Analysis and Selection of Prosodic Features for Language Identification

Conference Paper · January 2010

DOI: 10.1109/IALP.2009.34 · Source: IEEE Xplore

5 authors, including:

Raymond W. M. Ng
The University of Sheffield
32 PUBLICATIONS **122** CITATIONS
SEE PROFILE

Tan Lee
The Chinese University of Hong Kong
172 PUBLICATIONS **1,376** CITATIONS
SEE PROFILE

Cheung-Chi Leung
Institute for Infocomm Research
69 PUBLICATIONS **598** CITATIONS
SEE PROFILE

Bin Ma
Institute for Infocomm Research
214 PUBLICATIONS **2,022** CITATIONS
SEE PROFILE

Analysis and Selection of Prosodic Features for Language Identification

Raymond W. M. Ng*, Tan Lee*, Cheung-Chi Leung†, Bin Ma† and Haizhou Li†
*Department of Electronic Engineering
the Chinese University of Hong Kong, Hong Kong
Email: {wmng,tanlee}@ee.cuhk.edu.hk
†Institute for Infocomm Research, Singapore
Email: {ccleung,mabin,hli}@i2r.a-star.edu.sg

***Abstract*—Prosodic features are relatively simple in their structures and are believed to be effective in some speech recognition tasks. However, this kind of features is subject to undesirable bias factors, such as speaking styles. To cope with this, researchers have suggested various normalization and measure methods to the features, which makes the feature inventory very large. In this paper, we use a mutual information criterion to analyze and select a number of prosody-related features in a language identification (LID) task. Among twelve optimal features, eight of them are elaborated in this paper. The feature analysis metric, z-score, is shown to have a moderate to high correlation with LID accuracies. Feature selection proposed in this paper brings about the best performance among all prosodic LID systems to our knowledge. A further attempt in system fusion shows a 13% relative improvement the prosodic LID system brings to the conventional phonotactic approach to LID.**

***Keywords*-prosody; mutual information; language identification; feature analysis;**

I. Introduction

Prosodic features are the rhythmic and intonational properties in speech, examples are voice fundamental frequency (F0), F0 gradient, intensity and duration. They are relatively simple in structures, and are believed to be effective in some speech recognition tasks. In a perceptual study [1], syllable rhythm is shown to be both necessary and sufficient for discriminations between particular language pairs by human. On the other hand, prosodic features are known to convey various information such as lexical tones, speaking styles, emotional states, etc [2].

The general impression that prosodic features do not help in language identification (LID) is often the consequence of an oversimplified implementation in feature extraction. Muthusamy [3] indicates the feasibility for prosodic features to contribute to LID. It is shown in recent studies [4][5], that prosodic features alone can help in an LID task. With the emphasis of prosodic feature selection, in this paper we will report performance improvements our prosodic LID system attains. Although a prosodic LID system performs worse than the conventional ones using acoustic or phonotactic approaches, we will show that a prosodic LID system can contribute in LID by providing complementary information.

To make prosodic features useful in LID, two aspects need to be considered. First, the irrelevant factors, such as emotions, that are present in the features have to be removed. Second, there is no standard way to extract prosodic features. Along this thought, a feature selection mechanism would be desirable. Shriberg [6] suggested to process the prosodic features with various normalization techniques. A large number of resultant features are then used in support vector machine (SVM) training.

It would be inefficient to explore the use of features by running full classification repeatedly. In this paper, a mutual information based feature analysis and selection frontend is introduced. Such a frontend is application and classifier independent. It selects a concise set of optimal features for further system training in a classification task. In the following, the prosodic features are introduced in Section II. Section III discusses the analysis method. The analysis results, focusing on seven Asian languages, are shown in Section IV. In Section V, two experiments are used to show the LID performance improvements by using the feature analysis and selection method proposed. The contribution of prosodic LID to a conventional phonotactic approach would also be highlighted.

II. Prosodic feature extraction

Because prosodic features are believed to be carried by syllables in speech [4][5], segmentation is first done to obtain syllable-like units called pseudosyllables [4]. Automatic segmentation is implemented by a vocalic nuclei detection algorithm similar to [7]. A sonorant-band intensity contour is extracted, and a window post-processing method is used to locate the contour's local maxima, which are regarded as the nuclei of pseudosyllables. An example of a short utterance is shown in Figure 1. The vertical dotted lines mark six detected pseudosyllabic nuclei, based on which the continuous contours of F0 and intensity are segmented into syllable-level continuous contours, hereinafter referred to as *segment contours*. Extending the segment contour from one pseudosyllable to two gives a *pair contour*. The *triplet* and *utterance contour* cover even longer time span (Figure 2). These contours will be discussed further in Section II-B and II-C about normalization and regression.

