Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 277–291
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Modeling prosodic differences for speaker recognition

André Gustavo Adami *

Departamento de Informática, Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130 Caxias do Sul, RS 95070-560, Brazil

Received 23 January 2006; received in revised form 24 January 2007; accepted 4 February 2007

Abstract

Prosody plays an important role in discriminating 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. 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 an alternative approach that exploits 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 our approach, 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 energy contours. Using simple statistical models, we show that the statistical dependency between such discrete units can capture speaker-specific information. On the extended-data speaker detection task of the 2001 and 2003 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 more robust to communication channel effects than the state-of-the-art speaker recognition system. Since conventional speaker recognition systems do not fully incorporate different levels of information, we show that the prosodic features provide complementary information to conventional systems by fusing the prosodic systems with the state-of-the-art system. The relative performance improvement over the state-of-the-art system is about 42% and 12% for the extended-data task of the 2001 and 2003 NIST Speaker Recognition Evaluation, respectively.

Keywords: Automatic speaker recognition; Prosody; Speaker verification

1. Introduction

Prosody has long been acknowledged to discriminate speakers (Atal, 1972; Fant et al., 1991; van Dommelen, 1987). 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). 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) (Collier, 1975; Ladefoged, 1968). 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 (or F0, acoustic correlate of pitch) (Atkinson, 1978; Boves and Strik, 1988; Lehiste, 1970). The methods used to produce certain changes in pitch may be consistent for a speaker, but different across speakers (Fant et al., 1991).

Although perceptual changes in pitch and loudness can encode several prosodic phenomena, such prosodic features have not yet been fully exploited by speaker recognition systems. Most of such systems use distribution statistics (Markel et al., 1977) or the contours (Atal, 1972) of fundamental frequency and intensity (acoustic

* Tel./fax: +55 54 2182100.
E-mail address: agadami@ucs.br

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.02.005

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 of the typical independence assumption used in statistical modeling (Jelinek, 1997). In addition, the complex relationship between prosodic features is disregraded by approaches that 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 (Furui, 1997).

In this work, we propose the use of the rate of change of F0 and short-term energy contours to characterize speaker-specific information. This approach applies a modeling of both contours simultaneously to capture the speaking-style of a given speaker. Since recent work in speaker recognition (Andrews et al., 2001; Doddington, 2001; Navratil et al., 2003) has shown significant gains in accuracy and robustness through the inclusion of different levels of information, we also show that the proposed approach provides complementary information to conventional systems. In this work, experiments are conducted on the 2001 and 2003 NIST Speaker Recognition Evaluations (SRE) corpora using the NIST evaluation paradigm.

This paper is organized as follows. In Section 2, we present a brief review of previous prosody-based approaches to speaker recognition. In Section 3, we describe the proposed approach. Section 4 describes the 2001 and 2003 NIST Extended-data speaker detection task. In Section 5, we present the baseline systems. Section 6 presents the results of the proposed approach and the fusion with the baseline systems. In Section 7, we discuss the factors affecting the performance of the proposed approach.

2. Previous work on prosody-based speaker recognition

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.

The systems that use prosodic information can be divided into three classes: contour matching, contour statistics, and overall distribution statistics. In contour matching, systems use some distance between the contours of prosodic features to recognize a given speaker. Some of the representative works of this class include Doddington (1971) and Lummis (1973) where each one use the contour of pitch, intensity, and formant frequencies to perform recognition. One of the requirements of such approach is that the spoken message must be the same for training and testing. Adami et al. (2003) proposed a new approach based on the F0 contour template matching of frequent words to overcome the spoken message requirement. Such system yields a 13.3% EER on the 8-conversation training condition of extended-data one-speaker detection task of the 2001 NIST Speaker Recognition Evaluation (Martin, 2001).

In the systems based on the contour statistics, statistical measurements estimated from segments of speech are used to model speaker-specific information. Some of the works include Atal (1972) and Markel et al. (1977) who extracted the average of pitch and intensity of fixed-size segments to model speaker-specific information. The advantage of these systems over the contour matching ones is that the features can be extracted from unconstrained speech. Several approaches have been proposed using variable-length size segments for estimating the statistics. For example, Sonmez et al. (1998) and Kajarekar et al. (2003) use a F0 stylization algorithm and the speech pauses to detect a segment. The F0 stylization algorithm approximates the F0 contour using the smallest possible number of straight-line segments (which are used to extract the features), while preserving the speaker's intended pitch contour. Besides, the contour approximation reduces the noise introduced by the pitch tracker and micro-intonation effects that hide the speaker's intended pitch movements. Both systems model the distribution of the features (such as the median F0, slope, and duration) estimated from the segments to characterize speaker-specific information. Kajarekar reports a performance between 22% and 32% (approximately 20 min of training data for each speaker model) on the extended-data one-speaker detection task of the 2003 NIST SRE (Martin, 2003). Despite the use of segmentation, the estimated statistics from 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. In order to overcome such problem, Adami et al. (2003) propose a quantization of the slopes of the energy and fundamental frequency contours for each segment estimated by the F0 stylization algorithm (which bears most similarity to our approach) to build a discrete model for each speaker. Adami reports a performance of 14.1% on the extended-data one-speaker detection task of the 2001 NIST SRE. Even though the approach incorporates the temporal information of prosodic features, the goal of the F0 stylization algorithm is to estimate the speaker's intended pitch contour and not all the variations in the contour, which can be speaker dependent. In addition, the segments boundaries for both contours (F0 and energy) are defined by the F0 stylization algorithm, so the dynamics of the energy contour do not affect the segmentation.

In the overall distribution statistics, systems use the distribution statistics of prosodic features estimated over the entire training data to model speaker information. For example, Sönmez et al. (1997) proposed a probabilistic model (3-component lognormal mixture model) of pitch halving/doubling to characterize speaker-specific information. Carey et al. (1996) showed that the mean and variance of pitch and energy variance provide discriminatory information about the speaker when used individually. 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.

3. Prosody-based speaker recognition

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 (Atkinson, 1978; Boves and Strik, 1988; Lehiste, 1970; Werner and Keller, 1994). 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 (Peterson and Lehiste, 1960) the duration feature is not directly exploited in the speaker 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 of this work is that different speakers 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 characterize particular prosodic gestures are useful for describing speaker-specific information. That is, the prosodic information in an utterance is described as sequence of elementary patterns representing the joint state of the dynamics of F0 and intensity contours, and their respective duration.

3.1. Prosodic features estimation

Assuming that there are two types of F0 and intensity gestures, rising and falling, 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. Such gestures can be determined by estimating the rate of change of the feature of interest (F0 or intensity). Since unvoiced speech regions do not produce any F0 value, a fifth joint-state class is used to represent such regions. Then, the signal is segmented using the boundaries defined by the changes in the dynamics of both features. Each segment is labeled with one of the five joint-state classes, according to the state of both contours. 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.

The joint-state classes estimation is divided into five steps: (1) compute the rate of change for each contour, (2) detect the points where the dynamics of the contours change, (3) generate new segments using the detected points, (4) convert the segments into a sequence of symbols that represent the dynamics of both contours, and (5) integrate in the segment duration. Details of the estimation steps are described next.

