Continuous prosodic features and formant modeling with joint factor analysis for speaker verification

Najim DEHAK*+ , Patrick KENNY*, Pierre DUMOUCHEL*+

*Centre de recherche informatique de Montréal (CRIM)
{najim.dehak,patrick.kenny,pierre.dumouchel}@crim.ca
+École de technologie supérieure (ETS)

Abstract

In this paper, we introduced the use of formants contours with prosodic contours based on pitch and energy for speaker recognition. These contours are modeled on continuous manners by using the Legendre polynomials on basic unit which represents syllables. The parameters extracted from the Legendre polynomials coefficients plus the syllables duration are modeled with Gaussian Mixture Models (GMM). Factor analysis is used to treat the speaker and channel variability. The results obtained on the core condition of NIST 2006 speaker recognition evaluation show that the use of formant with prosodic information gives an absolute improvement of approximately 3% on equal error rate (EER) compared with the results obtained by prosodic informations alone. However when the formants and the prosodic system scores are fused with a state of the art cepstral joint factor analysis system, we obtain equivalent results to the results obtained when we fused system based on prosodic features alone with the same cepstral joint factor analysis system. This fusion gives a relative improvement of 8.0% (all trials) and 12.0% (English only) on EER compared to cepstral system alone.

Index Terms: Speaker recognition, prosodic features, formants, joint factor analysis.

1. Introduction

In the framework of speaker verification, the most popular modeling is based on Gaussian Mixture Models with short term cepstral features [1]. However in the last years, the speaker verification community was interested in the use of the speaker long-term characteristics. These characteristics can be modeled on cepstral level [4] or on prosodic level [2][3]. Speaker verification systems based on long-term characteristics are usually fused with short term cepstral systems [2][3].

In this paper, we present an extension of the work presented in [10]. This work consists on modeling long term prosodic speaker characteristics in continuous way rather than discrete. This modeling is based on pseudo syllable like basic unit. The pseudo syllables are detected and segmented in voiced part using energy contour. In each pseudo syllable, we carry out an approximation of the pitch and energy contour using the Legendre polynomials. The Legendre polynomials coefficients of the pitch and energy contours obtained plus the pseudo syllable durations are modeled by a GMM modeling. Factor analysis [8] is used to model the speaker and channel variability. The advantage of this approach is that it don't use any phonetic or word alignment compared to other method presented in [2][3].

The innovation presented in this paper is the modeling of F1 and F2 formant contours in the same way as the pitch and energy contours by using the Legendre polynomials. The use of formants is motived by two points: the first point is that the formants F1 and F2 respectively model the pharyngeal cavity and oral cavity of speaker. The second point is that the use of the formants will introduce a context to pseudo-syllables which will best modeled (the formants are usually used to detect and classify the vowels).

The formants were already used in speaker verification [15][16]. In [15], Mezghani *et al.* combine the formants with the cepstral coefficients (MFCC) and Tanabian *et al.* propose in [15] to model the formants trajectory using decision trees modeling and neural networks approach for speaker recognition.

The structure of this paper is as follow: Section 2 summarizes the Joint Factor Analysis Model; Section 3 defines the prosodic features and formants used in our system; experimental evaluation and results are presented in section 4; The fusion of prosodic system and cepstral system is given in Section 5; Section 6 concludes the paper and gives some perspectives.

2. Joint Factor Analysis as a model of prosody

Joint factor analysis is a model of speaker and session variability in GMM's. 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, covariance matrices, and weights of a mixture of C multivariate diagonal-covariance Gaussian densities defined in some continuous feature space with dimension F. The GMM for target speaker is obtained by adapting the parameters of the Universal Background Model (UBM) [1]. The UBM is trained using a large numbers of speakers utterances. In Joint Factor Analysis [8] [9], the basic assumption is that a speaker and channel-dependent supervector[1] M can be decomposed into a sum of two supervectors, s depends on the speaker and c depends to the channel

$$M = s + c \tag{1}$$

where s and c are normally distributed.