A. Frame-based and syllable-based features

These are features directly extracted from the segment contour in each pseudosyllable. F0, intensity and duration constitute the three types of such features, the details are included in Table VII. Five measures of F0 are shown in Figure 1. They are *nucleus F0*, *maximum F0*, *minimum F0*, *F0 span* and *F0 difference*. *Nucleus F0* is a frame-based F0 measurement at the syllable nucleus. *Maximum F0* and *minimum F0* are respectively the 95$^{\text{th}}$-percentile and 5$^{\text{th}}$-percentile values of the F0 segment contour. *F0 span* measures the range of F0 in the segment contour. *F0 difference* is the quotient of *F0 span* divided by the temporal offset of *maximum F0* from *minimum F0*. The same type of measures are also extracted from the intensity contour. As a result, there are five F0 and five intensity (EN) features.

There are three measures of duration: *Nuclei separation* is the separation between two consecutive nuclei. *Syllable width* is the width of a pseudosyllable measured from the intensity contour minimum on the left to that on the right. *Voiced ratio* is the ratio of the segment contour length to *syllable width*. Exceptionally long durations due to utterance breaks are detected by an outlier detection algorithm. In Figure 2, the first syllables of two detected utterances are marked with solid vertical lines, with which an utterance could be clearly defined.

B. Normalization

Raw measurements undergo two normalization methods: *Bias removal* (abbreviated as UB) is done by subtracting the mean. *Z-normalization* (abbreviated as Z) is the division of unbiased measure by the standard deviation. Also, the width of the normalization windows can vary. Three time spans are considered: (1) *Triplet*, covering three consecutive syllables, with the target syllable in the middle (2) *Utterance*(*Utt*) and (3)*File*, the longest available content in the file clip. These time spans can be referred to in Figure 2. Table VIII shows a summary of the normalization methods. For *syllable-based features*, a single value concludes the property of the pseudosyllable as a whole. Normalization over *triplet* is not done because of insufficient data for mean and variance calculations.

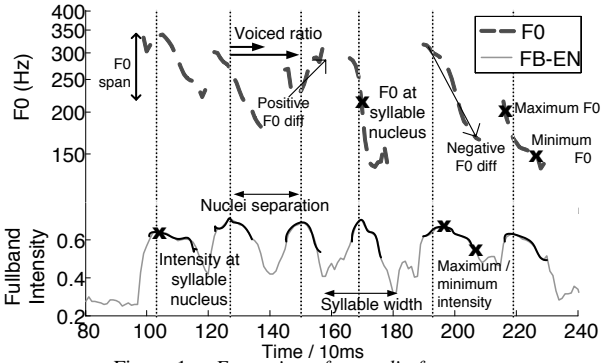

Figure 1. *Extraction of prosodic features*

C. Regression and residual features

F0 gradient is what motivates this kind of features. Lin [8] suggested up to the second-order coefficient from the polynomial regression of F0 contour provides language dependent information of the syllables. Thus in this study, first and second order regression coefficients are calculated from the F0 and from the intensity segment contour. $\boldsymbol{f} = [f_0, f_1, ..., f_{T-1}]$ represents a segment contour $\boldsymbol{f}$ with T frame-based measurements. In the general form, an M^{th}-order regression coefficient (a_M^*) is the highest-order coefficient in the least-square polynomial fit with an M^{th}-order polynomial,

$$a_M^* = \underset{a_M}{\operatorname{argmin}} \left|\left| \boldsymbol{f} - \sum_{m=0}^{M} a_m \boldsymbol{X}_m \right|\right|_2 \quad (1)$$

where $\boldsymbol{X}_m = [0^m, ..., (T-1)^m]$. $\sum_{m=0}^{M} a_m \boldsymbol{X}_m$ is the M^{th}-order regression curve. Motivated by the supra-tone units for tone modeling [9], regression is also performed on *pair contours* of F0 and intensity. Up to the fourth-order regression is done to capture the high order of curvature.