First step – compute the rate of change for each contour

The rate of change is approximated using delta features. The delta features have been used in automatic speech recognition systems to approximate the short-term dynamics of temporal trajectories (Furui, 1981; Hermansky, 1999). They are computed using a first-order orthogonal polynomial temporal fit of each contour over a finite length window (in time) (Soong and Rosenberg, 1988). Besides its extensive use, the delta features offer several advantages, such as, (1) the amount of detail of the rate of change is easily defined by increasing or decreasing the window length, and (2) they are simple to compute.

Since the F0 contour can have discontinuities due to unvoiced speech regions and halving or doubling effects (Sönmez et al., 1997), 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. Fig. 1 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 in the (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 caused by the doubling F0 effect in the region. The lognormal tied-mixture model proposed by Sönmez et al. (1997) is used to estimate the probabilities of halving and doubling F0 at the frame level.

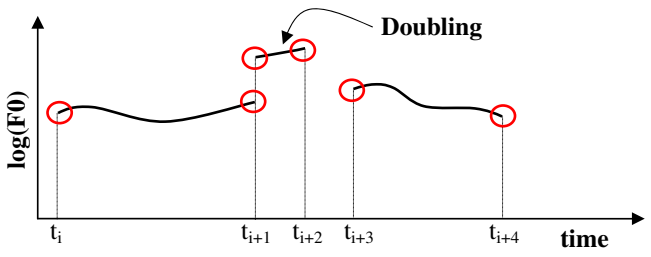

Fig. 1. Example of delta feature estimation of the F0 contour. The estimation is not performed across discontinuities (represented by circles) such as voiced/unvoiced speech regions (e.g., point at t_i, t_{i+2}, t_{i+3}, and t_{i+4}) and doubling/halving effects (e.g., point at t_{i+1}).

Second step – detect the points where the dynamics of both contours change

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 $0 < t_1^e < t_2^e < \cdots < t_k^e < N$, where

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

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 $\Delta f_0(t)$

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

where $\Delta f_{0,i}(t)$ is a continuous function valid over a subinterval of $\Delta f_0(t)$, $t_{i-1} \leqslant t \leqslant t_i$, 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 Fig. 1 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 $t_{i-1} < t_1^{f_{0,i}} < t_2^{f_{0,i}} < \cdots < t_l^{f_{0,i}} < t_i$, where

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

The filled circles in the F0 plot of Fig. 2 represent the critical points and the voiced regions boundaries estimated from the delta features (an approximation of the first derivative). In the energy plot, the filled circles represent only the critical points.

Third 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 Fig. 2 is not used in the segmentation because of unvoiced characteristic of such phone. Let $\{t_i^{\text{new}}\}_{i=0}^{m}$ be the collection of all valid points with $t_0^{\text{new}} < t_2^{\text{new}} < \cdots < t_m^{\text{new}} = 0$, $t_0^{\text{new}} = 0$ and $t_m^{\text{new}} = N$. The segment boundaries are defined as follows $(t_0^{\text{new}}, t_1^{\text{new}}), (t_1^{\text{new}}, t_2^{\text{new}}), \ldots, (t_{m-2}^{\text{new}}, t_{m-1}^{\text{new}}), (t_{m-1}^{\text{new}}, t_m^{\text{new}})$. The new segment boundaries are represented by vertical dotted bars in Fig. 2.

Fourth 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 joint-state classes that represent the dynamics of both contours within

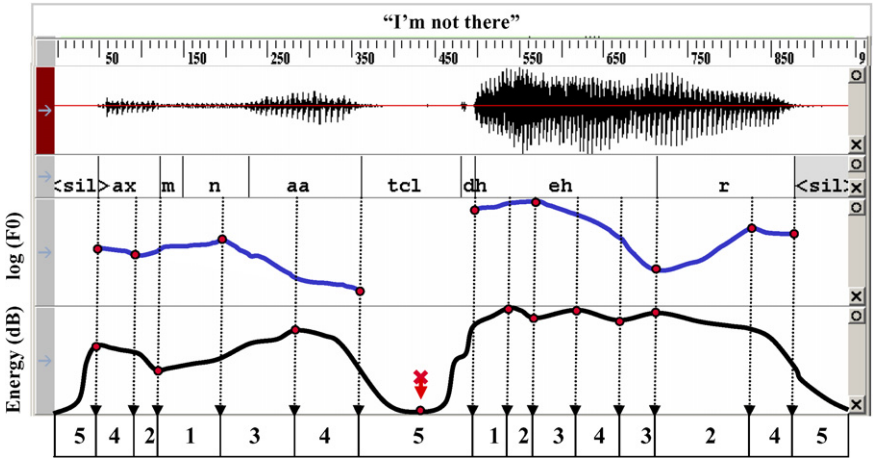

Fig. 2. Example of a joint-state class estimation for the sentence "I'm not there". The critical points are marked by filled circles. The vertical bars represent the segmentation boundaries generated from the critical points of both contours. At the bottom, the boxes represent the segmentation and their respective joint-state class.

the segment, as shown at the bottom of Fig. 2. For each segment, the classification rules are defined as follows:

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

Fifth step – integrate in the segment duration

Since the duration of each segment class varies within an utterance, as shown in Fig. 2, the duration of each segment is also incorporated into the sequence of joint-state classes by adding an extra label representing the segment duration. The duration, measured in number of analysis frames, is quantized into two levels: "Short" and "Long". 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 1–4) are smaller than the classes estimated from unvoiced speech regions (i.e., class 5). The threshold used to determine whether a segment is short or long is estimated from the median value using a held-out data set from Switchboard I. For voiced speech regions, "Short" is assigned to segment classes with duration shorter than eight 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. The addition of the duration of the example in Fig. 2 produces the following sequence: **5S 4S 2S 1S 3L 4L 5S 1S 2S 3S 4S 3S 2L 4S 5S**.

3.2. Speaker modeling and scoring

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 (Jelinek, 1997), *n*-gram modeling provides a viable and effective approach for modeling the speaker's usage of the proposed prosodic classes (Doddington, 2001; Reynolds et al., 2003).

Several approaches in speaker recognition (Andrews et al., 2002; Campbell et al., 2003; Doddington, 2001; Reynolds et al., 2003) 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 $P(s_{i-n+1},s_{i-n+2},\ldots,s_{i-1},s_i)$ over some training data is estimated by simply computing the frequencies of the sequence $s_{i-n+1},s_{i-n+2},\ldots,s_{i-1},s_i$ occurring in the training data. For example, in bigram models ($n=2$), the joint probability is estimated as follows:

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

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*" (Doddington, 2001).