In [8], Kenny *et al.* described how the speaker dependent supervector and channel dependent supervector can be represented in low dimensional spaces. In equation 1, the first term in the right hand side is modeled by assuming that the speaker supervector s for randomly chosen speaker is given by

$$s = m + vy + dz \tag{2}$$

[1] The GMM supervector is the concatenation of the GMM means.

Where $\boldsymbol{m}$ is the speaker and channel independent supervector (UBM means), $\boldsymbol{d}$ is diagonal matrix, $\boldsymbol{v}$ is a rectangular matrix of low rank and $\boldsymbol{y}$ and $\boldsymbol{z}$ are independent random vectors having standard normal distribution. 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 channel-dependent supervector $\boldsymbol{c}$ which represents the channel effect in an utterance, is assumed to be distributed according to

$$\boldsymbol{c} = \boldsymbol{u}\boldsymbol{x} \tag{3}$$

Where $\boldsymbol{u}$ is a rectangular matrix of low rank, $\boldsymbol{x}$ is distributed with standard normal distribution and this is equivalent to saying that $\boldsymbol{c}$ is normally distributed with mean 0 and covariance $\boldsymbol{u}\boldsymbol{u}^*$. The components of $\boldsymbol{x}$ are the channel factors in the Joint Factor Model based speaker verification.

In this paper we use the factor analysis model with formants and prosodic features in exactly the same way as we have used it in the past with cepstral features. We now describe how we calculate our prosodic feature vectors.

3. Feature extraction

Similar to [10], we extract log pitch, log energy and formant (F1,F2) at 10 ms intervals with the Praat package [7]. Pitch is obtaining using the autocorrelation method proposed in [7] and is defined only in voiced regions. For each utterance, the energy is normalized by subtracting the maximum of the same utterance.

3.1. Segmentation

A prosodic contour can extend over several syllables (depending on where devoicing occurs). We segment the contours into pseudo syllable-like regions in the same way as in [10][5]. This method is based on detecting the valley points of energy contour. In general, these valley points serve as segment boundaries but we impose a minimum duration constraint of 60 ms. This enables to calculated Legendre polynomial expansions with six terms. We used Legendre polynomials with order six because in practice, lower order Legendre polynomials do not adequately model the long prosodic segments.

3.2. Contours approximation and time normalization

In each pseudo syllable obtained, we carried out an approximation of the pitch, energy and formant (F1,F2) contours by taking the M leading terms in a Legendre polynomial expansion. That is, each contour $f(t)$ (where t represents time) is approximated as

$$f(t) = \sum_{i=0}^{M} a_i P_i(t) \tag{4}$$

where $P_i(t)$ is the ith Legendre polynomial and we took $M = 5$ in our implementation. The coefficients $a_0, \ldots, a_M$ serve to represent the contour $f(t)$.

However, in order for these coefficients to be comparable across segments, it is important to carry out a time normalization. All time segments must be scaled and mapped onto the same interval $[-1, +1]$. This technique of approximation of the prosodic contours was successfully used in speaker recognition [10], in quantitative phonetics [6] and in engineering applications [5].

In our experiments, for each pseudo syllable we used six coefficients to represent the pitch contour, six coefficients for energy contour, six coefficients for formant F1 contour and six coefficients for formant F2 contour. Adding the duration of the segment, we obtain a 25 dimensional feature vector for each segment. These are the feature vectors that we use for GMM and factor analysis modeling. Note that since the features represent pseudo syllables, there are far fewer of them than there 10 ms frames, as in conventional signal processing.

4. EXPERIMENTS

4.1. Database

The experiments are carried out in the core condition (all trials) of the NIST 2006 Speaker Recognition Evaluation [14]. This evaluation dataset contains 350 males and 461 females and 51448 test files. The world model UBM was learned on NIST 2004-2005 Speaker Recognition Evaluation databases. These same data are used thereafter to learn the factor analysis. Scores of decision are normalized using zt-norm normalization based on 100 impostors from NIST 2004-SRE database. The zt-norm technique proves to be useful in the factor analysis framework [12].

