\right)$$

$$E_i(\mathbf{X}) = \frac{1}{\gamma_i} \sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \boldsymbol{\Theta}^{UBM}, \mathbf{X}\right) \mathbf{x}_n$$

$$R_i(\mathbf{X}) = \frac{1}{\gamma_i} \sum_{n=1}^{N} p\left(\mathbf{x}_n, i \middle| \boldsymbol{\Theta}^{UBM}, \mathbf{X}\right) \mathbf{x}_n \mathbf{x}_n'$$

where, for the i^{th} mixture component of the target-speaker model, γ_i is the mixing weight, $E_i(\mathbf{X})$ is the mean, and $R_i(\mathbf{X})$ is the diagonal covariance matrix. The mixture density parameters of the UBM are represented by $\Theta^{UBM} = \{\varpi^{UBM}, \boldsymbol{\mu}^{UBM}, \Sigma^{UBM}\}$. In the second step, the mixture density parameter estimates of the target-speaker model are derived by unsupervised *maximum-a-posteriori* (MAP) re-estimation of the mixture density parameters from the UBM, as follows

$$\hat{\varpi}_i = \left[\alpha_i^w \frac{\gamma_i}{N} + (1-\alpha_i^w)\varpi_i\right]\tau$$
$$\hat{\boldsymbol{\mu}}_i = \alpha_i^m E_i(\mathbf{X}) + (1-\alpha_i^m)\boldsymbol{\mu}_i$$
$$\hat{\boldsymbol{\Sigma}}_i = \alpha_i^v R_i(\mathbf{X}) + (1-\alpha_i^v)(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2$$

where τ is a scale factor to ensure that the mixing weights sum to unity, and the parameters α_i^w, α_i^m, and α_i^v reflect the respective confidence in the parameter estimates $\gamma_i, E_i(\mathbf{X})$, and $R_i(\mathbf{X})$. That is, if a mixture component has a low probabilistic count of speaker's data, γ_i, then the new parameter estimate will rely on the parameter estimates of the UBM. The confidence parameters are defined as

$$\alpha_i^\rho = \frac{\gamma_i}{\gamma_i + r^\rho}, \qquad \rho \in \{w, m, v\}$$

where r^ρ is a fixed relevance factor for parameter ρ.

In this work, only the means of the mixture components are adapted from the UBM [59, 147]. The adaptation process is iterated only once. The relevance factor is set to 16.

3.3.3 Scoring

The likelihoods for the target-speaker model and UBM are accumulated using only the five best scoring components identified for the UBM on the test vectors [59]. This is motivated by two reasons. First, when a GMM with many mixture components is used to estimate the likelihood value for a feature vector, only a few of the mixtures contribute significantly to the likelihood value. This suggests that the likelihood value can be approximated using only the top C best scoring mixture components. Second, the adapted Gaussian components maintain a correspondence with the mixtures of the UBM, so that

the feature vectors close to a particular mixture in the UBM will also be close to the corresponding mixture in the target speaker model.

3.3.4 System Performance

The EER of the acoustic baseline for 8-conversation training condition on Switchboard I is 0.7%. Figure 3.4 shows the DET curve for the acoustic baseline system for different training conditions on the 2001 NIST SRE.

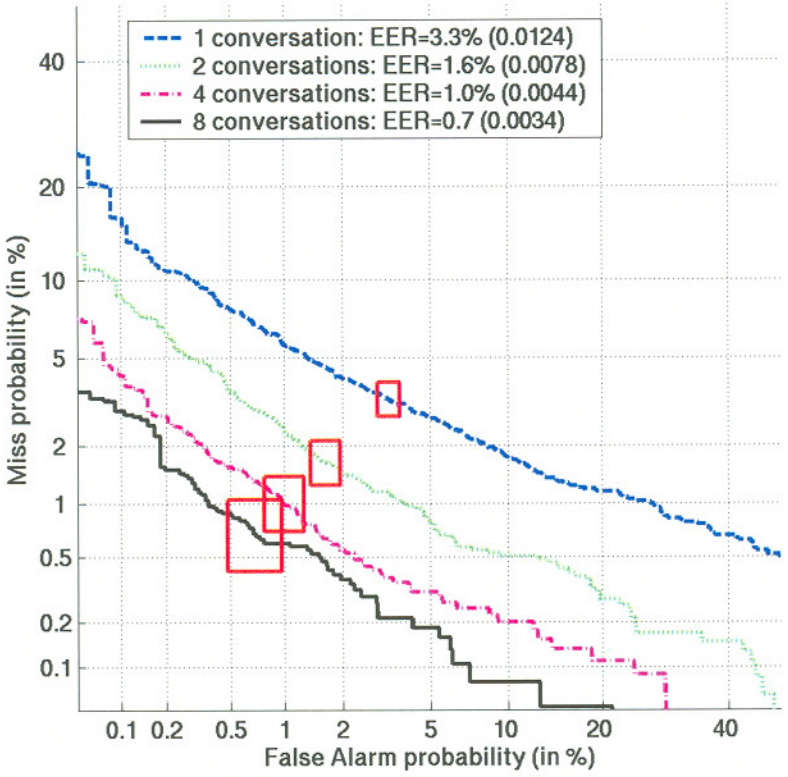

Figure 3.4: DET curves of the acoustic baseline system on the 2001 NIST SRE.

3.4 Prosodic Baseline System

The aim of the prosodic baseline system is to capture the characteristics of the F0 and short-term energy features distribution. This system is based on a likelihood ratio detector that uses GMMs for estimating the likelihoods as described in Section 3.3.2.

3.4.1 Feature Extraction

This section is divided into three parts. In the first part, we describe the process of fundamental frequency estimation. Then we describe the method used in this work to

estimate the fundamental frequency. Finally, the speech parameterization process is described.

Fundamental Frequency Estimation

The goal of fundamental frequency (F0) estimation is to obtain the period of the glottal excitation waveform. This waveform is generated by the periodic opening and closure of the vocal cords in the glottis while air is pushed through from the lungs. The periodic opening and closing happens for voiced speech only; otherwise, unvoiced speech is produced. Thus, the detection of voiced speech segments must be performed before estimating the F0. However, automatic F0 estimation is often a very difficult task for several reasons, such as:

- F0 changes with time;
- Glottal excitation waveform is not a perfect train of periodic pulses;
- Interactions between the vocal tract and the glottal excitation can alter the glottal excitation waveform (e.g., fast changes of the formant frequencies);
- Voicing is very irregular at voice onset and offset causing irregularities in the contour of adjacent periods;
- Unvoiced speech segments can be confused with low-level voiced speech;
- Pitch period can suddenly change and even halve (vocal fry), particularly at the end of stressed voiced speech segments.

Besides the inherent difficulties associated with F0 estimation, the accuracy can also be affected when the speech is transmitted through some transmission channel. For example, telephone systems introduce additive and convolutional noise to the speech and attenuate the F0 and formants.

Several approaches to automatic F0 estimation have been proposed [72]. They can be classified into three categories:

1. Time-domain techniques estimate the F0 directly from the speech waveform. These techniques search for patterns of gross speech waveform features (peak and valley, zero crossing, and auto-correlation measurements) from one period to the next. Despite the low computational complexity, these techniques are often noise sensitive.

2. Frequency-domain techniques use the property that periodicity in the time domain results in a series of impulses in the frequency domain at the F0 and its harmonics. These techniques are sensitive to the length of the analysis window so that their dynamic pitch range tends to be limited.
3. Hybrid techniques employ the previous two categories combined. For example, spectral flattening (so the effect of vocal tract is partially removed, leaving the glottal excitation spectrum) followed by autocorrelation.

The estimation of F0 usually is performed in three stages: pre-processing, F0-candidates estimation, and post-processing. The goal of the pre-processing stage is to remove any interfering signal component (e.g., background or channel noise, vocal tract influence, and DC offset), and to transform the signal to match the requirements of the method (e.g., low-pass filtering improves the F0 estimation because it removes the effects of the higher frequencies in the search for the F0 candidates). The goal of F0-candidates estimation stage is to generate a set of F0 candidates for the estimation of the true period. The last stage, post-processing, selects the best candidate and refines the F0 estimate.

Most of the methods for generating a set of F0 candidates employ the auto-correlation function [72, 149]. The reason is that the correlation of the speech signal will have a large peak corresponding to the pitch period. However, there are several problems related to methods based on the autocorrelation function. First, the function requires a long segment of speech to compute the correlation so that it can cover adequately the F0 range. For example, when the segment is short compared to the pitch period, rapid changes in formant frequencies can create peaks that are larger than the pitch; thus the procedure of selecting the highest peak as the pitch period will fail. Second, because of the large segments, rapid F0 changes can result in the loss of a clear peak in the correlation result.

Several methods that are not based on the auto-correlation method also present similar problems. For example, the cepstrum method uses the property that if a signal has periodicity then the cepstrum (the inverse Fourier transform of the short-time log magnitude spectrum) will have a peak at the fundamental period of that signal. However, the interval of speech which the spectrum must be computed is the same as that required for the auto-correlation function.

In this work, we use the RAPT – Robust Algorithm for Pitch Tracking proposed by Talkin [149] to estimate the F0 contour. Widely used for estimating F0, this method uses the normalized cross-correlation function (NCCF) to find a set of F0 candidates and dynamic programming to select the best F0 candidate or unvoiced hypotheses. The NCCF overcomes all of the shortcomings of the methods based on the auto-correlation function, i.e., it does not need a large segment of speech to estimate the correlation and is sensitive to rapid changes in the amplitude. This method is available through the get_f0 program in the ESPS/waves+ software package from Entropic Research Laboratory, Inc.[150].

RAPT - Robust Algorithm for Pitch Tracking

The algorithm performs a two-pass procedure based on the NCCF to find the F0 candidates. In the first pass, a set of coarse F0 candidates are estimated using the NCCF of a downsampled version of the speech signal. The downsampling is performed to reduce the computational cost of estimating the NCCF. First, the input signal is downsampled to a frequency F_{ds} determined by

$$F_{ds} = \frac{F_s}{round\left(\frac{F_s}{4F0_{\max}}\right)}$$

where F_s is the sampling frequency (Hz) of the original signal and $F0_{max}$ is the maximum F0 (Hz) to be searched. Before the downsampling, the signal is low-pass filtered using a symmetric finite impulse response (FIR) filter obtained by truncating the impulse response of an ideal $F_{ds}/2$ low-pass filter with a 5 ms duration Hanning window. Then, the NCCF is computed at all lags k for a speech frame from the downsampled signal with size w ms (also know as analysis window), every t ms (also know as analysis frame interval) using the following equation:

$$\phi_{i,k} = \frac{\sum_{j=m}^{m+n-1} s_j s_{j+k}}{\sqrt{e_m e_{m+k}}} \qquad m = iz;\ \ z = tF_{ds};\ \ n = wF_{ds};\ \ i = 0,1,\ldots,M-1$$

where s_m, $m=0, 1, 2,\ldots$, is a nonzero sampled speech signal with sampling frequency F_{ds}, i is the speech frame index for M speech frames, and

$$e_j = \sum_{l=j}^{j+n-1} s_l^2 .$$

Note that $\phi_{i,k}$ is bounded between –1 and 1, and $\phi_{i,k}$ is close to 1 for lags corresponding to integer multiples of the "period". In case of unvoiced regions, $\phi_{i,0}$ is 1 and $\phi_{i,k}$ is approximately zero for $k \neq 0$. All these properties hold independently of the amplitude of the speech signal. To reduce the computational cost, the NCCF is only estimated for a limited range of F0 values, [$F0_{min}$=50Hz, $F0_{max}$=500 Hz], so that $F_{ds}/F0_{max} \leq k \leq F_s/F0_{min}$. For each speech frame i, the maximum value of ϕ is defined as ϕ_{max}, and all local maxima that exceed 0.3 * ϕ_{max} are marked. Then, a more precise peak location and amplitude at F_s are obtained using parabolic interpolation on the three samples of ϕ defining each peak at F_{ds}. Finally, they are sorted by decreasing amplitude, and the top 19 candidates are saved. In the second pass, the NCCF function is computed using the original speech signal at F_s, but only for seven lags around the detected local maxima. Then, new local maxima are selected as the F0 candidates using the same process described in the first pass (the parabolic interpolation is not performed at this time).

Dynamic programming is used to obtain the optimum F0 and voicing estimation at each speech frame. Let I_i be the number of states proposed at frame i, where state 0 is unvoiced and the remainder ones are the F0 candidates. Thus, at each frame, I_i-1 possible fundamental frequencies (voiced states) and one unvoiced state will be proposed. For each frame i, a local cost value, $d_{i,j}$, for each state j and a transition cost, $\delta_{i,j,k}$, from the previous frame k are estimated for each state. There is also a cumulative penalty, $D_{i,j}$, associated with each state, which represents the best match with the state from the previous frame. The cumulative penalty is computed using the local cost and the smallest transition cost from the previous frame k using the following equation

$$\mathbf{D}_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} \{\mathbf{D}_{i-1,k} + \delta_{i,j,k}\}, \qquad 1 \leq j \leq I_i ,$$

with the initial conditions

$$D_{0,j} = 0, \qquad 1 \leq j \leq I_0; \qquad I_0 = 2 .$$

The state from the previous frame with the smallest transition to the current one is saved to define a trajectory to the best match in the previous frame. The optimum trajectory is defined by the trajectory with the smallest cumulative penalty. Since there is

an unvoiced state per frame, the voicing decision is also estimated in the trajectory. For a more comprehensive description of the F0 estimation algorithm, the reader is referred to Talkin's work [149].

Parameterization

The F0 and energy features are estimated every 10 ms from the speech signal. Since the distribution of F0 have a lognormal characteristic [132] and intensity is usually measured on the logarithmic scale, both features are represented on the logarithmic scale. Then, delta features are appended to the feature vector to characterize transitional information of each feature contour. The delta features are estimated over a 50 ms time interval using Equation 3.1. Therefore, the addition of the delta features generates a new 4-dimensional feature vector.

Since F0 cannot be estimated from unvoiced speech regions, only the feature vectors extracted for voiced speech regions are used in the speaker modeling. In addition, the speech activity detector used for the acoustic baseline system is employed to discard feature vectors extracted from silence and noisy regions.

The feature vectors at the beginning and ending of every voiced speech region are discarded to avoid discontinuities in the derivative computations. Given the discontinuities in the F0 contour created by the non-existence of F0 values for unvoiced speech region, the delta features around the boundary between voiced and unvoiced speech regions are not correctly estimated. Since delta features are estimated over a 50 ms time interval, the delta features in the first and last two feature vectors of each voiced speech region are discarded.

3.4.2 Speaker Modeling

Gaussian Mixture models are used to estimate the likelihoods for the UBM and the target speaker models, as described in Section 3.3.2. The UBM is a 512-component GMM trained with gender-balanced speech from cross-validation partitions not under test. Only the mean vectors of the target-speaker models are adapted from the UBM. The relevance factor is set to 16. The score is computed using only the five best scoring mixture components per frame.

3.4.3 System Performance

Figure 3.5 shows the DET curve for the prosodic baseline system on different training conditions on the 2001 NIST SRE. The EER of the prosodic baseline for 8-conversation training condition on Switchboard I is 15.2%.

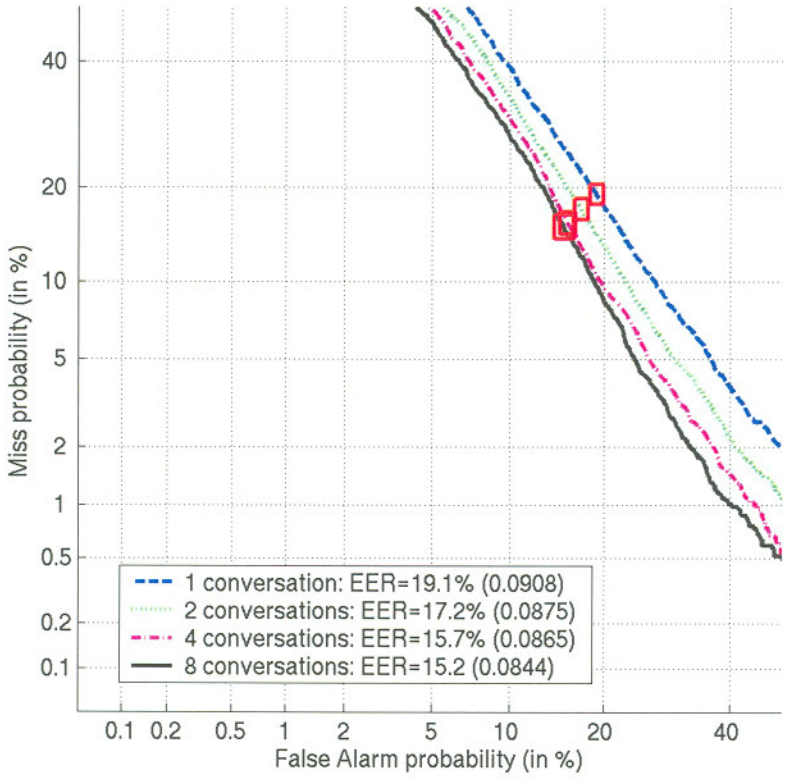

Figure 3.5: DET curves of the prosodic baseline system on the 2001 NIST SRE.

Given the small number of features (2 static and 2 delta, compared to the acoustic baseline's 19 mel-cepstrum coefficients plus 19 deltas), the limitation to only voiced frames, the noisiness of the pitch values due to halving and doubling of F0 estimates, and the simpler modeling technique (512 mixtures compared to the acoustic baseline's 2048 mixtures), the prosodic baseline provides a reasonable performance.

3.5 Fusion System

The goal of information fusion is to exploit the combination of different information sources such that the combination result is qualitatively and quantitatively (in terms of accuracy, reliability, or robustness) better than the individual sources [151-153]. Information fusion is analogous to the cognitive process used by humans to combine

information (e.g., sights, sounds, smells, tastes, and touch) from different sensors (e.g., ears, eyes, nose, and mouth) to make inferences about the external world, such as, assessment of the surrounding environment and identification of threats [151, 153, 154].

Since speaker recognition is based on a pattern recognition framework [45], the information fusion can be employed on several levels of the recognition process [151]. The fusion can take place at the raw data level (prior to feature extraction), at the feature extraction level (prior to classification), at the classification level (prior to decision), or at the decision level (after the decision has been made). The choice of when performing fusion depends on the type of sensor data and the data processing along the recognition process.

3.5.1 Classification-level Fusion System

In this work, a classification-level fusing system is used to combine different speaker recognition systems. In this type of fusion system, the likelihood scores produced by the likelihood ratio test from each system are combined into one score, which is then used by the decision module to make an acceptance or rejection decision. There are several reasons for fusing likelihood scores. First, most of the speaker recognition systems produce a likelihood score. Second, the dimensionality of the input data of the fusion system depends only on the number of systems that are going to be fused. Finally, since the fusion does not cause any change of the feature space or classification method, the fusion system is transparent to the fused systems.

Assuming that the relationship between different sources are nonlinear, the fusion system uses an artificial neural network (ANN) to combine the likelihood scores [15, 154, 155]. The advantages of the ANN approach are that it does not require any knowledge about the feature distribution, and the reliability of each information source is incorporated into the classification process. This approach treats the likelihood scores of all system as a feature vector, and the fusion problem as a pattern classification process [154]. It partitions the space into the regions corresponding to target and impostor speakers.

The neural network is a single-layer perceptron network [156, 157] depicted in Figure 3.6. The neural network has a layer consisting of inputs for each system, x_i, $i=1, 2,$

..., n, (and a bias term, $x_0=1$), no hidden layer, and an output layer with two sigmoidal output nodes: target, o_1, and impostor, o_2. Before feeding the likelihood scores to the network, the scores are normalized to zero mean and unit standard deviation using parameters derived from the training data.

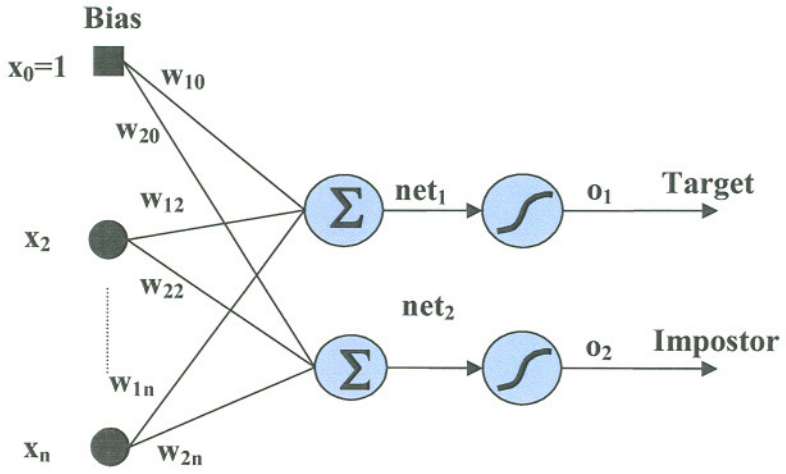

Figure 3.6: Single-layer perceptron network architecture.

A single-layer perceptron network takes a vector of likelihood scores as the input data, calculates a linear combination of these inputs, and then computes its outputs using a continuous function. More precisely, given inputs $\{x_0, x_1, x_2, \ldots x_n\}$, the network outputs are estimated using

$$o_j(x) = \phi\left(\sum_{i=0}^{n} w_{ji} x_i\right), \qquad j = 1,2$$

where j is the output node, w_{ji} the weight associated with the i^{th} input to the output node j, and $\phi(.)$ is the nonlinear continuous sigmoidal function or, alternatively, the logistic function,

$$\phi(y) = \frac{1}{1-e^{-y}}.$$

The backpropagation algorithm is used to learn the weights w_{ji} for the single-layer perceptron network. It employs gradient descent to minimize the squared error between the network output values and the target values t_k, $k=1, 2$, for these outputs. In this work, there are two possible types of target values. For the likelihood scores from the target trials, the target values are $t_1=1$ and $t_2=0$. Otherwise, the target values are $t_1=0$ and $t_2=1$.

In each iteration of the backpropagation algorithm, the weights are modified according to the following equation

$$w_{ji}(n) = w_{ji}(n-1) + \Delta w_{ji}(n)$$

where

$$\Delta w_{ji}(n) = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}(n-1).$$

Here $\Delta w_{ji}(n)$ is the weight update performed during the n^{th} iteration, η is a positive constant called the learning rate, $0 \leq \alpha \leq$ is a constant called the momentum, and δ_k is the error term for the output node *k* given by

$$\delta_k = o_k(1-o_k)(t_k - o_k)$$

The learning rate controls the adjustment applied to the weight. The momentum smoothes the weight estimate by controlling the influence of the weight update from the previous iteration on the current weight update. The learning rate η is set to 0.2 and the momentum α is set to 0.6. The training process iterates 20 times over the training data to estimate the input.

A single-layer perceptron network is trained for each number of training conversations from a given partition using the remaining partitions as the training data. Assume that we want to fuse the scores of a set of systems. For each partition, a neural network is trained for each evaluation condition (i.e., number of training conversations: 1, 2, 4, 8 and 16). The data used for training the neural network comes from the remaining partitions for the same evaluation condition. Therefore, a total of 30 neural networks (5 training conditions times 6 partitions) is trained for fusing systems on the 2001 NIST SRE, and 50 neural networks (5 training conditions times 10 partitions) for fusing systems on the 2002 NIST SRE.

Once a neural network is trained for a given partition and evaluation condition, the final score for each trial is estimated using the following equation

$$Score_{Fusion} = \frac{o_1 + (1-o_2)}{2},$$

where o_i, $i=1, 2$ are the values from the output nodes [156].

Chapter 4

Modeling Pitch and Intensity Dynamics

Most speaker and language recognition systems do not represent adequately the realization of prosodic phenomena when incorporating prosodic information. The incorporated prosodic information often disregards the temporal aspects or the relationship between prosodic features (e.g., F0, intensity, and duration). In this chapter, we describe a new approach for modeling the variations in the temporal aspects and interaction between F0 and short-term energy contours. We show that this approach can capture prosodic differences among speakers. Basically, the approach segments the speech signal at the points where there is a change in the dynamics of the F0 and short-term energy contours. Then, each segment is labeled with a class that represents the characteristics of dynamics of both contours within the segment. Finally, a simple statistical model is built from the sequence of segment classes. Section 4.1 introduces the assumptions of the approach for modeling prosodic differences. Sections 4.2 and 4.3 present two methods for segmenting the speech signal. Section 4.4 shows that the modeling approach provides complementary information to the acoustic and prosodic baseline systems. Section 4.5 demonstrates that, without any parameter tuning, the modeling approach provides similar performance on a different evaluation data. Section 4.6 analyzes the effects of amount of training data, speaker demographics, and handset on the performance of the modeling approach for speaker recognition. Finally, some extensions to the new prosodic modeling are presented in Section 4.7.

4.1 Prosodic Information Modeling

The acoustic realization of prosodic phenomena can be observed and quantified using F0, intensity, and duration. However, for the purpose of prosodic information representation, the focus of this work is on the modeling of patterns of variations in the F0 and intensity contours. One reason for modeling both features is that they have long been acknowledged to exhibit a high degree of interdependence [33, 39-41]. For example, F0 may be raised by increasing vocal fold tension, by increasing subglottal pressure, or by a combination of the two. Consequently, such parameters are going to contribute differently to the realization of different intonation, stress, or rhythm patterns. Since the problem with duration is not so much in measuring as in determining the points at which to perform the measurements [158], the duration feature is not directly exploited in the modeling. Besides, one of the goals of this work is to provide a representation of prosodic information that is language independent, i.e., a method that does not require knowledge about the segmental structure (e.g., phones and syllable) of a given language to describe prosodic information.

The hypothesis is that different speakers or languages may be characterized by different intonation, stress, or rhythm patterns produced by the changes in F0 and in intensity features. Therefore, the combination of F0 and intensity gestures (i.e., falling and rising) and duration that characterizes particular prosodic gestures are useful for describing speaker- and language-specific information. That is, the prosodic information in an utterance is described as sequence of elementary patterns representing the joint state of the dynamics (i.e., falling and rising) of F0 and intensity contours, and their respective duration. Assuming that there are two types of F0 and intensity gestures, the combination of the gestures from both features produces four possible joint-state classes: 1) rising F0 and rising intensity, 2) rising F0 and falling intensity, 3) falling F0 and rising intensity, and 4) falling F0 and falling intensity. Since unvoiced speech regions do not produce any F0 value, a fifth joint-state class is used to represent such regions.

Once the sequence of joint-state classes is estimated from the speech signal, several parameters can be computed from each segment defined by a joint-state class. Since the rhythm and tempo are related to duration patterns in speech, the duration of the segments is incorporated into the representation to capture such patterns. For example,

the duration of the joint-state class segments are shorter for a speaker with a fast speaking rate than a speaker with a slow speaking rate, since the former produces more rising and falling F0 gestures (i.e., pitch accents) than the latter.

4.2 F0-based Joint-state Class Estimation

In this section, we present an approach to estimate the joint state (i.e., rising, or falling) of the dynamics of the F0 and short-term energy contours conditioned to the changes in the F0 contour. This approach uses a piecewise linear fit of the F0 contour to segment and convert the speech signal into a sequence of joint-state classes.

The joint-state classes estimation is divided into four steps: 1) detect changes in the dynamics of the F0 contour, 2) segment both contours at the points (in the time dimension) of the detected changes in the dynamics of the F0 contour, 3) compute the rate of change of the energy contour for each segment, and 4) label each segment according to the dynamics of both contours. Details of the estimation steps are described next.

1st step – Detect changes in the dynamics of the F0 contour

A F0 close-copy stylization algorithm is used to detect the changes in the dynamics of the F0 contour. The close-copy stylization is a synthetic approximation of the F0 contour that is perceptually indistinguishable from the original [74, 159]. There are two reasons for using a stylization algorithm. First, the contour approximation reduces the noise introduced by the pitch tracker and micro-intonation effects that hide the speaker's intended pitch movements. Second, the algorithm approximates the F0 contour using the smallest possible number of straight-line segments, while preserving the speaker's intended pitch contour. Thus, the changes in the dynamics of the F0 contour are defined at the boundaries of the straight-line segments.

The F0 contour is stylized using a method based on a piecewise linear model, proposed by Sonmez et al. [74]. The stylized F0 contour from this method has been shown to be perceptually (listening tests) equivalent to actual F0 contours that contain micro-intonation. The method, which is performed separately on every conversation side, is divided into three steps. First, the F0 contour of a conversation side is estimated using Talkin's method [149] described in Section 3.4.1. Second, the F0 contour is filtered using

a median filter of order 5, i.e., the output at time n depends on the input at time n and the previous and the following 2 samples. The aim of filtering the F0 contour is to reduce the pitch tracker irregularities. Finally, a piecewise linear model is fitted to the median filtered F0 contour. For each voiced region, a piecewise linear model (splines method) is fit to the median filtered estimated (log) F0 values. The splines method approximates a function $f(x)$ using a number of straight lines $f_i(x)$, each valid over a sub-interval of the contour. Let $f(x)$ be the function that represents the voiced region contour, $\{(x_i, y_i)\}_{i=0}^{m}$ be a set of data points with $x_{i-1} < x_i$, for $i = 1, \ldots, m$, and $y_i = f(x_i)$. The parameters $\{x_i\}_{i=0}^{m}$ are referred to as the knots of the curve, and the x_0 and x_m are set to the beginning and ending points of the voiced region, respectively. The approximation function is given by

$$\tilde{f}(x) = \begin{cases} f_1(x) = a_1 x + b_1, & x_0 \le x \le x_1 \\ f_2(x) = a_2 x + b_2, & x_2 \le x \le x_3 \\ \vdots & \vdots \\ f_m(x) = a_m x + b_m, & x_{m-1} \le x \le x_m \end{cases}$$

were a_m and b_m are respectively the slope and the intercept of the line defined by (x_i, y_i) and (x_{i+1}, y_{i+1}). The number of data points is chosen proportional to the duration of the voiced region. The set of data points $\{(x_i, y_i)\}_{i=0}^{m}$ are estimated by minimizing the mean square error between the approximation function $\tilde{f}(x)$ and $f(x)$ on the voiced region. Figure 4.1 shows an example of F0 stylization of the utterance "I'm not there" (excerpt from Switchboard I), which has two voiced regions. The dotted curve in the F0 plot shows the piecewise linear approximation of the F0 contour for each voiced region. The vertical dotted-bars in the F0 plot of Figure 4.1 represent the estimated $\{x_i\}_{i=0}^{m}$ for each voiced region.

2nd step – Segment both contours at the detected changes in the F0 contour dynamics

Both contours are segmented at the knots of the curve estimated from all voiced regions. Let $\{b_i\}_{i=1}^{k}$, with $0 < b_1 < b_2 < \ldots < b_k < N$, be the union between the estimated knots x_i from all voiced regions, and $f_0(x)$ and $e(x)$ are respectively the F0 and energy contour

functions, whose domains are a closed and bounded interval [0, N]. Then the segments are defined on each of the intervals

$$(0,b_1),(b_2,b_3),\ldots,(b_{k-1},b_k),(b_k,N).$$

The vertical dotted-bars in the F0 plot of Figure 4.1 (representing the knots from each voiced speech region) are grouped together to generate the segment boundaries, $\{b_i\}_{i=1}^{k}$.