The speaker detection score is computed using a conventional log-likelihood ratio test between the target-speaker model and a speaker-independent model (also known as Universal Background Model – UBM) averaged over all *n*-gram types (Andrews et al., 2002; Doddington, 2001). Let $C_{\mathrm{Test}}(\phi)$ be the number of occurrences of the *n-gram* type ϕ in a test sequence, produced by a given speaker. The probability that the target speaker i generated the *n*-gram type ϕ is represented by $P_{\mathrm{TS}_i}(\phi)$. The probability that the *n*-gram type ϕ was produced by any speaker is represented by $P_{\mathrm{UBM}}(\phi)$. The averaged log-likelihood ratio LLR_i for the ith speaker is written as

$$\mathrm{LLR}_i=\frac{\sum_{\phi\in\Omega_{\mathrm{Test}}}C_{\mathrm{Test}}(\phi)[\log(P_{\mathrm{TS}_i}(\phi))-\log(P_{\mathrm{UBM}}(\phi))]}{\sum_{\phi\in\Omega_{\mathrm{Test}}}C_{\mathrm{Test}}(\phi)}$$

where Ω_{Test} represents the set of all possible *n*-gram types in the test sequence.

4. Evaluation task

The systems are evaluated on the extended-data one-speaker detection task proposed by NIST for the speaker recognition evaluations (SRE) in 2001 (Martin, 2001) and 2003 (Martin, 2003). Such task provides large amounts of training data to support the exploration and development of higher-level and more complex characteristics for speaker recognition. The goal of the 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 data for this task comprises of conversational, telephone speech from LDCs Switchboard corpora in a cross-validation procedure to obtain a large number of trials. The target speaker models were trained using 1, 2, 4, 8, or 16 conversation sides (approximately 2.5 min of speech per side). A complete conversation side was used for testing. The extended-data one-speaker detection task in the 2001 NIST SRE uses data from the Switchboard I corpus, and the 2003 NIST SRE uses data from the Switchboard II corpus (phases 2 and 3). The task in the 2001 NIST SRE consists of 483 speakers with 4105 target-speaker models and 57470 trials for the testing phase. In the 2003 NIST SRE, the task consists of 10932 target-speaker models and 156184 trials for the testing phase.

The performance measure used to evaluate the described systems is the equal error rate (EER). It represents the system performance when the false acceptance rate (accepting an impostor) is equal to the missed detection rate (rejecting a true speaker). In this work, we only report the results for 8-conversation training condition. A detection error trade-off (DET) curve is plotted to show system results at all operating points (Martin et al., 1997). The binomial test

for differences in proportion is used to check whether the difference between the EER of the systems is statistically significant (Gillick and Cox, 1989). Unless specified, the level of significance is set to $\alpha = 0.05$.

5. Baseline systems

Two baseline systems are described in this section. The first baseline system represents the most common approach for modeling prosodic information. The second system is the most successful acoustic system for speaker recognition. The motivation of the second system is to show that prosodic information can provide complementary information to conventional systems.

5.1. 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 (Reynolds et al., 2000b).

The F0 and energy features are extracted from the voiced regions every 10 ms. First, a speech activity detector (Reynolds et al., 1992) is used to locate the time intervals with speech activity. Then, the F0 and energy features are estimated using the Robust Algorithm for Pitch Tracking proposed by Talkin (1995). 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. Since F0 cannot be estimated from unvoiced speech regions, only the feature vectors extracted for voiced speech regions are used in the speaker modeling. 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. The first two frames and the last two frames of a voiced segment are discarded to avoid discontinuities in the derivative computations. Therefore, the addition of the delta features generates a new 4-dimensional feature vector.

The features are used to train a likelihood ratio detector consisting of a speaker-independent UBM and a speaker-dependent target speaker model. The UBM is a 512-component GMM trained with gender-balanced speech from cross-validation partitions not under test. The target speaker models are derived by adapting the UBM with the speaker's data.

In the testing phase, a likelihood ratio score is obtained as the ratio between the target speaker model and the UBM likelihood scores given a test message. The EER of the prosodic baseline for 8-conversation training condition is 15.2% on the 2001 NIST SRE and 17.1% on the 2003 NIST SRE, as shown in Fig. 3. The rectangles in Fig. 3 represent the 95% confidence interval around the EER point.

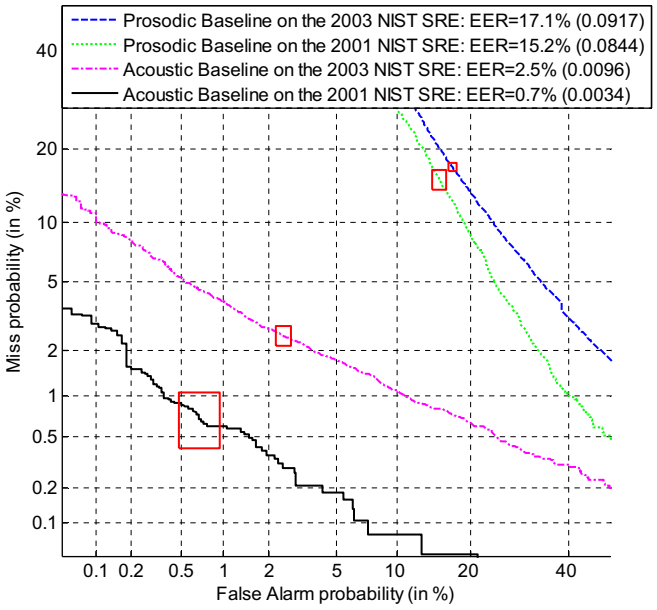

Fig. 3. DET curves of the baseline systems of the on the 2001 and 2003 NIST SRE. The curves are for the 8-conversation training condition.

5.2. Acoustic baseline system

Although this work proposes a prosody-based system, we want to show that our approach can also provide complementary information to acoustic-based systems. The most successful approach to text-independent speaker detection is based on a likelihood ratio detector that uses Gaussian mixture models for estimating the likelihoods (Reynolds et al., 2000b). Since this approach is the state-of-the-art on the NIST speaker recognition evaluation (Doddington et al., 2000), the parameters configuration of this system are the same as the ones used in the evaluations according to Reynolds et al. (2000a).

A 19-dimensional Mel-cepstral vector is extracted from the speech signal every 10 ms using a 20 ms window. Band-limiting is performed by only retaining the filterbank outputs from the frequency range 300–3138 Hz. Cepstral vectors are processed with RASTA filtering to mitigate linear channel bias effects. Delta features are then computed over a 50 ms window and appended to the cepstral vector producing a 38-dimensional feature vector. The feature vector stream is then processed through an adaptive, energy-based speech detector to discard low-energy vectors. A feature mapping technique is used on the 2003 NIST SRE to map the channel-dependent feature space into a channel-independent feature space.

The background model used for all targets is a gender-independent, handset-independent 2048 mixture GMM trained on 5.6 h of data selected from the 1996 NIST SRE. Target models are derived by Bayesian adaptation of the UBM parameters (only the mean vectors are adapted) using the designated training data.

The EER of the acoustic baseline for 8-conversation training condition is 0.7% on the 2001 NIST SRE and 2.5% on the 2003 NIST SRE, as shown in Fig. 3.

6. Results

In this section, we present the results of the proposed approach on the 2001 NIST SRE. Since the approach is optimized on the 2001 NIST SRE, the results on the 2003 NIST SRE are also presented.

6.1. Delta-window length

In most speech-based applications, the delta features are estimated over a typical time interval between 50 and 100 ms (Soong and Rosenberg, 1988). 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. Fig. 4 presents several sequences of joint-state classes and their performance using different delta window lengths for the utterance "I'm not there" depicted in Fig. 2. It 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.

