# Analysis and Selection of Prosodic Features for Asian Language Recognition

Raymond W. M. Ng[1], Tan Lee[1],
Cheung-Chi Leung[2], Bin Ma[2], Haizhou Li[2,3]
[1] Dept of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
[2] Institute for Infocomm Research, Singapore
[3] Dept of Computer Science and Statistics, University of Eastern Finland, Finland
[1]{wmng,tanlee}@ee.cuhk.edu.hk, [2]{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 feature attributes in a language identification (LID) task. The feature analysis metric, z-score, is shown to have a moderate to high correlation with LID accuracies. Feature attribute 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.*

## 1 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 (Ramus and Mehler 1999), syllable rhythm is shown to be both necessary and sufficient for the 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 (Fujisaki 2004).

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 (Muthusamy et al. 1994) indicates the feasibility for prosodic features to contribute to LID. It is shown in recent studies (Rouas 2007; Mary and Yegnanarayana, 2008), that prosodic features alone can help in an LID task. With the emphasis of prosodic feature selection, in this paper we will report the 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 (Shriberg et al. 2005) 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 2. Section 3 discusses the analysis method. The analysis results, focusing on seven Asian languages, are shown in Section 4. In Section 6, 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.

## 2 Prosodic Feature Extraction

Because prosodic features are believed to be carried by syllables in speech (Rouas 2007; Mary and Yegnanarayana 2008), segmentation is first done to obtain syllable-like units called pseudosyllables (Rouas 2007). Automatic segmentation is implemented by a vocalic nuclei detection algorithm proposed by Pfitzinger (Pfitzinger et al. 1996). 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 (EN) 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 2.2 and 2.3 about normalization and regression.

### 2.1 Frame-based and Syllable-based Attributes

Frame-based and syllable-based attributes are feature attributes directly extracted from the segment contour in each pseudosyllable. F0, intensity and duration constitute the three types of such attributes, the details are included in Table 1. Five measures of F0 are shown in Figure 1. They are *F0-nucleus*, *F0-maximum*, *F0-minimum*, *F0-span* and *F0-difference*. *F0-nucleus* is a frame-based F0 measurement at the syllable nucleus. *F0-maximum* and *F0-minimum* are respectively the 95th-percentile and 5th-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 *F0-maximum* from *F0-minimum*. The same types of measures are also extracted from the intensity contour. As a result, there are five F0 and five intensity (EN) attributes.

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

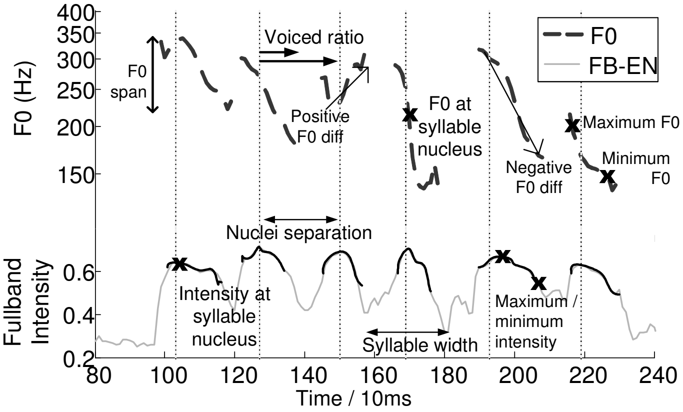

Figure 1. *Extraction of prosodic attributes*

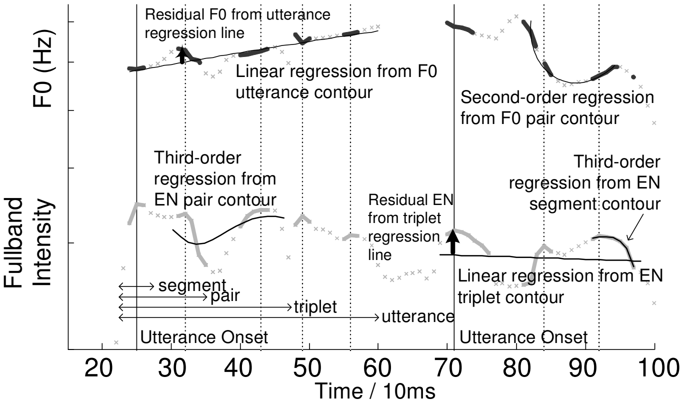

Figure 2. *Different time spans for normalization and regression*

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.

## 2.2 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*, delimited by the utterance breaks mentioned in Section 2.1, with its length vary from a few syllables (Figure 2) to a dozen; and (3) *File*, the longest available content in the file clip. These time spans can be referred to in Figure 2. Table 2 shows a summary of the normalization methods. For *syllable-based attributes*, a single value concludes the property of the pseudosyllable as a whole.