4.2. Results

We carried out two experiments with the aim to show the advantage of combining the formants with other prosodic contours in modeling of the speaker long-term characteristics.

- The first experiment consists in just using information concerning the contour of pitch and energy with the duration of the pseudo-syllables. The parameters used consist of six coefficients of Legendre polynomials for the pitch contour and six coefficients for energy plus the pseudo-syllable duration. The vector of parameters obtained has a dimension of 13. GMM is used to model these vectors of parameters. Factor analysis is also used to model the speaker and channel effect. In [10], many experiments were carried out to show the importance of using both speaker and channel factors to model these new prosodic features. We also proved the importance to combine information of pitch contour, energy contour and pseudo-syllables duration. In this experiment, the best results are obtained with an UBM which contains 512 Gaussians and factor analysis composed by 50 speaker factors and 20 channel factors for female and 40 speaker factors and 15 channel factors for male. The use of relatively few speaker and channel factors is justified by the fact that the eigenvalues of the speakers and channel decrease very quickly to zero. See [10] for more detail.
- In the second experiment, we used the features defined in section 3. The UBM used contains 1024 Gaussians (we used more of Gaussians because the formants gives a context to the pseudo-syllables which requires more Gaussians to better take account of into account these different context.). We obtained the best performance with factor analysis composed by 75 speaker factors and 35 channel factors for females and 50 speaker factors and 20 channel factors for males.

The results obtained in these two experiments are given in Table 1 and Table 2. These results show that contours of the formants F1, F2 bring additional information to contours of the pitch and energy. The use of the formants F1 and F2 gives a significant improvement of the performances, especially in the male case(an absolute improvement of approximately 3% in

Table 1: *Prosodic factor analysis system. 13 prosodic features (pitch and energy contours and pseudo-syllables duration. The NIST 2006 evaluation data. Results in equal error rate.*

	English	All trials
Male	15.6%	16.7%
Female	13,7%	15,9%
Both gender	14.5%	16.4%

Table 2: *Prosodic factor analysis system. 25 prosodic features (pitch, energy and formants F1, F2 contours and pseudo-syllables duration. The NIST 2006 evaluation data. Results in equal error rate.*

	English	All trials
Male	10.9%	11.6%
Female	12.9%	14.7%
Both gender	11.9%	13.4%

EER). The formants give information which better help to discriminate between the pseudo-syllables and thus to better model them. The DET curves [18] given in the Fig.1 show the results obtained by using the formant and the prosodic features together and also prosodic features only.

In the literature, the best results obtained by a prosodic systems are those produced by the SRI prosodic system. This system is based on Syllable-based Nonuniform Extraction Region Features (SNERFs) approach with SVM modeling [2]. The results obtained by this system are in the range 11.9% to 14.0% of EER on the English subset (both genders) of the NIST 2005 evaluation data [19].(There are not yet results published by SRI for their prosodic system on NIST2006 evaluation data). If we restrict ourselves to the English subset of the NIST 2006 evaluation dataset in the case of our formants and prosodic factor analysis system then our equal error rate is 11.9% (rather than 13.4% on the core condition as a whole). Our approach gives equivalent results to the results obtained by the SRI prosodic system. However, our system also gives results under other language conditions because our approach does not use any phonetic and word alignment compared to the SRI prosodic system. Considering its simplicity, the comparison seems to be quite favorable to our approach.

5. Fusion

We fused the scores obtained in the two preceding experiments with speaker verification system based on cepstral features. For each prosodic system, we have carried out linear combination of the scores with the cepstral system scores. The fusion weights are optimized directly in test data. The tests were carried out on the core condition of the NIST 2006 speaker recognition evaluation.