We also tried different combinations of delta window lengths for each contour, but the performance is only worse than the results presented in Fig. 4. 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.

6.2. Detection results on 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. 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 s (the detection of utterances is out of the scope of this work). Fig. 5 shows the DET curves of the bigram modeling of the joint-state classes and single-contour classes. The rectangles in Fig. 5 represent the 95% confidence interval around the EER point.

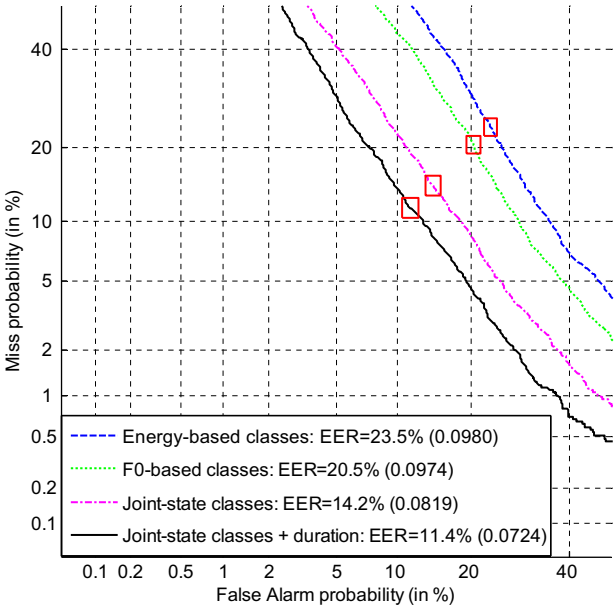

Fig. 5. DET curves for the joint-state class modeling.

The bigram modeling of the joint-state classes can capture more speaker-dependent information than the modeling of the contours alone. The EER of the systems based on the F0 and short-term energy contours are 20.5% and 23.5%, respectively. A comparison between these performances and the performance obtained from the joint-state classes modeling (14.2% EER) shows that the interaction between both contours carries speaker-specific information. 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 12% from the F0-based modeling performance. This result shows that even temporal patterns of the energy dynamics also convey speaker-dependent information.

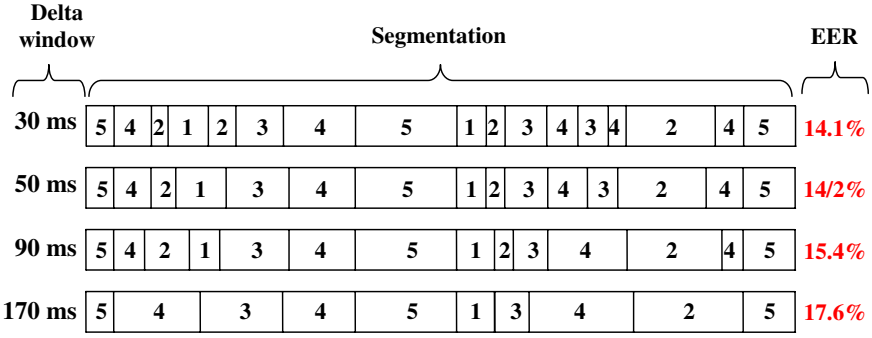

Fig. 4. Example of segmentation using joint-state classes of the utterance "I'm not there" using different delta window lengths.

The EERs of the joint-state modeling without and with the duration labels are 14.2% and 11.4%, respectively. The relative improvement over the prosodic baseline is 6% (without duration) and 25% (with duration). These results show that not only the sequence of joint-state classes is speaker dependent, but also that the duration provides speaker-dependent information.

The performance shows that the joint-state modeling based on the delta features provides a better speaker characterization than the modeling based on the F0 stylization algorithm used in previous work. The relative improvement in performance over the approach based on the F0 stylization algorithm to segment the signal in (Adami et al., 2003) is about 19% (joint-state modeling with the duration labels). The relative improvement is even greater (about 48%) when compared to the method proposed by Kajarekar et al. (2003). The improvement does not come as a surprise because of several reasons. First, the goal of a stylization algorithm is to estimate the speaker's intended pitch contour and not to capture the variations in the dynamics of both contours, which can be a speaker dependent characteristic. Second, the joint-state modeling provides a better modeling of the temporal aspects that determine certain prosodic phenomena than the feature distribution. Finally, both contours are segmented independently (in the method based on the F0 stylization algorithm, the segmentation for both contours come from the F0 contour only), which provides a better modeling of the dynamics of both contours and their relationship to produce prosodic phenomena.

The same system is used to conduct experiments on the 2003 NIST SRE corpora. The EER of the joint-state modeling with the duration labels is 14.2% on 8-conversation training condition. The performance of the joint-state classes shows a relative improvement of 17% over the prosodic baseline.

6.3. Speaker entropy

Using the data from Switchboard I corpus, we analyze the speaker information captured in the bigrams of the 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 (Cover and Thomas, 1991). Consider an information source that generates a sequence of symbols $\boldsymbol{X} = \{x_1, x_2, \ldots, x_n\}$ from a finite or countable infinite sample space S, according to some stochastic distribution law. The probability that $\boldsymbol{X}$ takes on the particular value x is written $P(x)$. The entropy of the source is defined as

$$H(X) = -\sum_{x \in X} P(x) \log(P(x)) \tag{2}$$

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 Eq. (1).

Fig. 6 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

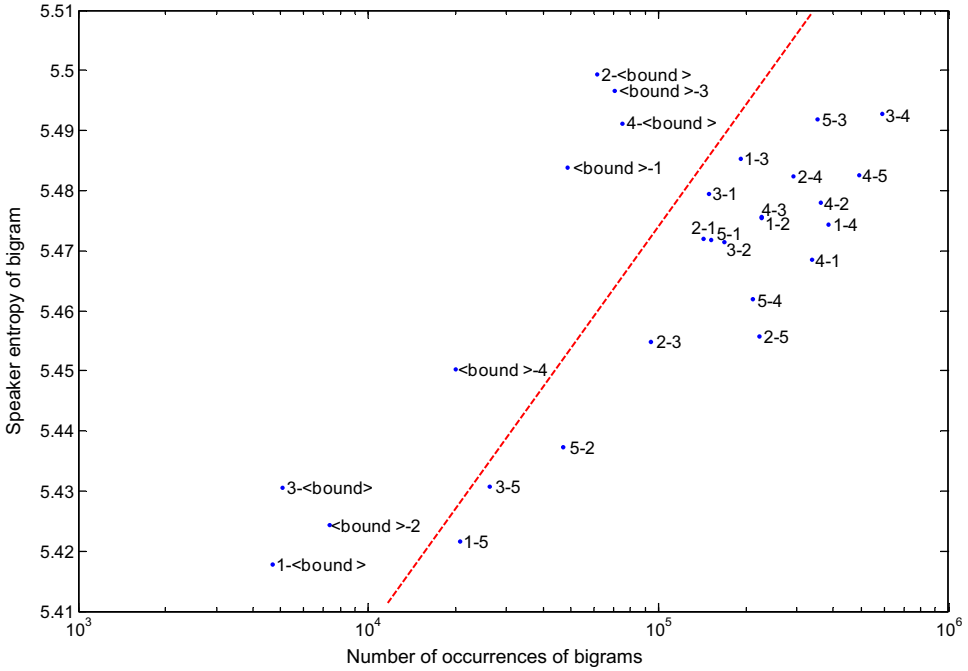

Fig. 6. Speaker entropy of joint-state class bigrams. The dashed line separates the bigrams types that include the ⟨bound⟩ symbol from the remaining bigram types.

