PROSODIC FEATURES AND FORMANT MODELING FOR AN IVECTOR-BASED LANGUAGE RECOGNITION SYSTEM

David Martínez, Eduardo Lleida, Alfonso Ortega, and Antonio Miguel

Aragon Institute for Engineering Research (I3A), University of Zaragoza, Spain

ABSTRACT

The prosody of a language is encoded in syllable length, loudness and pitch. These attributes make humans perceive rhythm, stress and intonation in speech. Depending on the language, these speech properties vary, making language classification possible. On the other hand, formants are the resonance frequencies of the vocal tract, depend heavily on the position adopted by the articulatory organs, and are especially useful to disambiguate vowel sounds. In this paper prosodic and formant information are combined to build a generative language identification system based on Gaussian models fed with iVectors. The system is evaluated on the NIST LRE09 database and the inclusion of formant information gives about 50% relative improvement for the 30 s task over a prosodic system without it. The fusion with a state-of-the-art acoustic system based on shifted delta cepstral coefficients (SDC) shows the complementarity of both approaches.

Index Terms— Language Identification, Prosody, Formants, iVectors, Joint Factor Analysis

1. INTRODUCTION

The three components of prosody are rhythm, stress, and intonation. The prosodic language identification (LID) approach is based on the modeling of this suprasegmental information. Traditionally pitch, energy, and duration of specific speech segments have been used for that goal [1, 2, 3]. One of the first attempts to automatically identify language from prosody is found in [3]. There several techniques are used to model pitch and energy contours, and linear prediction coefficients (LPC) are used to model also formants. In that work formant values are chosen for two reasons: *i)* it is known that human ear and brain make use of formant information to distinguish sounds, and *ii)* additive wideband noise has less effect in the peaks of the spectrum. We also know that, in addition to the fact that in different languages different vowel pronunciations are produced and then formants will be different (or said in a different way, the repertoire of voiced sounds is different for every language), the frequency of formants is heavily dependent on stress, and less (but also) significantly on duration [5]. This second phenomenon happens because of the influence of neighboring segments and is known as undershoot [6]. Since stress and duration behave different in different languages, formants should be a good feature to distinguish them.

It can be found that some authors obtain prosodic features with the help of an automatic speech recognition (ASR) system [2], but this makes the process computationally expensive, and in most works ASRs are avoided. In [1], pitch contours are approximated using Legendre polynomials over long temporal intervals, which seems to be logical and useful for prosody modeling. This approach has also been recently adopted for speaker identification (SID) [7, 9, 10], where pitch and also energy contours are approximated using linear combination of Legendre polynomials over syllable or syllable-like units. The regression coefficients together with durations of corresponding segments are the features describing the three characteristics of prosody. In [8], formants are used additionally to pitch, energy, and duration, and they are modeled in units representing syllables. In that work the use of formant information makes possible to reduce the error function with regard to the prosodic system without formants, but there are no further improvements when fusing a cepstral system with the prosodic system including formants, compared to the case of fusion of the ceptral system with prosodic features without formants.

One of the most popular prosodic approaches for SID was to use a standard JFA model [7, 8, 10]. Recently, the standard iVector approach [13], initially proposed to model MFCC features, was tested on polynomial coefficient prosodic features [11], showing remarkable performance on a SID task, comparable to that obtained using the JFA approach. Note that these approaches are applicable only to features that are always defined and are relatively low-dimensional, like the polynomial coefficient features described above. For more complex sets of features, another subspace modeling technique called the subspace multinomial model (SMM) [12] was introduced, which models the vector of weights from a background Gaussian mixture model (GMM) that takes into account probabilities of undefined values.

The iVector approach with prosodic features has also been tested for LID [4]. In that work, pitch and energy contours are modeled with Legendre polynomials in fixed regions of 200 ms, that approximate syllable-like units. Together with the number of voiced samples used to extract the contours, stress, rhythm and intonation are characterized. The iVector system as the one described in [14] is used as classifier. The prosodic system alone is still far from the state-of-the-art cepstral system of [14], but the fusion gives a promising improvement.

