Modeling Prosodic Features With Joint Factor Analysis for Speaker Verification

Najim Dehak, Pierre Dumouchel, and Patrick Kenny

***Abstract*—In this paper, we introduce the use of continuous prosodic features for speaker recognition, and we show how they can be modeled using joint factor analysis. Similar features have been successfully used in language identification. These prosodic features are pitch and energy contours spanning a syllable-like unit. They are extracted using a basis consisting of Legendre polynomials. Since the feature vectors are continuous (rather than discrete), they can be modeled using a standard Gaussian mixture model (GMM). Furthermore, speaker and session variability effects can be modeled in the same way as in conventional joint factor analysis. We find that the best results are obtained when we use the information about the pitch, energy, and the duration of the unit all together. Testing on the core condition of NIST 2006 speaker recognition evaluation data gives an equal error rate of 16.6% and 14.6%, with prosodic features alone, for all trials and English-only trials, respectively. When the prosodic system is fused with a state-of-the-art cepstral joint factor analysis system, we obtain a relative improvement of 8% (all trials) and 12% (English only) compared to the cepstral system alone.**

***Index Terms*—Joint factor analysis, Legendre polynomial, prosodic features, speaker recognition.**

I. Introduction

A LARGE majority of speaker verification systems are based on cepstral coefficients such as Mel frequency cepstral coefficients (MFCCs). Several models have been proposed for these kinds of features [1]–[3]; however, in the last ten years, we have observed the use of prosodic information for speaker verification [4]–[7]. Prosodic information characterizes the speaker's intonation and speaking style. An interesting characteristic of prosodic features, such as pitch and unit duration (for phonemes and syllables), is that they are less sensitive to channel effects than cepstral features. Prosodic systems are especially effective when large amounts of data are available to train speaker models [5], [8]. In these situations, systems that fuse both types of features (cepstral and prosodic features) give better results than the cepstral systems alone [7], [9].

Frequently used prosodic parameters are based on pitch and energy contours statistics. For example, in [4], Sönmez *et al.* show that pitch has a log normal distribution, and they propose a speaker verification system based on distances between pitch histogram values. The same authors propose a pitch contour stylization technique [10] based on the segmentation of the pitch contour. In each segment, they extract a set of parameters, such as the median, the slope of the pitch contour, and the segment feature duration. Each feature is modeled with a Gaussian distribution.

In [5], Adami *et al.* used an n-gram approach for modeling segments obtained by pitch and energy contour stylization. The objective of the n-gram approach is to model the speaker's speaking style. The authors also proposed the application of dynamic time warping between pitch contours extracted from two different recordings with the same context (same word or sentence). This approach gives better results but it requires both word alignment and word detection. The work presented in [7] by Kajarekar *et al.* introduces a novel approach called nonuniform extraction region features (NERFs). A NER is a region from the utterance between two consecutive pauses larger than a threshold. The pause duration threshold generally used is 500 ms. In [7], the authors extract a set of 32 features from each NER (although not all features can be extracted in all cases). This feature set corresponds to statistics of pitch contour evolution, and information concerning phone durations (or higher level units). The advantage of using NERFs comes from the extraction of long-term speaker characteristics. The suggested model for these features is a Gaussian mixture model. A drawback of this approach is that the features extracted by Kajarekar *et al.* [7] use information concerning phone durations in each NER, and this requires a phonetic alignment.

Another variant of NERFs uses syllables as the basic unit. This variant, named syllable-based nonuniform extraction region features (SNERFs), was introduced by Shriberg *et al.* and applied in [9] and [11]. The syllabic segmentation is obtained using a speech recognition system. This approach consists of discretizing the prosodic syllable features using several bins. The features are then modeled with a support vector machine (SVM) with an n-gram kernel. The results obtained with SNERFs [11] are given only on the English trials of the NIST speaker recognition evaluation dataset because the authors used an English speech recognition system. (The approach could be extended to other languages by using, in parallel, several speech recognition systems with various languages; an approach similar to parallel phone recognition language modeling (PPRLM) for language identification [12].) This approach represents the state-of-the-art in prosodic feature modeling, and fusing this system with a cepstral based system improves the performance of the latter [11].

Manuscript received February 15, 2007; revised April 20, 2007. This work was supported in part by the Natural Science and Engineering Reseach Council of Canada and in part by the Canadian Heritage Fund for New Media Research Networks. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Kay Berkling.

N. Dehak and P. Dumouchel are with the Centre de Recherche Informatique de Montréal (CRIM), Montréal, QC H3A 1B9, Canada and also with the École de Technologie Supérieure (ÉTS), Université du Québec, Montréal, QC H3O 1K3, Canada (e-mail: najim.dehak@crim.ca; pierre.dumouchel@etsmtl.ca).

P. Kenny is with the Centre de Recherche Informatique de Montréal (CRIM), Montréal, QC H3A 1B9, Canada (e-mail: patrick.kenny@crim.ca).

Digital Object Identifier 10.1109/TASL.2007.902758

The majority of the methods that we have mentioned are based on discrete modeling of pitch and energy contours. In this paper, we propose to use continuous modeling of these contours. The advantage of a continuous prosodic contour model is that continuous models already developed in the speaker recognition literature can be applied. In particular, factor analysis [2] can be used to treat the effects of the speaker and intersession variability in prosodic features. Continuous prosodic contour modeling based on Legendre polynomial expansions has been successfully used in the field of language identification [13] and in quantitative phonetics [14].

We extract pitch and energy at 10-ms intervals, and we break the contours into pseudosyllabic units (or *segments* for short). We approximate the pitch and energy contours in each segment by Legendre polynomial expansions. The Legendre polynomial coefficients for pitch and energy together with segment duration form the prosodic feature set. We calculate one feature vector for each pseudosyllable. We then model these features using Gaussian mixture models (GMMs) and compensate for speaker and session variability effects using joint factor analysis. The speaker factors play a crucial role here since the number of feature vectors corresponding to the given enrollment utterances (400 on average) may be too small for the classical maximum *a posteriori* (MAP) estimation to perform reliably. In our initial investigations, our segmentation into pseudosyllable units does not rely on the output of a speech recognition system as is the case with the SNERF approach, and the results obtained with our modeling are in the range of the results obtained with SNERF systems.