3rd step – Compute the rate of change of the energy contour within each segment

The rate of change of the energy contour within each segment is approximated by fitting a straight line to the energy samples within the segment. The dotted lines in the energy plot of Figure 4.1 represent the linear, least-squares fit estimated for each segment.

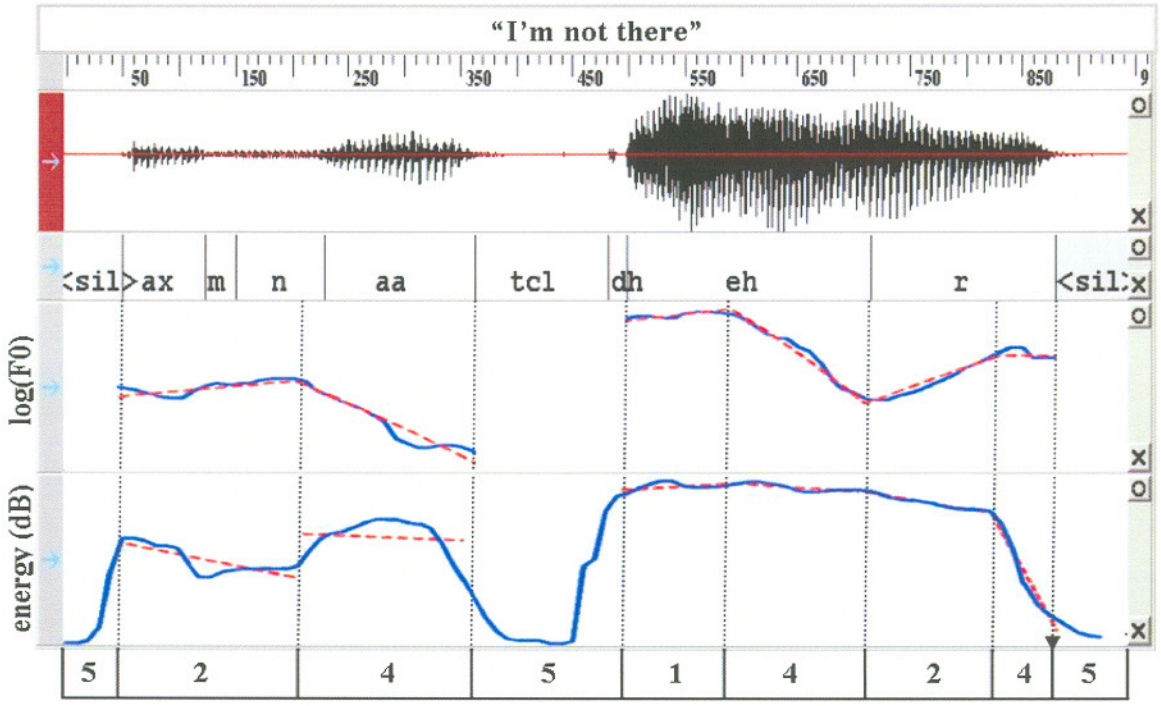

Figure 4.1: Example of joint-state class sequence estimation using F0 and energy contours of the utterance "I'm not there" from a conversation in Switchboard I.

4th step – Label each segment according to the dynamics of both contours

Each segment is labeled with a joint-state class according to the direction of the rate of change of both contours within the segment. The sign of the slope of the fitted line for each segment is used as the state of the F0 and energy contour over that segment. Thus, the state of a contour within a segment is defined as rising when the slope is positive, and falling when the slope is negative. Since the slope for unvoiced regions are not estimated,

these regions are converted into one specific class. Table 4.1 lists the 5 possible classes that reflect the joint-state of both contours within a segment. According to Table 4.1, the resulting sequence of joint-state classes for the utterance "I'm not there" is shown at the bottom of Figure 4.1.

Table 4.1: Joint-state classes of F0 and energy contours within a segment.

Class	Joint-state dynamics
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

4.2.1 Integrating in Duration Information

Since the duration of each segment class is the result of several factors, such as, speaking rate, rhythm, or stress, the duration of each segment is incorporated into the sequence of joint-state classes. For each joint-state class, an extra label representing the segment duration is added to the joint-state class symbol.

The duration, measured in number of analysis frames, is quantized into three levels: "Short," "Medium," and "Long". Two quantizers are estimated using the cumulative distribution function (CDF) of the duration of all segments. The 33rd and 66th percentiles of the CDF are selected as the two quantizers, q_1 and q_2. Then, a "Short" label is added to the segment class with duration equal or less than q_1 frames, a "Medium" label is added to the segment class with duration more than q_1 and equal or less than q_2 frames, and a "Long" label is added to the remainder segment classes. Using a held-out data set from Switchboard I, the estimated quantizers are 4 and 8 frames (or 40 and 80 ms, for a 10 ms analysis frame interval). The addition of the duration of the example in Figure 4.1 produces the following sequence: **5S 2L 4L 5L 1L 4L 2L 4M 5M**. Thus, the addition of the duration to the joint-state classes produces a new set of 15 classes to represent the dynamics of the F0 and energy contours.

4.2.2 Speaker Modeling and Scoring

There are several techniques for exploiting long term patterns in the sequence of joint-state classes. Some of the techniques include discrete hidden Markov model (DHMM),

binary-decision trees, and *n*-grams. In order to select the modeling technique, we run several preliminary experiments using modeling techniques such as DHMM [118], binary-decision trees [78], and *n*-grams [27]. The experiments were performed on the 2001 NIST SRE and each conversation side is represented by a sequence of 5 possible joint-state classes from Table 4.1 (duration is not incorporated into the classes). The modeling using a 10-state ergodic DHMM (2 states for each symbol) achieves the best performance (EER) for 8-conversation training, but it achieves the worst performance for 1-conversation training. As expected, the characteristic sparsity of the joint-state class representation affects more the methods that require higher number of parameters to describe a model. The results of binary-decision tree [78] and *n*-gram techniques using comparable configuration (both methods estimate probabilities of two symbols occurring in a given sequence) show similar performance under different training conditions. However, as the complexity (represented by the number of parameters) of the models is increased, the performance of the binary decision tree shows a similar result to the DHMM.

Given its consistent performance and simplicity, *n*-gram modeling is used to model the sequence of joint-state classes. Widely used in speech recognition systems [45], *n*-gram modeling provides a viable and effective approach for modeling the speaker's usage of the proposed prosodic classes [27, 160].

N-gram Modeling

Given that the speech signal is converted into a sequence of discrete symbols, a speaker model is defined as a probability distribution Pr(**S**) over sequences **S** that attempt to capture how likely a string **S** is generated by a given speaker. Let $\mathbf{S} = \{s_1, s_2, \ldots, s_m\}$ be a sequence of m symbols, where $s_i \in \Phi$, for $i = 1,2,\ldots,m$, and Φ denotes the set of all possible symbols. The probability distribution Pr(S) is expressed as

$$\Pr(S) = \prod_{i=1}^{m} P(s_{i-n+1}, s_{i-n+2}, \cdots, s_{i-1}, s_i),$$

where $P(s_{i-n+1}, s_{i-n+2}, \cdots, s_{i-1}, s_i)$ is the probability of n symbols happening in that order.

Several approaches in speaker recognition [27, 29, 156, 160] have been using the joint probability of symbols to estimate the probability distribution of a given sequence of

symbols. The maximum likelihood estimate of the probabilities $P(s_{i-n+1}, s_{i-n+2}, \cdots, s_{i-1}, s_i)$ over some training data is estimated by simply computing the frequencies of the sequence $s_{i-n+1}, s_{i-n+2}, \cdots, s_{i-1}, s_i$ occurring in the training data. For example, in bigram models, the joint probability is estimated as follows

$$P(s_{i-1}, s_i) \approx \frac{C(s_{i-1}, s_i)}{\sum_{j=1}^{m} C(s_{j-1}, s_j)} \tag{4.1}$$

where $C(\bullet)$ is the number of times that the parameters appear in the training data. This n-gram modeling approach is also referred to as "bag of n-grams" [27].

Likelihood Ratio Test Scoring

The speaker detection score is computed using a conventional log-likelihood ratio test between the target-speaker model and the UBM averaged over all n-gram types [27, 29]. Let S_{Test} be the sequence of symbols, which is referred to as test sequence, produced by a given speaker, which contains $|S_{Test}|$ symbols. The averaged log-likelihood ratio LLR_i for the i^{th} speaker is written as

$$\begin{aligned}
LLR_i &= \frac{\log\left(\frac{P_{TS_i}(S_{Test})}{P_{UBM}(S_{Test})}\right)}{|S_{Test}|} \\
&= \frac{\log\left(\frac{\prod_{j=1}^{|S_{Test}|} P_{TS_i}(s_{j-n+1}, s_{j-n+2}, \cdots, s_{j-1}, s_j)}{\prod_{j=1}^{|S_{Test}|} P_{UBM}(s_{j-n+1}, s_{j-n+2}, \cdots, s_{j-1}, s_j)}\right)}{|S_{Test}|} \\
&= \frac{\sum_{j=1}^{|S_{Test}|} \log\left(P_{TS_i}(s_{j-n+1}, s_{j-n+2}, \cdots, s_{j-1}, s_j)\right) - \sum_{j=1}^{|S_{Test}|} \log\left(P_{UBM}(s_{j-n+1}, s_{j-n+2}, \cdots, s_{j-1}, s_j)\right)}{|S_{Test}|} \\
&= \frac{\sum_{j=1}^{|S_{Test}|} \log\left(P_{TS_i}(s_{j-n+1}, s_{j-n+2}, \cdots, s_{j-1}, s_j)\right) - \log\left(P_{UBM}(s_{j-n+1}, s_{j-n+2}, \cdots, s_{j-1}, s_j)\right)}{|S_{Test}|}
\end{aligned}$$

where $P_{TS_i}(S_{Test})$ is the probability that the target speaker i generated the sequence S_{Test} and $P_{UBM}(S_{Test})$ is the probability that the sequence S_{Test} was produced by any speaker,

but the target speaker. Let Ω_{Test} be the set of all possible n-gram types in the sequence S_{Test}. Instead of summing over every n-gram, the summation can be rewritten so that the log-likelihood ratio is actually computed only for each n-gram type ϕ in the test sequence:

$$LLR_i = \frac{\sum_{\phi \in \Omega_{Test}} C_{Test}(\phi)\left[\log\left(P_{TS_i}(\phi)\right) - \log\left(P_{UBM}(\phi)\right)\right]}{\sum_{\phi \in \Omega_{Test}} C_{Test}(\phi)}$$

where $C_{Test}(\phi)$ is the number of occurrences of the n-gram type ϕ in the test sequence. The joint-probability estimates for a given target speaker and the UBM are given by:

$$P_{TS_i}(\phi) = \frac{C_{TS_i}(\phi)}{\sum_{\tau \in \Omega_{TS_i}} C_{TS_i}(\tau)}$$

$$P_{UBM}(\phi) = \frac{C_{UBM}(\phi)}{\sum_{\tau \in \Omega_{UBM}} C_{UBM}(\tau)}$$

where Ω_{TSi} and Ω_{UBM} represent the set of all possible n-gram types of the i^{th} target speaker (TS_i) and the UBM.

4.2.3 Speaker Detection Results

In this section, we present the performance of the joint-state classes on the 8-conversation training condition of the 2001 NIST SRE. Speaker models are represented using bigrams models of the sequence of joint-state classes estimated from the training data of the respective speaker. Since no special processing is performed for regions with halving and doubling pitch effects, the stylization algorithm can produce small segments for those regions during the fitting process. Therefore, all segments smaller than 30 ms are removed from the sequence of joint-state classes. To avoid the modeling of classes across utterances, we place **<bound>** symbols around each utterance. An utterance is defined as a period of time when one speaker is speaking and that there is no silent gap for more than 0.5 second (the detection of utterances is out of the scope of this work). Figure 4.2 shows the DET curves for the speaker modeling derived from the F0 and energy contours.

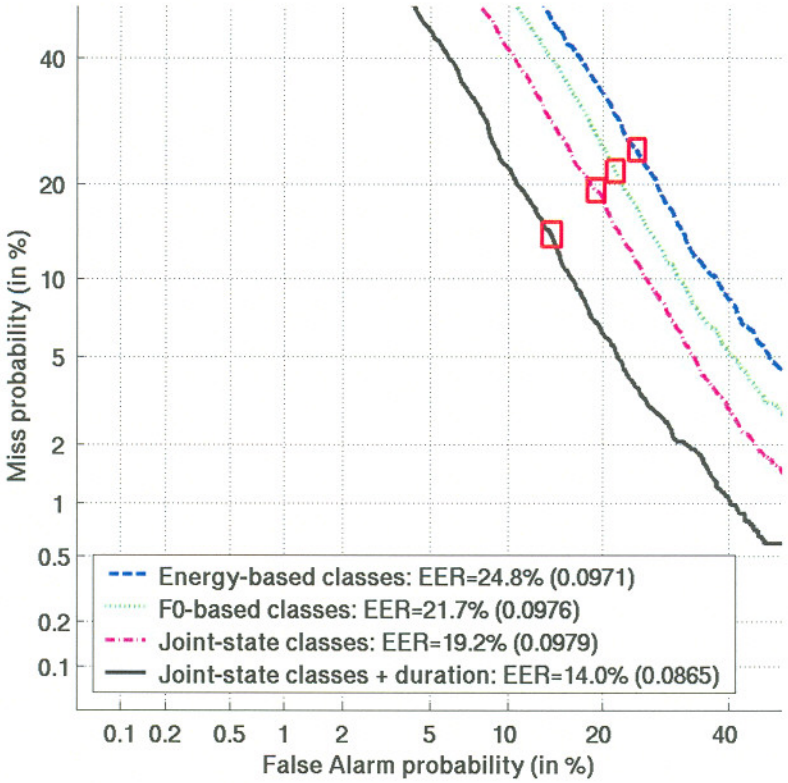

Figure 4.2: DET curves for the F0 and short-term energy alone and PWL-based joint-state classes.

The EER of the bigram modeling of joint-state classes is 19.2%. Compared to the 15.2% of the prosodic baseline, this result is very promising because the modeling is simpler, and the system only uses a sequence of five possible symbols (joint-state classes) of two streams (delta estimates of F0 and short-term energy) to represent the speech signal. The addition of the duration label to the joint-state class yields an EER of 14% (a 27% relative improvement over the joint-state classes without the duration label). This result shows that not only the sequence of joint-state classes is speaker dependent, but also that the duration provides speaker-dependent information. Note that the difference in performance between the 15 classes and the prosodic baseline is statistically significant.

The bigram modeling of the joint-state classes can capture more speaker-dependent information than the modeling of the contours alone. Using the segmentation obtained from the joint-state classes, we created a sequence of three possible classes (**r**ising, **f**alling, and **uv**oiced) for each contour. That is, each speaker model is estimated using a sequence of gestures from one of the contours. For example, the sequence of classes for the F0 contour in Figure 4.1 is: "**uv r f uv r f r f uv**". Then, using a bigram

modeling, we evaluate the performance of each stream separately. The EER of the systems based on the F0 and short-term energy contours are 21.7% and 24.8%, respectively. These results show the importance of the temporal information in representing differences in prosodic phenomena realization. A comparison between these performances and the performance obtained from the joint-state classes modeling shows that the interaction between both contours carries speaker-specific information. However, the joint-state modeling performance could be a result of the larger number of classes (5 joint-state classes versus 3 single-contour classes). Both single-contour classes are fused by averaging the likelihood scores for each trial. The EER of the fused scores is 22%, which shows that the joint-state class modeling captures the interaction between both contours.

Although environmental and channel effects restrict the use of energy features for speaker recognition, the energy-based modeling seems to provide discriminatory information about speakers. The performance of the energy-based modeling has a relative difference of 14% from the F0-based modeling. This result shows that even temporal patterns of the energy dynamics convey speaker-dependent information.

4.2.4 Discussion

The results show that the joint-state modeling of the F0 and short-term energy contours can provide other levels of information not previously exploited by the methods that employ the distribution of such prosodic features. Given that differences in the realization of prosodic phenomena are likely to produce differences in the sequence of joint-state classes, the joint-state classes provide the means for capturing prosodic differences among speakers.

Even though most prosodic phenomena are conveyed through changes in pitch and intensity, the proposed segmentation method favors the changes in the dynamics of the F0 contour in detriment of the changes in the dynamics of the energy contour. Figure 4.3 shows the segmentation on the energy contour and the linear fit used to estimate the rate of change within a segment. Changes in the energy trajectory (as pointed by the arrows in Figure 4.3) are not used in the estimation of the segments.

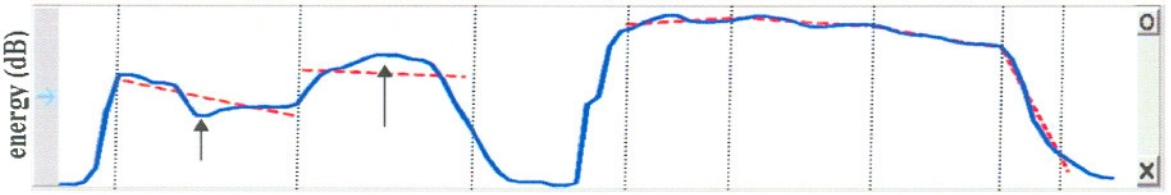

Figure 4.3: Short-term energy contour segmentation. The arrows point the changes in the contour dynamics that are not used in the joint-state class estimation.

Despite the distinct differences between voiced and unvoiced segments, the duration quantization scheme is applied to both types of segments. Since the segmentation is only performed over the voiced regions, the classes related to voiced regions present a smaller time interval than the class related to unvoiced segments. Besides, short pauses between words are labeled as unvoiced segments by the F0 estimation algorithm. Therefore, a more precise quantization approach is required to generate the labels for the duration of the segment classes.

4.3 Delta-based Joint-state Class Estimation

The approach described in this section uses the changes in the dynamics of both F0 and energy contours to segment the speech signal. In addition, the rate of change is used to detect changes in the dynamics of the prosodic feature contours.

The joint-state classes estimation is divided into four steps: 1) compute the rate of change for each contour, 2) detect the points where the dynamics of the contours changes, 3) generate new segments using the detected points, and 4) convert the segments into a sequence of symbols that represent the dynamics of both contours.

1st step – Compute the rate of change for each contour

The rate of change is approximated using delta features, defined in Equation 3.1. The delta features have been used in automatic speech recognition systems to approximate the short-term dynamics of temporal trajectories [51, 144]. Besides its extensive use, the delta features offer several advantages over the piecewise linear fit approach. First, the amount of detail of the rate of change is easily defined by increasing or decreasing the number of consecutive samples used in the delta estimation. Second, delta features are simpler to compute. Finally, it does not restrict the number of changes in the contour

dynamics within a voiced segment (the number of changes in the contour dynamics is bounded by the length of the voiced segment).

The F0 contour for a conversation side is estimated using Talkin's method [149], which is described in Section 3.4.1. Since the F0 contour can have discontinuities due to unvoiced speech regions and halving or doubling effects, the delta features estimation is performed differently for the F0 contour. First, delta features are only estimated for voiced speech regions. Second, delta features are not estimated across transitions between halving/doubling and normal F0. Figure 4.4 shows an example of discontinuities (represented by circles) of a F0 contour and the intervals used for estimating the delta features. In this example, the delta features are estimated for the two voiced speech regions located between (t_i, t_{i+2}) and (t_{i+3}, t_{i+4}) intervals. In the first voiced speech region (t_i, t_{i+2}), the delta features are estimated over two intervals, (t_i, t_{i+1}) and (t_{i+1}, t_{i+2}), due to the discontinuity created by the doubling F0 effect in the region.

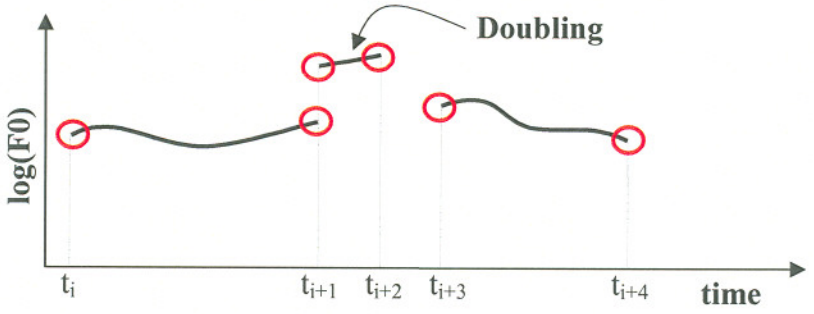

Figure 4.4: Example of delta feature estimation of the F0 contour. The estimation is not performed across discontinuities (represented by circles) such as voiced/unvoiced speech regions (e.g., point at t_i, t_{i+2}, t_{i+3}, and t_{i+4}) and doubling/halving effects (e.g., point at t_{i+1}).

A lognormal tied-mixture model is used to estimate the probabilities of halving and doubling F0 at the frame level. The doubling of F0 occurs when the system detects sub-harmonics multiples of the "true" pitch as the fundamental frequency. The halving of the F0 period can be either caused by the speaker (Shriberg et al. [161] show that glottalization or creaky voice is high correlated with halving) or by the pitch tracking system (pitch halving can occur when the F0 estimator detects longer pitch periods than the normal pitch). The estimation of the halving and doubling probabilities is based on the approach proposed by Sönmez [132]. Let f_0 be the true F0 contour of a given speech

sample, $\tilde{f}_0$ be the estimated F0 contour, which can have halving and doubling F0 values. The relationship between $\tilde{f}_0$ and f_0 can be defined as follows

$$\tilde{f}_0 = g(f_0) = g(\boldsymbol{f_0})$$

where g(.) is a probabilistic mapping

$$g(x) = \begin{cases} \frac{x}{2} & \text{with probability } \beta \\ x & \text{with probability } \alpha \\ 2x & \text{with probability } 1-\alpha-\beta \end{cases},$$

and $\log(f_0)$ has a normal distribution with mean μ and variance σ^2, $N(\mu, \sigma^2)$. This results in the following lognormal tied-mixture (LTM) model for the estimated F0:

$$\log(\tilde{f}_0) \approx LTM(\alpha, \beta, \mu, \sigma) = \beta N(\mu - \log(2), \sigma^2) + \alpha N(\mu, \sigma^2) + (1-\alpha-\beta) N(\mu + \log(2), \sigma^2)$$

where α is the probability of normal pitch and β is the probability of halving F0. The first term models halving F0, the second term models the normal F0, and the third term models the doubling F0. The parameter vector (α, β, μ, σ) is estimated using the expectation-maximization algorithm over a given conversation side. Then, the model is evaluated for each F0 sample to determine the posteriors for halving, normal, doubling F0.

Given the continuous characteristic of the short-term energy, the estimation of the delta features of energy contour is performed using the Equation 3.1. Since the short-term energy is computed in the F0 estimation algorithm, there is not extra processing for estimating the energy contour.

2nd step – Detect the points where the dynamics of both contours changes

The changes in the contour dynamics are defined by the points at which there is a voiced/unvoiced speech region transition and a change in the direction of the contour dynamics. A change in the direction of the contour dynamics occurs at the points where the first derivative is zero. These points are referred to as critical points of a function *f*. Let $e(t)$ be the continuous function of short-term energy with continuous first

derivative $\Delta e(t)$, whose domain is a closed and bounded interval [0, N]. The function $\Delta e(t)$ has exactly k critical points at $0 < t_1^e < t_2^e < \ldots < t_k^e < N$, where

$$\Delta e(t) = 0 \qquad \text{only for } t \in \left\{t_1^e, t_2^e, \ldots, t_k^e\right\}.$$

Given the discontinuities of the F0 contour, the detection of the changes in the F0 contour dynamics is performed somewhat differently from the detection in the energy contour. Since F0 can be only estimated from voiced speech regions, the detection is performed separately for each voiced speech region. Let $f_0(t)$ be a piecewise continuous function of F0 from a given voiced speech region with a piecewise continuous first derivative $\Delta f_0(t)$:

$$\Delta f_0(t) = \begin{cases} \Delta f_{0,1}(t), & t_0 \le t \le t_1 \\ \Delta f_{0,2}(t), & t_2 \le t \le t_3 \\ \vdots & \vdots \\ \Delta f_{0,i}(t), & t_{i-1} \le t \le t_i \end{cases}$$

where $\Delta f_{0,i}(t)$ is a continuous function valid over a sub-interval of $\Delta f_0(t)$, $t_{i-1} \le t \le t_i$, and t_0 and t_i are the beginning and ending point, respectively, of the voiced speech region. The number of continuous functions is the same as the number of discontinuities within $\Delta f_0(t)$ plus one. For example, the first voiced speech region (t_i,t_{i+2}) in Figure 4.4 can be described using two functions (there is only one discontinuity), whereas only one function is needed to describe the second voiced speech region (t_{i+3},t_{i+4}). Then, each function $\Delta f_{0,i}(t)$, whose domain is a closed and bounded interval [t_{i-1}, t_i], has exactly l critical points at $t_{i-1} < t_1^{f_{0,i}} < t_2^{f_{0,i}} < \ldots < t_l^{f_{0,i}} < t_i$, where

$$\Delta f_{0,i}(t) = 0 \qquad \text{only for } t \in \left\{t_1^{f_{0,i}}, t_2^{f_{0,i}}, \ldots, t_l^{f_{0,i}}\right\}.$$

The filled circles in the F0 plot of Figure 4.5 represent the critical points and the voiced regions boundaries. In the energy plot, the filled circles represent only the critical points.

3rd step – Generate new segments using the detected points

The segment boundaries are defined as the voiced speech region boundaries, the estimated critical points from the F0 and energy contours, and the beginning and ending points of the contours. Since F0 cannot be estimated from unvoiced speech regions, the

critical points from the energy contour that fall within an unvoiced speech region are not used in the segmentation. For example, the critical point in the */tcl/* phone of Figure 4.5 is not used in the segmentation because of unvoiced characteristic of such phone. Let $\{t_i^{new}\}_{i=0}^{m}$ be the collection of all valid points with $t_0^{new} < t_2^{new} < \ldots < t_m^{new}$, $t_0^{new} = 0$, and $t_m^{new} = N$. The segment boundaries are defined as follows

$$(t_0^{new}, t_1^{new}), (t_1^{new}, t_2^{new}), \ldots, (t_{m-2}^{new}, t_{m-1}^{new}), (t_{m-1}^{new}, t_m^{new})$$

The new segment boundaries are represented by vertical dotted bars in Figure 4.5.

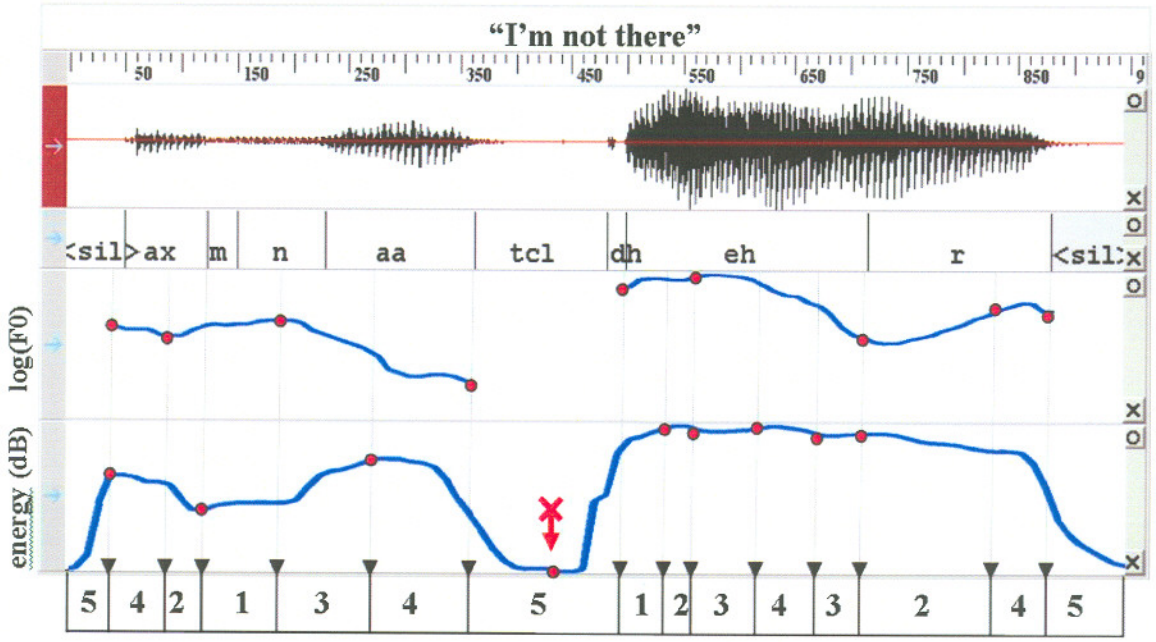

Figure 4.5: Example of a joint-state class estimation using delta features. The critical points are marked by filled circles. The vertical bars represent the segmentation boundaries generated from the critical points of both contours. At the bottom, the boxes represent the segmentation and their respective joint-state class.

4th step – Convert the segments into a sequence of symbols that represent the dynamics of both contours

Each segment is classified into one of the five classes that represent the dynamics of both contours within the segment, as shown in Table 4.1. For each segment, the classification rules are defined as follows

$$Class(i)=\begin{cases}1 & \text{if} \quad \Delta f_0(t)>0 \quad \wedge \quad \Delta e(t)>0 \\ 2 & \text{if} \quad \Delta f_0(t)>0 \quad \wedge \quad \Delta e(t)<0 \\ 3 & \text{if} \quad \Delta f_0(t)<0 \quad \wedge \quad \Delta e(t)>0 \\ 4 & \text{if} \quad \Delta f_0(t)<0 \quad \wedge \quad \Delta e(t)<0 \\ 5 & \text{if} \quad \neg\exists \Delta f_0(t) \end{cases} \qquad \text{for } \forall t \in \left(t_{i-1}^{new}, t_i^{new}\right).$$

Figure 4.5 shows, at the bottom, the segments and their respective joint-state class for the utterance "I'm not there".

4.3.1 How long should the delta window be?

In most speech-based applications, the delta features are estimated over a typical time interval between 50 and 100 ms [145]. In order to determine the adequate duration of the time interval, additional experiments were performed by varying the duration of the time interval between 30 ms and 170 ms. Figure 4.6 presents the DET curves for different time intervals (delta window) used in the delta features estimation.