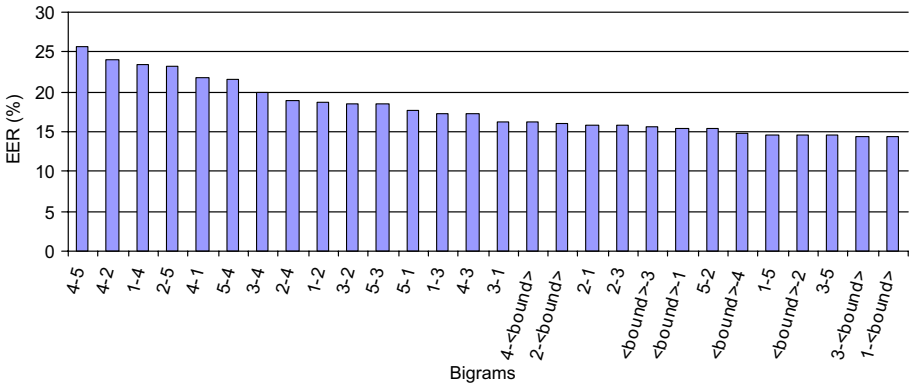

Fig. 7. Performance for the leave-one-out experiments. The x-axis represents the bigrams not used in the detection process.

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 s) as utterance boundaries, even though the pauses are not true utterance boundaries.

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 (Doddington, 2001; Xiang, 2003). 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. Fig. 7 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 Fig. 7 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", and so on).

The performances in Fig. 7 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 Fig. 7, 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.

6.4. System fusion

The scores from the systems were fused with a perceptron classifier (Campbell et al., 2003) using LNKnet.[1] The perceptron architecture chosen has no hidden layers and two output nodes. The number of input nodes is defined as the number of systems to be fused. Input values to the perceptron were normalized to zero mean and unit standard deviation using parameters derived from the training data. The perceptron weights were trained using the entire development data. The classifier corresponding to the number of training conversations is then used to fuse scores from systems. The fusion classifier is trained to minimize the detection cost (Martin, 2001, 2003). The hard decision from the perceptron was used as the hard decision for the submission. The score for the test file was set to $(s\_tgt + (1 - s\_non))/2$, where s_tgt and s_non are the perceptron scores for the target and nontarget classes, respectively.

6.4.1. Detection results

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 Fig. 8, we show the results of fusing the various systems on the 8-conversation training condition of the 2001 NIST SRE.

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 joint-state classes provides about 36% relative improvement over the joint-state classes. Indeed, these improvements show that there is speaker-specific information in the F0 and energy contours that is beyond their distribution statistics.

The performance of the fusion between the acoustic baseline, prosodic baseline, and the joint-state classes is

[1] http://www.ll.mit.edu/IST/lknet.

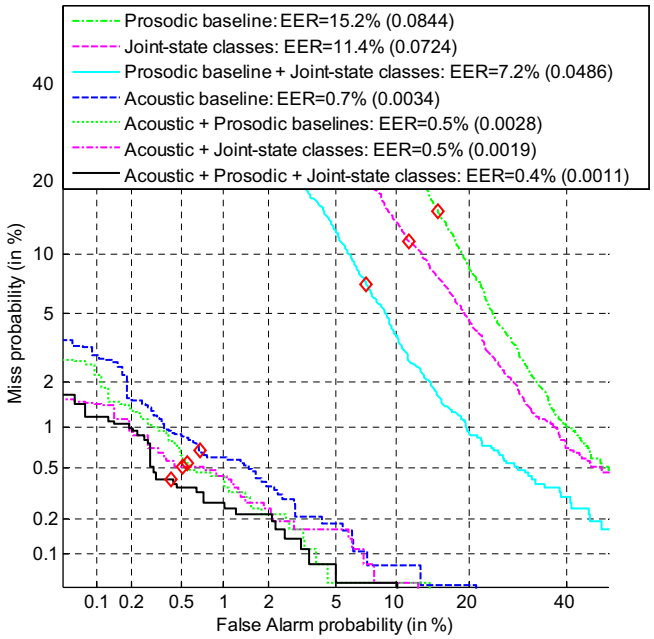

Fig. 8. DET curves for the fusion between prosody-based systems and the acoustic baseline on the 2001 NIST SRE corpora.

0.4%, an improvement of 42% 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 ($\alpha = 0.1$) over the acoustic baseline. Despite the better performance of the joint-state classes over the prosodic baseline, the fusion of either of the prosodic systems with the acoustic baseline achieves the same performance (0.5% EER). This result shows that improvement does not come from the individual performances, but from the amount of complementary information that the systems bring to the fusion. In addition, the performance of the fusion between all three systems shows that the complementary information of both prosody-based systems improves even further the performance of the acoustic baseline to 0.4%.

Fig. 9 shows the performances of the same systems on the 2003 NIST SRE corpora. 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 corpus.

7. 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).

7.1. Amount of training data

Speaker recognition systems that use long-term speech characteristics (e.g., phonemes, words, and prosodic

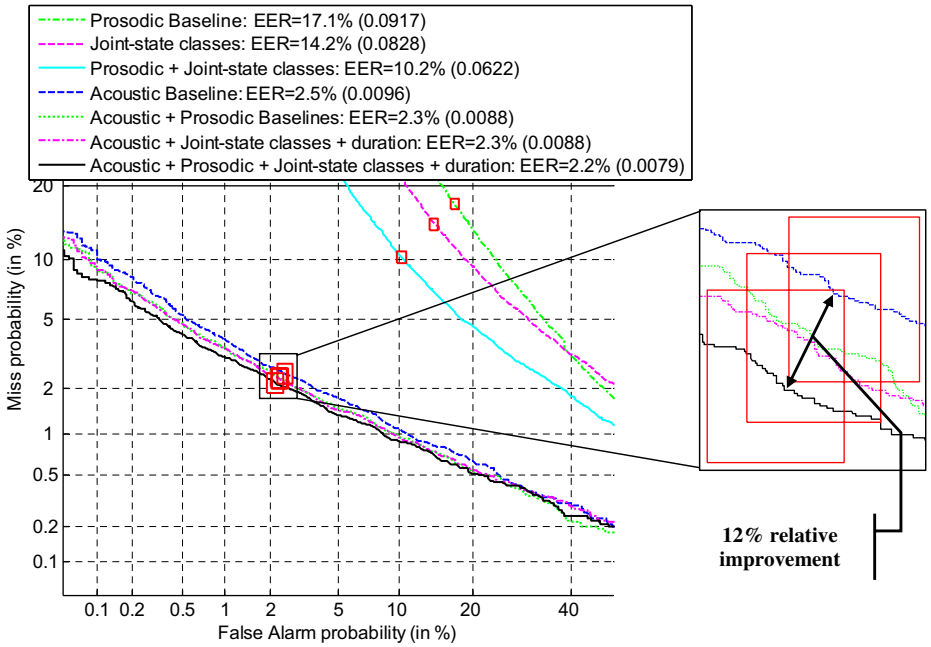

Fig. 9. DET curves of the baselines, joint-state classes, and the fusion between the systems on the 2003 NIST SRE. The small box on the right side shows a zoomed view of the performance curves that include the acoustic baseline.

features) are known for requiring large amounts of training data (Doddington, 2001; Sonmez et al., 1998; Weber et al., 2002). Fig. 10 shows the performance of the joint-state classes with duration system and the prosodic baseline on all training conditions.