The structure of this paper is as follows. Section II-A defines the prosodic features used in our system. Section II-B summarizes the joint factor analysis model. Experiments and results are presented in Section III. The fusion of the prosodic and cepstral systems is done in Section IV. Section V presents the conclusion.

II. Prosodic Features and Joint Factor Analysis

A. Feature Extraction

We extract log pitch and log energy values calculated at 10-ms intervals using the Praat package [15]. Pitch is calculated with the autocorrelation method proposed in [15] and is undefined in unvoiced regions. The Praat pitch extraction function settings are given in Table I. We used only the voiced part of the speech signal in our modeling. Log energy is normalized on an utterance basis by subtracting the maximum value for the whole utterance.

We now describe how pitch and energy contours (containing more than one syllable) are segmented into several pseudosyllables using only the energy contour.

1) Segmentation: In order to model the prosodic contours based on the syllable as a unit, we segment the long prosodic contours into syllable-like regions in the same way as in [13]. This method is based on detecting the valley points of the energy contour. In general, these valley points serve as segment boundaries, but we impose a minimum duration constraint of 60 ms. (This enables us to calculate Legendre polynomial expansions with six terms.) An example of log pitch and normalized log energy segmentation is given in Fig. 1.

TABLE I
Praat Pitch Extraction Arguments

Analysis widows	30 ms
Time step	10 ms
Pitch floor	75 Hz
Maximum number of pitch candidates	5
Very accurate	no
Silence threshold	0.03
Voicing threshold	0.6
Octave cost	0.01
Octave-jump cost	0.6
Voiced / unvoiced cost	0.14
Pitch ceiling	350 Hz

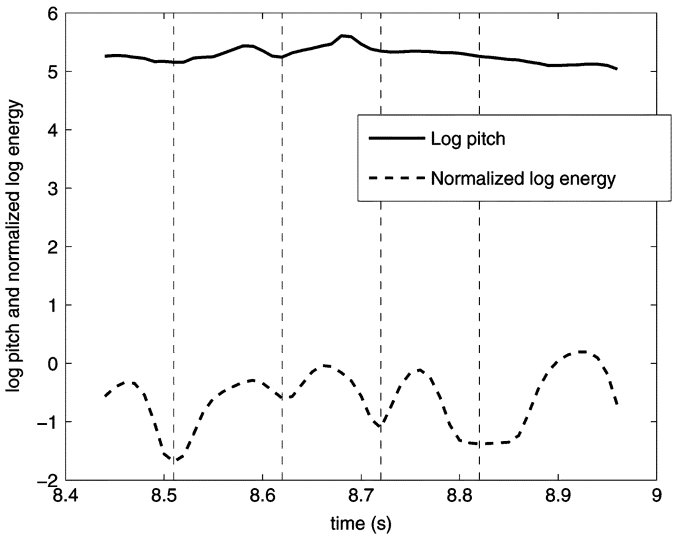

Fig. 1. Example of segmentation of the log pitch and normalized log energy contours extracted from voiced speech.

We will show in the next paragraph how the pitch and energy contours (based on pseudosyllable units) are approximated by Legendre polynomials.

2) Approximation and Time Normalization: In each segment obtained, we carried out an approximation of the pitch and energy contour by taking the M leading terms in a Legendre polynomial expansion. That is, each contour $f(t)$ (where t represents time) is approximated as

$$f(t) = \sum_{i=0}^{M} a_i P_i(t) \tag{1}$$

where $P_i(t)$ is the ith Legendre polynomial, and we set $M = 5$ in our implementation. Fig. 2 shows how Legendre polynomials (P_i) model a log pitch contour. Each coefficient models a particular aspect of the contour. For example, a_0 is interpreted as mean of the segment, a_1 is the slope, a_2 gives information about the curvature of the segment, and a_3, a_4, a_5 model the fine detail.

However, in order for these coefficients to be comparable across segments, it is important to carry out a time normalization. All the segments must be scaled and mapped onto the same interval $[-1, +1]$. This technique of approximation of the