Regressions of *triplet* and *utterance* contours are not intended to model the contour shape. They represent the sentence level intonation, providing another form of normalization to F0 and intensity. *Residual* F0 and intensity are calculated by subtracting the linear regression curve at nucleus from the F0/EN measurements at the same position, representing syllable-level fluctuations around the phrase curve (Figure 2).

D. Feature quantization and combination

With the extraction methods introduced above, there are totally 93 prosodic features. Table VII gives a summary of these features in terms of their types and measures. Table VIII summarizes different normalization methods for *frame-based* and *syllable-based* measures. It is typical to quantize the continuous prosodic features to discrete categories [4][6]. To make the fewest assumptions on the distribution of a prosodic feature, quantization assigns the features into equally populated bins. Thus, the prosodic representation for a syllable is a discrete feature vector with 93 elements. The

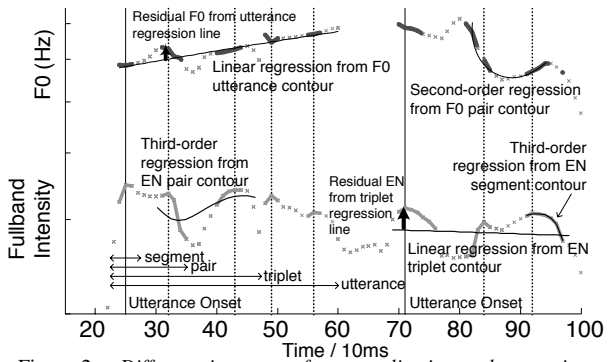

Figure 2. *Different time spans for normalization and regression*

cardinality for every discrete element can vary, depends on the quantization resolution. In this experiment, 3 different resolutions (3, 6 or 9) are tried. A trigram representation is constructed by taking Cartesian products, element-by-element, among the feature vectors of neighbouring syllables.

III. Mutual information criterion

The feature analysis in this paper follows a mutual information approach. The robustness of the method lies on the fact that it measures arbitrary dependencies between the analysis variables, such that it can be applied as a front-end process before classifier training. It is also suitable in classification tasks with complex decision boundaries [10].

Langauge detection is a typical langauge identification (LID) task. Given a segment of speech and a language hypothesis, a binary decision is made on the validity of the hypothesis. Both prosodic feature and language hypothesis validity are discrete, and we can model them with random variables. Consider F denoting the value of a prosodic feature, and let L indicate the validity of the language hypothesis (e.g. true = 1; false = 0). In a large corpus, the class distribution of a feature F is believed to contain information about L. This information amount can be quantified by mutual information with equation (2) [10].

$$I(L;F) = H(L) - H(L|F) \tag{2}$$

$H(L)$ and $H(L|F)$ are entropy terms quantifying uncertainty.

$$H(L) = -\sum_{l=\langle 0,1\rangle} P(l)\log P(l) \tag{3}$$

$$H(L|F) = -\sum_{f=1}^{K_F} P(f)\left(\sum_{l=\langle 0,1\rangle} P(l|f)\log P(l|f)\right) \tag{4}$$

K_F is the number of discrete categories (i.e. the quantization resolution) in a feature.

Among a group of features, an optimal feature F^*, in the sense of highest mutual information $I(L;F)$, is found by:

$$F^* = \underset{F}{\text{argmax}}\ I(L;F) \tag{5}$$

Because $I(L;F)$ have different order of magnitudes and dynamic ranges depending on K_F, Ng [11] proposed a feature comparison metric with the *z-normalized* mutual information:

$$F^* = \underset{F}{\text{argmax}}\ z = \underset{F}{\text{argmax}}\ \frac{I(L;F) - \text{E}_{S\in\mathbf{S}}[I(S;F)]}{\text{STD}_{S\in\mathbf{S}}[I(S;F)]} \tag{6}$$

$\mathbf{S}$ is a set containing a correct guess on the hypothesis validity L as well as many random guess on the validity S ($L \in \mathbf{S}, S \in \mathbf{S}$). $\text{E}_{S\in\mathbf{S}}[I(S;F)]$ and $\text{STD}_{S\in\mathbf{S}}[I(S;F)]$ are the mean and standard deviation respectively. It guarantees high information contents from F^* is only specific to L but not other S.