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 two conversation sides (approximately 5 min) for training. Therefore, the results show that considerable amount of data is required for estimating models using joint-state classes.

7.2. Telephone handset

The variation in the type of telephone handset is a major factor affecting the performance of systems that use telephone speech (Martin and Przybocki, 2000; Reynolds, 1997). 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 2003 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. Fig. 11 shows the performance for the systems on the matched- and mismatched-handset conditions.

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. The performance

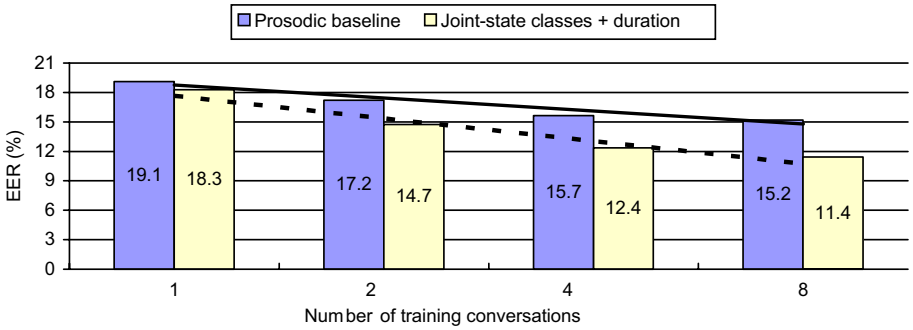

Fig. 10. Performance of prosody-based systems per number of training conversations. The dashed trend line depicts the performance improvement for the prosodic baseline and the dotted trend line depicts the improvement for the joint-state classes.

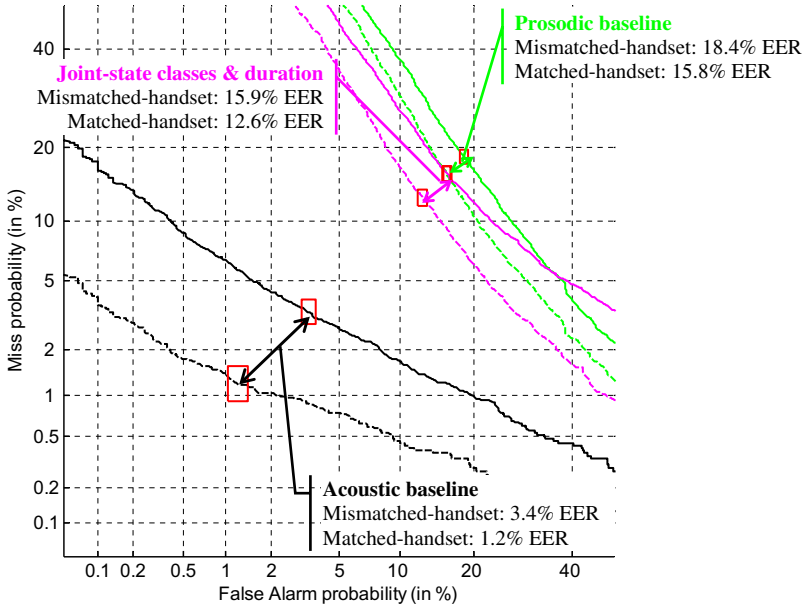

Fig. 11. Performance per handset condition of the baselines and the joint-state classes. The dashed and solid lines represent the performance of the systems under the matched- and mismatched-handset conditions, respectively.

per handset condition of the fusion between the acoustic baseline and the prosody-based systems is 3% for the mismatched-handset condition and 1.1% for the matched-handset condition. Despite the non-significant improvement over the matched condition, the performance of the mismatched-handset condition of the acoustic baseline is improved by at least 11%. This result shows that the improvement of the acoustic baseline performance is mostly due to the robustness of the prosody-based systems towards mismatched-handset condition.

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 still other factors, such as speaker gender or age, affecting the performance of the joint-state classes.

7.3. 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 (Klatt and Klatt, 1990). However, despite the high variability between genders, variability within gender can become a problem when the impostor has the same gender as the true speaker.

In this section, we analyze the systems performance using same-gender trials (the impostor and the true speaker have the same gender) on the 2001 and 2003 NIST SREs. To avoid the effects of different handsets between training and testing and lack of training data for model estimation, only the trials from matched-handset condition (as described in Section 7.2) of the 8-conversation training condition are analyzed. In the 2001 NIST SRE, there are 4305 female–female trials (1681 target trials and 2624 impostor trials) and 4261 male–male trials (1673 target trials and 2588 impostor trials). In the 2003 NIST SRE, there are 10 808 female–female trials (2988 target trials and 7820 impostor trials) and 9535 male–male trials (2575 target trials and 6960 impostor trials). Table 1 presents the performance for same-gender trials of the acoustic, baseline, and joint-state classes plus duration on the 8-conversation training condition.

The results of the acoustic baseline show that the there is no significant difference in the performance when comparing the trials within an evaluation. Previous work (Doddington et al., 2000; Martin and Przybocki, 2000) 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). However, this degradation happens for the 1-conversation training condition (on the 2001 NIST SRE, the EER for male–male is 2.3% and for female–female is 3.0%; on the 2003 NIST SRE, the EER for male–male is 2.4% and for female–female is 3.5%). It seems that the increase in the amount of training data (note that the amount of testing data for both conditions is kept the same) improves the model estimation.

Table 1
EER per gender of the baselines and the joint-state classes for 8-conversation training condition on the 2001 and 2003 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
2003	Female–female	1.2	14.3	12.9
	Male–male	1.2	18.4	14.1

The difference in performance of the prosody-based systems between female–female and male–male trials on both evaluations shows that there are others factors affecting the performance. One factor is the amount of training data. The average number of training feature vectors for the prosodic baseline and tokens for the joint-state classes approach is at least 13% higher for female speakers. Despite the difference in the amount of training and testing data, this is not the only reason for such difference in the performance. This can be seen in the results of the acoustic baseline, where there is no difference in performance across genders, but across evaluations.