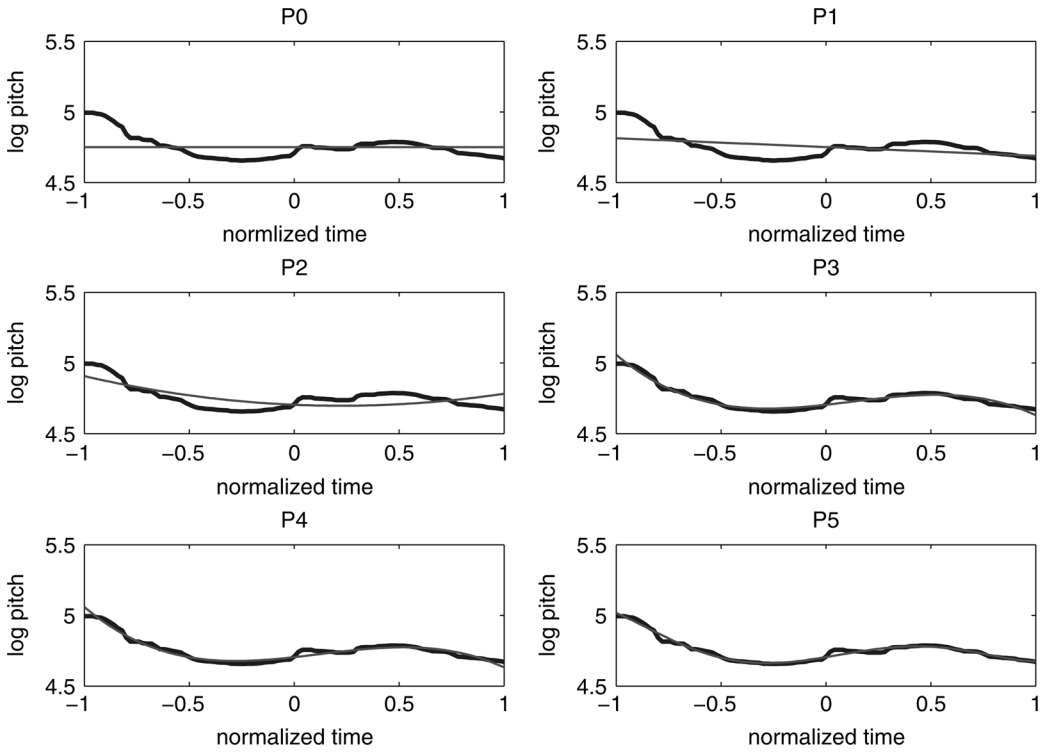

Fig. 2. Approximation of the log pitch contour using Legendre polynomials with different order.

prosodic contours has been successfully applied in quantitative phonetics [14] and in engineering applications [13].

For each segment, we used six coefficients to represent the pitch contour and six coefficients to represent the energy contour. These pitch and energy features, with the addition of the segment duration, produced a 13-dimensional feature vector for each segment. These are the prosodic feature vectors that we used for GMM and factor analysis modeling. Note that since we used only the voiced part of the speech signal and we imposed a pseudosyllable minimum duration of 60 ms, the total number of feature vectors within an utterance (an utterance is a 5-min telephone conversation) was much less than in the corresponding MFCC frames. (There is an average of 400 prosodic vectors per utterance.)

B. Joint Factor Analysis as a Model of Prosody

The joint factor analysis is a model of speaker and session variability in GMMs. Although it is traditionally used with cepstral-type features, it can be applied with any type of continuous features for which Gaussian mixture modeling is appropriate.

As usual, we assume that each speaker is represented by the means, covariances, and weights of a mixture of C multivariate diagonal-covariance Gaussian densities defined in a continuous feature space of dimension F. The GMM for a target speaker is obtained by adapting the parameters of a universal background model (UBM) trained using a large number of utterances. In joint factor analysis [16], [17], the basic assumption is that a speaker and channel-dependent supervector[1] $\boldsymbol{M}$ can be decomposed into a sum of two supervectors: a speaker supervector $\boldsymbol{s}$ and a channel supervector $\boldsymbol{c}$:

$$\boldsymbol{M} = \boldsymbol{s} + \boldsymbol{c} \tag{2}$$

where $\boldsymbol{s}$ and $\boldsymbol{c}$ are normally distributed. The motivation for assuming this type of decomposition is explained in [16] and [17]. Kenny *et al.* described how the speaker-dependent supervector and channel-dependent supervector can be represented in low-dimensional spaces. The first term in the right-hand side of (2) is modeled by assuming that if $\boldsymbol{s}$ is the speaker supervector for a randomly chosen speaker then

$$\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} \tag{3}$$

where $\boldsymbol{m}$ is the speaker- and channel-independent supervector (which can be taken to be the UBM supervector), $\boldsymbol{d}$ is a diagonal matrix, $\boldsymbol{v}$ (which contains the eigenvoice) is a rectangular matrix of low rank, and $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent random vectors having standard normal distributions. In other words, $\boldsymbol{s}$ is assumed to be normally distributed with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{v}\boldsymbol{v}^* + \boldsymbol{d}^2$. The components of $\boldsymbol{y}$ are the speaker factors. The speaker space is the affine space defined by translating the range of $v\boldsymbol{v}^*$ by m. If $(d = 0)$, then all speaker supervectors are contained in the speaker space and if $(d \neq 0)$, the term $\boldsymbol{d}z$ serves as a residual which compensates for the fact that it may not be possible in practice to estimate v reliably [17].

The channel-dependent supervector $\boldsymbol{c}$ which represents the channel effect in an utterance is assumed to be distributed according to

$$\boldsymbol{c} = \boldsymbol{u}x \tag{4}$$

[1]A GMM supervector is the concatenation of GMM means vectors.

TABLE II
RESULTS (ON EQUAL ERROR RATE) OBTAINED ON ALL TRIALS OF THE CORE CONDITION OF THE FEMALE SUBSET OF THE NIST 2006 EVALUATION DATASET USING PROSODIC JOINT FACTOR ANALYSIS WITH SEVERAL CONFIGURATIONS

	Speaker factors	Intersession factors	EER
1	50	20	**15.9%**
2	70	30	16.0%
3	90	40	16.2%
4	0	0	29.2%
5	50	0	28.0%
6	0	20	25.6%

where $\boldsymbol{u}$ is a rectangular matrix of low rank, $\boldsymbol{x}$ is also normally distributed. This is equivalent to saying that $\boldsymbol{c}$ is normally distributed with zero mean and covariance $\boldsymbol{uu}^*$. The components of $\boldsymbol{x}$ are the channel factors in the factor analysis model.

In cepstral factor analysis modeling, the term "channel variability" is used to represent the variability between several recording sessions of a given speaker because in the majority of the cases, this variability is caused by channel effects. However, for the high-level features as our modeling, the term "intersession variability" is probably more appropriate than channel variability.

For this paper, joint factor analysis with prosodic features is implemented essentially in the same way as standard joint factor analysis with cepstral features (only the features are different).