Table 1. *Summary to feature attribute extraction*

| Measure | | Type | Number of measures | Normalization methods | Total num of attributes |
|---|---|---|---|---|---|
| Frame-based | nucleus, max, min | F0 | 3 | 7 | 21 |
| | nucleus, max, min | Intensity (EN) | 3 | 7 | 21 |
| Syllable-based | span, difference | F0 | 2 | 5 | 10 |
| | span, difference | Intensity (EN) | 2 | 5 | 10 |
| | nuclei sep., syllable width | Duration | 2 | 5 | 10 |
| | voiced ratio | Duration | 1 | 1 | 1 |
| Regression | $1^{st}$,$2^{nd}$-order on segment contour | F0 / EN | 4 | 1 | 4 |
| | $1^{st}$-$4^{th}$-order on pair contour | F0 / EN | 8 | 1 | 8 |
| | $1^{st}$-order on triplet contour | F0 / EN | 2 | 1 | 2 |
| | $1^{st}$-order on utterance contour | F0 / EN | 2 | 1 | 2 |
| Residue | over triplet, over utterance | F0 | 2 | 1 | 2 |
| | over triplet, over utterance | Intensity (EN) | 2 | 1 | 2 |
| TOTAL: | | | | | 93 |

Normalization over *triplet* is not done because of insufficient data for mean and variance calculations.

### 2.3 Regression and Residue Attributes

F0 gradient is what motivates regression and residue attributes. Lin (Lin and Wang 2005) 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, the first and second order regression coefficients are calculated from the F0 and from the intensity (EN) segment contour. $\boldsymbol{f} = [f_0, f_1, \ldots, f_{T-1}]$ 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,

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

where $\boldsymbol{x}_m = [0^m, \ldots, (T-1)^{\mathrm{m}}]$. $\sum_{m=0}^{M} a_m \boldsymbol{x}_m$ is the $M$ th-order regression curve. Motivated by the supra-tone units for tone modeling (Lee and Qian 2007), regression is also performed on *pair contours* of F0 and EN. 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 EN. F0 and EN *residue* 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).

### 2.4 Feature Quantization and Combination

With the extraction methods introduced above, there are totally 93 prosodic attributes. Table 1 gives a summary of these attributes in terms of their types and measures. Table 2 summarizes different normalization methods for *frame-based* and *syllable-based* measures. It is typical to quantize the continuous prosodic attributes to discrete categories (Rouas 2007; Shriberg et al. 2005). To make the fewest assumptions on the distribution of a prosodic attribute, quantization assigns the attributes into equally populated bins. Thus, the prosodic

Table 2. *Normalization methods of frame-based and syllable-based attributes*

| Measure | Method | Window | Measure | Method | Window |
|---|---|---|---|---|---|
| Frame-based | Raw | n/a | Syllable-based | Raw | n/a |
| | Z* | file | | Z* | file |
| | | utterance | | | utterance |
| | | triplet | | | |
| | UB* | file | | UB* | file |
| | | utterance | | | utterance |
| | | triplet | | | |

* "Z" stands for Z-normalization, "UB" stands for bias removal

representation for a syllable is a discrete attribute vector with 93 elements. In this experiment, 3 different quantization resolutions (3, 6 or 9) are tested. A trigram representation is also constructed by taking Cartesian products, element-by-element, among the attribute vectors of neighbouring syllables.

## 3 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 (Battiti 1994).

Language detection is a typical language 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 attribute and language hypothesis validity are discrete, and we can model them with random variables. Consider *F* denoting the value of a prosodic attribute, and let *L* indicate the validity of the language hypothesis. In a large corpus, the class distribution of an attribute *F* is believed to contain information about *L*. This information amount can be quantified by mutual information with Eq. (2) (Battiti 1994).

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

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

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

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

*l* takes the value 1 or 0, where the value 1 indicates the validity of a language hypothesis *L* and vice versa for the value 0. *f* is the label of the discrete feature categories of an attribute *F*. For instance, with 9-bin unigram quantization, the value of *f* ranges from 1 to 9. $K_F$ , which is the quantization resolution in the attribute *F*, equals 9. *P*( *l* ) is the probability for a valid (or invalid) language hypothesis. *P*( *l* | *f* ) is the conditional probability given the value *f* of the discrete feature category in the attribute *F*.