Given the physiological and linguistic (e.g., vocabulary and speaking style) differences across speakers of the same gender, the age or regional differences between the impostor and true speaker can have a similar effect as the gender difference in the speaker detection performance. For example, impostor and target speaker with small age difference

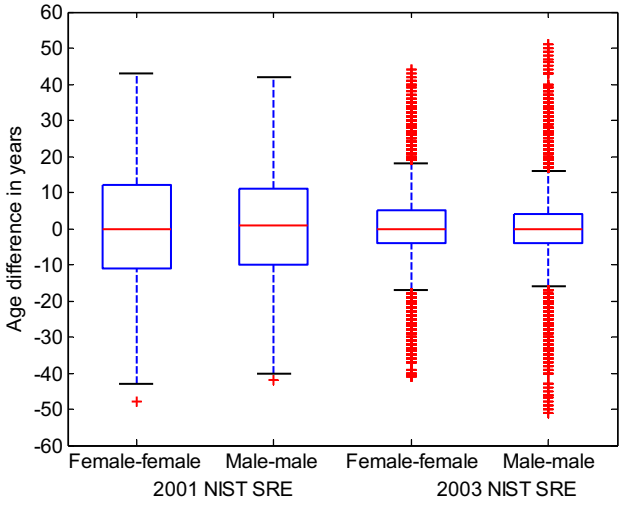

Fig. 12. Boxplots of the age difference between impostor and true speaker for same-gender impostor trials on the 2001 and 2003 NIST SREs.

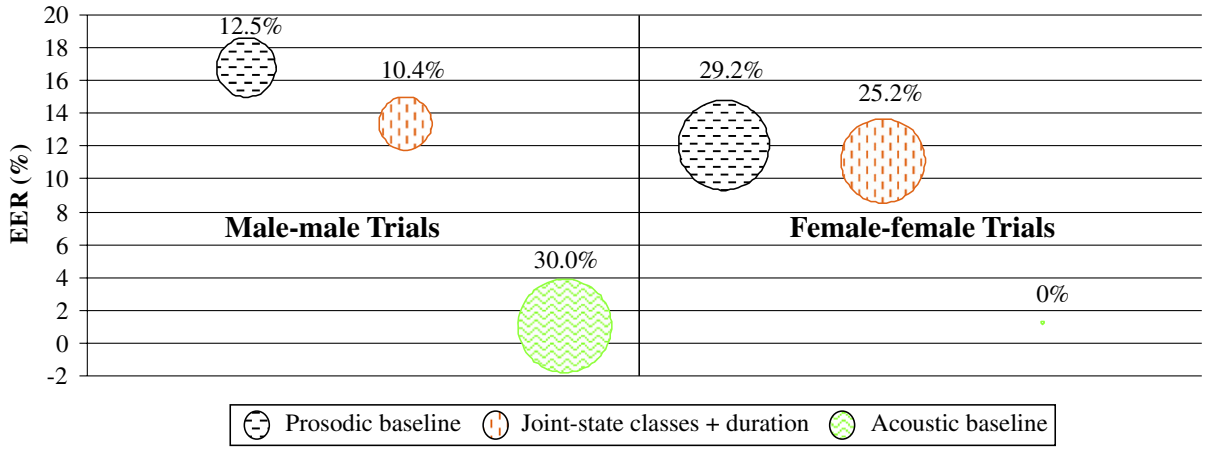

Fig. 13. Performance of the detection systems (center of each circle) on the 8-conversation training condition of the 2003 NIST SRE where the impostor has an age difference from the true speaker of 10 years or more. The radius of each circle represents the degradation in performance (the value above each circle) from the impostor trials with absolute age difference of more than 10 years to the impostor trials with age difference of 5 years or less.

can share more physiological or linguistic characteristics than speakers with large age differences. Therefore, we measured the performance of the systems based on the regional and age differences between the impostor and target speaker.

A comparison between the systems performance for speakers from different regions does not show any significant difference. The regional information about the speakers of the 2001 NIST SRE is the dialect region (i.e., South Midland, Western, North Midland, Northern, Southern, New York City, New England, and Mixed). The trials are about 50% for the speakers that share the same dialect region. The three systems do not show any significant difference between the two groups. In the 2003 NIST SRE, the regional information about the speakers is the state where the speaker was raised. However, given that most of the trials (about 90%) are from the speakers that share the same state, the performances could not provide a significant difference.

The age difference between the impostor and the true speaker affects the performance of the detection systems. Fig. 12 shows boxplots for same-gender trials on both evaluations. The boxplots show that the age difference between the impostor and true speaker on the 2001 NIST SRE is more spread than the 2003 NIST SRE. The median value of the absolute age difference for both gender conditions on the 2001 NIST SRE and 2003 NIST SRE is 11 years (mean is 12.7 years and standard deviation is 9.2 years) and 4 years (mean is 8.4 years and standard deviation is 9.4 years), respectively. In fact, the worse performance of the detection systems on the 2003 NIST SRE can be due to the smaller age difference between the impostor and the true speaker.

Since the age difference between the impostor and the true speaker affects the performance, the impostor trials were divided into three groups and the performance of the detection systems were analyzed. The range of the groups are, based on the median values of the distributions, defined as follows: absolute age difference equal or less than 5 years, more than 5 and equal or less than 10 years, and more than 10 years. The 2001 NIST SRE has approximately 50% of the impostor trials from the group with more than 10 years of absolute age difference, whereas the 2003 NIST SRE has approximately 50% of the impostor trials from the group with equal or less than 5 years absolute age difference. The performance for the group of the impostor trials with more than 10 years of absolute age difference on the 2003 NIST SRE is shown in Fig. 13 (similar pattern in performance is achieved in the 2001 NIST SRE). Note that the performances of the female–female trials are better than the male–male trials. This is likely result of the larger amount of training data produced by female speakers. However, the performances of the female–female trials of the prosody-based systems are more affected when the age difference is equal or less than 5 years, as shown in Fig. 13 by the percentages above the circles (also represented by the size of each circle). In the acoustic baseline, the effect is the opposite: the male–male trials are more affected by the small age difference.

Despite the effect of the small age difference, there may be still other factors affecting the systems performance. The degradation in performance of the acoustic baseline (as shown in Table 1) is partially explained by the small age difference (at least in the male–male trials).

8. Conclusions

This work presented a new approach to model prosodic information using the rate of change of F0 and short-term energy contours. Since different speakers may be characterized by different prosodic patterns (e.g., intonation, stress, and rhythm), such modeling was used to characterize speaker-specific information. We showed that the bigram modeling of the sequence of classes representing joint-state of both contours captures more speaker-dependent information than the system based on the distribution of the

F0, energy and their respective time derivatives. The proposed approach achieved a relative improvement over the baseline of 25% and 17% on the 2001 and 2003 NIST SRE, respectively. This 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. Since most of the study uses the evaluation data from the 2001 NIST SRE, we also show that all the results hold when using the evaluation data from the 2003 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 prosody-based systems.

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 36% relative improvement in performance on the 2001 NIST SRE, and 28% relative improvement on the 2003 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 (42% relative improvement). 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 2003 NIST SRE (12% relative improvement), which is the result of factors such as the smaller age difference between impostor and true speaker, and different handsets for training and testing.

In this work, we showed that the joint-state classes representing the dynamics of F0 and short-term energy contour are used for characterizing speaker-specific information. However, the relationship between prosodic phenomena and the joint-state classes was not established. 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. 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.