5.1. Baseline system

The speaker verification baseline system is the CRIM system used for NIST 2006 speaker recognition evaluation campaign [11]. The system is a factor analysis cepstral based system. It use 300 speaker factors and 75 channel factors. The UBM is trained with LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; the Fisher English Corpus, Part 1 and the NIST 2004 evaluation data. Factor analysis model is fitted in the LDC releases of Switchboard II, Phases 1, 2 and 3; Switchboard Cellular, Parts 1 and 2; the NIST 2004 and 2005 evaluation data; and the 2006 auxiliary microphone development data.

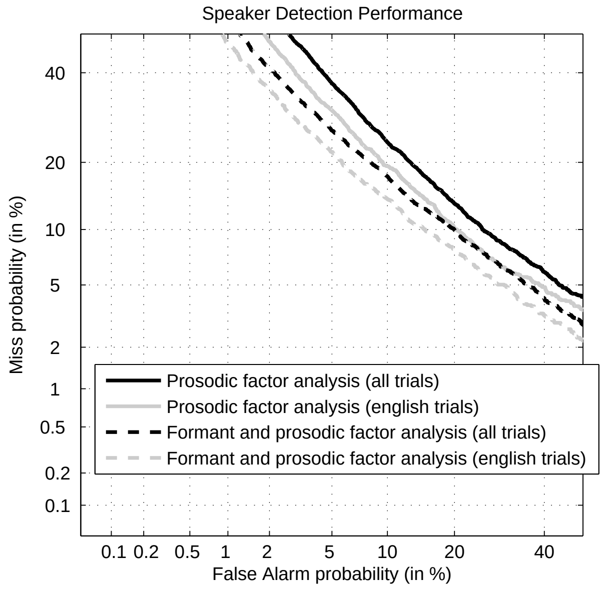

Figure 1: DET Curves showing results on the core condition of the NIST 2006 Evaluation data, English subset only and all trials. We used formant and prosodic features together and prosodic features alone.

The features are extracted on 25 ms Hamming window, Twelve mel frequency cepstral coefficients together with a log energy are calculated for every 10 ms. This 13-dimensional feature vector is subjected to feature warping [13] using a 3s sliding window. Delta coefficients are then calculated using a 5 frames window giving a 26-dimensional feature vector. The decision scores are normalized using zt-norm.

5.2. Results

The results of fusion are given in the Table 3. The fusion of the system based on the prosodic contours and formant with the baseline system produces results equivalent to the results obtained by the fusion of the system based on prosodic contours only with the baseline system. These results can be explained by the fact that ceptral features used in the baseline system probably model information concerning the formants. However the formants give significant improvement to the performances of the prosodic system as show in Table 1 and 2.

6. Conclusion and Perspectives

In this article, we propose to use the formants with prosodic contours for speaker recognition. This combination gives significant improvement of the results compared with the performance obtained just with the prosodic features. However the scores fusion between a cepstral system and system based on the formants and prosodic contours together gives equivalent results to the results obtained by the fusion of the same cepstral system with system based only on the prosodic parameters.

Table 3: *Fusing results between formant and prosodic factor analysis and baseline system and fusing results between baseline system and prosodic factor analysis. The NIST 2006 evaluation data. Core condition, all trials. Results in equal error rate.*

	English	All trials
Baseline systems	3.3%	5.0%
Baseline systems + system based on pitch,energy and pseudo-syllables duration	2.9%	4.6%
Baseline systems + system based on pitch,energy and formants F1, F2 contours and pseudo-syllables duration	2.9%	4.7%

As future works, it would be interesting to see the behavior of our modeling in the case when we have large amounts data to model the target speaker for example the 8 conversations task (a conversation is a recording of 5 minutes) of NIST2006 speaker recognition evaluation campaign.

The approach that we have proposed is based on using pseudo-syllables as a basic unit. This has the virtue of simplicity but other possibilities need to be explored such as SRI's non uniform extraction regions.(The non uniform extraction regions is a region from the utterance between two consecutive pauses which have lengths higher than a threshold; the pause threshold which is generally used is 500ms).