III. EXPERIMENTS WITH PROSODIC FEATURES

A. Database

We carried out our experiments on the core condition of the NIST 2006 speaker recognition evaluation (SRE) [18]. This evaluation set contains 350 males, 461 females, and 51 448 test utterances. For each target speaker model, a 5-min recording is available containing roughly 2 min of speech for a given speaker. We also tested our approach on the eight conversation training condition of the NIST 2006 SRE dataset [18]. This dataset contains 298 males, 402 females, and 32 509 test files. We used a UBM which contains 512 Gaussians, trained on the (13-dimensional) prosodic features extracted from the NIST 2004–2005 SRE datasets. The same data was also used to train the factor analysis model. In the factor analysis framework, it is necessary to use this kind of dataset to model intersession variability because each training speaker has to be recorded several times (ideally under a wide variety of recording conditions). Verification scores were normalized using zt-norm normalization with 100 t-norm models and 100 z-norm utterances from the NIST 2004 SRE. The zt-norm technique has proved to be useful in the factor analysis framework [16], [19].

B. Factor Analysis With Prosodic Features

The objective of the experiments carried out in this section is to find the best configuration of the factor analysis model (i.e., the optimal number of speaker and intersession factors) for the 13-dimensional prosodic features presented in Section II-A. The results obtained on the female subset of the core condition (all trials) of the NIST 2006 SRE dataset are summarized in Table II.

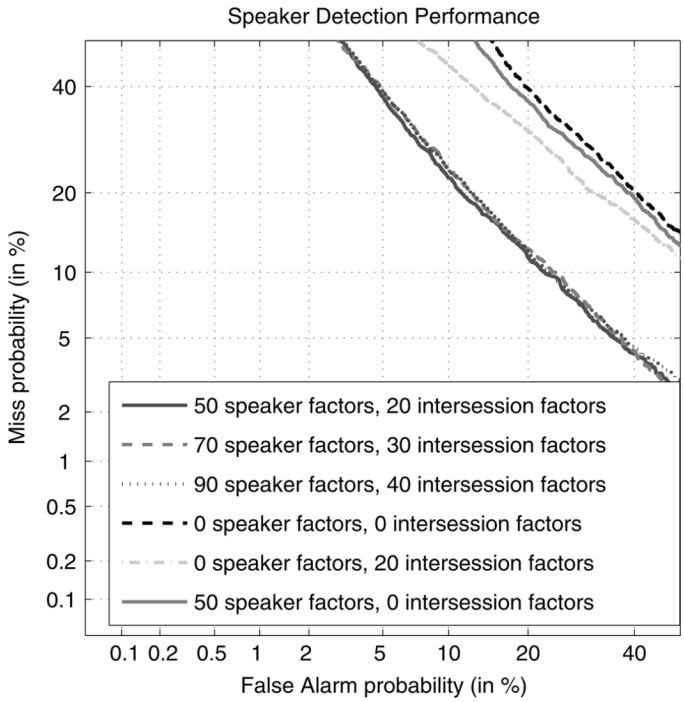

Fig. 3. DET curves showing the results on all trials of the core condition of the female subset of the NIST 2006 evaluation dataset using prosodic joint factor analysis with several configurations.

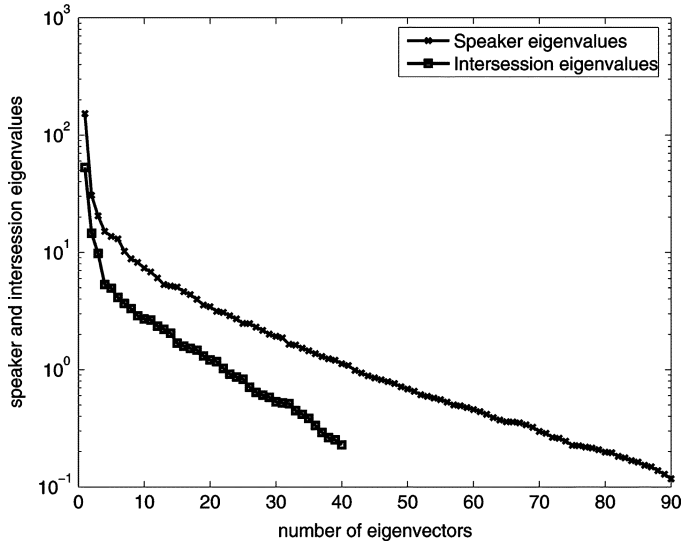

Fig. 4. Eigenvalues of $\boldsymbol{vv}^*$ (the speaker eigenvalues, upper curve) and the eigenvalues of $\boldsymbol{uu}^*$ (the intersession eigenvalues, lower curve) obtained by fitting the factor analysis model to the NIST2005 and 2004 SRE using prosodic features.

We found that the best configuration was 50 speaker factors and 20 intersession factors (see lines 1, 2, and 3 of Table II). The detection error tradeoff (DET) curves given in Fig. 3 show the results obtained in these experiments. An explanation for these results can be given by observing Fig. 4, which shows how the prosodic factor analysis model fits the training data (i.e., the female portion of the NIST 2004 and 2005 evaluation datasets).

Note that for both speaker and intersession eigenvalues, there is a very rapid decrease initially and an exponential decrease

thereafter. This suggests an explanation as to why large numbers of speaker and intersession factors are not helpful. Note also that the speaker variability (as measured by the sum of the eigenvalues) is much greater than intersession variability. This confirms that our prosodic features are less sensitive to intersession effects, but vary considerably from one speaker to another.

In order to show the effectiveness of the speaker and intersession factors, we carried out three experiments with and without speaker and intersession factors. The results are given in lines 4, 5, and 6 of Table II.