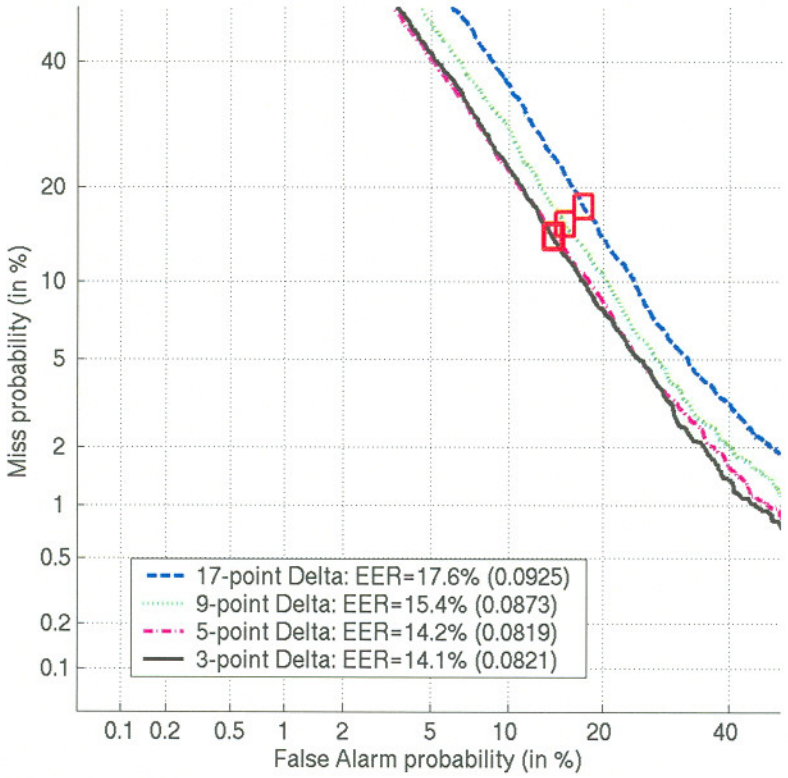

Figure 4.6: Speaker detection performance for bigram modeling of joint-state classes estimated using different time intervals for the delta features estimation: 30 ms (3-point window), 50 ms (5-point window), 90 ms (9-point window), and 170 ms (17-point window).

Figure 4.6 shows that the longer the time interval used to estimate the delta features, the less speaker information is captured by the joint-state classes modeling. Therefore, fast changes in the F0 and energy temporal trajectories carry more speaker-dependent information than relatively slower changes, in particular, changes within a 50 ms time interval. Figure 4.7 presents several sequences of joint-state classes using different delta window lengths for the utterance "I'm not there" depicted in Figure 4.1. It shows that the trajectory is described in more details as the delta window length is reduced.

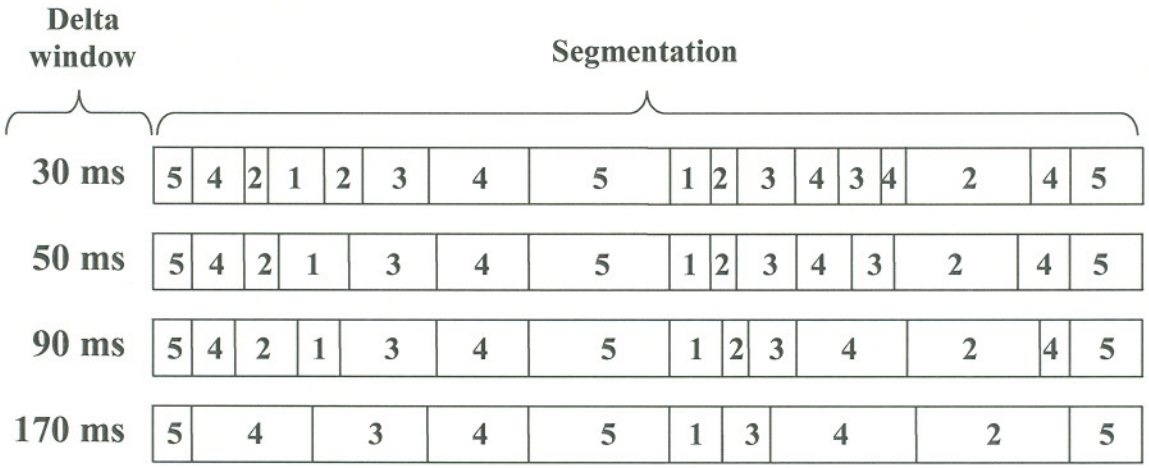

Figure 4.7: Example of segmentation using joint-state classes of the utterance "I'm not there" using different delta window lengths.

We also tried different combinations of delta window lengths for each contour, but the performance is only worse than the results presented in Figure 4.6. For example, we tried a 5-point window for the F0 contour and a 9-point window for the energy contour, and vice-versa. This shows the adequacy of the window length for both trajectories.

4.3.2 Integrating in Duration Information

The duration quantization method differs in two ways from the one described in Section 4.2.1. First, the duration quantization is different for voiced and unvoiced regions. Since the segmentation is performed on voiced speech regions, the joint-state classes estimated from these regions (i.e., classes 1 to 4) are smaller than the classes estimated from unvoiced speech regions (i.e., class 5). With shorter duration, the duration of joint-state classes from voiced speech segments are overestimated when grouped with unvoiced

segments. Second, we found that, because of the shorter segments, the quantization into 3 levels does not provide any extra information when compared to the quantization into 2 levels. Thus, the duration of segments from voiced and unvoiced regions are quantized into "Short" and "Long". The quantization value used to determine whether a segment is short or long is estimated from the median value using a held-out data set. For voiced speech regions, "Short" is assigned to segment classes with duration shorter than 8 frames (80 ms). For unvoiced speech region, "Short" is assigned to segment classes with duration less than 14 frames (140 ms). Therefore, the total number of possible symbols used to represent the prosodic information from a speech signal is 10.

4.3.3 Speaker Detection Results

The delta-based joint-state modeling is evaluated on the 2001 NIST SRE task, with the same configuration as defined in Section 4.2.2. Figure 4.8 shows the speaker detection performance for the bigram modeling of the delta-based joint-state classes and single-contour classes.

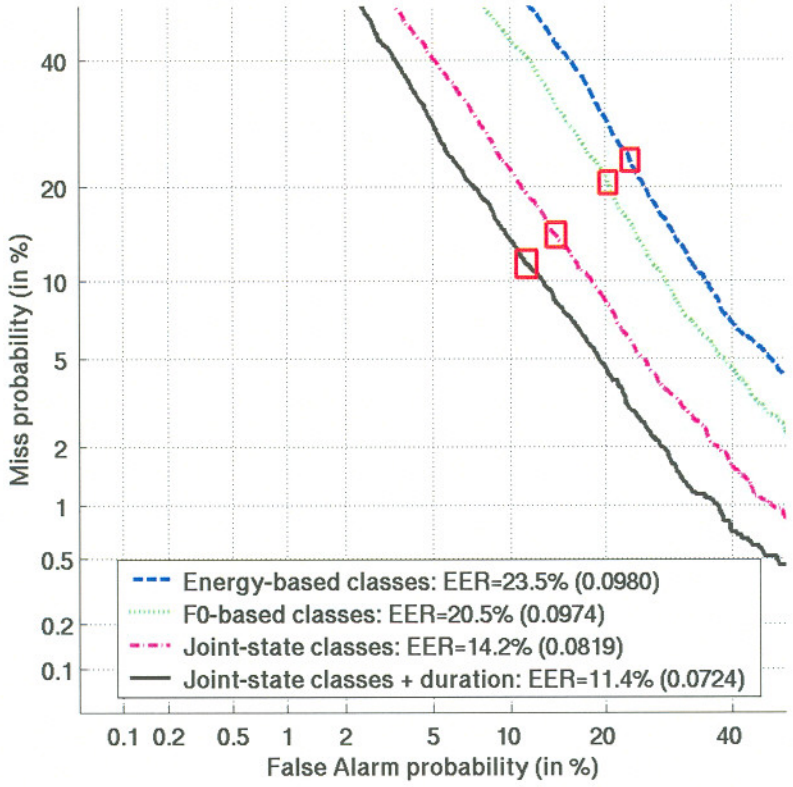

Figure 4.8: DET curves for the delta-based joint-state class modeling.

The performance of the energy-based (23.5%) and F0-based (20.5%) classes estimated from the delta features show an improvement over the piecewise linear fit. Such result does not come as a surprise because the goal of a stylization algorithm is to estimate the speaker's intended pitch contour and not all the variations in the contour. For this reason, the delta features estimated from time intervals shorter than 50 ms can capture these small variations that carry speaker-specific information. The segmentation of the energy contour using its critical points also provides improvement over the previous approach.

The EER of the delta-based joint-state modeling without and with the duration labels is 14.2% and 11.4%, respectively. These results show that independent estimation of the rate of change of the F0 and energy contours provide a better characterization of the speaker-specific information. The results of the delta-based approach show an improvement of 26% (without duration) and 18% (with duration) over the respective PWL-based approaches. The relative improvement over the prosodic baseline is 6% (without duration) and 25% (with duration).

4.3.4 Speaker Entropy

Using the data from Switchboard I corpus, we analyze the speaker information captured in the bigrams of the delta-based joint-state classes. The amount of information conveyed by a source of information can be measured using the information theory quantity of entropy. Entropy is a statistical measure of information or, in the information theory field, uncertainty [162]. Consider an information source that generates a sequence of symbols $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ from a finite or countable infinite sample space S, according to some stochastic distribution law. The probability that **X** takes on the particular value x is written P(x). The entropy of the source is defined as

$$H(X) = -\sum_{x \in X} P(x)\log(P(x)).$$

The entropy H(X) will be the highest when you know least about the next symbol and the lowest when you know most. Given that we want to measure the amount of speaker information per n-gram type, the symbols represent the speakers, and the source of information is the n-gram type. Then, the probability P(x) represents the probability that

some given n-gram type is produced by speaker x, which is approximated by the Equation 4.1.

Figure 4.9 shows a scatter plot of the speaker entropy for the bigrams of joint-state classes versus the number of occurrences of bigrams for 261 speakers. Most of the bigrams that include the boundary symbol **<bound>** (to mark the beginning or ending of an utterance) have higher entropy than the remaining bigrams with similar number of occurrences. That is, such bigrams do not provide as much information about the speaker who produced them as the bigrams that do not contain the boundary symbol <bound>. One of the reasons is that the method used to detect utterances does not perform a consistent detection. For example, due to the variability of the pause duration between words, the method is going to incorrectly detect some of the pauses (longer than 0.5 second) as utterance boundaries, even though the pauses are not true utterance boundaries.

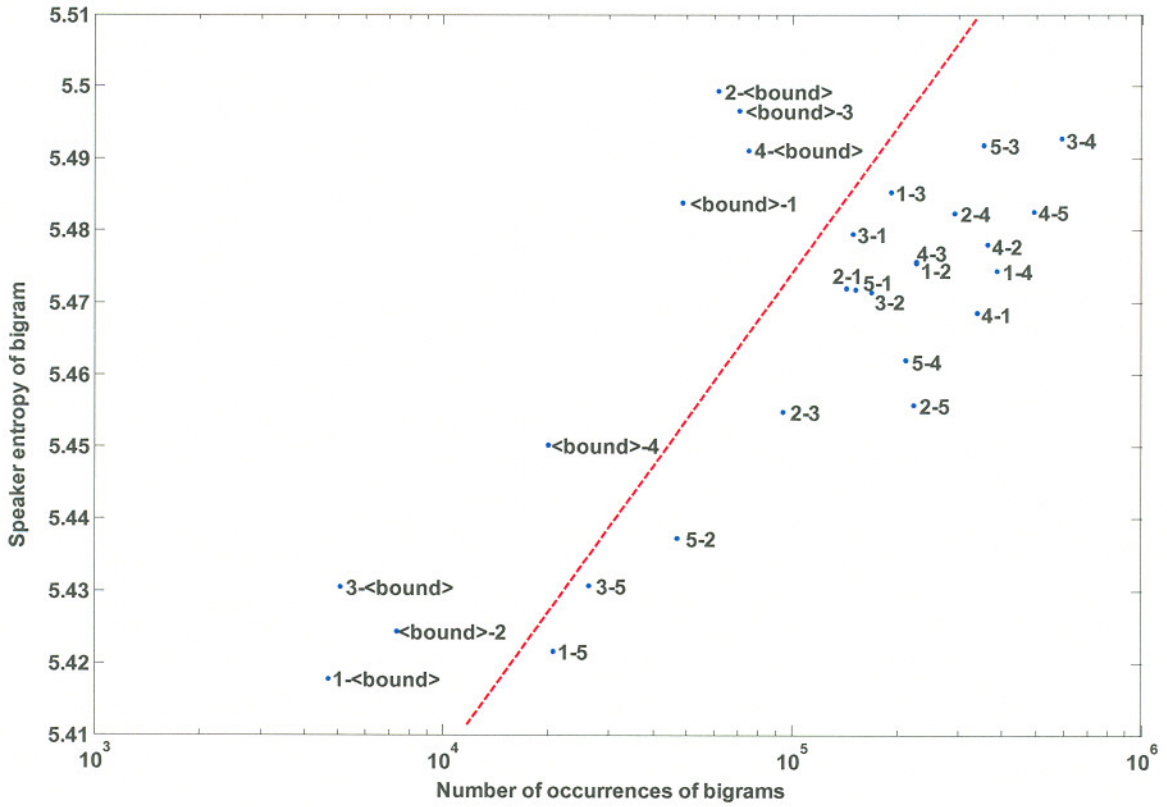

Figure 4.9: Speaker entropy of joint-state class bigrams. The dashed line separates the bigrams types that include the <bound> symbol from the remaining bigram types.

Even though low-entropy bigrams provide more speaker-specific information than high-entropy bigrams, it is expected that the speaker is consistent in producing such bigrams across different conversations [27, 163]. Therefore, low-entropy bigrams with high number of occurrences (e.g., "2-5" and "5-4" bigrams) can provide a more consistent modeling of speaker-specific information. Figure 4.10 shows the performances of the joint-state class bigram modeling on 8-conversation training condition (2001 NIST SRE) using a leave-one-out technique. The bigram in the *x*-axis represent the bigram type left out from the scoring process (e.g., the EER of joint-state classes bigram modeling without the bigram "4-5" is 25.7%). The performance degradation in Figure 4.10 is the highest for the experiments that do not use the low-entropy bigrams with high number of occurrences (e.g., "4-5", "4-2", "1-4", "2-4", and so on).

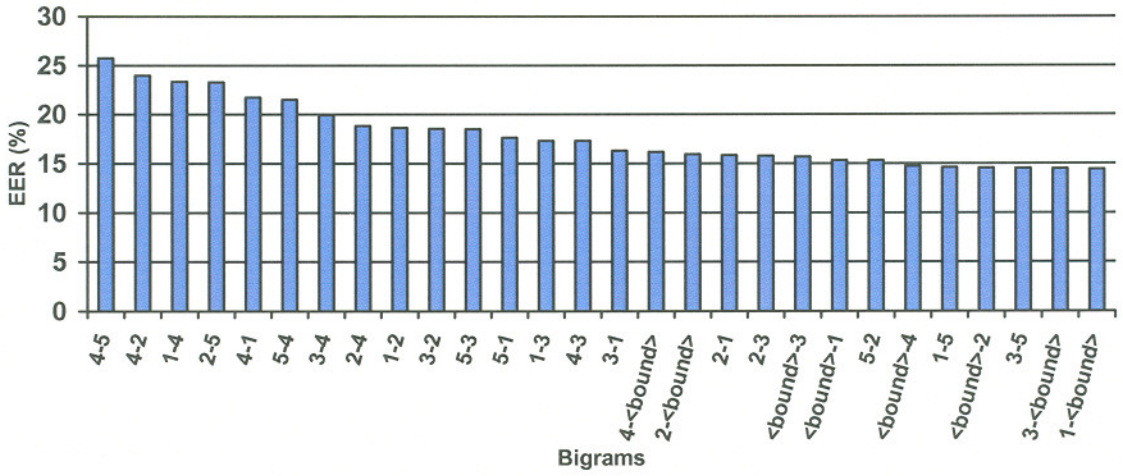

Figure 4.10: Performance for the leave-one-out experiments. The x-axis represents the bigrams not used in the detction process.

The performances in Figure 4.10 show that the removal of the joint-state class 4 (falling F0 and falling energy) from the scoring process have the worst performance degradation compared to the performance of the system without any bigram removal (14.2% EER) on the 2001 NIST SRE. Note that, in the first 10 worst performances in Figure 4.10, the joint-state class 4 is in 70% of the bigrams. This shows that there is a large variability in how frequently speakers produce a falling F0 and energy for different bigram contexts. For example, the standard deviation of the occurrence frequency of the bigram "4-5" for every speaker is 0.0181, whereas it is 0.0024 for a bigram "3-5" that does not affect the performance when removed.

4.4 Is joint-state class modeling really capturing complementary information?

Since the prosodic baseline system is modeling the absolute F0 and energy distribution, whereas the proposed 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 Figure 4.11, we show the results of fusing the various systems on the 8-conversation training condition of the 2001 NIST SRE using a single-layer perceptron fusion system described in Section 3.5.1.

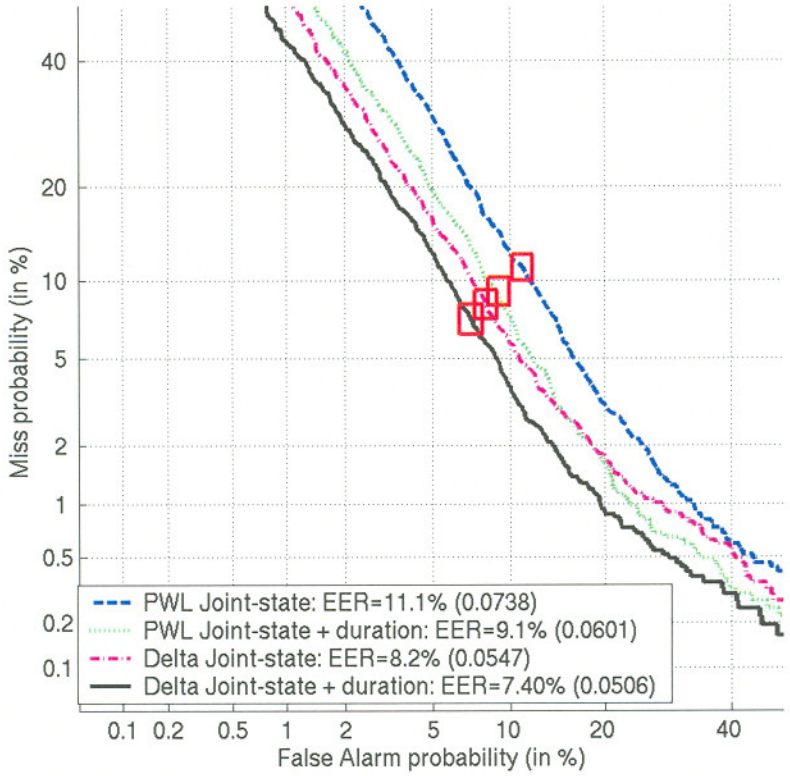

Figure 4.11: DET curves for the fusion between prosody-based systems and the prosodic baseline.

The improvement in performance of the fusion shows that the joint-state classes provide complementary information to the prosodic baseline. Note that the fusion between the prosodic baseline and both approaches (delta-based and PWL-based) provides a 35% relative improvement of the performance. Indeed, these improvements show that there is speaker-specific information in the F0 and energy contours that is beyond their distribution statistics.

Table 4.2 shows the performance of the fusion between the acoustic baseline with the prosodic baseline and delta-based joint-state classes on 8-conversation training condition. Note that the EER of the acoustic baseline for this condition is 0.7% with a DCF of 0.0034.

Table 4.2: Performances of the fusion with the acoustic baseline system.

Systems	**Fusion with acoustic baseline system**		
Prosodic baseline	✓		✓
Delta-based joint-state + duration		✓	✓
EER (DCF)	0.5 (0.0028)	0.5 (0.0018)	0.4 (0.0019)

The performance of the fusion between the acoustic baseline, prosodic baseline, and the delta-based joint-state classes is 0.4%, an improvement of approximately 40% over the acoustic baseline alone. This result additionally indicates that the prosodic features have complementary information to standard spectral information.

The fusion between the acoustic baseline and either of the prosodic systems obtains a 28% relative improvement (α=0.1) over the acoustic baseline. Despite the similarity in EER performance between the two fusions, the minimum DCF value shows that the joint-state classes modeling provides a lower false alarm rate than the prosodic baseline, i.e., less impostors are falsely recognized as true speakers. Figure 4.12 shows the contribution of the two types of errors (miss detection and false alarm) to the DCF value. Note that the fusion system is trained to minimize the DCF. The improvement of the DCF of the fusion between acoustic baseline (bar number 1 in Figure 4.12) and prosody-based systems (bar numbers 2 and 3 in Figure 4.12) comes mostly from the reduction of the false alarm error.

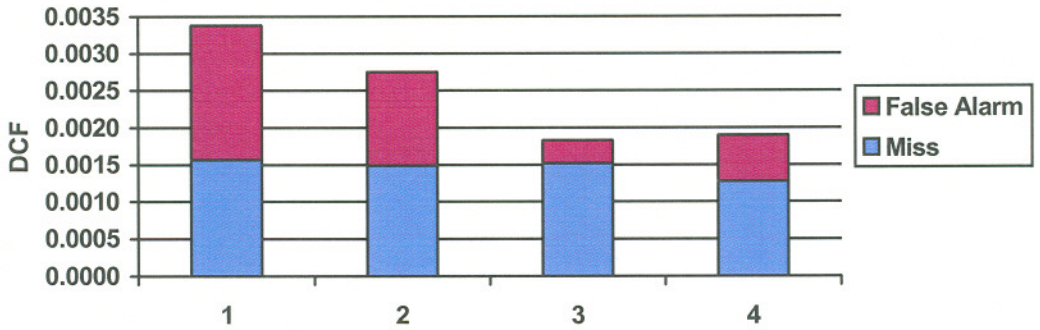

Figure 4.12: Contribution of the two errors types to the DCF value: 1 – acoustic baseline, 2 – fusion of acoustic and prosodic baselines, 3 – fusion of acoustic baseline and delta-based joint-state plus duration, and 4 – fusion of both baselines and delta-based joint-state plus duration.

4.5 Results on 2002 NIST Speaker Recognition Evaluation

Since most of the analysis is performed on the 2001 NIST SRE, we also run the same experiments, without any change in the configuration or parameter tuning, on the 2002 NIST SRE (Switchboard II phases 2 and 3). Figure 4.13 presents the performance on the 8-conversation training of the baselines, delta-based joint-state classes, and the fusion between the systems.

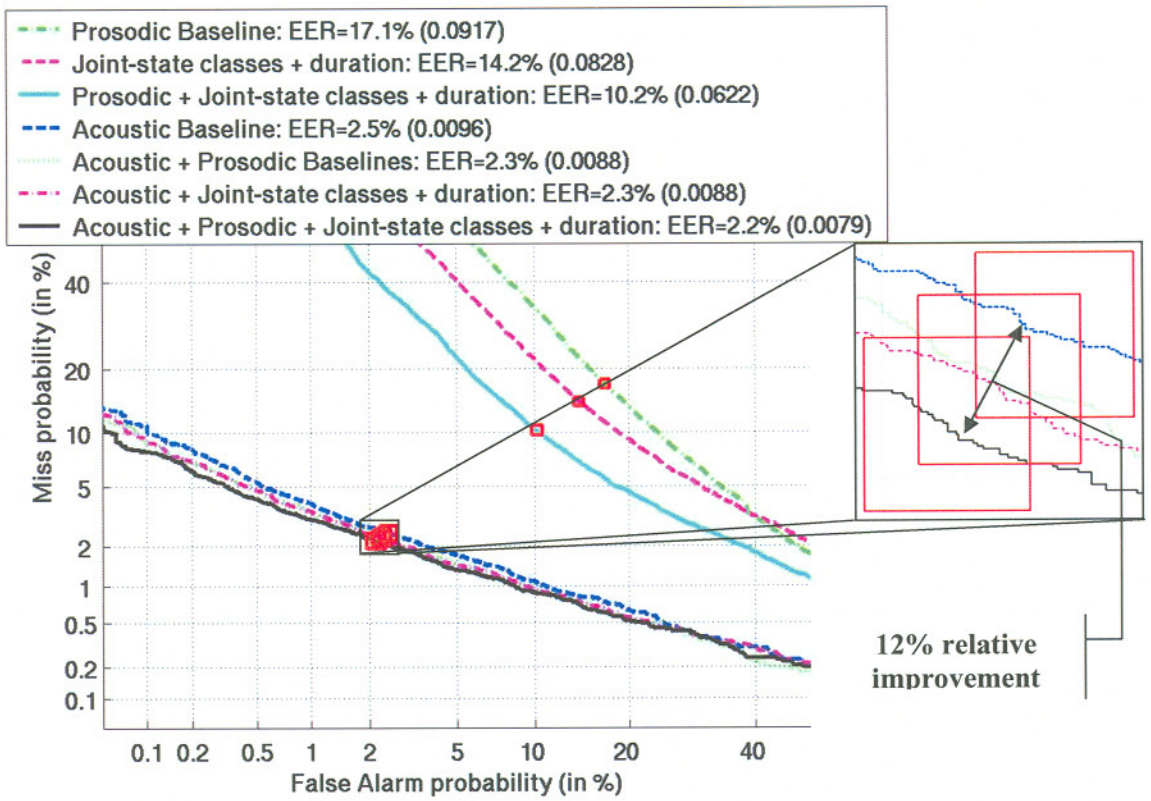

Figure 4.13: DET curves of the baselines, joint-state classes, and the fusion between the systems on the 2002 NIST SRE. The small box on the right side shows a zoomed view of the performance curves that include the acoustic baseline.

The performance of the individual systems and their fusion shows similar pattern when compared to the results for the 2001 NIST SRE. That is, the performance of the joint-state classes is better than the prosodic baseline, the fusions between the systems improve the performance of the prosodic and acoustic baseline, and the acoustic baseline has the best performance. The performance of the joint-state classes shows a relative improvement of 17% over the prosodic baseline. Note that the performance of the fusion between the acoustic baseline with either the prosodic baseline or the joint-state classes

does not yield a significant improvement over the performance of the acoustic baseline. However, the performance of the fusion of all three systems is significantly different (12% relative improvement) from the acoustic baseline performance. These results show that the complementary information characteristic of the joint-state classes holds for a different evaluation data.

4.6 Factors Affecting Performance

Several factors can affect the performance of a speaker recognition system. In this section, we analyze how the performance varies with respect to the amount of training data, handset mismatch between training and testing, and speaker demographics (e.g., age and gender).

4.6.1 Amount of Training Data

Speaker recognition systems that use long-term speech characteristics (e.g., phonemes, words, and prosodic features) are known for requiring large amounts of training data [27, 28, 74, 76, 78]. Figure 4.14 shows the performance of the joint-state classes plus duration system and the prosodic baseline on all training conditions.

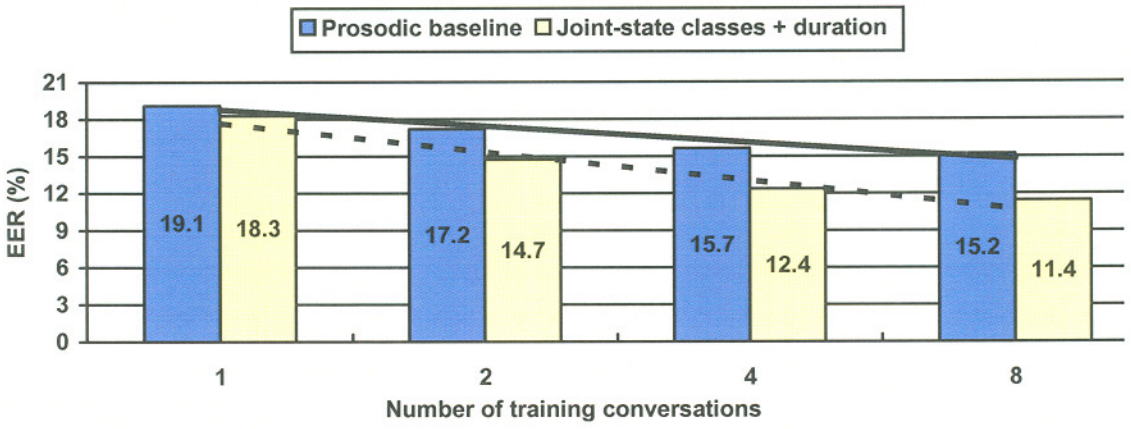

Figure 4.14: Comparison of performance for different number of training conversations. The dashed trend line depicts the performance improvement for the prosodic baseline and the dotted trend line depicts the improvement for the delta-based prosodic features.

Despite the 25% relative improvement of the joint-state based system over the baseline on 8-conversation training condition, there is no significant improvement for the 1-conversation training condition. The improvements over the baseline only become significant when there are at least 2 conversation sides (approximately 4 minutes) for

training. Therefore, the results show that considerable amount of data is required for estimating models using joint-state classes. Appendix A shows some statistics about the amount of training data generated per conversation side for both NIST evaluation data.

4.6.2 Telephone Handset

The variation in the type of telephone handset is a major factor affecting the performance of systems that use telephone speech [164, 165]. Since the 2001 NIST SRE does not provide enough target trials that use different handsets, we measured the performance for matched- and mismatched-handset conditions on the 2002 NIST SRE. The 8-conversation training condition has 28,907 test trials, which 50% of the target trials (not impostor trial) are matched. A matched target trial has the phone number of the test conversation occurring at least once in the speaker model training data. Figure 4.15 shows the performance for the systems on the matched- and mismatched-handset conditions.

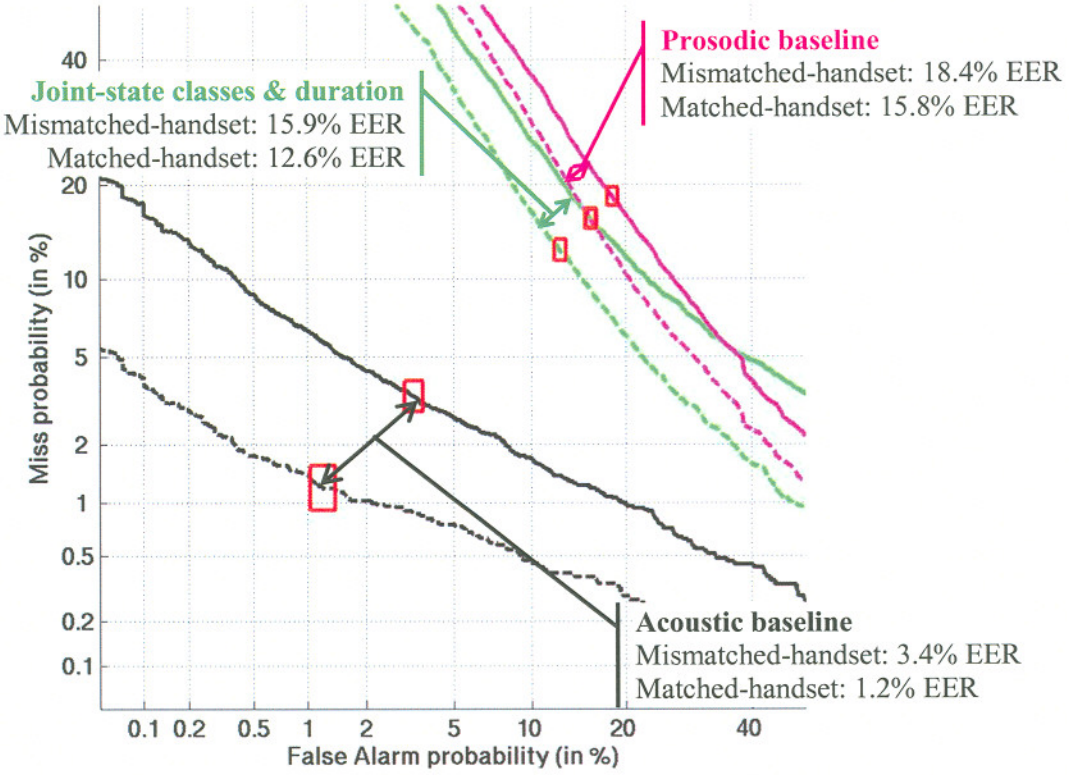

Figure 4.15: Performance per handset condition of the baselines and the joint-state classes. The dashed and solid lines represent the performance of the systems under the matched- and mismatched-handset conditions, respectively.