The rest of the paper remains as follows: in Section 2, the relation with previous works and the major contributions of this paper are presented; in Section 3 the prosodic features and formant extraction process is described; in Section 4, the generative Gaussian LID system based on iVectors is revised; in Section 5, the experimental setup and results are shown; in Section 6, conclusions are drawn.

2. RELATION TO PREVIOUS WORK AND MAJOR CONTRIBUTION

In this work a LID system based on prosodic features and formants with an iVector-based classifier is presented. The two most related works are [4], but there no formant information was considered, and [8], where the prosodic features and formant modeling is studied for a SID system. The major contribution presented is the modeling of prosodic features and formants for an iVector-based LID system, and how and which formants are the most useful for LID. It is shown that the existing gap in performance between cepstral and prosodic systems is reduced if formant contour modeling is added to the latter, and that F1 and F2 are the most discriminative formants for LID.

3. FEATURE EXTRACTION

3.1. Pitch, Energy and Formant Extraction

Our features carry information about the evolution of pitch, energy and formant central frequencies along time. To extract them we use The Snack Sound Toolkit [17]. They are obtained every 10 ms with 7.5 ms long windows. First, the pitch, energy and formant values are converted to log domain, to simulate human perception. Next, energy is normalized by subtracting its maximum value in the log scale. This makes it more robust to language-independent phenomena such as channel variations. The log pitch and log formant values are normalized by subtracting mean and dividing by standard deviation estimated over the corresponding file. Thus we avoid the dependence on the absolute pitch value of the speaker. The treatment of formants is similar to [8], but we convert them to log domain and study the influence from F1 to F4.

3.2. Segment Definition

After extracting pitch, energy and formant central frequencies for whole speech recordings, every recording is divided into segments where contours describing those features are created. In [10], different segment definitions were tested and segmentation based on syllables detected using an ASR system was found to perform best. Since the language is unknown in the case of LID, we want to avoid the use of ASR. In [4] it was shown that fixed-length segments was a good choice and gave better performance than segment boundaries determined by energy valleys. There, the signal was split into segments of 200 ms with a 50 ms shift. In the present work we also study 10 ms shifts that make possible to extract more information of the signal, and segments delimited by phoneme boundaries obtained from the BUT Hungarian phoneme recognizer [15] to see if the contours modeled within each phoneme are meaningful and contain more discriminative information than fixed segments.

3.3. Contour Modeling

For each segment, we drop all unvoiced frames for which no pitch was detected. Then pitch, energy and formant central frequencies are approximated by linear combination of Legendre polynomials as

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

where $f(t)$ is the contour being modeled and $P_i(t)$ is the i Legendre polynomial. Each coefficient a_i represents a characteristic of the contour shape: a_0 corresponds to the mean, a_1 to the slope, a_2 to the curvature, and higher order represents more precise detail of the contour. In our implementation, Legendre polynomials of order 5 give six coefficients for pitch, six for energy, and six for every additional formant. In Figure 1 a real F1 curve extracted from a 200 ms segment with 20 voiced frames is compared to its Legendre approximation.

Finally, the number of voiced frames used to calculate the polynomials is included to consider the segment duration. Thus a 13 dimension feature vector is obtained for the case with no formants, up to 37 dimensions if the first 4 formants (F1-F4) are included. Thus, we can consider that our features contain information of the three components of prosody: intonation in pitch, rhythm in duration, and stress in energy and in duration; and formant contours primarily carry information of vowel evolution. These are the features used to build our GMM universal background model (UBM). Supervectors of Baum-Welch statistics can then be estimated for each utterance, as in [13].

4. IVECTOR-BASED CLASSIFIER

4.1. Classifier

Once the iVectors for our training data are obtained, a linear generative classifier is trained as proposed in [14]. The distributions of iVectors for individual languages are modeled by Gaussian distributions with a single within-class (WC) full covariance matrix shared by all the languages.