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

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

Table 3. *z-score on frame-based unigram attributes with different quantization resolutions*

| Normalization | F0-nucleus | | | | F0-max | F0-min | EN-nucleus | | | | EN-max | EN-min |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | 3-bin | 6-bin | 9-bin | Avg.[#] | 9-bin | 9-bin | 3-bin | 6-bin | 9-bin | Avg.[#] | 9-bin | 9-bin |
| raw | 1.12 | 1.97 | 2.17 | 1.76 | 2.56 | 2.41 | 1.12 | 0.98 | 0.96 | 1.02 | 0.91 | 0.72 |
| Z-file | 3.31 | 3.71 | 4.04 | 3.68 | 4.40 | 3.91 | 3.48 | 3.59 | 3.44 | 3.50 | 3.11 | 3.07 |
| Z-utterance | 4.39 | 4.89 | 5.07 | 4.78 | 5.09 | 5.14 | 3.34 | 3.29 | 3.19 | 3.27 | 2.88 | 3.22 |
| Z-triplet | 6.44 | 6.57 | 6.66 | 6.56 | 6.24 | 6.58 | 2.93 | 3.20 | 3.12 | 3.08 | 2.71 | 3.30 |
| UB-file | 3.60 | 3.93 | 4.26 | 3.93 | 4.65 | 4.06 | 3.85 | 4.26 | 4.19 | 4.10 | 3.89 | 3.19 |
| UB-utterance | 4.32 | 4.73 | 4.82 | 4.62 | 5.26 | 5.23 | 4.24 | 4.73 | 4.83 | 4.60 | 4.60 | 3.37 |
| UB-triplet | 6.72 | 6.77 | 6.79 | 6.76 | 6.93 | 7.11 | 4.44 | 5.29 | 5.28 | 5.00 | 4.89 | 3.75 |

[#] “Avg.” refers to the averaged z-scores including 3-bin, 6-bin and 9-bin quantization (Ng et al. 2009)

[*] Attributes selected to train language recognizer have their *z*-scores underlined

Because *I*(*L*;*F*) has different order of magnitudes and dynamic ranges depending on $K_F$, Ng (Ng and Lee 2008) proposed a feature comparison metric with the *z-normalized* mutual information:

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

**S** is a set containing a correct guess on the hypothesis validity *L* as well as many random guess on the validity *S* ($L \in \mathbf{S}$, $S \in \mathbf{S}$). $\mathrm{E}_{S\in\mathbf{S}}[I(S;F)]$ and $\mathrm{STD}_{S\in\mathbf{S}}[I(S;F)]$ 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 attribute, and to three syllables when *F* is a trigram attribute. In the actual task of language recognition, the binary decision on language hypothesis is made to one segment of speech with over about 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 attributes.

## 4 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 Language Recognition Evaluation (LRE) development and evaluation corpora are studied (NIST 1996). In each language, there are 130 long utterances (~ 14400 pseudosyllables) for analysis.

The *z* analysis for every prosodic attribute is repeated for different language hypotheses and quantization resolutions. In Section 4.1, an analysis to the unigram attributes independent with target languages will be carried out. Some attributes will be selected. Language-dependent mix-bigram analysis will be presented in Section 4.2. Some cross-attribute pairs are also analyzed and will be introduced in Section 4.3. An optimal attribute gives large value of *z*. From past experience, we consider $z > 4$ as large.

### 4.1 Unigram Attributes

According to Table 1, the 93 prosodic attributes are divided into four measures: (I) *frame-based*, (II) *syllable-based*, (III) *regression* and (IV) *residue*. Included in Table 3, 4, 5 are the *z*-scores for different unigram attributes, averaged over seven target languages. Following the optimality conditions (Eq. (6)), some attributes win over the others. They will

Table 4. *z-score on syllable-based attributes with selected quantization resolutions*

| Normalization | F0-span | | F0-difference | | EN-span | | EN-difference | | nuclei sep. | | Syllable width | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| method | 9-bin | Avg.[#] | 9-bin | Avg.[#] | 9-bin | Avg.[#] | 9-bin | Avg.[#] | 9-bin | Avg.[#] | 9-bin | Avg.[#] |
| raw | 3.66 | 3.43 | 5.24 | 5.34 | 1.18 | 1.20 | 3.59 | 3.09 | 4.57 | 4.04 | 3.62 | 3.19 |
| Z-file | 4.67 | 4.44 | 4.36 | 3.84 | 2.37 | 1.99 | 1.28 | 1.03 | 3.86 | 2.96 | 0.93 | 0.75 |
| Z-utterance | 2.47 | 2.56 | 2.27 | 1.84 | 1.27 | 0.94 | 1.32 | 1.12 | 1.46 | 1.34 | 0.66 | 0.91 |
| UB-file | 5.01 | 4.82 | 4.59 | 4.39 | 0.40 | 0.25 | 2.38 | 1.82 | 4.07 | 4.00 | 3.36 | 3.46 |
| UB-utterance | 4.76 | 4.57 | 4.63 | 4.46 | 0.48 | 0.45 | 2.32 | 2.04 | 3.69 | 3.59 | 3.03 | 2.77 |