The significant difference in performance between matched- and mismatched-handset conditions shows that all systems are affected by the mismatched-handset condition. Note that the prosodic systems are more robust to mismatched-handset condition than the acoustic-based system. The performance degradation of the acoustic baseline, prosodic baseline, and the joint-state classes plus duration is 183%, 16%, and 26%, respectively.

A performance comparison between both evaluation data shows that the performance of the joint-state classes for the matched-handset condition (EER=12.6%) is significantly different from the performance obtained for the 2001 NIST SRE (EER=11.4%). This result shows that there are other factors, such as speaker gender or age, affecting the performance of the joint-state classes.

4.6.3 Speaker Demographics

Speaker gender is one of the most important sources of variability in speech. Certain acoustic characteristics of speech that give a voice its quality and individuality have contributions that range from those of speech production mechanism (i.e., differences between individual sound sources and the resonant frequencies of the vocal tract) to the effects of prosody and dialect. For example, female speakers have, on average, higher fundamental frequency than male speakers [94, 166]. Holmberg et al. [167] show that female speakers tend to have more gradual rises and falls in glottal flow than males. Using data from seven languages and dialects (including three dialects of English), Henton [168] shows that females appear to produce vowels in a manner that is more phonetically explicit than males. However, despite the high variability between genders, variability within gender can become a problem when the impostor has the same gender as the target speaker. Thus, we evaluate the performance of each system using same-gender trials (the impostor and the target speaker have the same gender) on the 2001 and 2002 NIST SREs.

First, we analyze the systems performance for the same-gender trials on the 1-conversation training. To avoid the effects of different handsets between training and testing, only the trials from matched-handset condition (as described in Section 4.6.2) are analyzed. In the 2001 NIST SRE, there are 6,250 female-female trials (1,875 target trials

and 4,375 impostor trials) and 6,948 male-male trials (1,844 target trials and 5,104 impostor trials). In the 2002 NIST SRE, there are 14,840 female-female trials (3,160 target trials and 11,680 impostor trials) and 12,126 male-male trials (2,506 target trials and 9,620 impostor trials). Table 4.3 presents the performance for same-gender trials of the acoustic, baseline, and joint-state classes plus duration on the 1-conversation training condition.

Table 4.3: EER per gender condition of the baselines and the joint-state classes for 1-conversation training condition on 2001 and 2002 NIST SREs.

NIST SRE	Trial	Acoustic baseline	Prosodic baseline	Joint-state classes + duration
2001	Female-female	3.0%	20.2%	18.5%
	Male-male	2.3%	19.6%	17.6%
2002	Female-female	3.5%	20.5%	18.2%
	Male-male	2.4%	22.1%	18.6%

As expected, the acoustic baseline performs much worse for female-female trials than male-male trials. Previous work [94, 165] has found that the performance of mel-cepstra based systems degrades with higher pitch frequency and with stronger "pitch mismatch" (i.e., pitch variation between enrolment and testing). In addition, the performance of female-female trials on 2002 NIST SRE is worse (approximately 33%) than the 2001 NIST SRE. Given the physiological and linguistic (e.g., vocabulary and speaking style) differences across speakers of the same gender, the age difference has a similar effect as the gender difference in the speaker detection performance, i.e., speakers with small age difference can share more physiological or linguistic characteristics than speaker with large age differences. Therefore, we measured the age difference of impostor trials on the 2001 and 2002 NIST SREs. The median value of the age difference for both gender condition on the 2001 NIST SRE and 2002 NIST SRE is 11 years and 4 years, respectively. Appendix B shows some statistics about age difference between impostor and target speaker. In fact, the worse performance of the acoustic baseline for the female-female trials can be explained by the small age difference of the impostor trials of the 2002 NIST SRE.

The prosodic baseline yields a worse performance for male-male trials only for the 2002 NIST SRE. Note that the performance for female-female trials is not affected by

the change of evaluation data. Contrary to the acoustic baseline, it seems that the small age difference of the 2002 NIST SRE affects the performance of the male-male trials.

The performance of the joint-state classes approach shows no significant difference between gender conditions on both NIST SREs. Only the performances of male-male trials show a small degradation (α=0.1) from the 2001 to the 2002 NIST SRE. However, these results can be consequence of the small amount of training data.

Since the performance of the joint-state classes approach considerably improves for 8-conversation training condition, we also analyze the speaker gender effects under such training condition. Note that there is no increase in the amount of testing data. In the 2001 NIST SRE, there are 4,305 female-female trials (1,681 target trials and 2,624 impostor trials) and 4,261 male-male trials (1,673 target trials and 2588 impostor trials). In the 2002 NIST SRE, there are 10,808 female-female trials (2,988 target trials and 7,820 impostor trials) and 9,535 male-male trials (2,575 target trials and 6,960 impostor trials). Table 4.4 presents the performance for same-gender trials of the acoustic, baseline, and joint-state classes plus duration on the 8-conversation training condition.

Table 4.4: EER per gender condition of the baselines and the joint-state classes for 8-conversation training condition on 2001 and 2002 NIST SREs.

NIST SRE	Trial	Acoustic baseline	Prosodic baseline	Joint-state classes + duration
2001	Female-female	0.5%	16.3%	10.8%
	Male-male	0.4%	16.3%	13.2%
2002	Female-female	1.2%	14.3%	12.9%
	Male-male	1.2%	18.4%	14.1%

The acoustic baseline does not show any significant difference between the gender conditions for the 2002 NIST SRE. However, the improvements from 1- to 8-conversation training conditions differ for 2001 and 2002 NIST SRE. The improvements obtained in the 2002 NIST SRE (66% for female-female trials and 50% for male-male trials) are smaller than the ones in the 2001 NIST SRE (83% for both gender conditions). Despite the increase in amount of training data, the small age difference seems to affect both same-gender trials.

The performances of the prosodic baseline show a similar pattern to the 1-conversation training condition performances: same performance for the 2001 NIST SRE

and male-male trials performance is worse than female-female trials for the 2002 NIST SRE. This result supports the claim that male-male trials performance is affected by the small age difference between impostor and true speaker. Interestingly, the performance of the female-female trials on the 2002 NIST SRE is better than the performance on the 2001 NIST SRE. One reason is that the increase in amount of training data provides a better modeling of female speakers than male speakers.

The performance of the joint-state classes approach shows a significant difference between female-female and male-male trials on both evaluations. One reason is that the female speakers produce more segment classes than males do (see Appendix A), which provides more training data for estimating a speaker model. On average, the number of joint-state classes per conversation side is about 13% higher for female speakers. However, the small age difference between impostor and true speaker on the 2002 NIST SRE affects more female-female trials than male-male trials.

4.7 Extensions to Prosodic Modeling

In this section, we describe two extensions to the joint-state classes estimation. The first approach deals with more than two levels for quantizing the rate of change of each contour. The second approach looks into the energy at different frequency bands.

4.7.1 Four-level Delta Quantization

The goal of increasing the number of quantization levels is to capture different levels of rising or falling movements. Based on intonation modeling literature [120], we quantize each contour into 4 possible levels: low rise, high rise, low fall, and high fall. The rise and fall levels are still defined by the critical points of the time-derivative (i.e., points at the zero-crossings of the rate of change). We use the median value of the rate of change to detect the type (i.e., low or high) of a rise or fall level. The median value is estimated for each contour from a held-out data set. Since the rate of change has approximately a normal distribution around zero, the median value is estimated from the absolute value of the rate of change. We use only one value for each contour to detect a high or low rate of change. Thus, each segment can be labeled with any one of the 17 possible labels (4 levels for the F0 contour times 4 levels for the energy contour plus unvoiced region).

Figure 4.16 shows the performances on 8-conversation training for the bigram modeling of the joint-state classes using two levels and four levels to describe the rate of change of both contours. The EER of the modeling of the joint-state classes using four-level quantization is 10.6% (a relative improvement of 25% over the two-level quantization). The four-level joint-state classes system shows a similar performance to the two-level joint-state plus duration classes system. Note that the four-level joint-state classes do not have the duration label as the two-level joint-state plus duration classes system.

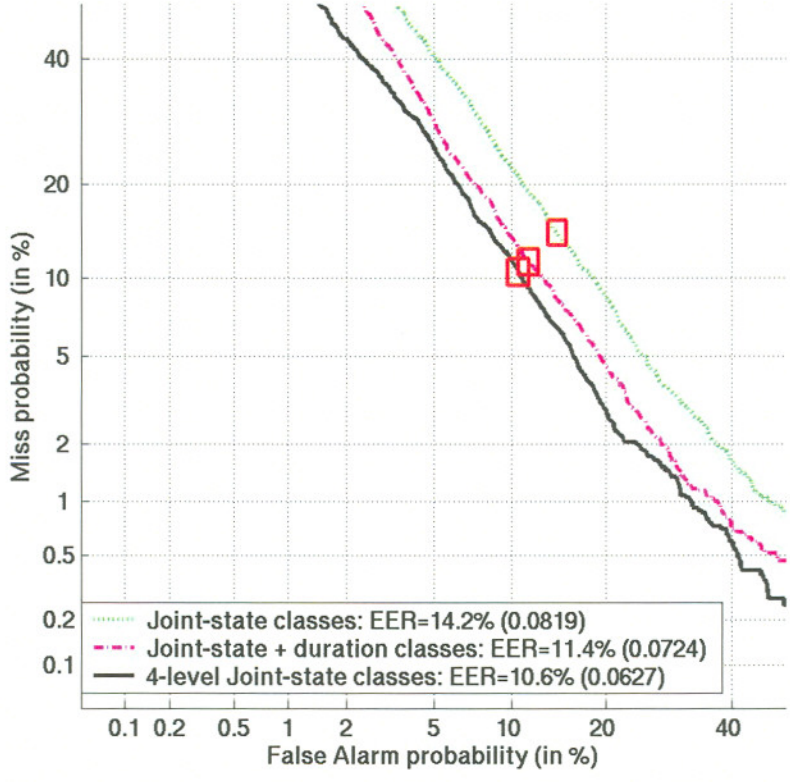

Figure 4.16: DET curves for joint-state classes based on two-level and four-level to describe the direction of the rate-of-change.

The detailed representation of the temporal trajectories provides a better modeling of the speaker-specific information. Figure 4.17 shows the scatter plot of speaker entropy for bigrams of both systems: 2-level and 4-level quantization. Note that the bigrams estimated from the 4-level quantization yields lower speaker entropy than the 2-level quantization. Consequently, the bigrams from 4-level quantization provide a better characterization of speaker-specific information.

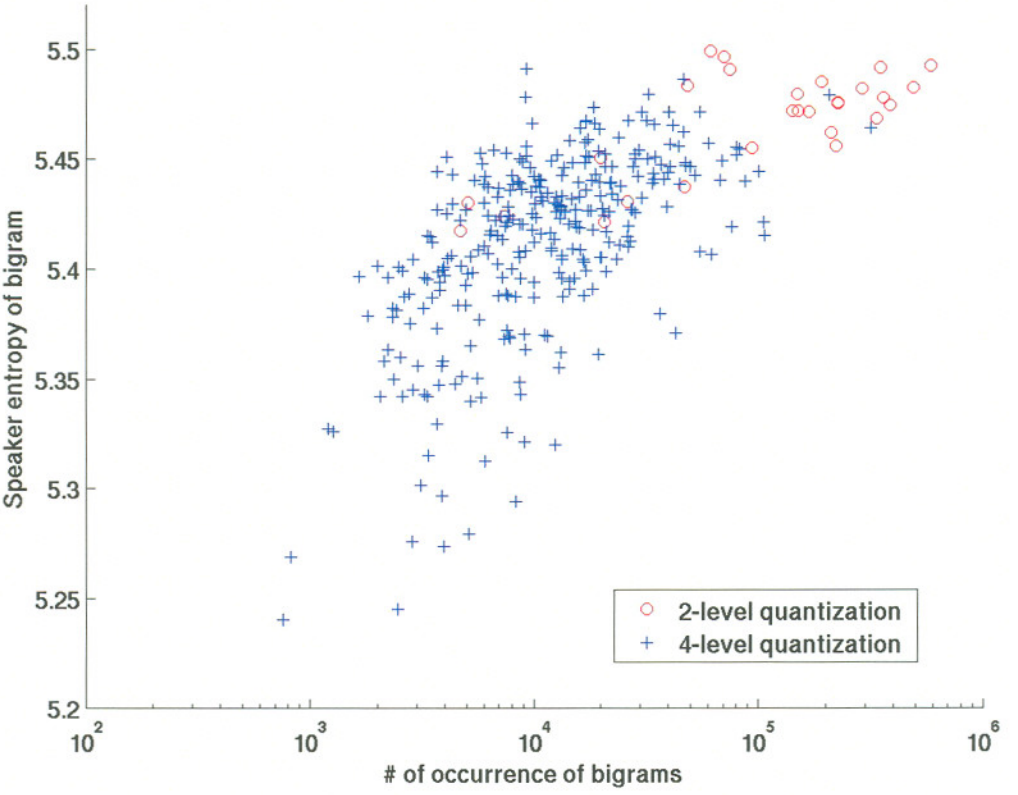

Figure 4.17: Speaker entropy of bigram for two-level and four-level delta quantization process.

The addition of duration label to the 4-level joint-state class does not yield any performance improvement on the 2001 NIST SRE. The performance of this system on 8-conversation training is 10.6% EER. Since each segment is the result of the critical points from both contours, this quantization process produces more segments, but with shorter duration. However, the duration of the segments does not seem to be consistent for the same speaker. Given the short delta window used for estimating the rate of change, small perturbations in the contours caused by the estimation method or noise can affect the rate of change estimation.

4.7.2 Sub-band based 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 sub-band based speaker detection system is depicted in Figure 4.18. We assume that the frequency bands are independent, so that it allows us to score and even combine different frequency bands. First, the frequency-localized temporal trajectories are estimated from the speech signal. Second, for each frequency-band, the sequence of joint-state classes is estimated using

the F0 and the respective temporal trajectory. Then, the fusion module selects and fuses the scores from the frequency bands.

There are several reasons for developing a sub-band modeling for speaker recognition. First, the sub-band approach allows us to deal with noise conditions (background and channel) that affect only part of the speech spectrum [169]. Second, there is variability in different frequency bands because of different phones, speakers, and channels [21, 170]. Third, since short-term energy conveys several levels of information, the analysis of the spectral domain allow us to extract only the relevant information (phone and speaker), and discard the irrelevant information (channel). For example, the region around 5-6 barks (approximately 500-600 Hz) contains the highest phone variability (consequence of the characteristics of voiced/unvoiced phones in this region where the voiced phones have high energy in this region and the unvoiced phones have low energy [171]), which is useful for speaker recognition [53].

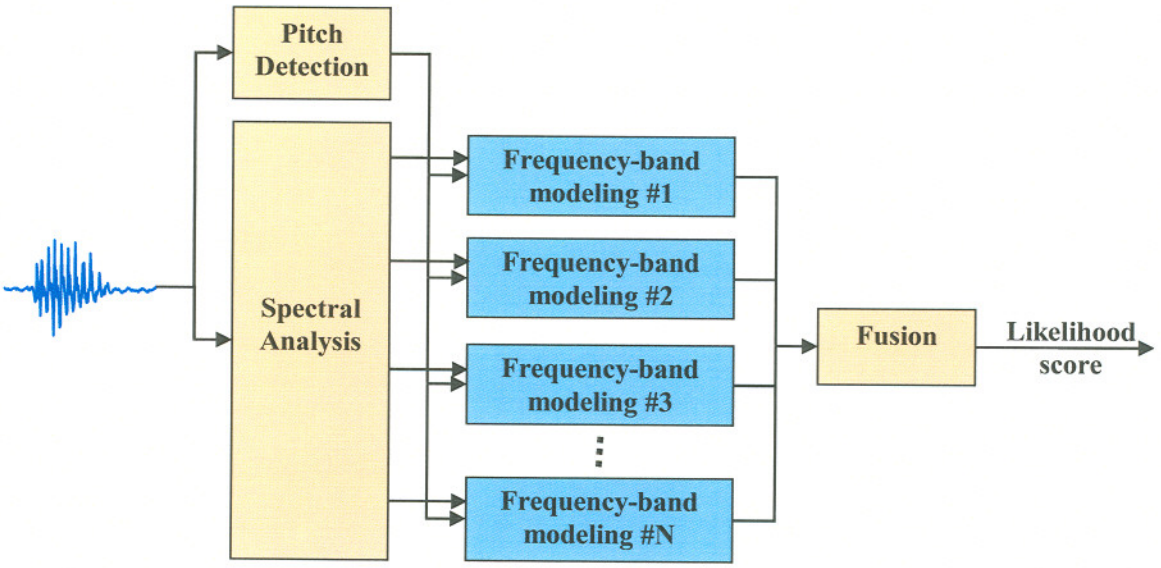

Figure 4.18: Sub-band based speaker detection system.

In our experiments, the system uses the temporal trajectories of 15 bark-scale critical-bands to estimate the sequence of joint-state classes. First, 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, the sequence of joint-state classes is estimated for each frequency band using the method described in Section 4.3. Then, a bigram model is estimated from each frequency band to characterize

a given speaker. When fusion is performed, the scores from the frequency bands are averaged to obtain the final score. Figure 4.19 shows the performance for each critical band on 2001 and 2002 NIST SREs, and the lower- and upper-cut-off frequencies of each critical-band.

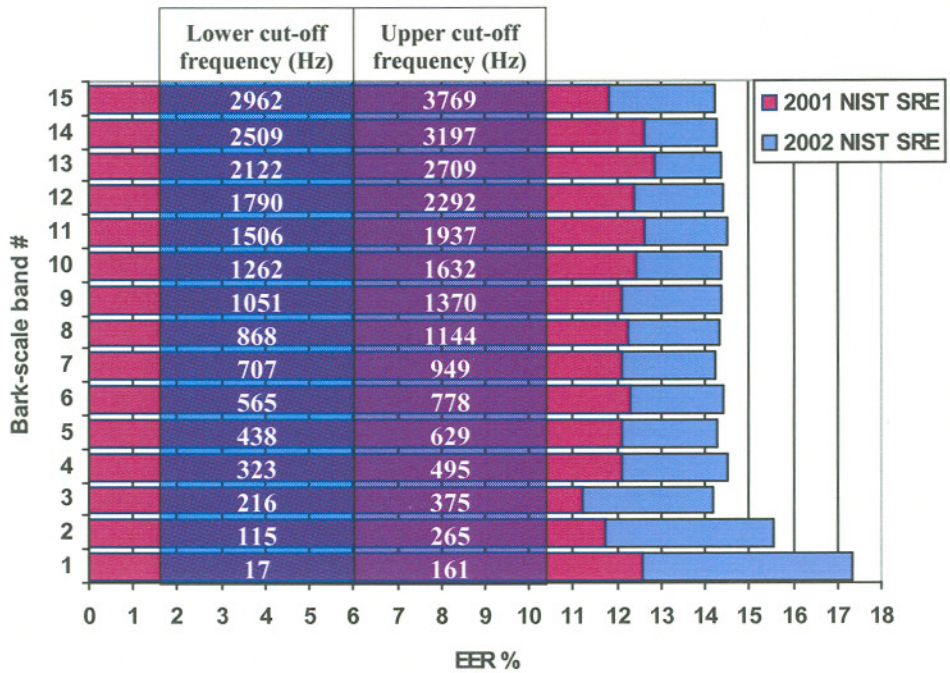

Figure 4.19: Performance for each bark-scale critical-band on the 2001 and 2002 NIST SREs when used to estimate the joint-state classes plus duration. The lower and upper cut-off frequencies are shown for each critical-band.

The performance of the bigram modeling of each frequency band shows that the first bands are more affected by the mismatched-handset condition than others. The performance of the first critical-band is significantly worse than the respective baseline (the joint-state classes plus duration that is based on the short-term energy contour) for both tasks. The second critical-band is also worse than the baseline for the 2002 NIST SRE. The reduced performance of the first two critical-band performance is expected because we are dealing with narrow-band telephone speech (300-3400 Hz) [170], and because channel variability is higher in lower bands whereas the speaker variability is higher in higher bands [21].

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 [172-175]. Lavner et al. [13] 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.

Since the performance 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 4.20 shows the fusion performances of some of the combinations.

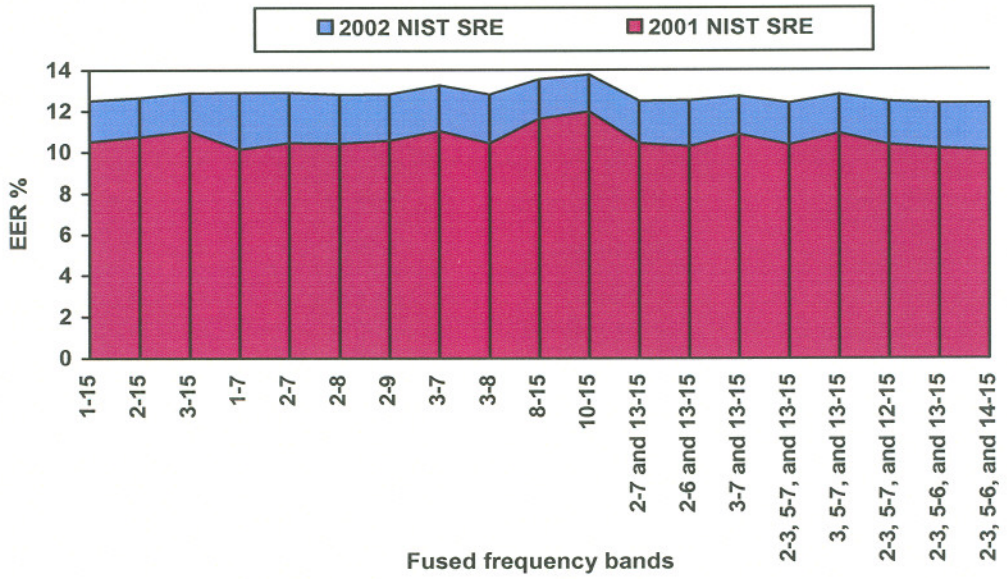

Figure 4.20: Performance of the fusion of different frequency bands on 2001 and 2002 NIST SREs.

The first fusion experiment combines the detection scores from all 15 frequency bands. The EER of this fusion is 10.5% for 2001 NIST SRE and 12.5% for 2002 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 5 possible 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 8^{th} to 15^{th} 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 [171].

Despite the insignificant difference between the performances on the two databases, the comparison across databases shows that the results are not random due to some nonzero correlation between the performances. Since the performances of the combinations on both NIST SREs seem to have some linear relationship, a significance test is used to check whether there is a correlation between the performances. The null hypothesis is that there is no correlation between the performances. With 19 samples and a correlation coefficient of 0.869, we computed a p-value smaller than 0.001. Thus, we conclude that the sample correlation is not due to chance, and the performances do have some correlation.

The best performance is achieved by the fusion of the 2^{nd}, 3^{rd}, 5^{th}, 6^{th}, 14^{th}, and 15^{th} 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 2002 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. [170]. They 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.

4.8 Summary

This chapter presented several approaches that use the dynamics of F0 and energy contours to capture prosodic differences between speakers. The approaches convert the speech signal into a sequence of discrete units that represent the dynamics of F0 and energy contour within a given segment. Then, using a simple statistical model, such sequence is used to discriminate among speakers.

The method presented in Section 4.1 uses segments estimated from the F0 contour to model the joint-state of the dynamics of the F0 and short-term energy contours. It segments both contours using a piecewise linear model of the F0 contour to detect the

segment boundaries. Then, each segment is classified into 5 classes according to the rate of change from both contours within the segment. The performance of this approach on the Switchboard I shows a comparable result to the prosodic baseline, which uses the F0, short-term energy and its first-derivatives. Then, the duration of the segment is integrated to the class category. The aim of integrating duration is to capture the speaking style (i.e., speaking rate, rhythm, and stress patterns) of the speaker. For example, a fast talker will have more pitch accents in an utterance that causes more frequently changes in both contours, thereby producing several small segments. The performance of integrating the duration is better (6% relative improvement) than the prosodic baseline. The results show that the temporal information of both contours can provide speaker-specific information. In addition, results from the modeling both contours individually show that the interaction between both contours can carry speaker-specific information. Despite the remarkable performance, this approach disregards that prosodic phenomena are also realized by changes in the energy contour.

In order to capture the variations in the dynamics of both contours, the method presented in Section 4.3 uses the changes in the dynamics from each contour to segment the speech signal. The delta features are used to detect the changes in the contour dynamics. The results show that the independent estimation of the dynamics of F0 and energy provide a better characterization of the speaker-specific information than using only the F0 contour to detect the segments. We also analyzed the performance of the proposed approach with respect to several factors: speaker demographics, amount of training data, and different handsets. We showed that this approach is more robust to mismatched-handset condition than the acoustic baseline. However, the sparsity associated with the long-term characteristic of the segment classes requires considerable amount of data (at least 4 minutes) for estimating the speaker models. The performance analysis across different tasks also showed that the small age difference between target speaker and impostor affects the systems performance.

We showed that the joint-state classes provide complementary information to conventional speaker detection systems. In a fusion experiment with the prosodic baseline, the performance improves 35% relative to the baseline. This result shows that the proposed approach captures information that is not represented in the prosodic

baseline. The fusion of the prosody-based systems with the state-of-the-art system (acoustic baseline) also yields an improvement in performance. This result agrees with the previous findings that prosodic information complements spectral information.

Some extensions to prosodic modeling were presented in this chapter. The first approach increases the number of quantization levels to 4 for describing the dynamics of each contour. The bigram modeling of the new representation yields a 25% improvement over the approach that uses 2-level to quantize the rate of change. However, the addition of duration does not yield any significant improvement. Given that the segmentation produces even smaller segments than the 2-level quantization, it seems that the segment duration is not consistent for the same speaker. The second approach extends the concept of modeling the dynamics of two different streams by replacing the short-term energy by frequency-band energy. The motivation of this approach is that the 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 best performance is achieved by fusing the scores from 6 frequency bands: 4 below 600 Hz and 2 above 2500 Hz. Note that such frequency ranges have been long acknowledge to carry more speaker-specific information than the remainder frequencies.

Chapter 5

Phonetic-based Modeling of Pitch and Intensity Dynamics

The literature has shown that modeling of a sequence of phones or phonemes can capture differences in phonetic realization or lexicon [16, 28, 29, 77-79, 160, 176]. However, when a person produces speech, the native language imposes limits on variation in the integration of the segmental and suprasegmental levels [48], as discussed by Lehiste [39]:

> *"Languages seem to differ with respect to the distribution of the fundamental frequency contour over the voiced portion of the syllable. [A deaf subject] produced the word feel with a fundamental frequency movement that continued into the final /l/; the result sounded nonnatural and nonnative."*

The necessity and nature of temporal integration exemplified in the quotation from Lehiste shows that the loudness as well as pitch will have to be correctly aligned with the segmental stream. Therefore, we integrate the joint-state classes from the F0 and energy contours with segmental information to create a new set of classes that can capture the interaction between the prosody and the segmental information. Section 5.1 and 5.2 describe, respectively, methods to estimate the new set of classes using segmental information obtained from a large-vocabulary continuous speech recognition system and a phone recognizer. Section 5.3 shows the performance of the fusion between the phonetic-based classes and the baselines. Section 5.4 shows the results on the 2002 NIST

SRE. Section 5.5 analyzes the effects of amount of training data, speaker gender, and handset on the performance of the phonetic-based classes.

5.1 Phoneme-based Joint-state Modeling

This section describes the phoneme-based joint-state modeling approach. This approach produces a sequence of classes estimated from a time-alignment between a sequence of joint-state classes (as described in Chapter 4) and a sequence of English phonemes. First, the phoneme sequence and the joint-state classes of F0 and energy contours are estimated from the speech signal. Then, both sequences are time-aligned. Second, a new sequence of segments is estimated using the boundaries from both sequences. Finally, each new segment is labeled with a concatenation of the phoneme class and joint-state class that correspond to time interval of the new segment. Figure 5.1 shows an example of phoneme-based joint-state modeling. The segments labeled as silence are not used for modeling. The number of phoneme-based joint-state classes depends on the number of phoneme classes.

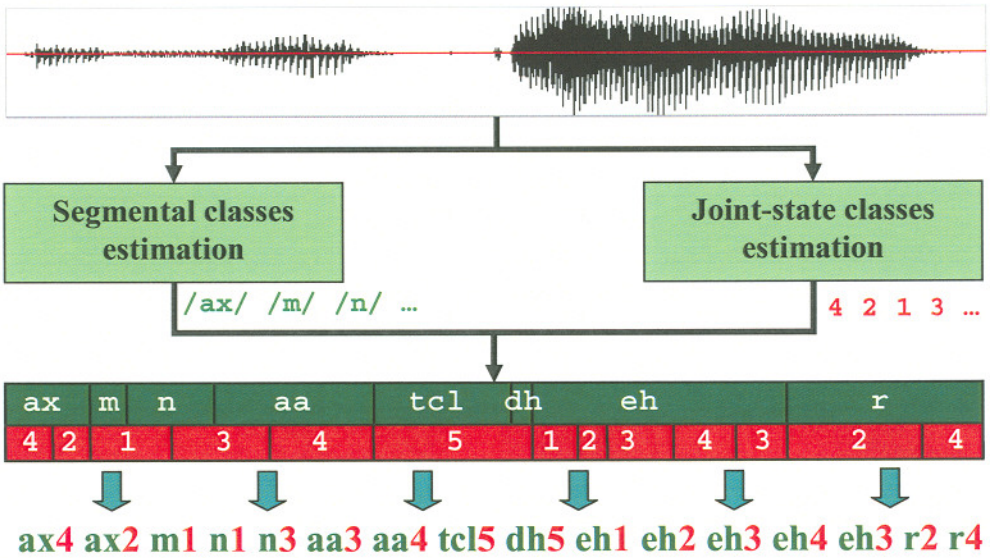

Figure 5.1: Diagram of the speech signal segmentation process using prosodic and segmental information.