In mutual information analysis, L corresponds to one syllable when F is a unigram feature, and to three syllables when F is a trigram feature. In the actual task of language recognition, the binary decision on language hypothesis is made to one segment of speech with over a hundred syllables (in the 30-sec test condition). Nevertheless, we will show in the following, the mutual information analysis can help language recognition by selecting optimal features.

IV. Analysis

Seven Asian languages are chosen as the target languages in this study, namely Farsi, Hindi, Japanese, Korean, Mandarin, Tamil and Vietnamese. 30-second utterances from NIST 1996 LRE development and evaluation corpora are studied. In each language, there are 130 long utterances (≈14400 pseudosyllables) for analysis.

The z analysis for every prosodic feature is repeated for different language hypotheses and quantization resolutions. In Section IV-A, an analysis independent with target languages and quantization resolutions will be carried out. In Section IV-B, eight prosodic features are chosen from the 93 prosodic features. Language dependent analysis will be done. An optimal feature gives large value of z. From past experience, we consider $z > 4$ as large.

A. Language and resolution-independent analysis

According to Table VII, the 93 prosodic features are divided into four measures: (I) *frame-based* (II) *syllable-based* and (III) *regression* and (IV) *residual*. Included in Table I and II are the averaged z's over seven target languages in three quantization resolutions (3, 6 and 9) for some of the features. Following the optimality conditions (6), some features win over the others. They will be chosen for further analysis.

Table I
z-score on some frame-based and syllable-based features for F0 and EN

Feature name	Normalization method	z for F0-type	z for EN-type
Nucleus	raw measurements	1.76	1.02
	Z-utterance	4.78	3.27
	Z-triplet	6.56	3.08
	UB-utterance	4.62	4.60
	UB-triplet	6.76	5.00
Difference	raw measurements	5.34	3.09

Table I is about frame-based and syllable-based F0 and EN features. Raw EN is very vulnerable to channel effects and raw F0 reflects individual characteristics. These features have low z scores. Normalization in triplet level generally gives larger z as opposed to that in utterance level. UB-triplet normalization is more suitable than Z-normalization for both nucleus F0 and nucleus EN. These conclusions can be extended to maximum and minimum features as well. For syllable-based features, F0 difference (raw,z=5.34), nuclei separation (UB-file,z=4.00) and voiced ratio (raw,z=4.54) are selected.

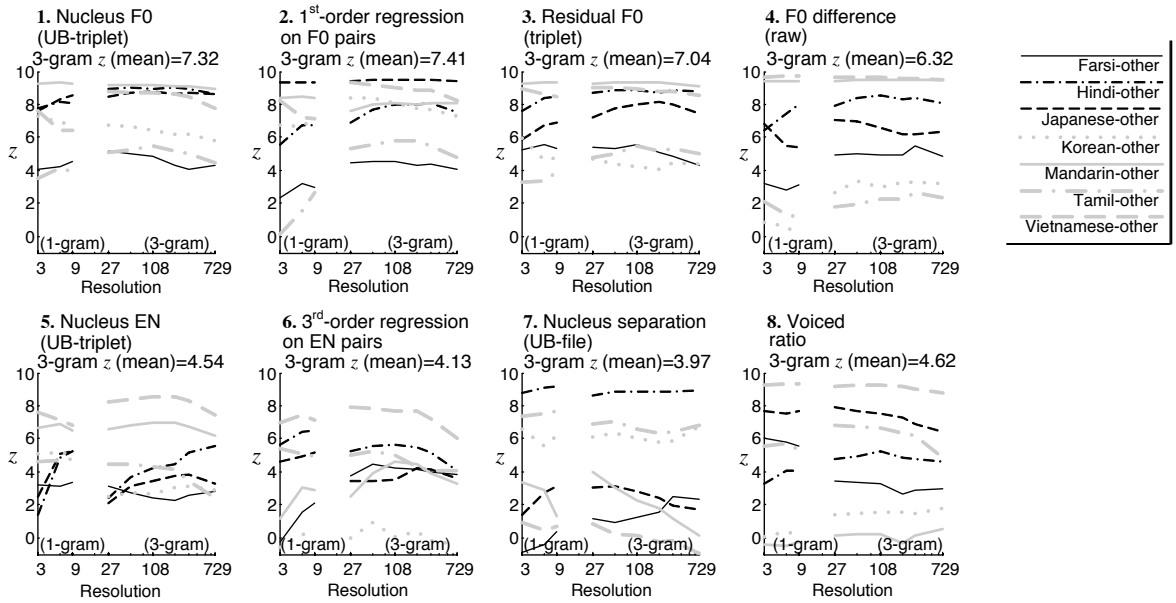

Figure 3. *Plots of z scores for eight selected prosodic features. X-axis: quantization resolution. Y-axis: z scores. The two segments in every plot correspond to unigram and trigram configuration respectively*