In [17] the authors used quite similar features as us but for a language identification task, and they model these features using a continuous HMM (rather than a memoryless GMM as in our case) to capture longer term prosodics. Their results using this HMM are better then using only a GMM [5][17] which suggests that using HMMs might also be a good idea for speaker recognition.

7. Acknowledgments

This work was funded in part by the Canadian Heritage Fund for New Media Research Networks.

8. References

[1] Reynolds, D.A., Quatieri, T.F., and Dunn R.B., Speaker Verification using Adapted Gaussian Mixture Models,Digital Signal Processing, pp 19-41, 2000.

[2] Shriberg, E., Ferrer, L., Kajareka, S., Venkataraman, A., and Stocke, A., Modeling prosodic feature sequences for speaker recognition, Speech Communication, pp 455-472, 2005.

[3] Kajarekar, S., Ferrer, L., Sönmez, K., Zheng, J., Shriberg, E., and Stolcke, A., Modeling NERFs For Speaker Recognition, Proc. Odyssey 2004, pp- 51-56, Toledo, Spain, jun 2004.

[4] Baker, B., Sridharan, S., Speaker Verification using Hidden Markov Models in Multilingual Text-constrained Framework, Proc. Odyssey 2006, San juan, Puerto Rico, jun 2006.

[5] Lin, C-Y. and Wang H-C. , Language Identification Using Pitch Contour Information, Proc. ICASSP, Philadelphia, pp 601-604, march 2005.

[6] Grabe, E., Kochanski, G. and Coleman, J., Quantitative modelling of intonational Variation, Proc. Speech Analysis and Recognition in Technology, Linguistics and Medicine", 2003.

[7] Boersma, P. and Weenink, D., Praat: doing phonetics by computer, http://www.praat.org/

[8] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel P., Factor Analysis Simplified, Proc. ICASSP 2005, Philadelphia,pp 637-640, march,2005. Avaible at http://www.crim.ca/perso/patrick.kenny/

[9] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel P., Improvements in Factor Analysis Based Speaker Verification, Proc. ICASSP, Toulouse, France, pp 113-116, may 2006.

[10] Dehak, N., Dumouchel, P. and Kenny, P., Modeling prosodic features with joint factor analysis for speaker verification, submitted to *IEEE Trans. Audio Speech and Language Processing*. Avaible at www.crim.ca/perso/najim.dehak/public/IEEE_TRAN.pdf

[11] Kenny, P. and Yin, S-C., The CRIM system for the 2006 NIST speaker recognition evaluation, The 2006 NIST Speaker Recognition Workshop, San Juan, Porto Rico, jun 2006.

[12] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel P., Joint Factor Analysis versus Eigenchannels in Speaker Recognition, to appear in *IEEE Trans. Audio Speech and Language Processing*, may 2007, http://www.crim.ca/perso/patrick.kenny/.

[13] Pelecanos, J. and Sridharan, S., Feature warping for robust speaker verification, Proc. Speaker Odyssey, Crete, Greece, pp 213-218, jun 2001.

[14] http://www.nist.gov/speech/tests/spk/index.htm.

[15] Mezghani, A., and O'Shaughnessy, D., Speaker Verification Using a New Representation Based on a Combination of MFCC and Formants, IEEE Canadian Conference on Electrical and Computer Engineering, Saskatoon, SK, May 2005.

[16] Tanabian, M-M., Tierney, P., Zahirazami, B., Automatic speaker recognition with formant trajectory tracking using CART and neural networks, IEEE Canadian Conference on Electrical and Computer Engineering, Saskatoon, SK, May 2005.

[17] Lin, C-Y. and Wang H-C., Language Identification Using Pitch Contour Information In The Ergodic Markov Model, Proc. ICASSP, Toulouse France, may 2006.

[18] Martin, A., Doddington, G., Kamm, T., Ordowski, M., Przybocki, M., The DET Curve in Assessment of Detection Task Performance, EUROSPEECH 1997.

[19] http://www.speech.sri.com/projects/verification/sri-sre06-presentation.ppt.