The sequence of phonemes is generated using a large-vocabulary continuous speech recognition (LVCSR) system. The phoneme sequence and timing information are generated by the SRI's large vocabulary conversational speech recognition system [177]. Besides the word sequence, the system also outputs a time-aligned sequence of 43

possible phonemes from a dictionary and a "silence" label. The word error rate of this system is 30.2% on Switchboard I. Since the LVCSR system can generate 43 phonemes, the phoneme-based joint-state modeling has 215 classes, which is the product between 43 phoneme classes and 5 joint-state classes.

Despite the errors from the LVCSR and the F0 estimation systems, no particular processing is performed during the alignment of both sequences to deal with any type of error from both systems. Therefore, errors from the LVCSR and F0 estimation systems are propagated to the alignment process. For example, an error in the detection of the phoneme boundaries affects the duration information; a missed detection of voicing causes an alignment between an unvoiced label and a phoneme that is produced during voicing. To deal with some of the errors caused by the systems, after the alignment, the new segments with duration lower than a certain threshold are discarded.

5.1.1 Integrating in Duration Information

The duration information can be estimated from three possible sources: phoneme class, joint-state class, or the generated segment from the alignment of the sequence of phonemes and joint-state classes. The first source assumes that every phoneme has some intrinsic duration associated with it that is independent of the prosodic level, despite the context influence in duration (e.g., shortening of vowels before voiceless obstruents in English). The second source assumes that the prosodic information affects the segmental stream. The third source assumes that both streams (suprasegmental and segmental) affect each other during integration. Our preliminary experiments show that the last two sources perform similarly. The latter source was chosen as the basis for integrating in duration information.

The segment duration is quantized into two levels: Short and Long. A quantizer is estimated a priori by selecting the median duration of all segments. Short is assigned to segments shorter than 6 frames (60 ms). For example, the sequence presented in Figure 5.1 is labeled as follows: **ax4S ax2S m1S n1S n3S aa3L aa4L tcl5S dh5S eh1S eh2S eh3S eh4S eh3S r2L r4S**. The addition of the duration label increases the set of classes to 430 elements.

5.1.2 Speaker Detection Results

Before speaker modeling, all the segments with duration shorter than 30 ms are discarded. A bigram modeling is used to model the sequence of phoneme-based joint-state classes. The scoring process uses only the bigrams that occur more than 500 times in the background and target-speaker models. Doddington [27] shows that the removal of infrequent *n*-grams improves the detection performance. Figure 5.2 shows the speaker detection performance for 8-conversation training on the 2001 NIST SRE for the modeling of the sequence of phonemes alone and the phoneme-based joint-state classes.

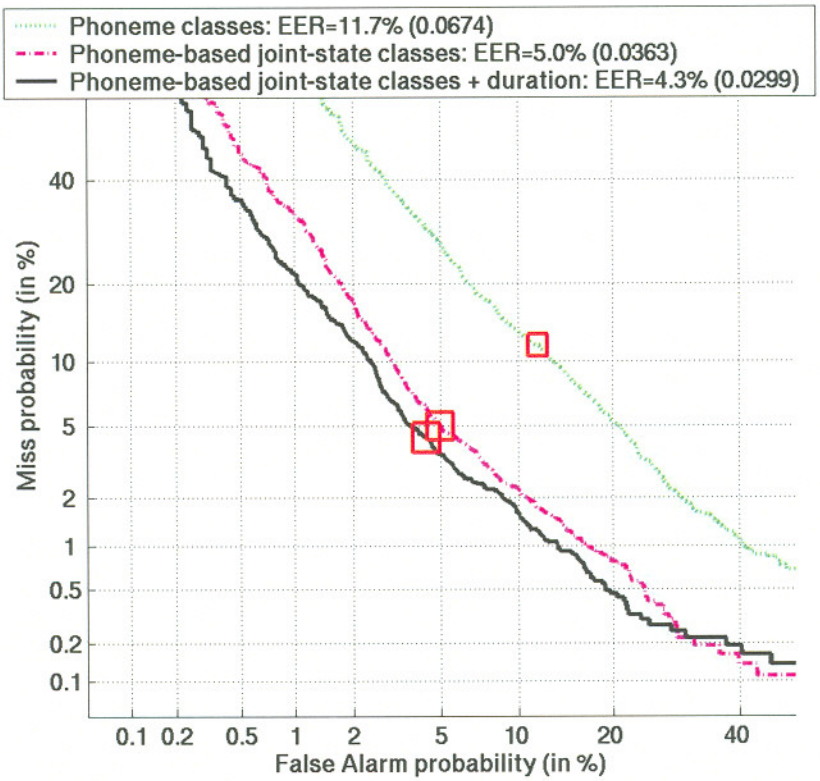

Figure 5.2: DET curve for the phoneme-based joint-state modeling.

The EER of the system using only bigrams of phonemes is 11.7%, and of the system using phoneme-based joint-state class modeling is 5%. The significant difference (α=0.01) in performance shows that the phoneme-based joint-state classes are capturing the interaction between phonemes and the F0 and energy contours. To verify that this claim is true, we also run another experiment that fuses the scores from the systems that use the phoneme classes (EER = 11.7%) and joint-state classes (EER = 14.2%) to model speaker-specific information. The EER of the fusion (at score level) of both systems is

6.7%, which is 34% worse than the modeling of both streams together. Indeed, the new modeling approach captures information that only exists in the integration between the segmental and prosodic streams.

The integration of the duration label to the classes yields an EER of 4.3%. The relative improvement over the modeling without duration is 14%. One reason for the small improvement is the limited amount of training data to estimate the bigram models. Note that the addition of duration label doubles the number of classes, but not the amount of training data. Therefore, the accuracy of the speaker models are going to be affected by the small amount of training data.

5.2 Phone-based Joint-state Modeling

There is no doubt that the output from LVCSR systems can be used to capture idiolectal differences between speakers [27]. However, LVCSR systems usually recognize only a limited set of canonical pronunciations for each word, which are defined in the system dictionary. Such limitation has motivated researchers to use language-dependent phone recognizers to obtain the sequence of sounds produced by the speaker for pronunciation modeling [16, 29, 78, 79, 176]. Differently from LVCSR systems, the output of phone recognizers is not constrained to a sequence of likely phones. Besides, phone recognizers are much simpler (with respect to the processing requirement and implementation complexity) than LVCSR systems. In this section, we examine the modeling of joint-state classes with the phonetic sequence generated by the TRAPS phone recognizer. The TRAPS phone recognizer is described in Section 5.2.1, and the results of the phone-based joint-state modeling are presented in Section 5.2.2.

5.2.1 TRAPS Phone Recognizer

The TRAPS phone recognizer uses a hybrid hidden Markov model (HMM) followed by a Viterbi search to convert the speech signal into a sequence of phones. The HMM is a network of context-independent phones, which are modeled using 3 emitting states. The probability densities in each state are modeled using a TRAPS (TempoRAl PatternS) posterior probability estimator [178-180]. The sequence of phones is obtained after the

Viterbi search, i.e., no language modeling is performed. The phone set is the same as the one used by the LVCSR system.

The TRAPS estimator focuses on the temporal characteristics of the speech signal rather than the spectral characteristics. This approach uses a collection of multilayer perceptrons (MLPs) to estimate the posterior probability of phone classes from the temporal patterns of critical-band energies. Then, the posteriors are subsequently combined (using another "merging" MLP) to produce a global estimate of the posterior probabilities. Figure 5.3 shows the diagram of the TRAPS estimator used for estimating the posterior probabilities of phone classes.

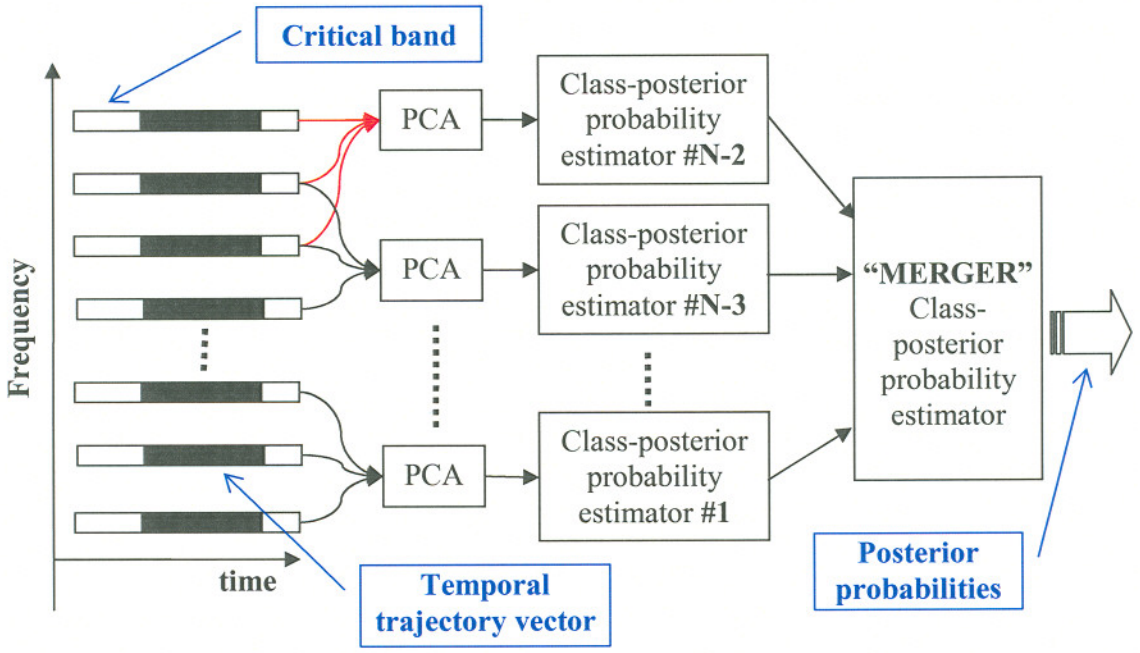

Figure 5.3: Diagram of a TRAPS-based posterior probability estimator for *N* critical-bands.

The input features to each class-posterior estimator are three adjacent 0.5-second temporal patterns of log compressed Bark-scaled critical-band energies [181, 182] projected in a lower dimension using Principal Component Analysis (PCA). Since 15 Bark-scale critical bands are estimated from the speech signal, there are 13 class-posterior estimators with 100 neurons (sigmoid activation function) in the hidden layer, and 44 neurons (softmax activation function) in the output layer. The merging MLP has similar architecture as the class-posterior estimators, except that it has 500 neurons in the hidden layer. The band-classifier MLPs and the merging MLP are trained using back-

propagation with a cross-entropy error criterion. The MLPs are trained using 68 hours of speech data excerpt from CallHome [183] (3 hours), Switchboard I & II (61 hours), and Switchboard Cellular (4 hours) [184].

5.2.2 Speaker Detection Results

The same configuration described in Section 5.1.2 is used to evaluate the speaker detection systems on the 8-conversation training condition of the 2001 NIST SRE. The DET curves for the system using phone-based joint-state modeling on 8-conversation training are shown in Figure 5.4.

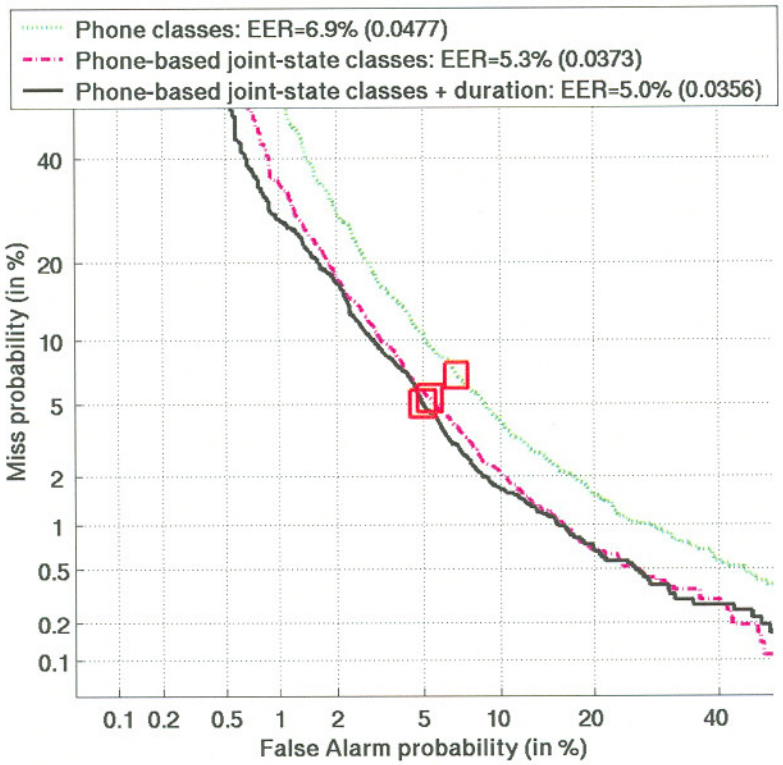

Figure 5.4: DET curve for the phone-based prosody modeling.

The performance obtained by just modeling the sequence of phones (6.9% EER) shows that the speaker's pronunciation provides a better speaker characterization than the speaker's intended selection of phonemes (11.7% EER) obtained by the LVCSR system. Appendix C shows similar results on the 2001 NIST SRE for a phone-based joint-state classes modeling that uses a conventional phone classifier (mel-cepstra-based HMM).

The improvement of the phone-based joint-state modeling over the phone sequence modeling is 23%. Even though the sequence of phones provides a better

modeling of the speaker's pronunciation patterns, the unconstrained characteristic of phone recognizers can be the source of several problems in the phone-based joint-state modeling, such as:

- Missed detection of phones and low accuracy of the phone timing information affect alignment between the phone classes and the joint-state classes. The timing information error **"a)"** in Figure 5.5 causes two errors: a class **aa5** that aligns an unvoiced segment with a voiced phone, and a duration reduction of the class **tcl5**. The missed detected phone */dh/* denoted by **"b)"** in Figure 5.5 causes two errors: a class **ey5** that aligns an unvoiced segment with a voiced phone, and a duration increase in of the class **tcl5**.
- Inconsistencies in the phone recognition (e.g., the recognition of a given phone varies across recordings) produce inconsistent phone-based joint-state classes.

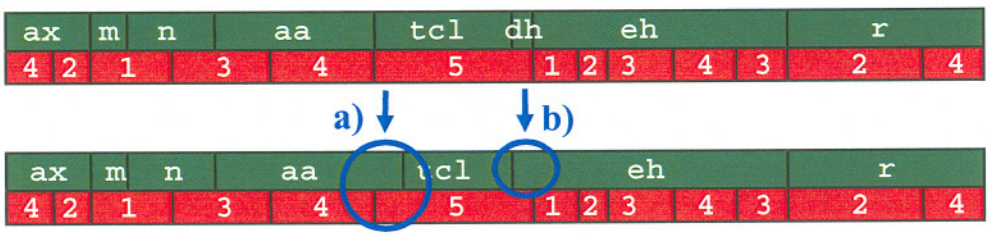

Figure 5.5: Examples of misalignment problems caused by errors from the TRAPS-based phone recognizer: a) timing information of phone boundaries and b) missed detection of phones. The correct transcription and alignment are shown at the top and the TRAPS transcription and respective alignment are shown at the bottom.

One of the consequences of misalignment between the phone sequence and joint-state classes is the generation of extra segments. In the Switchboard I, each conversation side produces, on average, about 1,532 phone classes and 2,909 phone-based joint-state classes. Note that the number of phone-based joint-state classes produced per conversation side is about 90% greater than the number of phone classes produced per conversation side. In addition, the number of phoneme-based joint-state classes produced per conversation side is only 27% greater than the number of phoneme classes produced per conversation side. Such large difference between the percentages of produced classes per conversation side suggests that the phone-based joint-state class estimation have more misalignment problems than the phoneme-based joint-state class estimation.

The integration of the duration label to the phone-based joint-state classes yields an EER of 5%. This result confirms the misalignment problem in the phone-based joint-state class estimation. That is, the misalignment between both streams generates more and smaller segments than the phoneme-based estimation. Consequently, the duration label provides little or no information about the speaker because most of the classes are labeled as "Short".

5.3 Fusion Results

Using the fusion system described in Section 3.5.1, we fused the scores of the phone-based and the phoneme-based joint-state classes systems with the acoustic baseline system. Figure 5.6 shows the performances of the acoustic baseline and the fusion of the prosodic systems (the duration information is incorporated in the phonetic-based joint-state classes) with the acoustic baseline for 8-conversation training condition on the 2001 NIST SRE.

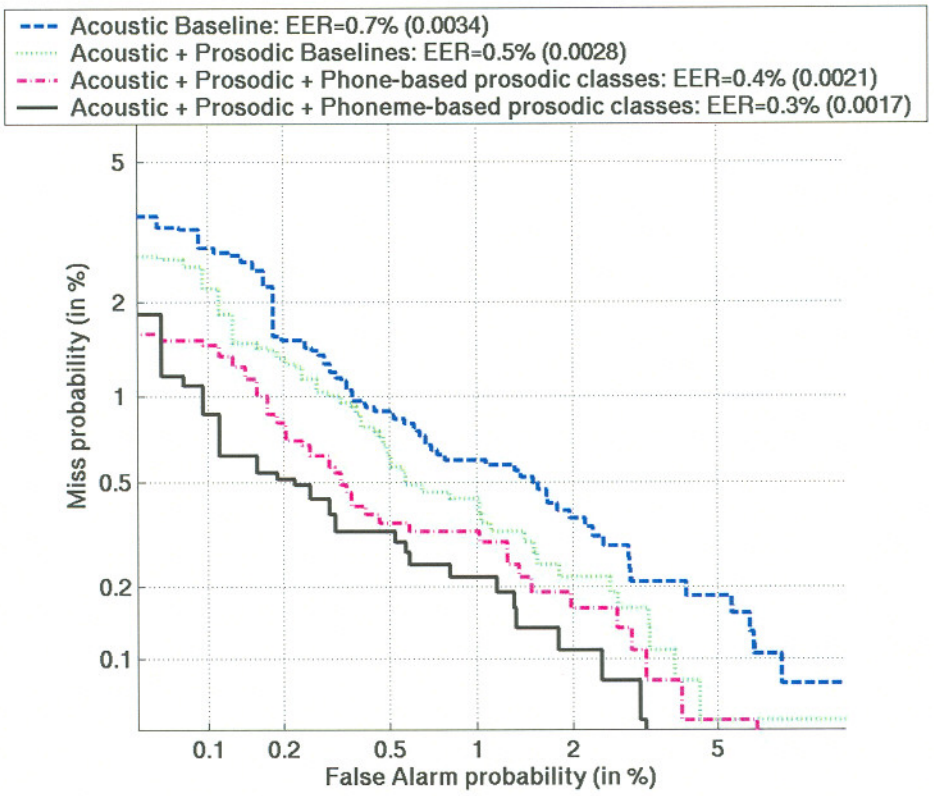

Figure 5.6: DET curve of the fusion between the acoustic and prosodic systems.

fusion system that includes the phone-based prosodic classes (2.2% EER) and 32% for the fusion system that includes the phoneme-based prosodic classes (1.7% EER).

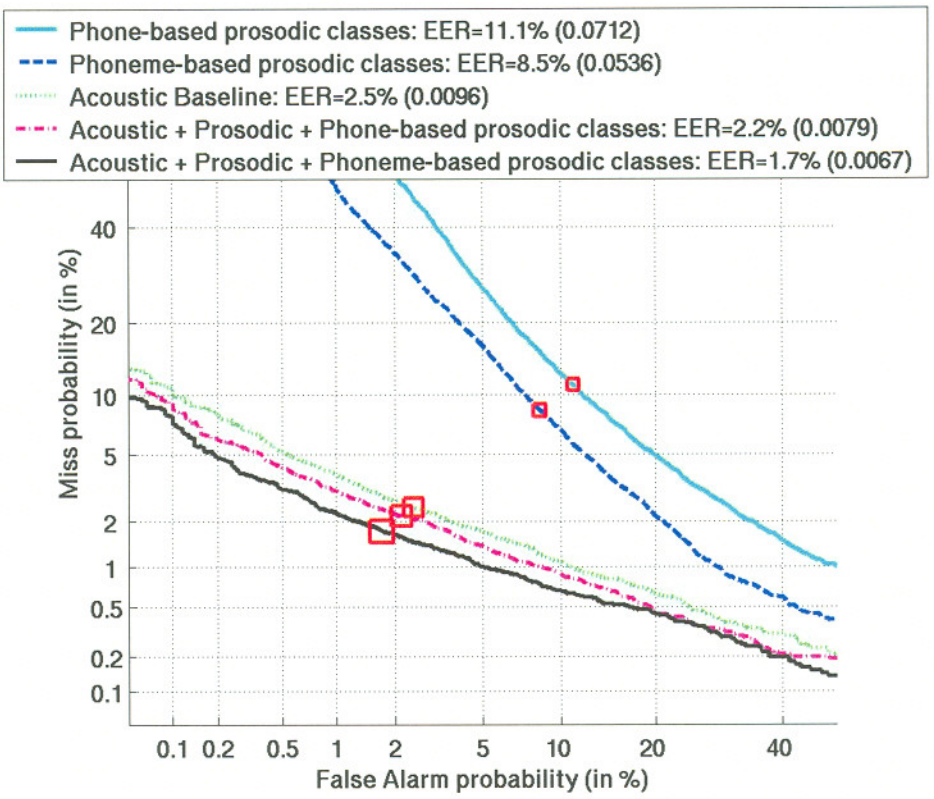

Figure 5.8: DET curves of the baselines, phonetic-based joint-state classes, and the fusion between systems on the extended-data speaker detection task in 2002 NIST SRE.

Despite the superior performance of the phoneme-based joint-state classes (EER=8.5%) over the joint-state classes (EER=14.2%), the fusion of the acoustic baseline with either systems yields the same improvement over the baseline. This result shows that the improvements in performance of the fused systems do not depend on the performance of the systems alone, but in the amount of complementary information that the fused systems have.

The performances on the 2002 NIST SRE are considerably worse than the performances obtained on the 2001 NIST SRE. Some of the factors affecting the performance are analyzed in the following section.

5.5 Factors Affecting Performance

In this section, we analyze the factor affecting the performance of the phonetic-based systems. The analyzed factors are the same ones described in Section 4.6: the amount of training data, handset mismatch between training and testing, and same-gender trials.

5.5.1 Amount of Training Data

Table 5.1 shows the performances on 1- and 8-conversation training conditions for the systems using the phoneme and phone classes alone, and the new segment classes (phoneme-based and phone-based joint-state classes) on the 2001 NIST SRE. All the joint-state classes include the duration label. Note that most of the systems perform the same or worse than the delta-based joint-state classes for 1-conversation training condition. This result shows that larger the set of classes used to characterize speaker-dependent information, more training data is required to accurately estimate the model. The results for 8-conversation training condition and relative improvement confirm the requirement of considerable amount of training data for speaker modeling.

Table 5.1: Comparison of performance of the phonetic-based systems for 1- and 8-conversation training conditions on the 2001 NIST SRE.

System	1-conversation training EER	8-conversation training EER	Relative improvement
Delta-based joint-state classes	18.3%	11.4%	38%
Phoneme classes	29.2%	11.7%	60%
Phoneme-based joint-state classes	18.2%	4.3%	76%
Phone classes	19.8%	6.8%	66%
Phone-based joint-state classes	15.8%	5.0%	68%

On 8-converstation training condition, only the performance of phoneme classes is similar to the performance of the delta-based joint-state classes. One reason could be that speaker-specific information is better represented by a longer n-gram than bigram. However, a trigram modeling of the phoneme sequence proposed by Andrews et al. [77] achieves a 13% EER on the 8-conversation training condition. Andrews's work refutes the longer n-gram hypothesis.

5.5.2 Telephone Handset

Figure 5.9 shows the performances for the phonetic-based systems alone and the fusion between the phonetic-based joint state modeling and baselines per handset condition on the 2002 NIST SRE.

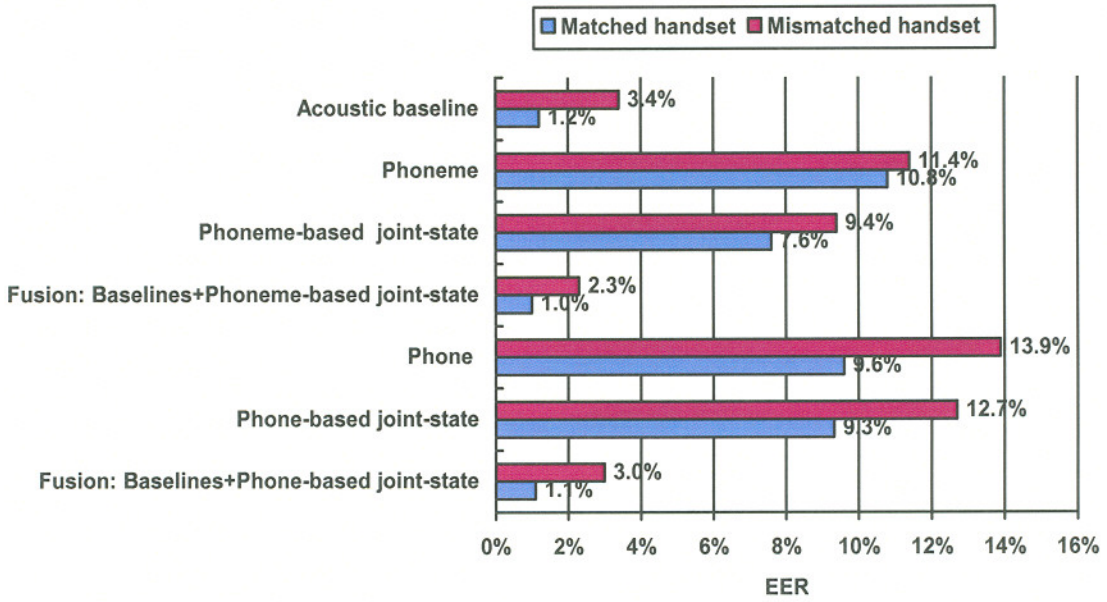

Figure 5.9: Handset condition performance of the fusion between the baselines and the phonetic-based prosodic classes.

The performances of the phoneme- and phone-based systems for both handset conditions show that the phoneme-based system is not affected as much as the phone-based system by the mismatched-handset condition. Using only the segmental information, the degradation in performance (from 10.8% to 11.4% EER) for the phoneme classes is not significant, whereas the relative degradation (from 9.6% to 13.9% EER) for the phone classes is about 45%. Consequently, any problem in the estimation of segmental and joint-state classes is likely to affect the alignment of both streams. The phoneme-based and phone-based joint-state classes have a relative degradation in performance of 24% and 37%, respectively. The robustness of the phoneme-based approach is reflected into the performance of the fusion with the baselines. The fusion of the phoneme-based joint-state classes with the baselines achieves a 32% relative improvement over the performance of the acoustic baseline under mismatched-handset condition, i.e., from 3.4% to 2.3% EER.

5.5.3 Speaker Demographics

Table 5.2 presents the performance for same-gender trials of the phonetic-based systems for 8-conversation training condition on the 2001 and 2002 NIST SREs.

Table 5.2: The EER per gender condition of phonetic-based systems for 8-conversation training condition on 2001 and 2002 NIST SREs.

Systems	2001 NIST SRE		2002 NIST SRE	
	Male-male	Female-female	Male-male	Female-female
Phoneme classes	12.6%	11.6%	11.3%	11.2%
Phone classes	6.9%	6.8%	10.1%	10.2%
Phoneme-based joint-state classes + duration	3.8%	4.9%	8.9%	6.9%
Phone-based joint-state classes + duration	5.1%	5.2%	10.7%	9.6%

The performance of the phoneme and phone class modeling does not show any significant difference between same-gender trials. However, the comparison of the results on the 2002 NIST SRE with respect to the 2001 NIST SRE shows that the performance of male-male trials for phoneme classes improves. Despite the reduction in the amount of training data for each model (approximately 29%), the phoneme modeling maintains or improves the performance across different evaluation data. This demonstrates the robustness of the LVCSR system in the phoneme sequence estimation. In addition, it seems that the phoneme class modeling is not affected by the small age difference between impostor and true speaker.

The performance of the phonetic-based joint-state classes for both same-gender trials degrades from the 2001 to 2002 NIST SRE. Given the large set of classes (430 elements), the reduced amount of training data per model affects the *n*-gram modeling. In the 2002 NIST SRE, a conversation side produces, on average, 29% fewer phoneme-based classes (phoneme and phoneme-based joint-state) and 35% fewer phone-based classes (phone and phone-based joint-state) than a conversation in the 20021 NIST SRE.

The performances of the phonetic-based joint-state classes show a significant difference between same-gender trials on the 2002 NIST SRE. Since the difference of the amount of training data per gender is very small (about 6%), it seems that the small age difference between impostor and true speaker affects more the performance of male-male

trials. In fact, the degradation of the performance from the 2001 to the 2002 NIST SRE is higher for male-male than female-female trials.

5.6 Summary

In this chapter, we presented an approach to model the interaction between prosodic and segmental information using a sequence of phonetic classes and the dynamics of F0 and short-term energy contours. This approach performs a time-alignment between both information streams to generate a sequence of symbols that represent the dynamics of prosodic features conditioned on the phonetic class (phoneme or phone).

Two approaches using the sequence of phones or phonemes were described. The main difference between the modeling of both streams is that the sequence of phonemes represents the intended linguistic message (due to the corrections made by the LVCSR), and the sequence of phones approximates the sequence of sounds realized by the speaker. Using the 2001 NIST SRE, we showed that integrating the joint-state of F0 and energy contours to the segmental information provides a better characterization of speaker-specific information. The small improvements obtained by the addition of the duration label to the phonetic-based joint-state classes illustrate the data requirement for n-gram modeling. The performances for the 2001 and 2002 NIST SRE show that the success of the phonetic-based joint state modeling depends on several factors, such as the robustness of the sub-systems (phone recognizer, LVCSR system, or joint-state class estimation algorithm), correct alignment of the prosodic and phonetic streams, consistency in the phonetic sequence estimation, and reasonable amounts of training data.

Finally, the fusion between the baselines and the phonetic-based joint-state classes shows that the proposed representation also provides complementary information. The fusion of phoneme-based joint-state classes and baselines yields at least a 32% relative improvement over the acoustic baseline on both NIST evaluations. The fusion with phone-based joint-state classes yields at least a 12% relative improvement over the acoustic baseline.

Chapter 6

Prosodic Modeling for Language Detection

Since the joint-state classes are used to characterize prosodic phenomena, this chapter demonstrates that such classes can also be used to capture language-specific information. In this chapter, we describe few modifications of the joint-state class estimation for the language detection task. Then, the joint-state class modeling is evaluated on the 2003 NIST Language Recognition evaluation paradigm. The chapter is structured as follows. Section 6.1 introduces the language detection framework of the 2003 NIST Language Recognition Evaluation (LRE). Section 6.2 describes two language detection baselines and their performances on the 2003 NIST LRE. Section 6.3 describes the modifications on the joint-state classes modeling and results on the 2003 NIST LRE. Section 6.4 shows that the joint-state classes can also provide complementary information to the baseline systems.