Table II
z-score on some regression and residual features

Extraction method	z for F0-type	z for EN-type
1st-order on segment	6.39	3.41
2nd-order on segment	5.10	4.84
1st-order on pair	6.11	4.99
2nd-order on pair	4.89	5.17
3rd-order on pair	4.77	3.84
Residual on triplet	6.66	5.35
Residual on utterance	5.36	4.77

We then look at regression and residual features in Table II. It can be found that for F0 segment contour, linear regression already gives high language distinguishability (z=6.39). For EN segment contour, second-order regression are preferred (z =4.84). This is reasonable as EN segment contours are always a concave curve as opposed to the rather linear F0 segment contour (Figure 1). The same conclusion is arrived for pair contour. First-order regression are chosen for F0 pair contour. For EN pair contour, experience from preliminary studies shows third-order regression should be chosen, despite the relatively low z here.

B. Language-dependent analysis with trigram features

Through the above analysis, 12 optimal features are selected. Due to the identical nature of some features, only eight of these features (Table IX) are included below for trigram analysis. Figure 3 plots the z-scores of these features (**1.**-**8.**) in different resolutions. The averaged z scores are also included in the figure. These scores for trigram features can be interpreted in the same way as in previous tables because z is a universal metric. z values for almost all trigram features are all larger than their unigram counterparts.

When the hypothesis language is Vietnamese, all but one durational trigram feature (**7.**) gives high z-scores. When we look into every feature type, F0 trigram features (**1.**-**4.**) exhibit high z-scores when the hypothesis language is Mandarin, Hindi, Japanese or Vietnamese. Some F0 features (**1.**,**2.**) are helpful to the detection of Korean. Intensity features are fairly important to the language detection of Hindi and Tamil. Nucleus separation trigram (**7.**) is important to Hindi, Tamil and Korean detection, while voiced ratio trigram (**8.**) is important to Vietnamese, Japanese, Tamil and Hindi. The general trend is that F0 features are effective for tone and pitch-accent languages. Further phonetic analysis is necessary for the prosodic characteristics in these target languages, especially for intensity and durational features.

On the effect of quantization resolutions, for trigram features (line segments over 27-729 on the x-axes) the general decreasing trend of z towards the right end indicates sparsity effects. $6\times6\times6$=216 is considered sufficient for trigram configuration. For unigram features (line segments over 3-9 on the x-axes), nine-level quantization resolution is considered appropriate.

V. Experiments

From the proposed 93 features, seven language hypotheses in nine quantization resolutions, there gives 5859 configurations of unigram features. The motivation of the mutual information analysis is to make feature selection possible as a frontend process. To justify this, we perform two sets of experiments. In Section V-A, we choose hundreds of features, calculate their z-scores, perform LID tests and

obtain the Pearson correlation between the two figures. In Section V-B, we deploy the feature selection method and evaluate its effectiveness in three langauge recognition tasks. In all LID tests described below, the bag-of-sounds paradigm in [12] is applied for language classifier training.

A. Correlations between z and LID-results

15 language pairs among English, French, German, Japanese, Mandarin and Spanish are considered. NIST LRE corpora are used. 945 prosodic features (or combinations of them) are studied. For each prosodic feature, a z-score is calculated and an LID test is done using the corresponding feature. Reminded that z analysis is done on NIST LRE 96 data set only, while LID test data comes solely from NIST LRE 03 evaluation data set. Table III shows a summary. Moderate to high correlations are observed for unigram (0.59) and trigram (0.66) features. As a front-end feature selection which only makes use of syllable-level statistics, this correlation provides significant amount of information about the features.