References

Adami, A., Mihaescu, R., Reynolds, D.A., Godfrey, J., 2003. Modeling Prosodic Dynamics for Speaker Recognition. ICASSP, Hong Kong, pp. 788–791.

Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, J.J., 2001. Phonetic, Idiolectal and Acoustic Speaker Recognition, 2001: A Speaker Odyssey, Crete, Greece, pp. 55–63.

Andrews, W.D., Kohler, M.A., Campbell, J.P., Godfrey, John J., Hernandez-Cordero, J., 2002. Gender-dependent Phonetic Refraction for Speaker Recognition. ICASSP, Orlando, FL, pp. 149–152.

Atal, B.S., 1972. Automatic speaker recognition based on pitch contours. J. Acoust. Soc. Amer. 52 (6), 1687–1697.

Atkinson, J.E., 1978. Correlation analysis of the physiological factors controlling fundamental voice frequency. J. Acoust. Soc. Amer. 63 (1), 211–222.

Boves, L., Strik, H., 1988. The fundamental frequency-subglottal pressure ratio in speech. J. Acoust. Soc. Amer. 84 (S1), S82.

Campbell, J.P., Reynolds, D., Dunn, R.B., 2003. Fusing High- and Low-level Features for Speaker Recognition, EUROSPEECH, Geneva, Switzerland, pp. 2665–2668.

Carey, M.J., Parris, E.S., Lloyd-Thomas, H., Bennet, S., 1996. Robust Prosodic Features for Speaker Identification. ICSLP, pp. 1800–1803.

Collier, R., 1975. Physiological correlates of intonation patterns. J. Acoust. Soc. Amer. 58 (1), 249–255.

Cover, T.M., Thomas, J.A., 1991. Elements of Information Theory. John Wiley & Sons, Inc.

Doddington, G., 1971. A method for speaker verification. J. Acoust. Soc. Amer. 49 (1), 139(A).

Doddington, G., 2001. Speaker recognition based on idiolectal differences between speakers. EUROSPEECH, Aalborg, Denmark, pp. 2521–2524.

Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A., 2000. The NIST speaker recognition evaluation – overview, methodology, systems, results, perspective. Speech Commun. 31 (2-3), 225–254.

Fant, G., Kruckenberg, A., Nord, L., 1991. Prosodic and segmental speaker variations. Speech Commun. 10, 521–531.

Furui, S., 1981. Comparison of speaker recognition methods using statistical features and dynamic features. IEEE Trans. Acoust. Speech Signal Process. 29 (3), 342–350.

Furui, S., 1997. Recent advances in speaker recognition. Pattern Recognition Lett. (18), 859–872.

Gillick, L., Cox, S.J., 1989. Some Statistical Issues in the Comparison of Speech Recognition Algorithms, ICASSP. IEEE, Glasgow, Scotland, pp. 532–535.

Hermansky, H., 1999. Mel cepstrum, deltas, double-deltas – What else is new. In: Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland.

Jelinek, F., 1997. Statistical Methods for Speech Recognition. MIT Press, Cambridge.

Kajarekar, S. et al., 2003. Speaker Recognition Using Prosodic and Lexical Features. ASRU, St. Thomas, U.S., Virgin Islands, pp. 19–24.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Amer. 87 (2), 820–857.

Ladefoged, P., 1968. Linguistic aspects of respiratory phenomena. Ann. New York Acad. Sci. 155, 141–151.

Lehiste, I., 1970. Suprasegmentals. MIT Press, Cambridge, MA.

Lummis, R.C., 1973. Speaker verification by computer using speech intensity for temporal registration. IEEE Trans. Audio Electroacoust. AU-21 (2), 80–89.

Markel, J.D., Oshika, B.T., Gray Jr., A.H., 1977. Long-term feature averaging for speaker recognition. IEEE Trans. Acoust. Speech Signal Process. 25 (4), 330–337.

Martin, A., 2001. NIST 2001 Speaker Recognition Evaluation Plan.
Martin, A., 2003. NIST 2003 Speaker Recognition Evaluation Plan.
Martin, A., Przybocki, M., 2000. The NIST 1999 speaker recognition evaluation – an overview. Digital Signal Process. 10 (1–3), 1–18.
Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., 1997. The DET Curve in Assessment of Detection Task Performance. EUROSPEECH, Rhodes, Greece, pp. 1895–1898.
Navratil, J., Jin, Q., Andrews, W., Campbell, J.P., 2003. Phonetic Speaker Recognition Using Maximum-Likelihood Binary-Decision Tree Models. ICASSP, Hong Kong, pp. 796–799.
Peterson, G.E., Lehiste, I., 1960. Duration of syllable nuclei in English. J. Acoust. Soc. Amer. 32 (6), 693–703.
Reynolds, D.A., 1997. HTIMIT and LLHDB: Speech Corpora for the Study of Handset Transducer Effects. ICASSP, Detroit, pp. 1535–1538.
Reynolds, D. et al., 2003. The SuperSID Project: Exploiting High-level Information for High-accuracy Speaker Recognition. ICASSP, Hong Kong, pp. 784-787.
Reynolds, D.A., Rose, R.C., Smith, M.J.T., 1992. PC-based TMS320C30 implementation of the Gaussian mixture model text-independent speaker recognition system, International Conference on Signal Processing Applications and Technology, Cambridge, MA, pp. 967–973.
Reynolds, D.A., Dunn, R.B., McLaughlin, J.J., 2000a. The Lincoln Speaker Recognition System: NIST Eval2000. ICSLP, Beijing, China, pp. 470–473.
Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000b. Speaker verification using adapted mixture models. Digital Signal Process. 10, 19–41.
Sönmez, K., Heck, L., Weintraub, M., Shriberg, E., 1997. A Lognormal Tied Mixture Model of Pitch for Prosody-Based Speaker Recognition. EUROSPEECH, Rhodes, Greece, pp. 1391–1394.
Sonmez, K., Shriberg, E., Heck, L., Weintraub, M., 1998. Modeling Dynamic Prosodic Variation for Speaker Verification. ICSLP, Sydney, Australia, pp. 3189–3192.
Soong, F.K., Rosenberg, A.E., 1988. On the use of instantaneous and transitional spectral information in speaker recognition. IEEE Trans. Acoust. Speech Signal Process. 36 (6), 871–879.
Talkin, D., 1995. A robust algorithm for pitch tracking (RAPT). In: Kleijn, W.B., Paliwal, K.K. (Eds.), Speech Coding and Synthesis. Elsevier, New York.
van Dommelen, W.A., 1987. The contribution of speech rhythm and pitch to speaker recognition. Lang. Speech 30 (4), 325–338.
Weber, F., Manganaro, L., Peskin, B., Shriberg, E., 2002. Using Prosodic and Lexical Information for Speaker Identification. ICASSP, Orlando, FL, pp. 141–144.
Werner, S., Keller, E., 1994. Prosodic aspects of speech. In: Keller, E. (Ed.), Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges. John Wiley, Chichester, New York, pp. 23–40.
Xiang, B., 2003. Text-independent speaker verification with dynamic trajectory model. IEEE Signal Process. Lett. 10 (5), 141–143.