6.1 Language Detection Task

The goal of the language detection task is to determine whether or not a test segment of speech is from the target language. It is assumed that the test segment has only speech from one language. The language detection task has a similar formulation as the speaker detection task, i.e., hypothesis testing of two mutually-exclusive hypothesis: the target language is present, or not. Therefore, the likelihood ratio test is used to make a decision.

In this work, the systems are evaluated using the 2003 NIST Language Recognition Evaluation paradigm [185]. The evaluation data is from the Language Data

Consortium's (LDC) CallFriend corpus, which is a collection of unscripted conversations for 12 languages recorded over digital telephone. The languages are: Arabic (conversational Egyptian), English (American), Farsi, French (Canadian), German, Hindi, Japanese, Korean, Mandarin, Spanish (Latin America), Tamil, and Vietnamese. The test segments can last nominally 3 seconds, 10 seconds, and 30 seconds. The test data comprises 80 test segments of each of the three test durations, for each of the 12 target languages.

Similar to the speaker detection task, the system performance is measured using the expected cost of making a detection decision, as described in Section 3.1.1. The detection decision uses the missed detection probability (rejecting the correct language) and the false acceptance probability (detecting the wrong language for a given test segment). The costs of miss detection and false alarm are equal to 1 and the prior probability of the target language is 0.5. The error probabilities are plotted as DET curves to show the system results.

6.2 Baseline Systems

We used two systems to show that the prosody-based speech representation can provide complementary information. Both systems use a sequence of phones produced by phone recognizers to model the phonological properties of a spoken language. Among the various phonological factors, the sequence of phones can capture the phonotactics constraints and phone usage.

6.2.1 Single Phone-recognizer based Language Detection System

This system builds language models for all target languages using the sequence of phones from a single phone recognizer. Since the same phone recognizer is used across different languages, this type of approach eliminates the need for training corpora from each target language [5]. The TRAPS phone recognizer, described in Section 5.2.1, is used to estimate the sequence of English phones from a given utterance. Its phone set has 43 possible symbols.

The language detection task is divided into two phases: training and testing. In the training phase, a trigram model for each target language is estimated based on the

sequences of phones produced by the TRAPS phone recognizer for the respective language. In the testing phase, a likelihood score is computed for every language model given the phone sequence estimated from the test message. Since it is allowed to use the knowledge of the whole set of target languages, each score is normalized using the scores estimated from all language models. The diagram of the language detection system is shown in Figure 6.1.

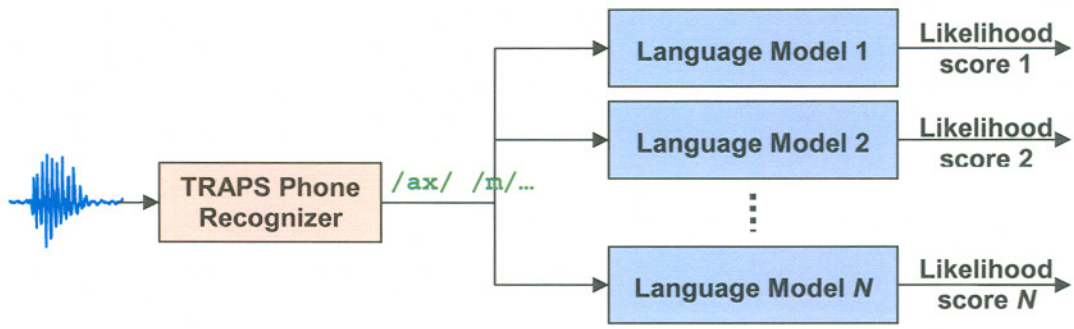

Figure 6.1: Diagram of an *N*-language detection system based on a single phone-recognizer.

Figure 6.2 shows the DET curves of the single-phone recognizer-based language detection system under the three evaluation conditions.

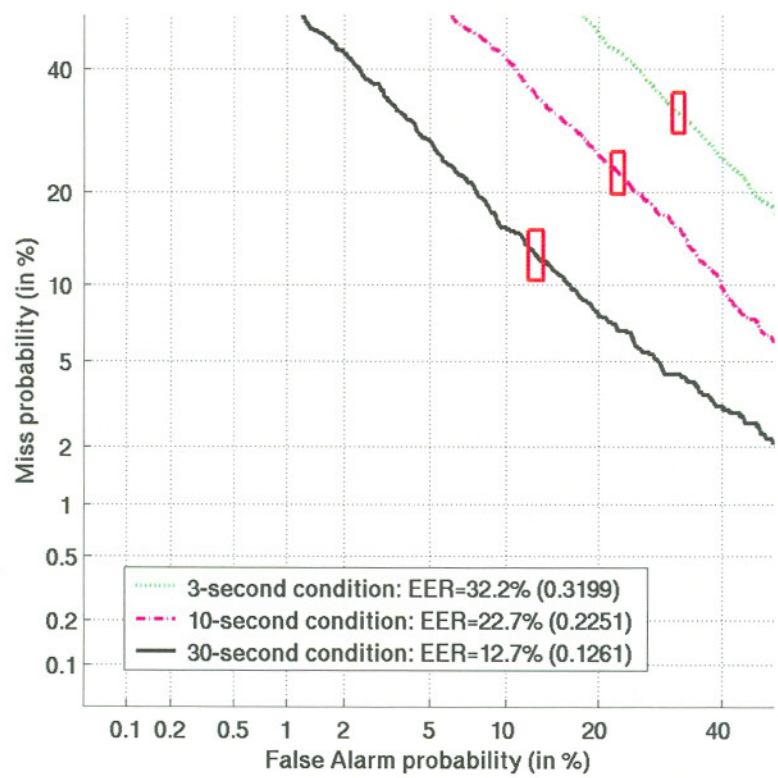

Figure 6.2: DET curve for the single phone-recognizer based language detection system.

6.2.2 Parallel Phone-recognizers based Language Detection System

Similar to the work done by Yan [15], this system uses six language-dependent phone recognizers (English, German, Hindi, Japanese, Mandarin Chinese, and Spanish) to obtain a time-aligned sequence of phones that are used to model each language [5, 14, 186]. As shown in Figure 6.3, the system is divided into three parts: front-end, language-dependent score generator, and score fusion.

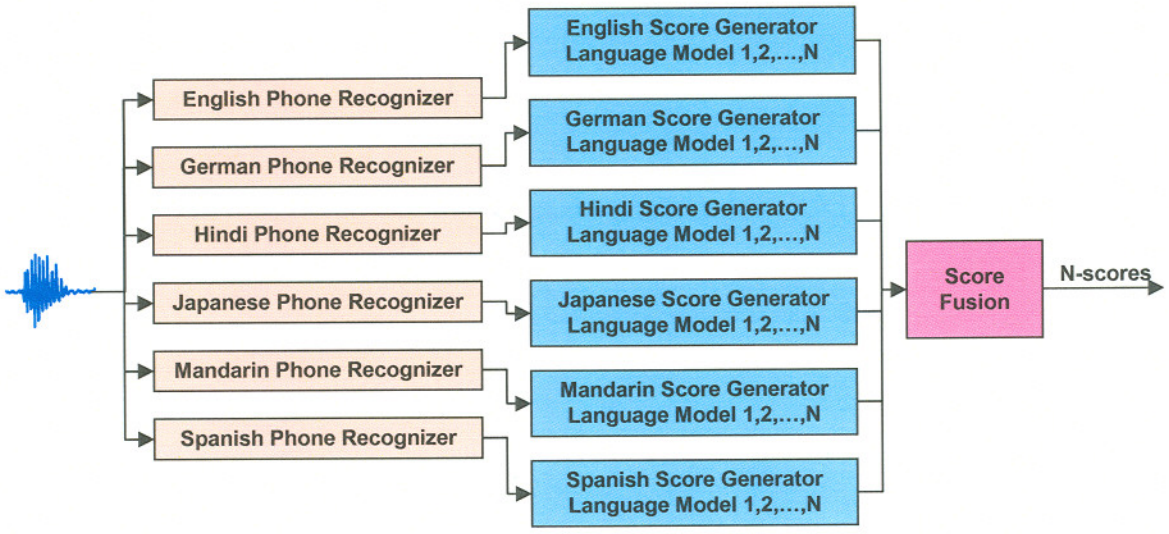

Figure 6.3: Diagram of an *N*-language detection system based on multiple phone-recognizers.

The front-end is composed by six language-dependent phone recognizers trained on the OGI Multi-language Telephone Speech corpus [137]. Each phone recognizer is a hidden Markov model trained on the respective language using Mel-frequency cepstral coefficients as the input features. Then, each HMM is retrained on CallFriend training data [185]. The phone set used by the recognizers has 221 phones (43 for English, 40 for German, 40 for Hindi, 29 for Japanese, 39 for Mandarin, and 30 for Spanish).

The language-dependent score generator estimates a likelihood score for each target language. The time-aligned phone string from each phone recognizer is scored using 12 language models trained on the language that corresponds to the phone recognizer. For example, the sequence of phones coming out from the English phone recognizer is scored against 12 language models (one model for each target language) trained on the English phone sequence. The language models are trigram models estimated from the sequence of phones recognized from the CallFriend training data.

Finally, the score fusion takes the likelihood scores from all score generators and produces the final set of scores. Since the detection task has 12 languages and the system uses 6 language-dependent phone recognizers, the score generators produce 72 scores (i.e., 6 scores for each one of the 12 languages). The classifier is a 3-layer feed-forward neural network (multi-layer perceptron neural network) trained on the CallFriend training and development scores. The neural network has a layer consisting of inputs for the 72 scores from the language-dependent score generator, 25 neurons with sigmoid activation function in the hidden layer, and 12 neurons (one for each language) with sigmoid activation function in the output layer. Figure 6.4 shows the DET curves of this system on the three conditions.

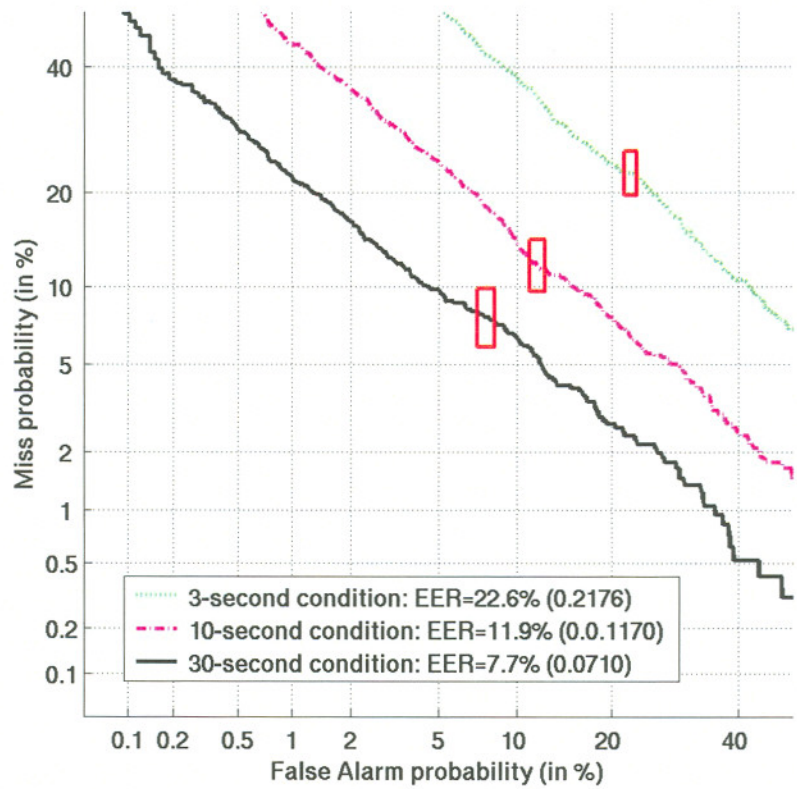

Figure 6.4: DET curve for the multiple phone-recognizers based language detection system.

6.3 Modeling Language Information using Prosodic Features

Since the speaker detection and language detection tasks share several similarities, we developed a language detection system that uses the joint-state classes to model the target languages. The estimation of the joint-state classes follows the same approach described in Section 4.3.

The time interval used to estimate the time derivatives is longer for language detection than for speaker recognition. Figure 6.5 shows the EER of the trigram modeling of the joint-state classes plus duration[1] using different time intervals for the delta features computation. The best performance of the language detection system uses time intervals between 50 ms and 90 ms to estimate the delta features. The difference between the results within this interval and the results of the remainder time intervals is statistically significant. Note that the time interval that yields the best performance for speaker detection is equal or less than 50 ms. The time interval differences between language and speaker detection shows that the details in contours account more for the speaker than language information. Thus, the experiments from now on use 90 ms delta window for estimating the joint-state classes.

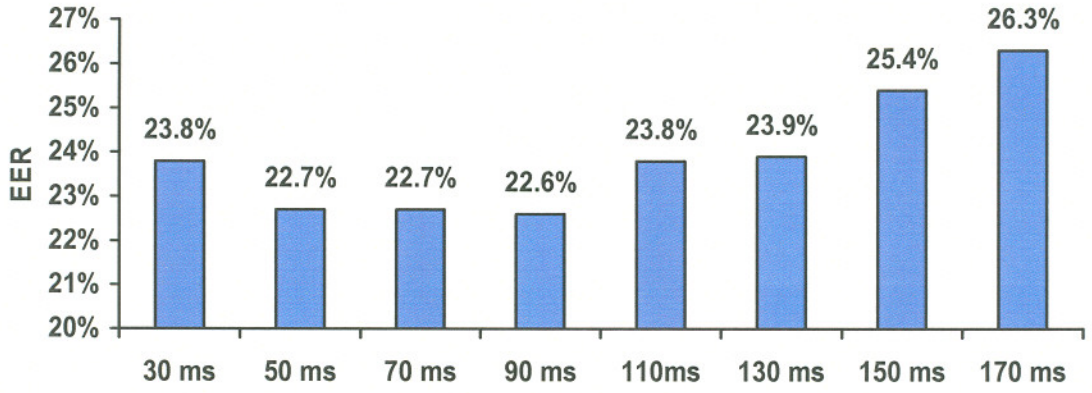

Figure 6.5: EER for different delta window lengths.

Figure 6.6 shows the performance on the three test conditions of the trigram modeling of the joint-state classes estimated using a 90 ms delta window. First, the performance of the joint-state classes seems to be less affected by the reduction of the test segment duration than the phone-based approaches. One reason is that the trigram models are better estimated than the phone-based models, because of the small number of joint-state classes. Even though the performance of the joint-state class based system is worse than the performances of the phone-based systems, it is still an encouraging result because the joint-state class based system uses a sequence of just 10 possible symbols to

[1] The values used to quantize the duration are the same as the speaker, i.e., "Short" is assigned to segments with duration less than 8 frames for voiced frames, and for segments with duration less than 14 frames for unvoiced segments.

model each language, whereas the single phone-recognizer based system uses 43 symbols and the parallel phone-recognizer based system uses 221 symbols.

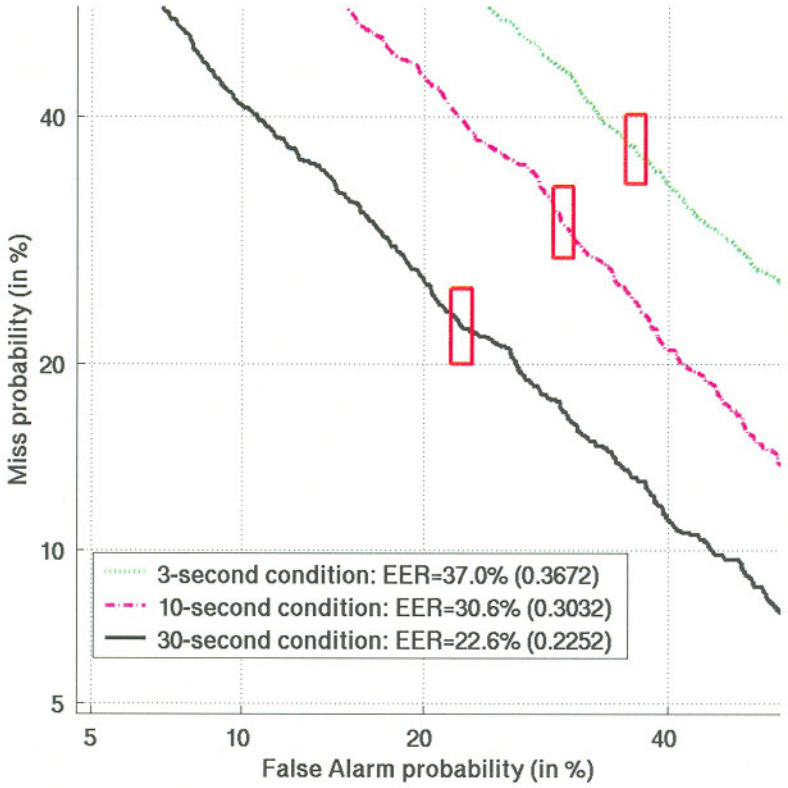

Figure 6.6: DET curve for the prosodic segments based language detection system.

Figure 6.7 shows the performance (DCF) of all three systems for each language. The results show that the joint-state classes system performs about the same as the phone-based systems for some languages. The performances of the joint-state class based system for Arabic, Vietnamese, and French languages does not show any significant difference with respect to the performances of the single phone-recognizer based system.

The joint-state classes system does not show consistent results among the tonal languages (Vietnamese and Mandarin). Contrary to the results for the Vietnamese language, the joint-state class based system performs worse for Mandarin than the phone-based systems. These results can be a consequence of how much language information can be modeled by the joint-state classes system for each one of these two languages, i.e., the dynamics between pitch and intensity play a more important role for Vietnamese than Mandarin. Such claim can be checked by fusing the prosody-based system with the phone-based systems, which is described in Section 6.4.

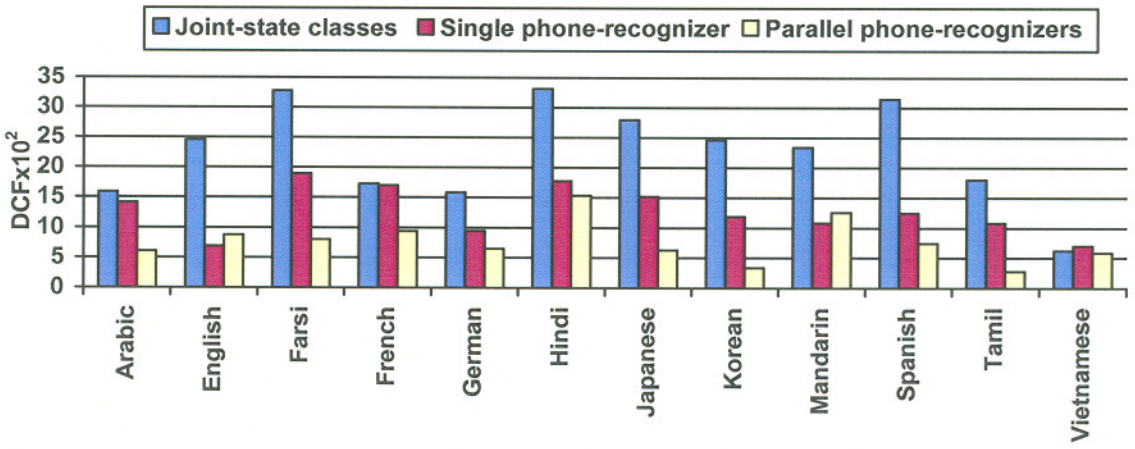

Figure 6.7: The decision cost per language for the baselines and the joint-state classes systems.

The analysis of the output of the joint-state class system showed that several languages are commonly misrecognized as each other (Appendix D). Despite the good performance of German in comparison to the other languages, German is frequently misrecognized as English, and vice-versa. This misrecognition can be result of the fact that both languages come from the same family and share prosodic characteristics. Surprisingly, other languages like Farsi and Arabic are frequently misrecognized as coming from English. The misrecognition between languages also occurs for Japanese and Korean. One reason for the misrecognition is that Korean has a similar prosodic behavior to Japanese [187]. Mandarin is also confused with Japanese, which is a consequence for using similarly the pitch. These results are also observed in language identification experiments with human subjects. Muthusamy [7] reports a confusion between the tonal languages (Mandarin and Vietnamese), pitch accent language (Japanese), and Korean.

6.4 Systems Fusion

We fuse the scores to verify whether the joint-state classes approach can provide complementary information to the baseline systems. The fusion process averages the normalized scores from each system. First, the scores of each system are mean and variance normalized. The mean and variance are estimated from the target language scores of the respective segment (normalization over different test segments is not allowed). Then, the scores are fused by averaging them. Figure 6.8 shows the DET curves

for the fused systems on the 30-second condition. The results show that fusion significantly (α=0.01) improves the performance of both phone-based systems (17% for the single phone-recognizer based system and 14% for the parallel phone-recognizers based system). Indeed, the joint-stat class based system has complementary information with respect to the phone-based systems.

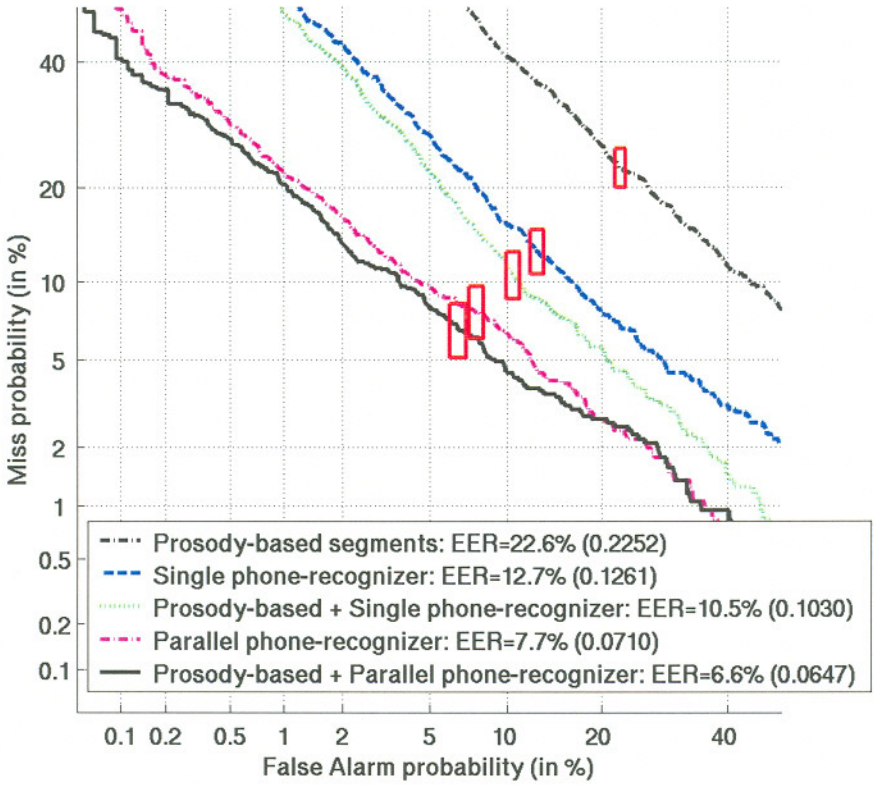

Figure 6.8: DET curves of the phone- and prosody-based systems fusion on the 30-second condition.

Table 6.1 presents the performance for each language of both fused systems and the respective relative difference in performance. As we expected, the tonal languages Mandarin and Vietnamese showed improvements in performance. A consistent improvement also happened for other languages: Arabic, Farsi, French, and German. Surprisingly, Korean and Japanese performances improve for the fusion of joint-state classes and the single phone-recognizer based system, but not for the parallel phone-recognizers based system. We can speculate that, because of the large number of speech patterns represented by the phones from the six language-independent recognizers, the parallel phone-recognizers based system is already capturing some prosodic information.

The performance of the fusion between joint-state classes and single phone-recognizer based systems for English, Hindi, and Spanish deteriorates or does not show any significant difference. The performances for Spanish and Hindi come from the poor performance of the joint-state class based system on these languages. The confusion matrix of the joint-state class based system (see Appendix D) shows that the missed detection rate is high (approximately 80%) for both languages. This also affects other languages, like English, because the poor model can cause more false alarms.

Table 6.1: Performance (DCF) of the fused systems per language. The relative difference columns show the relative difference in performance with respect to the phone-based system performance before the fusion (NS stands for non significant difference).

Language	Prosody+English Recognizer	Relative Difference	Prosody+Multiple Recognizers	Relative Difference
Arabic	0.1085	23.3%	0.0483	20.6%
English	0.0739	NS	0.0767	12.9%
Farsi	0.1466	22.5%	0.0631	21.8%
French	0.1165	31.4%	0.0824	12.6%
German	0.0869	8.4%	0.0500	22.8%
Hindi	0.1926	-8.6%	0.1580	NS
Japanese	0.1187	21.8%	0.0648	NS
Korean	0.0841	29.5%	0.0420	-25.4%
Mandarin	0.0949	12.5%	0.0938	25.6%
Spanish	0.1455	-16.4%	0.0756	NS
Tamil	0.0824	24.1%	0.0239	NS
Vietnamese	0.0131	81.4%	0.0097	83.4%

6.5 Summary

In this chapter, we presented a modification of the joint-state modeling of the dynamics of F0 and short-term energy contours to characterize language-specific information. This approach is based on the notion that different languages differ in the realization of prosodic phenomena. Such differences can be detected by modeling the dynamics of F0 and energy contours. Since the spoken language is the relevant information, the modeling was modified so that the speaker-specific information is suppressed from the joint-state class estimation. We showed that increasing the time interval used for estimating the rate of change can improve the modeling of language-specific information. We found that estimating the rate of change over time intervals between 50 ms and 90 ms yields the best

performance of the joint-state classes on the 2003 NIST Language Recognition evaluation.

The joint-state class based modeling often misrecognizes languages that shares prosodic characteristics. Some cases include the duo German and English, and the trio Korean, Japanese, and Mandarin. Given that humans rely on prosodic information to discriminate languages, this misidentification is also reported for language identification experiments that involve human subjects.

We showed that the prosody-based modeling can provide complementary information to the conventional systems. Results show that the prosody-based modeling can yield a relative improvement of 17% when fused with one of the most successful language detection system.

Chapter 7

Conclusions

This thesis presented new approaches to model prosodic information using the rate of change of F0 and short-term energy contours. Since different speakers and different languages may be characterized by different prosodic patterns (e.g., intonation, stress, and rhythm), such modeling was used to characterize speaker- and language-specific information.

In Chapter 4, we described two methods to convert the speech signal into a sequence of classes that represent the dynamics (i.e., rate of change) of both F0 and short-term energy contours. The first method uses the segments obtained from the F0 contour to segment the speech signal. A F0 stylization algorithm is used to segment the F0 contour, where each F0 segment represents a prosodic gesture (rising or falling) of pitch. Then, each speech segment is classified into 5 classes according to the rate of change of both contours within the segment. Using the 2001 NIST evaluation paradigm, we showed that the bigram modeling of the sequence of classes representing joint-state of both contours captures more speaker-dependent information than modeling the contours individually. Since the duration of speech segments can vary across classes, an extra label representing a three-level quantization of the segment duration is added to the joint-state class. The performance of the bigram modeling of the segment classes and duration is better than the prosodic baseline system, which is based on the statistics of the distributions of F0, short-term energy, and their respective time derivatives. This result shows that there is speaker-specific information conveyed in the temporal aspects of F0

and energy contours, and in the manner in which both features interact with each other to produce certain prosodic phenomena. The second method extends the first one by making the segmentation dependent on both contours, and using delta features to detect segment boundaries. These modifications reduce the complexity of the system by removing pre-processing of the stylization algorithm, and provide the means of controlling the amount of detail (by reducing or increasing the time interval used for the delta features estimation) in estimating the contour dynamics. Since the speech signal is segmented using the changes in the dynamics from both contours, the length of the segments is smaller than the F0-based segmentation method. Therefore, the duration quantization is modified so that two labels (short and long) are used to represent the segment duration. Despite that this approach has a reduced set of discrete units (10 symbols: 5 segment classes times 2 duration labels) compared to the previous approach (15 symbols: 5 segment classes times 3 duration labels), the performance of the bigram modeling of the delta-based approach is about 19% better than the performance of the F0-based approach. This result shows that incorporating the changes in the dynamics of the energy contour in the segmentation and using delta parameters to approximate the rate of change provide a better characterization of speaker-specific information. Since most of the study uses the evaluation data from the 2001 NIST SRE, we also show that all the results also hold when using the evaluation data from the 2002 NIST SRE. The same systems were evaluated using the new evaluation data, i.e., no modification was performed on the systems configuration.

We also analyzed the effect of several factors on the performance of the speaker detection systems. Given the long-term characteristics of the joint-state classes, the speaker detection systems based on joint-state classes requires considerable amount of training data for an adequate speaker modeling. The effect of different handsets used for training and testing does not affect the joint-state classes as much as it affects the acoustic baseline. Contrary to the prosodic and acoustic baseline, the age difference between impostor and true speaker affects more the only female trials than only male trials for the joint-state classes.

Another goal of the proposed approach is to provide complementary information to conventional systems. The fusion between the prosodic baseline and the joint-state

classes system yields a 35% relative improvement in performance on the 2001 NIST SRE, and 21% relative improvement on the 2002 NIST SRE. This shows that the proposed approach provides speaker-specific information that is not captured by the statistics of the F0 and energy distributions. The fusion between the acoustic baseline and the prosodic systems also improves the performance of the acoustic baseline on the 2001 NIST SRE. The improvement of the state-of-the-art system shows that prosodic systems provide complementary information. Despite the remarkable results on the 2001 NIST SRE, the fused systems yield a reduced improvement over the acoustic baseline on the 2002 NIST SRE, which is the result of the factors such as the smaller age difference between impostor and true speaker, and different handsets for training and testing.

Two extensions to the joint-state class modeling approach were also presented in Chapter 4. The first approach uses 4 quantization levels to describe the dynamics of the F0 and energy contours. The performance of such approach is better than the 2-level approach, except when duration label is added to the classes. This result does not come as a surprise because as the number of quantization levels increases, more details of the contour are captured by the segmentation. Consequently, the segmentation process produces more, but smaller, segments such that the duration information provides little or no information about the speaker. The second approach extends the concept of modeling the dynamics of two different streams by modeling the interaction between F0 and different frequency bands. The goal is to exploit the variability of different frequency bands due to phone, speaker, and channel information. The performance of the bigram modeling of the individual frequency bands yields similar performance to the short-term energy. Given the several sources of information affecting different parts of the speech spectrum, we showed that a fusion of only 6 frequency bands (below 600 Hz and above 2500 Hz) yields a better performance than the short-term energy.