- Line 4 of Table II corresponds to an experiment where we did not use the speaker and intersession factors ($u = 0, v = 0, d \neq 0$). This is quite equivalent to the standard GMM-UBM approach for speaker verification [1]. The results in line 1 and 4 of Table II show that when we did not use speaker and intersession factors, there is a very large degradation in performance [15.9% equal error rate (EER) with speaker and intersession factors versus 29.2% EER without speaker and intersession factors].
- Line 5 of Table II corresponds to an experiment which consists of using only 50 speaker factors and no intersession factors ($u = 0, v \neq 0, d \neq 0$). This modeling is a combination of eigenvoice MAP and classical MAP. The purpose of this experiment is to show the importance of the intersession factors. The results given in line 1 and 5 of Table II show that the use of intersession factors improves the performance from an EER of 28.0% (without intersession factors) to 15.9% (with intersession factors).
- Line 6 of Table II corresponds to an experiment which consists of using classical MAP adaptation for enrollment, intersession factors, but no speaker factors ($u \neq 0, v = 0, d \neq 0$). The purpose of this experiment is to verify the contribution of the speaker factor component. The results given in line 1 and 6 of Table II show that the use of speaker factors improves the performances from an EER of 25.6% (without speaker factors) to 15.9% (with speaker factors).

We conclude from the last experiment that the speaker factors play an important part in enrolling target speakers. An explanation of the result is that in our approach, we have few feature vectors to estimate a target speaker model (an average of 400 vectors per enrollment). It is important to note that in classical MAP adaptation, only the Gaussians observed in the enrollment data are adapted. (This is because in traditional MAP adaptation, the GMM supervector covariance matrix is assumed to be diagonal. There is no correlation between the Gaussians in GMM model.) However, in factor analysis modeling, the GMM supervector covariance matrix was given by $\boldsymbol{v}\boldsymbol{v}^* + \boldsymbol{d}^2$ with diagonal matrix $\boldsymbol{d}$ and low rank rectangular matrix $\boldsymbol{v}$. The matrix $\boldsymbol{v}$ takes into account the correlations between the Gaussians in a speaker model. (Gaussians which are not observed in the enrollment data are also adapted by using statistics of the other Gaussians.) The number of speaker factors whose values have to be estimated in enrollment is much less than the number of parameters estimated in classical MAP adaptation. Thus, the method is effective even with very small amounts of enrollment data.

The use of the intersession factors proves to be important in our approach because they model session variability (see lines 1 and 5 of Table II). Our prosodic factor analysis system gives the best results when we use both the speaker and intersession factors. The following section shows the effectiveness of Legendre polynomials for modeling the prosodic contours and the importance of the information given by energy for speaker modeling.

TABLE III
RESULTS (ON EQUAL ERROR RATE) OBTAINED ON ALL TRIALS OF THE CORE CONDITION OF THE FEMALE SUBSET OF THE NIST 2006 EVALUATION DATASET USING JOINT FACTOR ANALYSIS WITH SEVERAL TYPES OF PROSODIC FEATURES

	Features	EER
1	slope + curvature + duration	22.9%
2	pitch contour + duration	20.2%
3	pitch and energy contours + duration	**15.9%**

C. Importance of Energy, Duration, and Pitch

In order to compare with other approaches to prosodic feature extraction and modeling, we performed three experiments on the female subset of the NIST 2006 evaluation data (core condition, all trials), varying the feature set as follows.

- In the first experiment, we computed, for each segment, the slope and curvature of the pitch and energy contours as well as duration of the segment as features in a manner similar to [5]. Note that the slope and curvature correspond to the coefficients a_1 and a_2 of (1). The result is given in line 1 of Table III.
- In the second experiment, we used as segment features the Legendre polynomial coefficients of the pitch contour (all six coefficients) and the duration of the segment. The energy contour was not used. This modeling was similar to [4]. Line 2 of Table III gives the results of this experiment.
- In the last experiment, we used all 13 prosodic features, as described in Section II-A. The result of this experiment is given on line 3 of Table III.

Our best performance with the various prosodic feature sets was obtained with the full 13-dimensional feature set (see line 3 of Table III). The energy contour clearly adds a substantial amount of information to the pitch contour (a 4% absolute reduction in EER, comparing the results of the lines 2 and 3 of Table III). The same conclusion is found in SNERF modeling [11]. Shriberg *et al.* found that using information about the pitch, energy, and duration of different units give the best performance. We can see in Table III that the slope and curvature representation of the pitch and energy contour is not as good as using all of the Legendre polynomial coefficients, (comparing the results for line 1 and 3 of Table III).

D. Results for Both Genders

We tested the factor analysis model with the 13 prosodic coefficients for both genders on the core condition of the NIST 2006 speaker recognition evaluation dataset. We used the same factor analysis configuration for each gender (50 speaker factors and 20 intersession factors). The UBM size is 512 Gaussians for each gender. The decision scores are normalized with zt-norm. The results obtained (under the two conditions: English only, and all trials) are given in Table IV.

TABLE IV
RESULTS (ON EQUAL ERROR RATE) OBTAINED WITH GENDER-DEPENDENT PROSODIC FACTOR ANALYSIS ON THE CORE CONDITION OF THE NIST 2006 EVALUATION DATASET

Gender	English	All trials
Female	13.6%	15.9%
Male	16.4%	17.3%
Both genders	14.6%	16.6%

TABLE V
RESULTS (ON EQUAL ERROR RATE) OBTAINED WITH GENDER-DEPENDENT PROSODIC FACTOR ANALYSIS ON THE EIGHT CONVERSATION TRAIN AND ONE CONVERSATION TEST CONDITION OF THE NIST 2006 EVALUATION DATASET

Gender	English	All trials
Female	9.3%	10.4%
Male	10.2%	12.8%
Both genders	9.8%	11.3%

The results show that these prosodic features give better results for the female cases compared to the male ones. The opposite is also true for our cepstral-based factor analysis system. Ferrer *et al.* have recently published EERs in the range 12.3%–14.2% on the English subset (both genders) of the NIST 2006 evaluation data obtained with three systems based on the SNERF approach with an SVM classifier [20]. If we restrict ourselves to the English subset of the NIST 2006 speaker recognition evaluation dataset then our equal error rate is quite similar, namely 14.6% (rather than 16.6% on the core condition as a whole).