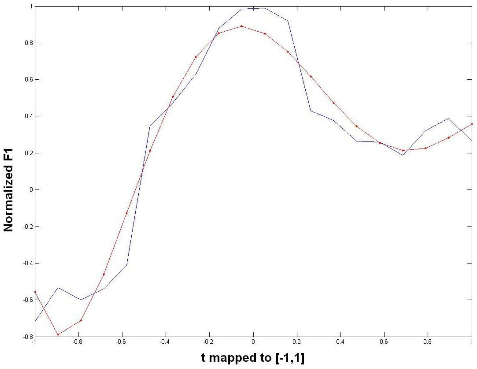

Fig. 1. *Comparison of actual F1 curve (straight blue) with Legendre approximation of order 5 (dotted red).*

For an iVector $\mathbf{w}$ corresponding to a test utterance, the loglikelihood for each language is

$$\ln p(\mathbf{w}|l) = -\frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}+\mathbf{w}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_l-\frac{1}{2}\boldsymbol{\mu}_l^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_l+const,$$

where $\boldsymbol{\mu}_l$ is the mean vector for language l, $\mathbf{\Sigma}$ is the common covariance matrix, and $const$ is a language- and iVector-independent constant irrelevant for making decisions. The quadratic term $\mathbf{w}^T\mathbf{\Sigma}^{-1}\mathbf{w}$, which is constant over classes, would be also irrelevant, if the log-likelihoods were directly used to obtain posterior probabilities of classes. However, since the likelihoods are used only as input features to the calibration backend, it makes a difference in our system, as explained in [14].

4.2. Fusion and Calibration Backend

For calibration, a Gaussian backend followed by discriminative multiclass logistic regression is used to postprocess scores obtained from the described classifiers. Note that the Gaussian backend is essentially the same model as our generative classifier. However, its inputs are the scores from the classifiers described above rather than the iVectors. It is trained on the separate development dataset to obtain well-calibrated scores. When fusing multiple systems, a separate Gaussian backend is trained for each subsystem and outputs of the Gaussian backends are fused by multiclass logistic regression. MultiFocal toolkit has been used to implement the backend [18].

5. EXPERIMENTS AND RESULTS

5.1. Test Data

Our results are reported for the closed-set tasks of 3, 10 and 30 seconds of the NIST LRE09 evaluation [16]. The data comprises 31178 recordings of 23 target languages. Results are reported in terms of optimal C_{avg} (C_{avg*}), which is an error metric defined in [16].

Task	Pitch+Energy+Duration	+F1	+F2	+F3	+F4
3s	32.26	28.68	**26.39**	27.32	28.55
10s	22.23	17.27	**14.57**	16.02	17.64
30s	15.01	9.94	**7.50**	9.11	10.56

Table 1. $C_{avg*} \times 100$ *for the prosodic and formant systems with 200 ms fixed segments and 50 ms shift. In the first column we have 13 dimension features and in every column a formant is added without removing the previous one. Since 6 coefficients per formant are added, in the last column we have 37 dimension features.*

5.2. Training and Development Data

Our training dataset comes from the following databases: CALLFRIEND, NIST LRE03-05-07-11, OHSU, SRE04-06-08 and VOA3. The data comprises 54 languages, which are all used to train our UBM. For training iVector extractor matrices T and the Gaussian classifier, we use data of only the 23 target languages.

A separate dataset was used for training the fusion/calibration backend, which includes data from NIST LRE07-11 and VOA3, not included in the training dataset.

5.3. Results with Prosodic and Formant Features

5.3.1. Influence of the formants

The configuration taken as baseline for the experiments presented in this work comes from the best results obtained in [4], that is, UBM with 2048 components, fixed-length regions of 200 ms and 50 ms shift, and instead of 400 dimension iVectors, 600 dimension iVectors are taken. In table 1, the comparison of the system without formants and with formants F1 to F4 is presented. See how the inclusion of only F1 decreases significantly the error rate, and when adding F1 and F2, the best results are obtained. The further addition of F3 and F4 is not beneficial compared to the case with only F1 and F2. Theoretically, F1 to F3 are primarily acoustic correlates of vowel height, place of articulation and rounding, and are known as vowel formants, because they make possible vowel discrimination, whereas F4 and higher formants are often said to be acoustic correlates of the speakers' vocal tract characteristics [19]. Moreover, F4 and higher formants appear to have little useful perceptual effect when their central frequency is changed [20]. In addition, estimations of F3, and also of higher formants, are often difficult owing to the low energy in their frequency ranges [20], and hence, it is logical that they do not contribute so reliably to the classification. Given these facts, it can be said that F1 and F2 were expected to contribute the most to the discrimination of languages and that is what it happens in our experiments. For the rest of the experiments, our features will include pitch, energy, duration, F1 and F2.