In Chapter 5, we presented an approach to model the dynamics of F0 and short-term energy together with segmental information. The goal is to capture the interaction between segmental and prosodic information that is speaker dependent. This approach uses the segmentation from the F0 and short-term energy contours and some segmental information to create new segment classes. Such classes represent the dynamics of prosodic features conditioned on the phone or phoneme type (segmental information).

We showed that integrating the joint-state of F0 and energy contours to the segmental information can provide a better modeling of speaker-specific information. Results on the 2002 NIST SRE show that the alignment of both streams plays an important role in the performance. Most of the alignment problems are due to the lack of robustness and inconsistent results in the estimation of the segmental stream.

The fusion between the baselines and the phonetic-based joint-state classes shows that the proposed representation provides complementary information. The fusion of phoneme-based joint-state classes and baselines yields at least a 32% relative improvement over the acoustic baseline on both NIST evaluations. The fusion with phone-based joint-state classes yields at least a 12% relative improvement over the acoustic baseline.

Since the structure of prosody is not exactly the same for every language, we described in Chapter 6 how the joint-state classes are used to discriminate languages. The goal is to capture the most common prosodic patterns produced by a given language. Since the relevant information is the spoken language, the modeling was modified so that the speaker-specific information is suppressed from the joint-state class estimation. We showed that increasing the time interval used for estimating the rate of change improves the modeling of language-specific information. The trigram modeling of the joint-state classes based on the rate of change estimated over time intervals between 50 ms and 90 ms yields the best performance. We also showed that the joint-state class modeling can provide complementary information to the conventional systems. Results show that the prosody-based modeling can yield a relative improvement of 17% when fused with one of the most successful language detection system.

7.1 Future Work

In this thesis, we showed that the joint-state classes representing the dynamics of F0 and short-term energy contour are used for characterizing speaker- and language specific-information. However, the relationship between prosodic phenomena and the joint-state classes was not established in this work. Further research is required to study the relationship between certain patterns in the sequence of joint-state classes and the prosodic phenomena, such as intonation, stress, or rhythm.

The results presented in this work are only applied to the speaker and language detections tasks. However, these are not the only applications that can get benefit from prosodic information. It will be interesting to investigate how the joint-state classes can be used to provide prosodic information to different speech-based applications, such as speech segmentation into topic and sentences, emotion recognition, speech synthesis, and dialect recognition.

Despite the *n*-gram modeling showed to be an efficient method for modeling the joint-state classes, bigram modeling does not exploit longer dependencies between classes. The development of different approaches that can explore complex dependencies between the joint-state classes seems to be the next step. However, such development must take into account the data sparsity of the joint-state classes. The data sparsity also encourages the investigation of adaptation techniques to overcome such problem.

The segmentation technique can also be used to exploit the interaction of different information streams besides pitch and intensity. Traditionally, different information streams have been modeled as coming from independent sources of information. The proposed segmentation technique provides the framework to investigate and exploit the temporal aspects and interaction between different information streams.

In the phonetic-based joint-state modeling, both streams are integrated by performing a time-alignment of both streams. However, estimation errors coming from either of the streams affect the accuracy of such alignment. Instead of performing a time alignment of both streams, a method that takes into account the information being aligned would reduce the errors from the stream estimation. For example, the alignment method would never put together an unvoiced segment from the prosodic stream with a vowel from the phonetic stream.

Appendix A

Speaker Recognition Evaluation: Amount of Training Data Factor

One of the factors affecting the performance in the speaker recognition evaluation is the amount of training data, as shown in Section 4.6.1. In this appendix, we present some statistics about the amount of training data generated per conversation side for both NIST Speaker Recognition evaluations.

Figure A.1 shows the statistics of the short-term feature vectors generated per conversation side for each gender for the prosodic baseline. Since F0 is one of the dimensions of the short-term feature vector, the models are estimated using only the feature vectors from voiced speech regions.

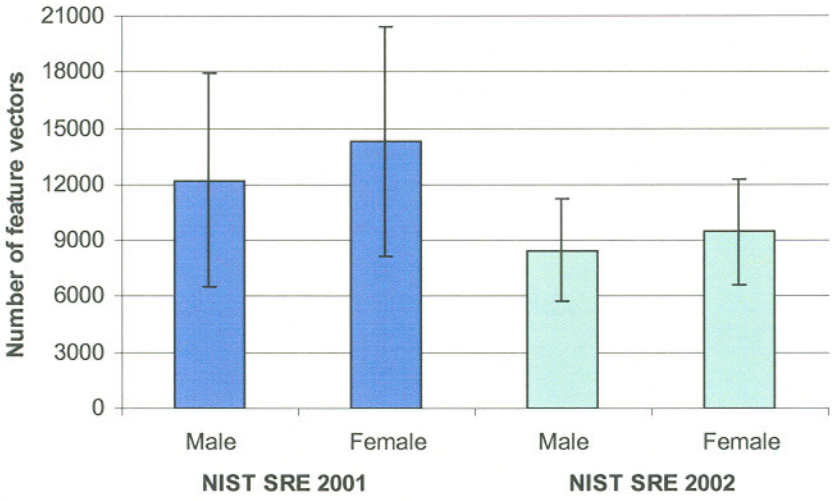

Figure A.1: Statistics of the feature vectors generated per conversation side for both NIST evaluations. The average number of joint-state classes is represented by a bar and the standard deviation by the length of a vertical line.

On average, conversation sides from female speakers produce more feature vectors than male speakers: approximately 17% and 12% for the evaluation data of the 2001 and 2002 NIST SRE, respectively. Another difference is that the number of feature vectors per conversation side decreases from the 2001 NIST SRE evaluation to the 2002. On average, the conversation sides from the 2002 NIST SRE produce 32% fewer feature vectors the conversation sides from the 2001 NIST SRE. Such difference in the number of feature vectors across evaluations is due to the use of a speech activity detector [143] to select the speech regions for further processing in the 2002 NIST SRE. Such processing is not performed in the 2001 NIST SRE because speech/silence segmentation was already performed in the Switchboard I corpus by the Institute for Signal and Information Processing (ISIP) at Mississippi State University [188].

Figure A.2 shows the statistics of the phoneme and phone class symbols generated per conversation side for each gender and NIST evaluation. Note that both genders have a balanced number of class symbols per conversation sides in both NIST evaluations. However, the number class symbols per conversation across evaluations decreases for both sequences. On average, the conversation sides from the 2002 NIST SRE produce 32% fewer class symbols than the ones from 2001 NIST SRE. Such difference is also result from the use of the speech activity detector mentioned before.

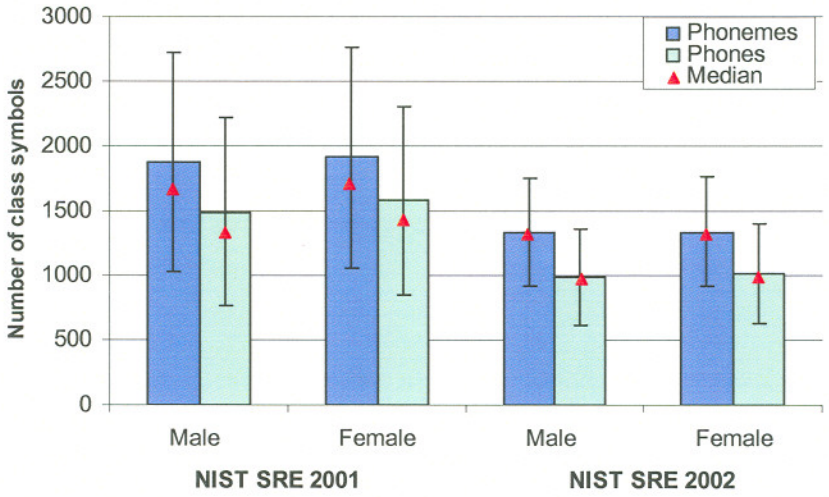

Figure A.2: Statistics of the phoneme/phone classes generated per conversation side for both NIST evaluations. The average number of phoneme/phone classes is represented by a bar, the standard deviation by the length of a vertical line, and the median value by a triangle.

Another difference is that the average number of phoneme class symbols is larger than the number of phone class symbols per conversation. In the 2001 NIST SRE, a

conversation side has about 19% fewer phone class symbols than phoneme class. The addition of the speech activity detector in the 2002 NIST SRE increases the difference between the number of phone and phoneme class symbols to 25%. One reason is that the errors in the recognition together with the unconstrained sequence of phones can produce short segments (shorter than 30 ms, according to the configuration described in Section 5.1.2) that are removed for the speaker modeling. In addition, given the long-term characteristic of the features used by the phone recognizer (0.5 second temporal trajectory), phones with short duration are likely not recognized.

Figure A.3 shows the statistics of the joint-state class symbols, described in Section 4.3, generated per conversation side for each gender and NIST evaluation. Similar to the short-term feature vectors used for the prosodic baseline, the conversation side from female speakers generates more joint-state class symbols (approximately 13%) than male speakers'. Note that the number of class symbols per conversation side is similar for both evaluations. Differently from the feature vectors estimation for the prosodic baseline, the joint-state class approach does not use a speech activity detector to select the speech regions for processing. This shows that the speech activity detector is indeed responsible for the fewer features vectors per conversation on the 2002 NIST SRE.

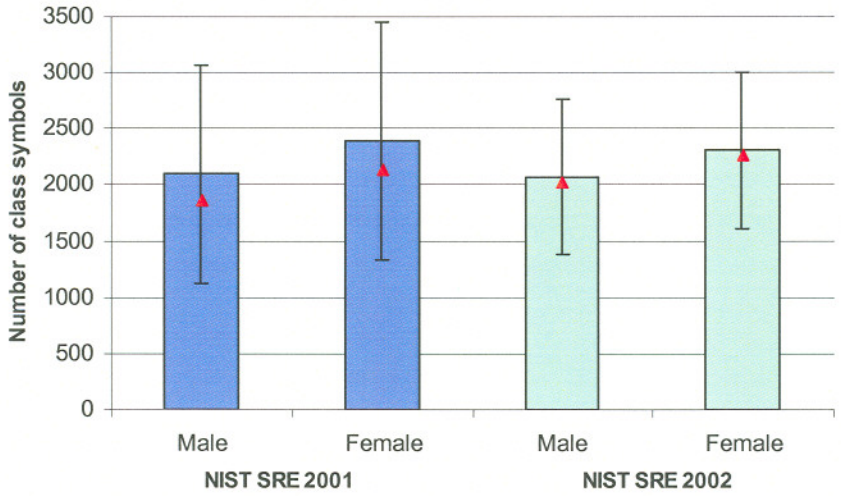

Figure A.3: Statistics of the joint-state classes generated per conversation side for both NIST evaluations. The average number of joint-state classes is represented by a bar, the standard deviation by the length of a vertical line, and the median value by a triangle.

Figure A.4 shows the statistics of the phonetic-based joint-state class symbols, described in Chapter 5, generated per conversation side for each gender and NIST evaluation. Besides the characteristic difference between the numbers of class symbols per gender, the integration between joint-state classes and phone classes produces a larger number of class symbols than the integrating between joint-state classes and phoneme classes.

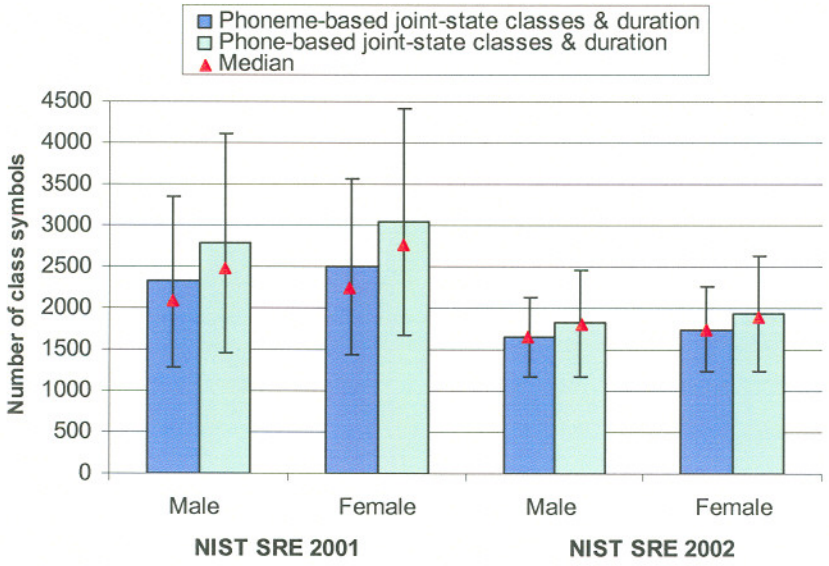

Figure A.4: Statistics of the phonetic-based (phoneme/phone) joint-state classes generated per conversation side for both NIST evaluations. The average number of phonetic-based joint-state classes is represented by a bar, the standard deviation by the length of a vertical line, and the median value by a triangle.

Appendix B

Speaker Recognition Evaluation: Age Factor

Figure B.1 and Figure B.2 show the distribution of speakers per age group from the 8-conversation training condition for 2001 and 2002 NIST SREs, respectively. The caller age histogram of the 2002 NIST SRE in Figure B.2 shows skewness to the right because most of the callers are college students.

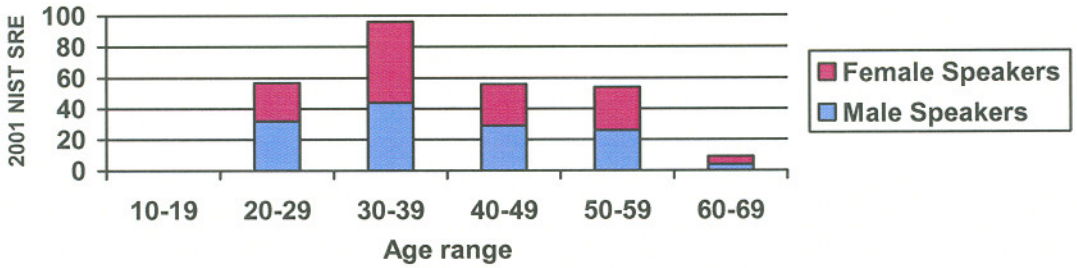

Figure B.1: Number of callers per age range from the 8-conversation training condition of the 2001 NIST Speaker Recognition Evaluation extended-data speaker detection task.

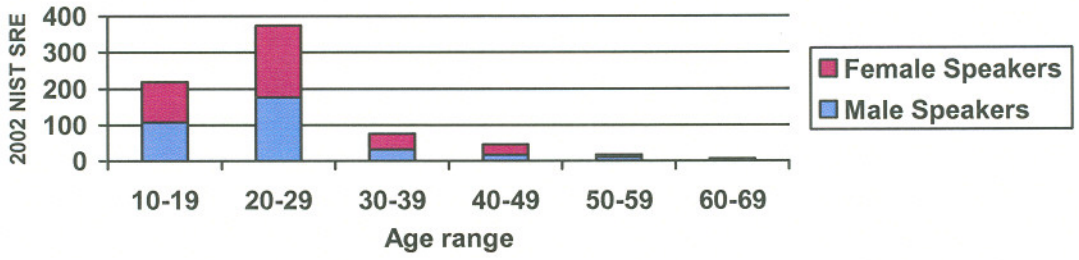

Figure B.2: Number of callers per age range from the 8-conversation training condition of the 2002 NIST Speaker Recognition Evaluation extended-data speaker detection task.

Figure B.3 shows the boxplot of the age difference between true speaker and impostor for same-gender trials on the 2001 and 2002 NIST SRE. Male-male and female-female trials show similar characteristics within each evaluation. However, the boxplot shows that the age difference of the 2001 NIST SRE trials is more variable than those of the 2002 NIST SRE trials.

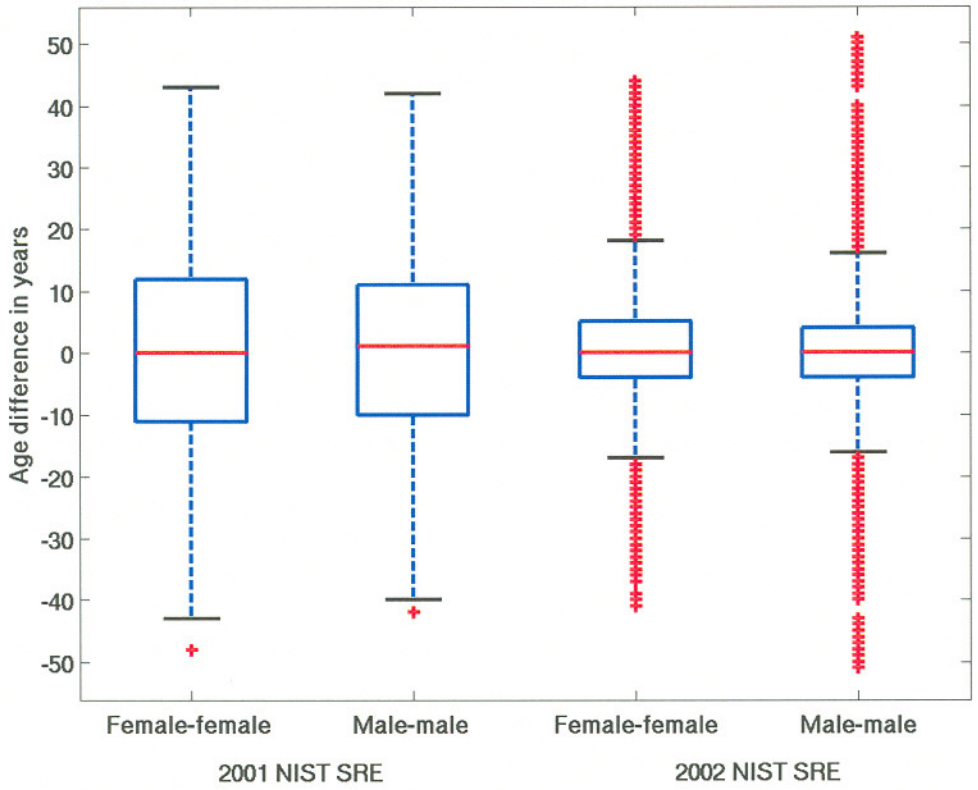

Figure B.3: Boxplot of the Age difference between the true speaker and impostor for same-gender trials. Outliers are represented by a '+' symbol.

Table B.1 shows some statistics about the absolute age difference between true speaker and impostor for same-gender trials with the speaker models from 8-conversation training condition.

Table B.1: Absolute age difference statistics for same-gender trials for 8-conversation training condition.

	2001 NIST SRE		2002 NIST SRE	
	Male-male	**Female-female**	**Male-male**	**Female-female**
Number of trials	4396	4545	6960	7820
Median	11	11	4	5
Mean	12.4	13.0	8.0	8.7
Standard deviation	9.1	9.3	9.6	9.2

Appendix C

Phone-based Joint-state Class Modeling using a Conventional Phone Recognition System

In Section 5.2, TRAPS, a long-term speech features based phone recognizer, was used to estimate the sequence of phones from a given conversation side. In this appendix, we show the speaker detection results for a system that uses a conventional phone recognizer based on short-term features.

The sequence of phones is computed using a gender-dependent phone recognition system. The phone recognizer uses the front-end of Zissman's Parallel Phone Recognition with Language Modeling (PPRLM) system [4]. Such system employs a hidden Markov model (HMM) followed by a Viterbi search to recognize the phonemes. The HMM is a network of context-independent phones, which are modeled using 3 emitting states. The probability densities in each state are modeled using a 6-component GMM. The observations are 12 Mel-cepstral coefficients (energy coefficient is discarded) and 13 delta-parameters of the cepstral coefficients. The recognizer was trained on OGI Multi-Language Telephone Speech Corpus [137]. Since our goal is to obtain a sequence of phones, the language modeling is not performed. This phone recognizer outputs a time-aligned sequence of 44 possible phones plus a silence label.

The same configuration as described in Section 5.2.2 is used to evaluate the speaker detection performance. The DET curves for 8-conversation training for speaker detection using the phone-based prosody modeling are shown in Figure C.1. Note that

these results do not show a significant difference when compared to the results for the experiments using TRAPS.

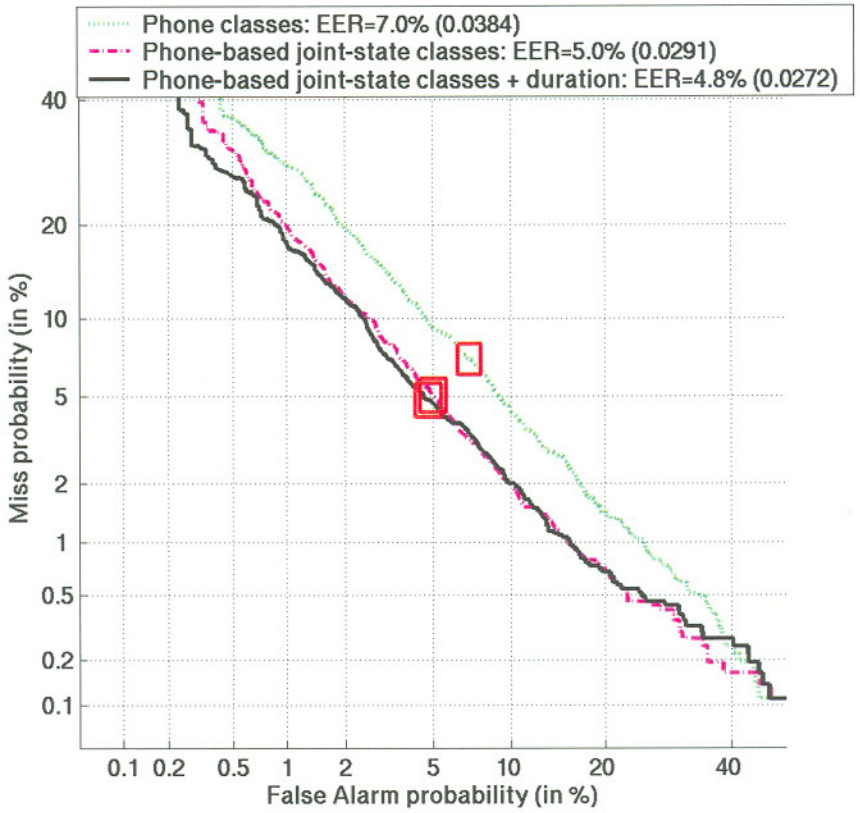

Figure C.1: DET curve for the phone-based (PPRLM) joint-state class modeling.

Appendix D

Confusion Matrices for the Language Detection Systems

This appendix shows the confusion matrices of the language detection systems described in Chapter 6 (Tables D.1 to D.3) and their fusions (Tables D.4 and D.5) for the 30-second test condition. There are 80 test segments for each language.

Table D.1: Confusion matrix of the joint-state class based system on 30-second test condition.

		Input Language											
		Arabic	**English**	**Farsi**	**French**	**German**	**Hindi**	**Japanese**	**Korean**	**Mandarin**	**Spanish**	**Tamil**	**Vietnamese**
Recognized Language	**Arabic**	45	11	5	7	7	4	3	3	1	4		1
	English	9	20	10	4	10	3	1	2	4	2	1	
	Farsi	3	9	20	9	4	6	5	4	4	5		
	French	3			32	2	1	1	1		3		
	German	5	15	5	6	44	12	3	4	4	5	1	
	Hindi	1	2	7	8	4	19	5	7	4	10	7	
	Japanese	2	11	1	3	3	8	28	17	13	13	2	
	Korean	5	8	9	2	3	4	16	31	5	7	4	1
	Mandarin	1	1	2	2	1	2	7	2	32	8	4	3
	Spanish		2	5	5	1	12	6	5	3	16	4	3
	Tamil	3		12	1		6	5	3	7	7	53	4
	Vietnamese	3	1	4	1	1	3		1	3		4	68

Table D.2: Confusion matrix of the single phone-recognizer based system on 30-second test condition.

		Input Language											
		Arabic	English	Farsi	French	German	Hindi	Japanese	Korean	Mandarin	Spanish	Tamil	Vietnamese
Recognized Language	Arabic	36	1	2	2	3							
	English	1	61	4	7	7	1			3			
	Farsi	4		36	4	1			1	2		1	
	French	1	3		32	1	1			1	1		
	German	3	1	2	2	54		1		2	2		
	Hindi	4		3	3	4	46	3	2	3	6	1	
	Japanese	9		4	1	1	4	49	17	5	3		
	Korean	4	6	6	9	4	5	12	55	5	2	4	
	Mandarin		3	9	5	2	2	2	1	47	2		1
	Spanish	6	1	3	13	1	5	4	1	3	58		1
	Tamil	10		3		1	15	7	3	1	4	67	1
	Vietnamese	2	4	8	2	1	1	2		8	2	7	77

Table D.3: Confusion matrix of the parallel phone-recognizers based system on 30-second test condition.

		Input Language											
		Arabic	English	Farsi	French	German	Hindi	Japanese	Korean	Mandarin	Spanish	Tamil	Vietnamese
Recognized Language	Arabic	71		1	2						3	2	
	English	2	68	8	6	6	6		1	2	2		9
	Farsi			56			1				1		
	French	2			56						2		
	German				1	68		1					
	Hindi			2		1	46				2		
	Japanese		1	3	1	2	7	72	8	1	1	5	
	Korean		1				2	1	65				
	Mandarin	2	9	8	7	3	4	4	4	77		1	4
	Spanish	1		2	7		10	2			69		
	Tamil						2		2			72	
	Vietnamese	2	1				2						67

Table D.4: Confusion matrix of the fusion between single phone-recognizer based system and joint-state class based system on 30-second test condition.

		Input Language											
		Arabic	English	Farsi	French	German	Hindi	Japanese	Korean	Mandarin	Spanish	Tamil	Vietnamese
Recognized Language	Arabic	53	1				2	1			2		
	English	2	71	2	6	3	3	3		2	9		
	Farsi	6		58	6	1	3	3	7	4	3		
	French				48		1	3	3	1	5		
	German	6	1	4	7	74	5	1	2	4	4		
	Hindi	2	1	4	1		44	6	2		6	4	
	Japanese	2	1		2		3	35	6		3		
	Korean	1		1	1		2	13	49	1	1		1
	Mandarin	2	3				1	2	3	62	1		
	Spanish	1	1	6	8	1	4	8	3	3	44	5	2
	Tamil	4		3			12	5	5	2	2	70	
	Vietnamese	1	1	2	1	1				1		1	77

Table D.5: Confusion matrix of the fusion between parallel phone-recognizers based system and joint-state class based system on 30-second test condition.

		Input Language											
		Arabic	English	Farsi	French	German	Hindi	Japanese	Korean	Mandarin	Spanish	Tamil	Vietnamese
Recognized Language	Arabic	72		2	2	1	1				1	3	
	English	2	70	8	7	7	7		1	2	4		6
	Farsi			57	1		1				1		
	French	1			59		1				2		
	German				1	66							
	Hindi			3			48	1			2		
	Japanese		1	2	1	2	5	76	9	1	3	3	
	Korean	1	1				2	2	67				
	Mandarin	3	8	4	3	4	2		3	76		1	4
	Spanish	1		3	6		9	1			67		
	Tamil						3			1		73	
	Vietnamese			1			1						70

Bibliography

[1] D. O'Shaughnessy, Speaker Recognition, *IEEE ASSP Magazine*, 3, (1986), 4-17.

[2] S. Furui, Recent Advances in Speaker Recognition, *Pattern Recognition Letters*, 18, (1997), 859-872.

[3] J. P. Campbell, Speaker Recognition: A Tutorial, *Proceedings of the IEEE*, 85, 9, (1997), 1437-1462.

[4] M. A. Zissman, Comparison of Four Approaches to Automatic Language Identification of Telephone Speech, *IEEE Transactions on Speech and Audio Processing*, 4, (1996), 31-44.

[5] Y. K. Muthusamy, E. Barnard, and R. A. Cole, Reviewing Automatic Language Identification, *IEEE Signal Processing*, 11, (1994), 33-41.

[6] A. Schmidt-Nielsen and T. H. Crystal, Human vs. Machine Speaker Identification with Telephone Speech, in *Proceedings of ICSLP '1998* (Sydney, Australia, 1998), pp. 221-224.

[7] Y. K. Muthusamy, N. Jain, and R. A. Cole, Perceptual Benchmarks for Automatic Language Recognition, in *Proceedings of ICASSP '1994* (Adelaide, Australia, 1994), pp. 333-336.

[8] D. van Lancker, J. Kreiman, and T. D. Wickens, Familiar Voice Recognition: Patterns and Parameters, Part II: Recognition of Rate-altered Voices, *Journal of Phonetics*, 13, (1985), 39-52.

[9] D. van Lancker, J. Kreiman, and T. D. Wickens, Familiar Voice Recognition: Patterns and Parameters, Part I: Recognition of Backward Voices, *Journal of Phonetics*, 13, (1985), 19-38.

[10] A. E. Thymé-Gobbel and S. E. Hutchins, On Using Prosodic Cues in Automatic Language Identification, in *Proceedings of ICSLP '1996* (Philadelphia, 1996), pp. 1768-1771.

[11] A. Cutler, The Comparative Perspective on Spoken-language Processing, *Speech Communication*, 21, 1-2, (1997), 3-15.

[12] F. Ramus and J. Mehler, Language Identification with Suprasegmental Cues: A Study Based on Speech Resynthesis, *Journal of the Acoustical Society of America*, 105, 1, (1999), 512-521.

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

[14] M. A. Zissman and E. Singer, Automatic Language Identification of Telephone Speech Messages using Phoneme Recognition and N-gram Modeling, in *Proceedings of ICASSP '1994* (Adelaide, Australia, 1994), pp. 305-308.

[15] Y. Yan, E. Barnard, and R. A. Cole, Development of an Approach to Automatic Language Identification based on Phone Recognition, *Computer Speech and Language*, 10, 1, (1996), 37-54.

[16] T. Schultz, Q. Jin, K. Laskowski, A. Tribble, and A. Waibel, Speaker, Accent, and Language Identification Using Multilingual Phone Strings, in *Proceedings of Second International Conference on Human Language Technology Conference '2002*, 2002), pp. NA.

[17] M. Adda-Decker, F. Antoine, P. B. d. Mareüil, I. Vasilescu, J. Vaissière, E. Geoffrois, and J.-S. Liénard, Phonetic Knowledge, Phonotactics and Perceptual Validation for Automatic Language Identification, in *Proceedings of ICPhS '2003* (Barcelona, Spain, 2003), pp. 747-750.

[18] D. A. Reynolds, Speaker Identification and Verification using Gaussian Mixture Models, *Speech Communication*, 17, (1995), 91-108.

[19] F. Weber, B. Peskin, M. Newman, A. Corrada-Emmanuel, and L. Gillick, Speaker Recognition on Single- and Multispeaker Data, *Digital Signal Processing*, 10, (2000), 75-92.

[20] D. A. Reynolds, R. B. Dunn, and J. J. McLaughlin, The Lincoln Speaker Recognition System: NIST Eval2000, in *Proceedings of ICSLP '2000* (Beijing, China, 2000), pp. NA.