Thus, the results obtained by our prosodic system are comparable to the results obtained with other systems based on the SNERF approach. The advantage of our approach is that the segmentation into pseudosyllabic units is carried out in an unsupervised manner by using only the energy contour. On the other hand, a speech recognition system is needed for the SNERF approach. Although the results obtained by these systems are reported only on the English trials, our system is not limited by language restrictions.

In order to compare with other approaches, Queensland University of Technology (QUT) gave us the results of their prosodic system,[2] which is based on an approach similar to that proposed by Adami *et al.* [5]. On the core condition of NIST 2006 speaker recognition evaluation dataset, EERs of 21.1% (English trials) and 22.5% (all trials) were obtained. It is clear that our approach produces better results.

E. Prosodic Factor Analysis on the Eight Conversation Train and One Conversation Test Task

We tested the factor analysis model with the 13 prosodic coefficients for both genders on the eight conversation train and one conversation test condition of the NIST 2006 speaker recognition evaluation dataset. We used the same factor analysis configuration for each gender as used on the last experiment (50 speaker factors, 20 intersession factors). The gender-dependent UBM contains 512 Gaussians. The decision scores are normalized with zt-norm. The results obtained (under the two conditions: English only, and all trials) are given in Table V.

[2][Online]. Available: http://research.ee.sun.ac.za/srefusion/index.php/QUT

We see an improvement of the results obtained by our prosodic system in this task compared to the core condition of the same dataset. However, the results of the prosodic systems based on a SNERF approach [20] are in the range 4.8%–5.2% on the eight conversation train and one conversation test condition of NIST 2006 speaker recognition evaluation dataset (English trials). If we compare these results with the results obtained with our prosodic system (9.8% in EER), it is clear that the prosodic systems based on SNERF approach give better results under this condition of NIST 2006 evaluation. A probable explanation to this phenomenon is that it is necessary to increase the number of the speaker factors and intersession factors to better model the speaker variability and the intersession variability. In cepstral factor analysis system, the best results are obtained when a large number of speaker factors is used in [16, Table VII]. Exploring this will require using large quantities of training data for the prosodic factor analysis model since the target speaker model adaptation tends to saturate quickly with the actual corpus (speaker eigenvalues decrease very quickly to zero; see Fig. 4). Unfortunately, at this point, we do not have sufficient data to train this large number of speaker factors.

IV. FUSION OF PROSODIC AND CEPSTRAL FEATURES

Prosodic systems are usually combined with a baseline cepstral system. To combine our prosodic system with a cepstral system, we carried out a linear combination of the scores of these two systems with equal fusion weights for both systems (prosodic system and baseline system). This fusion technique (often referred to as *naive Bayes*) was compared with other techniques in [21]. The results show that this approach gives equivalent performance to those obtained with the neural network. Since the fusion weights are assumed equal, it is not necessary to use development data to optimize them. In our experiments, it was impractical to divide the training corpus into training and development corpus, because we have insufficient data to train the factor analysis model with prosodic features. (We only used the voiced speech part and we imposed a minimum duration of 60 ms for the pseudosyllables. The total number of observations vectors used to train the UBM and the factor analysis is about three million observations for each gender). The tests were carried out in the core condition and in the eight conversation train and one conversation test condition of NIST 2006 speaker recognition evaluation (all trials and English trials). The following section describes the baseline system.

A. Baseline System

The speaker verification baseline system is the CRIM system used for NIST 2006 speaker recognition evaluation campaign. The system uses factor analysis on cepstral based system. It uses 300 speaker factors and 75 intersession factors. The UBM, which contains 2048 Gaussians, was trained with Linguistic Data Consortium (LDC) releases of Switchboard II, Phases 1, 2, and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1 and Part 2; and NIST 2004 evaluation data. The factor analysis model was trained on the LDC releases of

TABLE VI
FUSION RESULTS (ON EQUAL ERROR RATE) BETWEEN THE BASELINE SYSTEM AND PROSODIC FACTOR ANALYSIS ON THE CORE CONDITION OF THE NIST 2006 EVALUATION DATASET

Speaker verification systems	English	All trials
Baseline system	3.3%	5.0%
Baseline system + prosodic factor analysis	2.9%	4.6%

TABLE VII
FUSION RESULTS (ON EQUAL ERROR RATE) BETWEEN THE BASELINE SYSTEM AND PROSODIC FACTOR ANALYSIS ON THE EIGHT CONVERSATION TRAIN AND ONE CONVERSATION TEST OF THE NIST 2006 EVALUATION DATASET

Speaker verification systems	English	All trials
Baseline system	1.4%	2.0%
Baseline system + prosodic factor analysis	1.5%	2.2%

Switchboard II, Phases 1, 2, and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 and 2005 evaluation data.

The features are extracted using a 25-ms Hamming window. Twelve Mel frequency cepstral coefficients together with log energy are calculated every 10 ms. This 13-dimensional feature vector is subjected to feature warping [22] using a 3-s sliding window. Delta coefficients are then calculated using a five-frame window giving a 26-D feature vector. The resulting decision scores using these features are normalized using zt-norm.

B. Fusion Results

The fusion results of the baseline system and our prosodic factor analysis system are given in Table VI. These results show that, on the core condition of NIST 2006 speaker recognition evaluation dataset, our prosodic features bring additional information to the cepstral parameters. The combination of the two systems gives an 8% and 12% relative reduction in EER for all trials and English trials, respectively. On the eight conversation train and one conversation test condition of NIST 2006 speaker recognition evaluation, naive Bayes fusion between prosodic system and baseline system does not improve the performance (see Table VII). An explanation of these performances is that the results obtained with cepstral factor analysis system are extremely good and it would be very hard to improve them. This low EER may be due to the fact that a part of the data used on the eight conversation train and one conversation test condition of the NIST 2006 SRE came from the NIST 2005 SRE dataset, and in addition, factor analysis was also trained on NIST 2005 SRE. If we compare the cepstral factor analysis results (EERs of 1.4% on English trials and 2.0% on all trials) with those of the MIT-LL/IBM [23] obtained by fusing several cepstral and high-level systems (EERs of 1.5% on English trials and 2.6% on all trials), it is clear that the cepstral factor analysis performance is better than the fusion of several systems.