It can be seen in table 1 that by adding F1 and F2 the relative improvement with regard to the case with only pitch, energy, and duration, is of 18.20% for the 3 s condition, of 34.46% for the 10 s condition, and a 50.03% for the 30 s condition.

Task	50 ms	10 ms	phoneme boundaries
3s	26.39	**24.18**	30.92
10s	14.57	**12.48**	19.60
30s	7.50	**5.78**	12.02

Table 2. $C_{avg*} \times 100$ *for the prosodic system adding F1 and F2. Comparison of fixed-length regions of 200 ms with 50 and 10 ms shift, and regions delimited by phonemes.*

5.3.2. Influence of the segment definition

In our baseline, Legendre polynomials are computed in fixed-length windows of 200 ms shifted every 50 ms. In table 2 results can also be seen for a shift of 10 ms and with regions delimited by the boundaries of the phonemes recognized by the BUT Hungarian phoneme recognizer. In the case of 10 ms shift, the goal is to see if more useful information can be extracted from the signal, as it happens with acoustic LID systems trained with SDC. And the reduction in the error metric is important.In the case of using phonemes as delimiters, fewer segments are obtained, mainly because these segments are not overlapped, and the system alone is not expected to be better than the one with fixed-length and overlapped regions. However it is interesting to see that the performance is not dramatically reduced and that due to its different way of delimiting regions, it can help further in a fusion with the other approaches, as it is checked in the next section. The goal is to look for regions more meaningful than the fixed ones, like phonemes, that can be more correlated with the contours modeled on them.

5.4. Fusion with Acoustic iVector-Based System

5.4.1. Acoustic system

A state-of-the-art acoustic system is built in the same fashion as in [14]. It uses the same configuration (SDC 7-1-3-7, 2048 Gaussians, 600-dimension iVectors) and it is trained with the same data as the prosodic and formant system. Therefore, the improvements obtained from fusing both systems can be only attributed to the complementarity of prosodic, formant and cepstral features.

5.4.2. Fusion results

Table 3 shows the results for the state-of-the-art acoustic system, and three fusions with prosodic and formant systems. The first, with the prosodic system without formants, the second with the best prosodic and formant system, that includes F1 and F2 and windows are shifted 10 ms, and the third includes two prosodic and formant systems, the previous one and the one with F1 and F2 and regions delimited by phoneme boundaries. It can be seen that the first two formants do not only increase the system performance of the prosodic system alone, but also give a better fusion with the acoustic system. In addition, the phoneme boundaries help more in short files. In longer files, although phoneme boundaries can give additional information, this is not useful for discriminating among languages any more, because the information gathered by the iVector seems to have the same discriminative effect in the classification. This third fusion gives a 14.92% relative improvement for the 3 s condition, a 17.79 % for the 10 s condition, and a 12.71 % for the 30 s condition, over the acoustic system alone.

Condition	Ac	Ac+PwoF	Ac+PwF	Ac+PwF+PwFph
3 s	14.48	13.50	12.57	**12.32**
10 s	4.44	3.92	3.71	**3.65**
30 s	1.81	1.74	1.59	**1.58**

Table 3. $C_{avg*} \times 100$ *for the generative iVectors-based acoustic system, and 3 different fusions with the prosodic and formant systems.*
Ac: acoustic system
PwoF: prosodic system without formants
PwF: prosodic system with F1 and F2 and 10 ms window shift
PwFwph: prosodic system with F1 and F2 and regions delimited by phoneme boundaries

6. CONCLUSIONS

Prosodic features that model rhythm, stress and intonation are combined with formant information to build a LID system that is tested over the NIST LRE09 database. In order to capture these properties of speech, pitch, energy, and formant central frequencies are extracted from the audio signal and approximated with Legendre polynomials in 200 ms regions. Also the number of voiced frames within this region is used as a feature. The classification is made with a generative iVector-based system followed by a Gaussian backend. Finally, scores are linearly transformed for calibration and fusion. The novelty of the work is the study of formant information and its influence on the results, showing that the most useful formants for LID are F1 and F2. This result is in agreement with the theoretical background that states that F1, F2 and also F3 are correlated with the acoustic articulation of voiced sounds. However, the estimate of F3 by current algorithms is noisy and does not help for this task. It is also shown that the window shift used to extract coefficients is very influential on the results, and a 10 ms shift seems to be a good choice. The gap between acoustic and prosodic systems is significantly reduced with the addition of formant information, and the fusion of both approaches gives further improvements over the acoustic one. Nevertheless, we think that there is still work to be done on prosodic systems that can bring further improvements to this approach of LID.