[21] S. Kajarekar, *Analysis of Variability in Speech with Applications to Speech and Speaker Recognition*, (Ph.D. thesis, OGI School of Science and Engineering at OHSU, 2002), pp. 116.

[22] B. A. Hanson, T. H. Applebaum, and J. C. Junqua, Spectral Dynamics for Speech Recognition under Adverse Conditions, in *Automatic Speech and Speaker Recognition*, F. K. Soong, C. H. Lee, and K. K. Paliwal, Eds.: Kluwer Academic Publishers, 1966.

[23] B. H. Juang, Speech Recognition in Adverse Environments, *Computer Speech and Language*, 5, (1991), 275-294.

[24] C. Avendano and H. Hermansky, On the Properties of Temporal Processing for Speech in Adverse Environments, in *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics '1997* (New Paltz, New York, 1997), pp. NA.

[25] F. Cummins, F. Gers, and J. Schmidhuber, Language Identification from Prosody without Explicit Features, in *Proceedings of EUROSPEECH '1999* (Budapest, Hungary, 1999), pp. 371-374.

[26] J. Farinas and F. Pellegrino, Automatic Rhythm Modeling for Language Identification, in *Proceedings of EUROSPEECH '2001* (Aalborg, Scandinavia, 2001), pp. 2539-2542.

[27] G. Doddington, Speaker Recognition based on Idiolectal Differences between Speakers, in *Proceedings of EUROSPEECH '2001* (Aalborg, Denmark, 2001), pp. 2521-2524.

[28] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg, Using Prosodic and Lexical Information for Speaker Identification, in *Proceedings of ICASSP '2002* (Orlando, FL, 2002), pp. 141-144.

[29] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernandez-Cordero, Gender-dependent Phonetic Refraction for Speaker Recognition, in *Proceedings of ICASSP '2002* (Orlando, FL, 2002), pp. 149-152.

[30] B. S. Atal, Automatic Speaker Recognition Based on Pitch Contours, *Journal of the Acoustical Society of America*, 52, 6, (1972), 1687-1697.

[31] W. A. van Dommelen, The Contribution of Speech Rhythm and Pitch to Speaker Recognition, *Language and Speech*, 30, 4, (1987), 325-338.

[32] G. Fant, A. Kruckenberg, and L. Nord, Prosodic and Segmental Speaker Variations, *Speech Communication*, 10, (1991), 521-531.

[33] S. Werner and E. Keller, Prosodic Aspects of Speech, in *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Fututre Chalenges*, E. Keller, Ed. Chichester, New York: John Wiley, 1994, pp. 23-40.

[34] K. Mori, N. Toba, T. Harada, T. Arai, M. Komatsu, M. Aoyagi, and Y. Murahara, Human Language Identification with Reduced Spectral Information, in *Proceedings of EUROSPEECH '1999* (Budapest, Hungary, 1999), pp. 391-394.

[35] J. Vaissière, Language-independent Prosodic Features, in *Prosody: Models and Measurements*, A. Cutler and D. R. Ladd, Eds. Berlin: Springer-Verlag, 1983, pp. 53-56.

[36] D. Hirst and A. Di Cristo, A Survey of Intonation Systems, in *Intonation Systems: A Survey of Twenty Languages*, D. Hirst and A. Di Cristo, Eds. Cambridge, UK: Cambridge University Press, 1998, pp. 1-44.

[37] P. Ladefoged, Linguistic Aspects of Respiratory Phenomena, *Annals of New York Academy of Sciences*, 155, (1968), 141-151.

[38] R. Collier, Physiological Correlates of Intonation Patterns, *Journal of the Acoustical Society of America*, 58, 1, (1975), 249-255.

[39] I. Lehiste, *Suprasegmentals*. Cambridge, Massachussets: MIT Press, 1970.

[40] J. E. Atkinson, Correlation Analysis of the Physiological Factors Controlling Fundamental Voice Frequency, *Journal of the Acoustical Society of America*, 63, 1, (1978), 211-222.

[41] L. Boves and H. Strik, The Fundamental Frequency-Subglottal Pressure Ratio in Speech, *Journal of the Acoustical Society of America*, 84, S1, (1988), S82.

[42] J. D. Markel, B. T. Oshika, and A. H. Gray Jr., Long-term Feature Averaging for Speaker Recognition, *IEEE Transactions Acoustic, Speech, and Signal Processing*, 25, 4, (1977), 330-337.

[43] M. Savic, E. Acosta, and S. K. Gupta, An Automatic Language Identification System, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 817-820.

[44] K. Fukunaga, *Introduction to Statistical Pattern Recognition*: Academic Press, 2nd ed., 1990.

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

[46] D. Reynolds, An Overview of Automatic Speaker Recognition Technology, in *Proceedings of ICASSP '2002* (Orlando, Florida, 2002), pp. 4072-4075.

[47] J. J. Wolf, Efficient Acoustic Parameters for Speaker Recognition, *Journal of the Acoustical Society of America*, 51, 6, (1972), 2044-2056.

[48] F. Nolan, *The Phonetic Bases of Speaker Recognition*. New York: Cambridge University Press, 1983.

[49] G. R. Doddington, Speaker Recognition - Identifying People by Their Voices, *Proceedings of the IEEE*, 11, (1985), 1651-1664.

[50] A. E. Rosenberg and F. K. Soong, Recent Research in Automatic Speaker Recognition, in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 701-737.

[51] S. Furui, Comparison of Speaker Recognition Methods using Statistical Features and Dynamic Features, *IEEE Transactions on Acoustic, Speech, and Signal Processing*, 29, 3, (1981), 342-350.

[52] N. Morgan, H. Hermansky, and C. Wooters, Continuous Speech Recognition Using PLP Analysis with Multilayer Perceptrons, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 49-52.

[53] S. Kajarekar and H. Hermansky, Speaker Verification Based on Broad Phonetic Categories, in *Proceedings of 2001: A Speaker Odyssey '2001* (Crete, Greece, 2001), pp. 201-206.

[54] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, N.J.: Prentice-Hall, 1978.

[55] B. S. Atal, Effectiveness of Linear Prediction Characteristics of Speech Wave for Automatic Speaker Identification and Verification, *The Journal of the Acoustical Society of America*, 55, (1974), 1304-1312.

[56] H. Hermansky, C. Avendano, S. van Vuuren, and S. Tibrewala, Recent Advances in Addressing Sources of Non-Linguistic Information, in *Proceedings of Tutorial and Research Workshop on Robust Speech Recognition for Unknown Communication Channels '1997* (France, 1997), pp. 103-106.

[57] S. Furui, Cepstral Analysis Technique for Automatic Speaker Verification, *IEEE Transactions on Acoustic, Speech and Signal Processing*, 29, 2, (1981), 254-272.

[58] H. Gish and M. Schmidt, Text-independent Speaker Identification, *IEEE Signal Processing Magazine*, 11, 4, (1994), 18-32.

[59] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, Speaker Verification Using Adapted Mixture Models, *Digital Signal Processing*, 10, (2000), 19-41.

[60] A. E. Rosenberg and M. R. Sambur, New Techniques for Automatic Speaker Recognition, *IEEE Transactions on Acoustic, Speech and Signal Processing*, 23, 2, (1975), 169-176.

[61] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B.-H. Juang, A Vector Quantization Approach to Speaker Recognition, in *Proceedings of ICASSP '1985* (Tampa, FL, 1985), pp. 387-390.

[62] K.-P. Li and G. Hughes, Talker Differences as They Appear in Correlation Matrices of Continuous Speech Spectra, *Journal of the Acoustical Society of America*, 55, 4, (1974), 833-837.

[63] B. S. Atal, Automatic Recognition of Speakers from Their Voices, *Proceedings of the IEEE*, 64, 4, (1976), 460-475.

[64] M. Savic and S. K. Gupta, Variable Parameter Speaker Verification System Based on Hidden Markov Modeling, in *Proceedings of ICASSP '1990* (Albuquerque, NM, 1990), pp. 281-284.

[65] D. E. Sturim, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, Speaker Verification using Text-Constrained Gaussian Mixture Models, in *Proceedings of ICASSP '2002* (Orlando, Florida, 2002), pp. 677-680.

[66] S. Sarma and V. Zue, A Segment-based Speaker Verification System using SUMMIT, in *Proceedings of EUROSPEECH '1997* (Rhodes, Greece, 1997), pp. 843-846.

[67] R. Faltlhauser and G. Ruske, Improving Speaker Recognition Performance using Phonetically Structured Gaussian Mixture Models, in *Proceedings of EUROSPEECH '2001* (Aalborg,m Denmark, 2001), pp. 751-754.

[68] A. Park and T. J. Hazen, ASR Dependent Techniques for Speaker Identification, in *Proceedings of ICSLP '2002* (Denver, CO, 2002), pp. 1337-1340.

[69] U. V. Chaudhari, J. Navratil, S. H. Maes, and R. Gopinath, Transformation Enhanced Multi-grained Modeling for Text-independent Speaker Recognition, in *Proceedings of ICSLP '2000* (Beijing, China, 2000), pp. 298-301.

[70] G. Doddington, A Method for Speaker Verification, *Journal of the Acoustical Society of America*, 49, 1, (1971), 139(A).

[71] R. C. Lummis, Speaker Verification by Computer Using Speech Intensity for Temporal Registration, *IEEE Transactions on Audio and Electroacoustics*, AU-21, 2, (1973), 80-89.

[72] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, A Comparative Performance Study of Several Pitch Detection Algorithms, *IEEE Transactions Acoustic, Speech, and Signal Processing*, 24, 5, (1976), 399-418.

[73] H. Levitt and L. R. Rabiner, Analysis of Fundamental Frequency Contours in Speech, *Journal of the Acoustical Society of America*, 49, 2, (1971), 569-582.

[74] K. Sonmez, E. Shriberg, L. Heck, and M. Weintraub, Modeling Dynamic Prosodic Variation for Speaker Verification, in *Proceedings of ICSLP '1998* (Sydney, Australia, 1998), pp. 3189-3192.

[75] K. Bartkova, D. Le Gac, D. Charlet, and D. Jouvet, Prosodic Parameter for Speaker Identification, in *Proceedings of ICSLP '2002* (Denver, Colorado, 2002), pp. 1197-1200.

[76] A. Adami, R. Mihaescu, D. A. Reynolds, and J. Godfrey, Modeling Prosodic Dynamics for Speaker Recognition, in *Proceedings of ICASSP '2003* (Hong Kong, 2003), pp. 788-791.

[77] W. Andrews, M. A. Kohler, and J. P. Campbell, Phonetic Speaker Recognition, in *Proceedings of EUROSPEECH '2001* (Aalborg, Denmark, 2001), pp. 2517-2520.

[78] J. Navratil, Q. Jin, W. Andrews, and J. P. Campbell, Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models, in *Proceedings of ICASSP '2003* (Hong Kong, 2003), pp. 796-799.

[79] Q. Jin, J. Navratil, D. Reynolds, J. Campbell, W. Andrews, and J. Abramson, Combining Cross-Stream and Time Dimensions in Phonetic Speaker Recognition, in *Proceedings of ICASSP '2003* (Hong Kong, 2003), pp. 800-803.

[80] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sonmez, E. Shriberg, A. Stolcke, H. Bratt, and R. R. Gade, Speaker Recognition Using Prosodic and Lexical Features, in *Proceedings of ASRU '2003* (St. Thomas, U.S. Virgin Islands, 2003), pp. 19-24.

[81] L. Ferrer, H. Bratt, V. R. R. Gadde, S. S. Kajarekar, E. Shriberg, K. Sönmez, A. Stolcke, and A. Venkataraman, Modeling Duration Patterns for Speaker Recognition, in *Proceedings of EUROSPEECH '2003* (Geneva, Switzerland, 2003), pp. 2017-2020.

[82] B. Peskin, J. Navratil, J. Abramson, D. Jones, D. Klusacek, D. Reynolds, and B. Xiang, Using Prosodic and Conversational Features for High-performance Speaker Recognition: Report from JHU WS'02, in *Proceedings of ICASSP '2003* (Hong Kong, 2003), pp. 784-787.

[83] K. R. Farrell, R. J. Mammone, and K. T. Assaleh, Speaker Recognition using Neural Networks and Conventional Classifiers, *IEEE Transactions on Speech and Audio Processing*, 2, 1, (1994), 194-205.

[84] A. L. Higgings, L. Bahler, and J. E. Porter, Voice Identification using Nearest-neighbor Distance Measure, in *Proceedings of ICASSP '1993* (Minneapolis, MN, 1993), pp. 375-378.

[85] S. P. Kishore and B. Yegnanarayana, Speaker Verification: Minimizing the Channel Effects using Autoassociative Neural Network Models, in *Proceedings of ICASSP '2000* (Istanbul, Turkey, 2000), pp. 1101-1104.

[86] S. A. Zahorian, Reusable Binary-paired Partitioned Neural Networks for Text-independent Speaker Identification, in *Proceedings of ICASSP '1999* (Poenix, AZ, 1999), pp. 849-852.

[87] A. G. Adami and D. A. C. Barone, A Speaker Identification System Using a Model of Artificial Neural Networks for an Elevator Application, *Information Sciences*, 138, 1-4, (2001), 1-5.

[88] Y. Bennani and P. Gallinari, On the use of TDNN-extracted Features Information in Talker Identification, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 385-388.

[89] J. Oglesby and J. S. Mason, Radial Basis Function Networks for Speaker Recognition, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 393-396.

[90] T. Ganchev, D. K. Tasoulis, M. N. Vrahatis, and N. Fakotakis, Locally Recurrent Probabilistic Neural Network for Text-Independent Speaker Verification, in *Proceedings of EUROSPEECH '2003* (Geneva, Switzerland, 2003), pp. 1673-1676.

[91] A. B. Poritz, Linear Predictive Hidden Markov Models and the Speech Signal, in *Proceedings of ICASSP '1982* (Paris, France, 1982), pp. 1291-1294.

[92] A. E. Rosenberg, C. Lee, and S. Gokcen, Connected Word Talker Verification using Sub-word Unit Hidden Markov Models, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 381-384.

[93] D. Reynolds and R. C. Rose, Robust Text-independent Speaker Identification using Gaussian Mixture Speaker Models, *IEEE Transactions on Speech and Audio Processing*, 3, 1, (1995), 72-83.

[94] G. R. Doddington, M. A. Przybocki, A. F. Martin, and D. A. Reynolds, The NIST Speaker Recognition Evaluation - Overview, Methodology, Systems, Results, Perspective, *Speech Communication*, 31, 2-3, (2000), 225-254.

[95] A. E. Rosenberg, J. DeLong, C. H. Lee, B. H. Juang, and F. K. Soong, The Use of Cohort Normalized Scores for Speaker Verification, in *Proceedings of ICSLP '1992* (Banff, Canada, 1992), pp. 599-602.

[96] S. van Vuuren, *Speaker Verification in a Time-Feature Space*, (Ph. D. thesis, Oregon Graduate Institute of Science and Technology, 1999), pp. 161.

[97] J. Navratil, Spoken Language Recognition - A Step Toward Multilinguality in Speech Processing, *IEEE Transactions on Speech and Audio Processing*, 9, 6, (2001), 678-685.

[98] M. Sugiyama, Automatic Language Recognition using Acoustic Features, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 813 - 816.

[99] D. Cimarusti and R. B. Ives, Development of an Automatic Identification System of Spoken Languages: Phase I, in *Proceedings of ICASSP '1982* (Paris, France, 1982), pp. 1661-1663.

[100] J. T. Foil, Language Identification Using Noisy Speech, in *Proceedings of ICASSP '1986* (Tokyo, Japan, 1986), pp. 861-864.

[101] F. J. Goodman, A. F. Martin, and R. E. Wohlford, Improved Automatic Language Identification in Noisy Speech, in *Proceedings of ICASSP '1989* (Glasgow, Scotland, 1989), pp. 528 - 531.

[102] R. B. Ives, A Minimal Rule AI Expert System for Real-Time Classification of Natural Spoken Languages, in *Proceedings of 2nd Annual Artificial Intelligence and Advanced Computer Technology Conference '1986* (Long Beach, CA, 1986), pp. 337-340.

[103] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr., Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features, in *Proceedings of ICSLP '2002* (Denver, CO, 2002), pp. 89-92.

[104] Q. Dan and W. Bingxi, Discriminative Training of GMM for Language Identification, in *Proceedings of Workshop on Spontaneous Speech Processing and Recognition '2003* (Tokyo, Japan, 2003), pp. MAP8.

[105] M. A. Zissman, Automatic Language Identification using Gaussian Mixture and Hidden Markov Models, in *Proceedings of ICASSP '1993* (Minneapolis, MN, 1993), pp. 399-402.

[106] S. Nakagawa, Y. Ueda, and T. Seino, Speaker-independent, Text-independent Language Identification by HMM, in *Proceedings of ICSLP '1992* (Banff, Canada, 1992), pp. 1011-1014.

[107] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, Modeling Prosody for Language Identification on Read and Spontaneous Speech, in *Proceedings of ICASSP '2003* (Hong Kong, 2003), pp. 40-43.

[108] A. S. House and E. P. Neuburg, Toward Automatic Identification of the Language of an Utterance. I. Preliminary Methodological Considerations, *Journal of the Acoustical Society of America*, 62, 3, (1977), 708-713.

[109] K.-P. Li and T. J. Edwards, Statistical Models for Automatic Language Identification, in *Proceedings of ICASSP '1980* (Denver, CO, 1980), pp. 884-887.

[110] F. Pellegrino, Automatic Language Identification: An Alternative Approach to Phonetic Modeling, *Signal Processing*, 80, 7, (2000), 1231-1244.

[111] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller Jr., Language Identification using Gaussian Mixture Model Tokenization, in *Proceedings of ICASSP '2002* (Orlando, FL, 2002), pp. 757-760.

[112] N. Parlangeau-Vallès, F. Pellegrino, and R. André-Obrecht, Investigating Automatic Language Discrimination via Vowel System and Consonantal System Modeling, in *Proceedings of ICPhS '1999* (San Francisco, CA, 1999), pp. 141-144.

[113] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, A Segment-Based Approach to Automatic Language Identification, in *Proceedings of ICASSP '1991* (Toronto, Canada, 1991), pp. 353-356.

[114] K. Kirchhoff and S. Parandekar, Multi-Stream Statistical N-Gram Modeling with Application to Automatic Language Identification, in *Proceedings of EUROSPEECH '2001* (Aalborg, Denmark, 2001), pp. 803-806.

[115] T. J. Hazen and V. W. Zue, Recent Improvements in an Approach to Segment-Based Automatic Language Identification, in *Proceedings of ICSLP '1994* (Yokohama, Japan, 1994), pp. 1883-1886.

[116] J. Farinas, F. Pellegrino, J.-L. Rouas, and R. André-Obrecht, Merging Segmental and Rhythmic Features for Automatic Language Identification, in *Proceedings of ICASSP '2002* (Orlando, FL, 2002), pp. 753-756.

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

[118] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Prentice-Hall, NJ, 1993.

[119] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-time Processing of Speech Signals*. New York: Macmillan Publishing Company, 1993.

[120] P. Ladefoged, *A Course in Phonetics*, 3rd ed. Fort Worth: Harcourt Brace Jovanovich College Publishers, 1993.

[121] J. N. Holmes, Mechanisnms and Models of Human Speech Production, in *Speech Synthesis and Recognition*, *Aspects of Information Technology*, A. C. Downtown and R. D. Dowsing, Eds. Berkshire: Van Nostrand Reinhold, 1988.

[122] J. Godfrey and J. N. Brodsky, Acoustic Correlates of Emphasis, *Journal of the Acoustical Society of America*, 80, (1986), S49(A).

[123] I. Titze, *Principles of Voice Production*. Englewood Cliffs: Prentice-Hall, 1994.

[124] R. Silipo and S. Greenberg, Automatic Transcription of Prosodic Stress for Spontaneous English Discourse, in *Proceedings of ICPhS '1999* (San Francisco, 1999), pp. 2351-2354.

[125] R. Silipo and S. Greenberg, Prosodic Stress Revisited: Reassessing the Role of Fundamental Frequency, in *NIST Speech Transcription Workshop*. College Park, MD, 2000.

[126] F. Nolan and K. Farrar, Timing of F0 Peaks and Peak Lag, in *Proceedings of ICPhS '1999* (San Francisco, CA, 1999), pp. 961-964.

[127] R. J. Baken, *Clinical Measurement of Speech and Voice*. London: Taylor and Francis Ltd, 1987.

[128] M. Barlow and M. Wagner, Measuring the Dynamic Encoding of Speaker Identity and Dialect in Prosodic Parameters, in *Proceedings of ICSLP '1998* (Sydney, Australia, 1998), pp. 81-84.

[129] P. Taylor, The Rise/Fall/Connection Model of Intonation, *Speech Communication*, 15, 1-2, (1994), 169-186.

[130] K. E. A. Silverman, *The Structure and Processing of Fundamental Frequency Contours*, (Ph.D. thesis, University of Cambridge, 1987), pp. 116.

[131] D. R. Ladd and K. E. A. Silverman, Vowel Intrinsic Pitch in Connected Speech, *Phonetica*, 31, (1984), 31-40.

[132] K. Sonmez, L. Heck, M. Weintraub, and E. Shriberg, A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition, in *Proceedings of EUROSPEECH '1997* (Rhodes, Greece, 1997), pp. 1391-1394.

[133] *NIST 2001 Speaker Recognition Evaluation Plan*, 03/01/2001, http://www.nist.gov/speech/tests/spk/2001/doc. Date viewed: 04/01/2003.

[134] *NIST 2003 Speaker Recognition Evaluation Plan*, 02/10/2003, http://www.nist.gov/speech/tests/spk/2003/doc/2003-spkrec-evalplan-v2.2.pdf. Date viewed: 04/01/2004.

[135] M. J. Carey, E. S. Parris, H. Lloyd-Thomas, and S. Bennet, Robust Prosodic Features for Speaker Identification, in *Proceedings of ICSLP '1996*, 1996), pp. 1800-1803.

[136] T. J. Hazen and V. W. Zue, Segment-Based Automatic Language Identification, *Journal of the Acoustical Society of America*, 101, 4, (1997), 2323-2331.

[137] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, The OGI Multi-language Telephone Speech Corpus, in *Proceedings of ICSLP '1992* (Banff-Alberta, Canada, 1992), pp. 895-898.

[138] A. E. Rosenberg and S. Parthasarathy, Speaker Background Models for Connected Digit Password Speaker Verification, in *Proceedings of ICASSP '1996* (Atlanta, GA, 1996), pp. 81-84.

[139] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, The DET Curve in Assesment of Detection Task Performance, in *Proceedings of EUROSPEECH '1997* (Rhodes, Greece, 1997), pp. 1895-1898.

[140] L. Gillick and S. J. Cox, Some Statistical Issues in the Comparison of Speech Recognition Algorithms, in *Proceedings of ICASSP '1989* (Glasgow, Scotland, 1989), pp. 532-535.

[141] *NIST 2002 Speaker Recognition Evaluation Plan*, 03/04/2002, http://www.nist.gov/speech/tests/spk/2002/doc. Date viewed: 04/01/2004.

[142] J. J. Godfrey, E. C. Holliman, and J. McDaniel, SWITCHBOARD: Telephone Speech Corpus for Research and Development, in *Proceedings of ICASSP '1992* (San Francisco, CA, 1992), pp. 517-520.

[143] D. A. Reynolds, R. C. Rose, and M. J. T. Smith, PC-based TMS320C30 Implementation of the Gaussian Mixture Model Text-independent Speaker Recognition System, in *Proceedings of International Conference on Signal Processing Applications and Technology '1992* (Cambridge, MA, 1992), pp. 967-973.

[144] H. Hermansky, Mel cepstrum, deltas,double-deltas,.. - What else is new?, in *Robust Methods for Speech Recognition in Adverse Conditions*. Tampere, Finland, 1999.

[145] F. K. Soong and A. E. Rosenberg, On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition, *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36, 6, (1988), 871-879.

[146] H. Hermansky and N. Morgan, RASTA Processing of Speech, *IEEE Transactions on Speech and Audio Processing*, 2, 4, (1994), 578-589.

[147] R. B. Dunn, D. Reynolds, and T. F. Quatieri, Approaches to Speaker Detection and Tracking in Conversational Speech, *Digital Signal Processing*, 10, 1-3, (2000), 93-112.

[148] A. P. Dempster, N. M. Laird, and D. B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society*, 39, (1977), 1-38.

[149] 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.

[150] Enthropic Research Laboratory, ESPS Toolkit 5.0, Washington, DC,1993.

[151] D. L. Hall, *Mathematical Techniques in Multi-sensor Data Fusion*. Norwood, MA: Artech House Inc., 1946.

[152] B. V. Dasarathy, *Decision Fusion*: IEEE Computer Society, 1993.

[153] D. L. Hall and J. Llinas, An Introduction to Multisensor Data Fusion, *Proceedings of the IEEE*, 85, 1, (1997), 6-23.

[154] M. Pavel and H. Hermansky, Information Fusion by Human and Machine, in *Proceedings of The First European Conference on Signal Analysis and Prediction '1997* (Prague, Czech Republic, 1997), pp. 350-353.

[155] H. Hermansky, S. Tibrewala, and M. Pavel, Towards ASR on Partially Corrupted Speech, in *Proceedings of ICSLP '1996* (Philadephia,PA, 1996), pp. 462-465.

[156] J. P. Campbell, D. Reynolds, and R. B. Dunn, Fusing High- and Low-level Features for Speaker Recognition, in *Proceedings of EUROSPEECH '2003* (Geneva, Switzerland, 2003), pp. 2665-2668.

[157] C. M. Bishop, *Neural Networks for Pattern Recognition*, First ed: Clarendon Press, Oxford, 1997.

[158] G. E. Peterson and I. Lehiste, Duration of Syllable Nuclei in English, *Journal of the Acoustical Society of America*, 32, 6, (1960), 693-703.

[159] S. G. Nooteboom, The Prosody of Speech: Melody and Rhythm, in *The Handbook of Phonetic Sciences*, W. J. Hardcastle and J. Laver, Eds. Oxford: Basil Blackwell Limited, 1997, pp. 640-673.

[160] D. Reynolds, W. Andrews, J. P. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition, in *Proceedings of ICASSP '2003* (Hong Kong, 2003), pp. 784-787.

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

[162] T. M. Cover and J. A. Thomas, *Elements of Information theory*: John Wiley & Sons, Inc., 1991.

[163] B. Xiang, Text-independent Speaker Verification with Dynamic Trajectory Model, *IEEE Signal Processing Letters*, 10, 5, (2003), 141-143.

[164] D. A. Reynolds, HTIMIT and LLHDB : Speech Corpora for the Study of Handset Transducer Effects, in *Proceedings of ICASSP '1997* (Detroit, 1997), pp. 1535-1538.

[165] A. Martin and M. Przybocki, The NIST 1999 Speaker Recognition Evaluation - An Overview, *Digital Signal Processing*, 10, 1-3, (2000), 1-18.

[166] D. H. Klatt and L. C. Klatt, Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers, *Journal of the Acoustical Society of America*, 87, 2, (1990), 820-857.

[167] E. B. Holmberg, R. E. Hillman, and J. S. Perkel, Glottal Airflow and Transglottal Air Pressure Measurements for Male and Female Speakers in Soft, Normal and Loud Voice, *Journal of the Acoustical Society of America*, 84, 2, (1988), 511-529.

[168] C. Henton, Acoustic Variability in the Vowels of Female and Male Speakers, *Journal of the Acoustical Society of America*, 91, 4, (1992), 2387.

[169] S. Sharma, *Multi-Stream Approach to Robust Speech Recognition*, (Ph.D. thesis, Oregon Graduate Institute of Science and Technology, 1999), pp. 135.

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

[171] H. Hermansky and N. Malayath, Spectral Basis Functions from Discriminant Analysis, in *Proceedings of ICSLP '1998* (Sydney, Australia, 1998), pp. 1379-1382.

[172] I. Pollack, J. M. Picket, and W. H. Sumby, On the Identification of Speakers by Voice, *Journal of the Acoustical Society of America*, 26, 3, (1954), 403-406.

[173] A. J. Compton, Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally, *Journal of the Acoustical Society of America*, 35, 11, (1963), 1748-1752.

[174] S. Hayakawa and F. Itakura, Text-dependent Speaker Recognition using the Information in the Higher Frequency, in *Proceedings of ICASSP '1994* (Adelaide, Australia, 1994), pp. 137-140.

[175] S. Furui and M. Akagi, Perception of Voice Individuality and Physical Correlates, *Journal of the Acoustical Society of Japan*, J66-A, (1985), 311-318.

[176] Q. Jin, T. Schultz, and A. Waibel, Speaker Identification Using Multilingual Phone Strings, in *Proceedings of ICASSP '2002* (Orlando, FL, 2002), pp. 145-148.

[177] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauche, C. Richey, E. Shriberg, K. Sonmez, F. Weng, and J. Zheng, The SRI March 2000 Hub-5 Conversational Speech Transcription System, in *Proceedings of NIST Speech Transcription Workshop '2000* (College Park, MD, 2000), pp. NA.

[178] H. Hermansky and S. Sharma, Temporal Patterns (TRAPS) in ASR of Noisy Speech, in *Proceedings of ICASSP '1999* (Phoenix, Arizona, USA, 1999), pp. 289-292.

[179] H. Hermansky and S. Sharma, TRAPs Classifiers of Temporal Patterns, in *Proceedings of ICSLP '1998* (Sydney, Australia, 1998), pp. 1003-1006.

[180] 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 *Proceedings of ICASSP '2000* (Istanbul, Turkey, 2000), pp. 1117-1120.

[181] P. Jain and H. Hermansky, Beyond a Single Critical-band in TRAP based ASR, in *Proceedings of EUROSPEECH '2003* (Geneva, Switzerland, 2003), pp. 437-440.

[182] P. Schwarz, P. Matejka, and J. Cernocky, Recognition of Phoneme Strings using TRAP Technique, in *Proceedings of EUROSPEECH '2003* (Geneva, Switzerland, 2003), pp. 825-828.

[183] Linguistic Data Consortium, *CALLHOME American English Speech*, http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97S42. Date viewed: 04/01/2004.

[184] J. P. Campbell and D. Reynolds, Corpora for the Evaluation of Speaker Recognition Systems, in *Proceedings of ICASSP '1999* (Phoenix, Arizona, 1999), pp. 2247-2250.

[185] A. Martin and M. Przybocki, NIST 2003 Language Recognition Evaluation, in *Proceedings of EUROSPEECH '2003* (Geneva, Switzerland, 2003), pp. 1341-1344.

[186] Y. Yan, An Approach to Automatic Language Identification Based on Language-dependent Phone Recognition, in *Proceedings of ICASSP '1995* (Detroit, MI, 1995), pp. 3511-3514.

[187] G. Cinque, A Null Theory of Phrase and Compound Stress, *Linguistic Inquiry*, 24, (1993), 239-298.

[188] Institute for Signal and Information Processing, *Switchboard Resegmentation Project*, 01/29/2003, http://www.isip.msstate.edu/projects/switchboard/. Date viewed: 04/01/2004.