The DET curves [24] given in Fig. 5 show the performance of the baseline system and fused systems on the English subset of the core condition of NIST 2006 Evaluation data as well as on all trials.

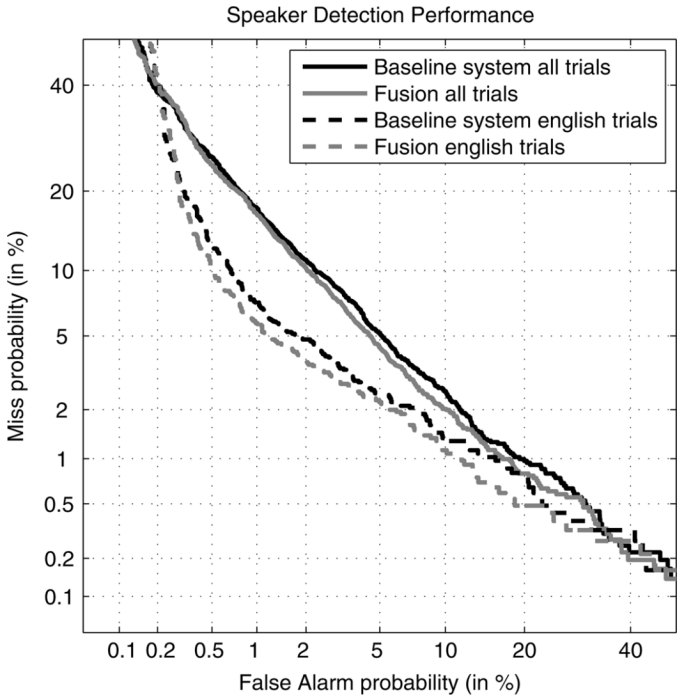

Fig. 5. DET curves showing the fusion results between the baseline system and prosodic factor analysis on the core condition of the NIST 2006 Evaluation dataset.

V. CONCLUSION AND PERSPECTIVES

Although the most successful approach to speaker recognition relies on short-term spectral features such as MFCCs, it has long been recognized that prosodic contours contain complementary information which is much more likely to be robust to the intersession effects which make the speaker recognition problem so challenging. In order to exploit prosodic information, many systems have been developed which use sophisticated modeling techniques such as n-gram modeling of stylized pitch contours [5], or complex language-dependent features, which can only be extracted with the aid of a speech recognizer [7], [11]. However, recent work in language identification [13] and quantitative phonetics [14] has shown that a simple approach to prosodic feature extraction, namely fitting pitch and energy contours with Legendre polynomial expansions, can be very effective. In this paper, we have explored the application of this type of prosodic feature extraction to speaker verification, and we have shown how a prosodic feature-based system fuses well with a state-of-the-art cepstral system; giving a relative EER reduction of 12% on the NIST 2006 set (core condition, English trials). This type of performance improvement is comparable with the best results that have been obtained using any type of prosodic feature modeling. An interesting characteristic of our modeling is that the prosodic features performed better on females than on males (the opposite is true of our cepstral-based system).

A key aspect of the coefficients in the Legendre polynomial expansion is that they define a continuous rather than a discrete feature set. Thus, they are amenable to modeling with the methods that have already been developed for modeling cepstral features in state-of-the-art speaker recognition systems, such as Gaussian mixture modeling [1] and factor analysis [17]. Given the Legendre coefficients, no new algorithmic development is

required to implement a prosodic factor analysis speaker recognition system. All that is needed is to run a series of experiments to determine how best to configure the factor analysis model: the number of Gaussians in the universal background model, the number of speaker factors, and the number of intersession factors. Our experiments showed that both speaker factors and intersession factors play a useful role. Speaker factors are helpful because the number of prosodic feature vectors available for enrolling a target speaker is relatively small. (There is only one feature vector per pseudosyllable, rather than one vector per 10 ms for cepstral features.) Intersession factors are useful because the Legendre coefficients are not entirely robust to session variability. We intend to explore these issues more fully in future work. It would be interesting to test these prosodic features on the auxiliary microphone tasks of the NIST speaker recognition campaign (where channel conditions in enrollment and testing are radically different).

The results of our experiment on eight conversation train and one conversation test of the NIST 2006 SRE are not satisfactory compared with those obtained on the core condition. A probable explanation to this performance is that it is necessary to increase the number of the speaker factors and intersession factors to better model the speaker variability and intersession variability. Exploring this will require using large quantities of training data which we do not have at this point.

In extracting prosodic features, we have used the pseudosyllable as the basic unit. Although this has the virtue of simplicity, other possibilities need to be explored such as the NERF and SNERF approaches [7], [11] (with or without word conditioning). In [25], the authors used features quite similar to us but the task was language identification, and they modeled their features using a continuous HMM (rather than a memoryless GMM as in our case) to capture longer term prosodics. Their results using the HMM are better than using only a GMM [13], [25], which suggests that using HMMs might also be a good idea for speaker recognition.

ACKNOWLEDGMENT

The authors would like to thank B. Baker and R. Vogt from Queensland University of Technology (QUT) for giving them the results of their prosodic systems. They would also like to thank the anonymous reviewers and M. Roch for their comments that helped to improve the content of this paper.

REFERENCES

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Process.*, vol. 10, no. 1–3, pp. 19–41, 2000.

[2] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Factor analysis simplified," in *Proc. ICASSP'05*, Philadelphia, PA, Mar. 2005, pp. 637–640.

[3] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *IEEE Int. Workshop Neural Netw. Signal Process.*, Sydney, Australia, 2000, vol. 2, pp. 775–784.