7. ACKNOWLEDGMENTS

This work extends the one developed during an internship of David Martínez at SRI International. We thank the fruitful discussions held there that have made this paper possible.

This work was funded by the Spanish Government and the European Union (FEDER) under project TIN2011-28169-C05-02.

8. REFERENCES

[1] Chi-Yueh Lin, and Hsiao-Chuan Wang, “Language Identification Using Pitch Contour Information”, *Proc. ICASSP 2005*, Philadelphia.

[2] L. Mary, B. Yegnanarayana, “Extraction and Representation of Prosodic Features for Language and Speaker Recognition”, *Speech Comm.* 50 (2008).

[3] J. T. Foil, “Language Identification Using Noisy Speech”, *Proc. ICASSP 1986*, Tokyo.

[4] D. Martínez, L. Burget, L. Ferrer, and N. Scheffer, “iVector-Based Prosodic System for Language Identification”, *Proc. ICASSP 2012*, Kyoto.

[5] T. Gay, “Effect of Speaking Rate on Vowel Formant Movements”, *The Journal of the Acoustical Society of America*, vol. 63, pp. 223-230.

[6] L. Björn, “ Spectrographic Study of Vowel Reduction”, *The Journal of the Acoustical Society of America*, vol. 35, pp. 1773-1781.

[7] N. Dehak, P. Demouchel, P. Kenny, “Modeling Prosodic Features with Joint Factor Analysis for Speaker Verification”, *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, Sept. 2007.

[8] N. Dehak, P. Kenny, and P. Dumouchel, “Continuous Prosodic Features and Formant Modeling with Joint Factor Analysis for Speaker Verication”, *Proc. Interspeech 2007*, Antwerp.

[9] L.Ferrer, et al., “A Comparison of Approaches for Modeling Prosodic Features in Speaker Recognition”, *Proc. ICASSP 2010*, Dallas.

[10] M. Kockmann, L. Burget, J. Černocký, “Investigations into Prosodic Syllable Contour Features for Speaker Recognition”, *Proc. ICASSP 2010*, Dallas.

[11] M. Kockmann, et al., ”iVector Fusion of Prosodic and Cepstral Features for Speaker Verification”, *Proc. Interspeech 2011*, Florence.

[12] M. Kockmann, et al., “Prosodic Speaker Verification Using Subspace Multinomial Models with Intersession Compensation”, *Proc. Interspeech 2010*, Makuhari.

[13] N. Dehak, et al., “Front-End Factor Analysis for Speaker Verification”, *IEEE Trans. on Audio, Speech and Language Proc.*, vol. 19, May 2011.

[14] D. Martínez, O. Plchot, L. Burget, O. Glembek, P. Matějka, “Language Identification in iVectors Space”, *Proc. Interspeech 2011*, Florence.

[15] P. Schwarz, “Phoneme Recognition Based on Long Temporal Context”, Ph.D. Thesis, Brno University of Technology, 2009.

[16] “The 2009 NIST Language Recognition Evaluation Plan (LRE09)”, http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.

[17] K. Sjölander, “The Snack Sound Toolkit”, http://www.speech.kth.se/snack/.

[18] N. Brümmer, “FoCal Multi-class: Toolkit for Evaluation, Fusion and Calibration of Multi-class Recognition Scores - Tutorial and User Manual”-. http://sites.google.com/site/nikobrummer/focalmulticlass.

[19] K. Stevens, “Acoustic Phonetics”, Cambridge, MA: The MIT Press, 1999.

[20] J. Benesty, et al.,“Springer Handbook of Speech Processing”, Springer Verlag 2008.