Table III
z-score calculation and LID test

Feature used	Nucleus F0 (Z-file),Nucleus EN (Z-file), 1-deg regression of F0 segment, Nuclei separation (Raw,Z-file), Voiced ratio, Residual F0 over utterance line
Resolutions	(1-gram) 3 ,6 ,9 (3-gram) 27,54,108,216,324,729
Configurations	(1-gram) 315 (3-gram) 630
Correlations	(1-gram) **0.59** (3-gram) **0.66**

B. Applying analysis results in LID test

Finally, we present the results in three LID experiments. The first one is a pairwise language identification using 45-second speech of 6 selected languages from the Oregon Graduate Institute Telephone Speech (OGI-TS) corpus [13]. In the experiment, 50 speakers per language are used for training and 36 speakers for testing. Training and testing set are mutually exclusive in terms of speakers and contents. In Table IV, the LID results from our prosodic LID system are compared with the studies of Rouas [14] and Lin [15] with an identical task. Features after selection are used (Table IX). Both of the quoted researches focus on the explicit use of prosodic features in language identification.

In the second experiment, we extend from the rather small prosodic corpus to a typical language detection task with NIST LRE corpora. NIST LRE 96 development and evaluation sets are used for training and NIST LRE 03 data set is used for testing. To illustrate the significance of feature analysis, we design a control experiment. There are two testing conditions: "no feature analysis" (the control) and "with feature analysis". In the "no feature analysis" condition, features are selected with two rules: (I) use Z-utterance normalization whenever possible; (II) use first order regression for energy pair segment. These are the rules we have previously applied in feature selection. In both conditions, 11 features are used (Table IX). Using the bag-of-sound paradigm in LID[12], the two conditions give rise to vector space models of the same dimension (1737 terms). The results are compared with Mary [5] with an identical task in Table V. It can be seen that feature selection gives a lower equal error rate (EER). Looking into particular language hypothesis, we can see the correspondence between the high z and low EER in Vietnamese. In Section IV, the optimal features are shown to help more in pitch-accent languages (e.g. Japanese), but in fact there is EER reduction in every target language after feature selection.

The last experiment shows the performance improvement that a prosodic system can bring to a phonotactic LID system. NIST LRE 2009 test data is used. It is a large data set with conversational telephone speech and telephone bandwidth broadcast speech. The LID task is also language detection, in total there are 23 language hypotheses. To show the contribution of prosodic features, a linear score fusion between the prosodic LID system (weight=0.1) and a phonotactic system [16] (weight=0.9) is done. The prosodic system uses the feature set "with feature selection" (Table IX). A 13% relative EER reduction from 7.14% to 6.21% is achieved.

Table IV
Pairwise LID identification rates with OGI-TS

	(Rouas[14])	(Lin[15])	Our system "with feature selection"
Identification rate	65.01%	80.13%	85.45%

Table V
Language recognition EER with NIST LRE 03 (30-sec)

Target language	(Mary[5])	"no feature selection"	"with feature selection"
Farsi	n/a	27.50%	25.00%
Japanese	n/a	21.20%	17.40%
Vietnamese	n/a	10.09%	7.50%
7 Asian Languages	n/a	21.24%	18.26%
Full Set*	32.00%	21.83%	19.67%

* Full Set EER is the EER average of 12 language hypotheses in LRE2003

Table VI
Phonotactic and prosodic LID system fusion

	Phonotactic system [16]	Prosodic system	System fusion
EER	7.14%	22.71%	6.21%

VI. Conclusion

Through this paper, it is suggested that a careful selection of features and an appropriate analysis method will make prosodic features more useful than it is generally thought. To our knowledge, the prosodic LID system proposed in this paper achieves the best LID results among all prosodic LID systems. Fusion with phonotactic LID systems is proven successful. Some language dependent analysis shed lights on further studies on the prosodic characteristics in languages. While the analysis are specific to prosodic features in LID,

the paradigm of analysis are general and can be replicated in other classification tasks.

ACKNOWLEDGMENT

This research is partially supported by Earmarked Research Grants (Ref: CUHK 414108, 413507) from the Hong Kong Research Grants Council.

REFERENCES

[1] F. Ramus and J. Mehler, "Language identification with suprasegmental cues: A study based on speech resynthesis," *J. Acoust. Soc. Am.*, vol. 105, no. 1, pp. 512-521, 1999.

[2] H. Fujisaki, "Information, Prosody, and Modeling - with Emphasis on Tonal Features of Speech," in *Proc. Speech Prosody*, 2004, pp. 1-4.

[3] Y.K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 33-41, Oct. 1994.

[4] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904-1911, 2007.