[4] K. Sönmez, L. Heck, M. Weintraub, and E. Shriberg, "A lognormal tied mixture model of pitch for prosody-based speaker recognition," in *Proc. Eurospeech*, Rhodes, Greece, Sep. 1997, pp. 1391–1394.

[5] A. G. Adami, R. Mihaescu, D. A. Reynolds, and J. J. Godfrey, "Modeling prosodic dynamics for speaker recognition," in *Proc. ICASSP'03*, Hong Kong, 2003, pp. 788–791.

[6] S. Kajarekar, L. Ferrer, A. Venkataraman, K. Sönmez, E. Shriberg, A. Stolcke, H. Bratt, and R. R. Gadde, "Speaker recognition using prosodic and lexical features," in *Proc. IEEE ASRU*, Dec. 2003, pp. 19–24.

[7] S. Kajarekar, L. Ferrer, K. Sönmez, J. Zheng, E. Shriberg, and A. Stolcke, "Modeling NERFs for speaker recognition," in *Proc. Odyssey'04*, Toledo, Spain, Jun. 2004, pp. 51–56.

[8] L. Ferrer, H. Bratt, S. Kajarekar, E. Shriberg, K. Sönmez, K. Stocke, and A. Venkataraman, "Modeling duration patterns for speaker recognition," in *Eurospeech*, Geneva, Switzerland, 2003, pp. 2017–2020.

[9] E. Shriberg, L. Ferrer, A. Venkataraman, and S. Kajarekar, "SVM modeling of SNERF-Grams for speaker recognition," in *Proc. ICSLP'04*, Jeju Island, Korea, Oct. 2004, pp. 1409–1412.

[10] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub, "Modeling dynamic prosodic variation for speaker verification," in *Proc. ICSLP'98*, Sydney, Australia, Aug. 1998, pp. 2631–2634.

[11] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stocke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, pp. 455–472, 2005.

[12] M. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 54–58, Jan. 1996.

[13] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information," in *Proc. ICASSP'05*, Philadelphia, PA, Mar. 2005, pp. 601–604.

[14] E. Grabe, G. Kochanski, and J. Coleman, "Quantitative modelling of intonational variation," in *Proc. Speech Anal. Recognition Technol., Ling., Med.*, 2003, pp. 45–57.

[15] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer." [Online]. Available: http://www.praat.org/

[16] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[17] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[18] [Online]. Available: http://www.nist.gov/speech/tests/spk/index.htm

[19] R. Vogt, B. Baker, and S. Sridharan, "Modelling session variability in text-independent speaker verification," in *Proc. Interspeech*, Lisbon, Portugal, 2005, pp. 3117–3120.

[20] L. Ferrer, E. Shriberg, S. Kajarekar, and K. Sönmez, "Parametrezation of prosodic feature distributions for SVM modeling in speaker recognition," in *Proc. ICASSP'07*, Honolulu, HI, 2007, pp. 233–236.

[21] W. M. Campbell, D. A. Reynolds, and J. P. Campbell, "Fusing discriminative and generative methods for speaker recognition: Experiments on switchboard and NFI/TNO field data," in *Proc. Speaker Odyssey*, Toledo, Spain, June 2004, pp. 41–44.

[22] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, Crete, Greece, Jun. 2001, pp. 213–218.

[23] W. M. Campbell, D. E. Sturim, J. Navratil, W. Shen, and D. A. Reynold, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition," in *Proc. ICASSP'07*, Honolulu, HI, 2007, pp. 217–220.

[24] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," *Proc. Eurospeech*, vol. 4, pp. 1895–1898, 1997.

[25] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information in the ergodic markov model," in *Proc. ICASSP'06*, Toulouse, France, May 2006, pp. 193–196.

Najim Dehak received the M.S. degree in pattern recognition and artificial intelligence applications from the Université de Pierre et Marie Curie, Paris VI, France, in 2004 and the engineer degree in artificial intelligence from the Université des Sciences et de la Technologie d'Oran, Oran, Algeria, in 2003. He is currently pursuing the Ph.D. degree at the École de Technologie Supérieure (ETS), Université du Québec, Montréal, QC, Canada.

He is with the Centre de Recherche d'Informatique de Montréal (CRIM), Montréal. His research interests are speaker modeling and recogntion.

Pierre Dumouchel received the B.Eng. degree from McGill University, Montréal, QC, Canada, and the M.Sc. and Ph.D. degrees from the INRS-Télécommunications, Montréal.

He is Scientific Vice-President at the Centre de Recherche Informatique de Montréal (CRIM) and a Full Professor at the École de Technologie Suprieure (ETS), Université du Québec, Montréal. He was Vice-President of Research and Development at CRIM from 1999 to 2004. Before, he assumed the role of Principal Researcher of the CRIM's Automatic Speech Recognition Team and was a Scientific Columnist at Radio-Canada, the French Canadian National Radio. He has more than 20 years of expertise in speech recognition research, eight years in managing a research team, and three years in managing the Research and Development unit of CRIM. His research has resulted in many technology transfers to such companies as Nortel, Locus Dialog, Canadian National Defence, Le Groupe TVA, as well as many small and medium-sized enterprises, as such as Ryshco Media. His research interests are in searching by transduction and automatic adaptation to new environments. He favored applications of speech recognition for the hard-of-hearing and audiovisual film indexation.

Patrick Kenny received the B.A. degree in mathematics from Trinity College, Dublin, U.K., and the M.Sc. and Ph.D. degrees, also in mathematics, from McGill University, Montréal, QC, Canada.

He was a Professor of Electrical Engineering at INRS-Télécommunications, Montréal, from 1990 to 1995 when he started up a company (Spoken Word Technologies) to spin off INRS's speech recognition technology. He joined the Centre de Recherche Informatique de Montréal (CRIM), Montréal, in 1998, where he now holds the position of Principal Research Scientist. His current research interests are concentrated on Bayesian speaker and channel adaptation for speech and speaker recognition.