[5] L. Mary, B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker reocgnition," *Speech Commun.*, vol. 50, no. 10, pp. 782-796, 2008.

[6] E. Shriberg et al., "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, vol. 46, no. 3-4, pp. 455-472, 2005.

[7] H.R. Pfitzinger, S. Burger, and S. Heid, "Syllable detection in read and spontaneous speech," in *Proc. ICSLP*, 1996, pp. 1261-1264.

[8] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information," in *Proc. ICASSP*, 2005, pp. 601-604.

[9] T. Lee and Y. Qian, "Tone modeling for speech recognition," *Advances in Chinese Spoken Language Processing*, Eds.: C.-H. Lee et al., World Scientific Publishing, Singapore, 2007.

[10] R. Battiti, "Using Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Trans. Neural Networks*, vol. 5, no. 4, pp. 537-550, 1994.

[11] R.W.M. Ng and T. Lee, "Entropy-based analysis of the prosodic features of Chinese dialects", in *Proc. ISCSLP*, 2008, pp. 65-68.

[12] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 271-284, Jan. 2007.

[13] Y.K. Muthusamy, R.A. Cole, and B.T. Oshika, "The OGI Multilanguage telephone speech corpus," in *Proc. ICSLP*, 1992, pp. 895-898.

[14] J.-L. Rouas et al., "Modeling prosody for language identification on read and spontaneous speech," in *Proc. ICASSP*, 2003, pp. 40-43.

[15] C.-Y. Lin and H.-C. Wang, "Language identification using pitch contour information in the ergodic markov model," in *Proc. ICASSP*, 2006, pp. 193-196.

[16] H. Li et al., "IIR system description for the 2009 NIST Language Recognition Evaluation," Materials from *NIST LRE 2009 workshop*.

APPENDIX

Table VII
Summary to feature extraction and combination

Type	Measure		To be further normalized
F0	Frame-based:	Nucleus	yes
		Maximum	yes
		Minimum	yes
	Syllable-based:	Span	yes
		Difference	yes
	Regression:	1st-order on segment contour	no
		2nd-order on segment contour	no
		1st-order on pair contour	no
		2nd-order on pair contour	no
		3rd-order on pair contour	no
		4th-order on pair contour	no
		1st-order on triplet contour	no
		1st-order on utterance contour	no
	Residual:	over triplet regression line	no
		over utterance regression line	no
EN	Frame-based:	Nucleus	yes
		Maximum	yes
		Minimum	yes
	Syllable-based:	Span	yes
		Difference	yes
	Regression:	1st-order on segment contour	no
		2nd-order on segment contour	no
		1st-order on pair contour	no
		2nd-order on pair contour	no
		3rd-order on pair contour	no
		4th-order on pair contour	no
		1st-order on triplet contour	no
		1st-order on utterance contour	no
	Residual:	over triplet regression line	no
		over utterance regression line	no
Dur	Syllable-based:	Nucleus separation	yes
		Syllable Width	yes
		Voiced ratio	no

Table VIII
Normalization methods of frame-based and syllable-based features

Measure	Method	Window	Measure	Method	Window
Frame-based	Raw	n/a	Syllable-based	Raw	n/a
	Z-normalized(Z)	file		Z-normalized(Z)	file
		utterance			utterance
		triplet			
	Bias removal(UB)	file		Bias removal(UB)	file
		utterance			utterance
		triplet			

Table IX
List of optimal features in analysis, and two feature sets in LID

Type	Measure	for analysis	"no feature selection for LID"	"with feature selection for LID"
F0	Nucleus	**1.**UB-triplet	Z-utt	UB-triplet
	Difference	**4.**Raw	Z-utt	UB-file
	Regression on segment	(not shown)	1st-order	1st-order
	Regression on segment	(not shown)	2nd-order	2nd-order
	Regression on pair	**2.**1st-order	1st-order	1st-order
	Residual feature	**3.**over triplet regression line	over utterance regression line	over triplet regression line
EN	Nucleus	**5.**UB-triplet	Z-utt	UB-triplet
	Difference	(not shown)	Z-utt	UB-file
	Regression on segment	(not shown)	2nd-order	2nd-order
	Regression on pair	**6.**3rd-order	1st-order	3rd-order
DUR	Nucleus separation	**7.**UB-file	Z-utt	UB-file
	Voiced ratio	**8.**raw ratio	(not used)	(not used)