Table 5. *z-score on regression and residue attributes*

| Regression | on F0-segment | | on F0-pair | | on EN-segment | | on EN-pair | | Residue | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| method | 9-bin | Avg.[#] | 9-bin | Avg.[#] | 9-bin | Avg.[#] | 9-bin | Avg.[#] | method | 9-bin | Avg.[#] |
| 1st-order | 6.89 | 6.39 | 6.35 | 6.11 | 3.90 | 3.41 | 5.15 | 4.99 | over F0-utterance | 5.39 | 5.36 |
| 2nd-order | 5.28 | 5.10 | 5.20 | 4.89 | 4.94 | 4.84 | 5.05 | 5.17 | over F0-triplet | 6.74 | 6.66 |
| 3rd-order | n/a | n/a | 4.98 | 4.77 | n/a | n/a | 4.13 | 3.84 | over EN-utterance | 4.87 | 4.77 |
| 4th-order | n/a | n/a | 5.15 | 4.89 | n/a | n/a | 4.66 | 4.24 | over EN-triplet | 5.56 | 5.34 |

[#] “Avg.” refers to the averaged z-scores including 3-bin, 6-bin and 9-bin quantization (Ng et al. 2009)

[*] Attributes selected to train language recognizer have their *z*-scores underlined

be chosen for the language recognition experiment in Section 6. For readers' comprehension, these chosen attributes will have their *z*-scores underlined from Table 3 to Table 7.

### 4.1.1 Resolution

The *z*-scores of unigram attributes with 9-bin quantization are consistently higher than the “Avg.” columns, which are the averaged *z*-scores including 3-bin, 6-bin and 9-bin quantization (Ng and Lee 2009). Therefore, 9-bin quantization will be done to all unigram attributes.

### 4.1.2 Frame-based Attributes

Table 3 is about frame-based F0 and EN attributes. Raw EN is very vulnerable to channel effects and raw F0 reflects individual characteristics. These attributes have low $z$ scores. Normalization in triplet level generally gives larger $z$ as opposed to those in utterance level and file level. UB-triplet normalization is more suitable than Z-normalization for both F0-nucleus and EN-nucleus. These conclusions can be extended to maximum and minimum attributes as well. F0-maximum, F0-minimum, EN-maximum will be chosen in the subsequent training of language recognizer. EN-minimum will not be chosen as it is vulnerable to various acoustic noise ($z$ =3.75).

### 4.1.3 Syllable-based Attributes

Table 4 compares three pairs of similar syllable-based attributes, and their normalization methods. F0-span is preferred to F0-difference. EN-difference is preferred to EN-span. Nuclei separation is preferred to syllable width. UB-file is found to be the most suitable normalization method for all chosen attributes.

### 4.1.4 Regression and Residue Attributes

In Table 5, it can be found that for F0 segment and pair contour, linear regression already gives high language distinguishability ($z$ =6.89, 6.35). For EN segment contour, second-order regression is preferred ($z$ =4.94). This is reasonable as an EN segment contour is always a

Table 6. *z-score on frame-based, regression and residue trigram attributes with different target languages*

| Attribute name | | Target language | | | | | | | Average over 7 languages |
|---|---|---|---|---|---|---|---|---|---|
| | | Farsi | Hindi | Japanese | Korean | Mandarin | Tamil | Vietnam. | |
| F0 | nucleus (UB-triplet) | 5.07 | 8.91 | 8.93 | 6.15 | 9.20 | 5.06 | 8.20 | 7.36 |
| | regression (1st-order on segment) | 6.80 | 7.18 | 8.59 | 5.44 | 9.16 | 3.56 | 9.38 | 7.16 |
| | regression (2nd-order on segment) | 5.32 | 1.83 | 6.25 | 4.29 | 5.80 | 3.04 | 9.37 | 5.13 |
| | regression (1st-order on pair) | 4.30 | 8.12 | 9.57 | 7.63 | 7.96 | 3.64 | 8.29 | 7.07 |
| | residue (over triplet) | 6.02 | 8.82 | 8.12 | 3.82 | 9.31 | 4.56 | 8.70 | 7.05 |
| EN | nucleus (UB-triplet) | 2.91 | 4.51 | 4.40 | 3.61 | 6.76 | 4.27 | 8.31 | 4.97 |
| | regression (2nd-order on segment) | 5.77 | 1.31 | 6.05 | 1.25 | 2.37 | 6.29 | 8.61 | 4.52 |
| | regression (1st-order on pair) | 3.68 | 2.31 | 5.62 | 7.68 | 6.82 | 6.06 | 8.31 | 5.78 |
| | residue (over triplet) | 4.53 | 3.20 | 4.36 | 4.66 | 7.91 | 4.42 | 8.28 | 5.34 |

Table 7. *z-score on syllable-based trigram attributes with different target languages*

| Attribute name | | Target language | | | | | | | Average over 7 languages |
|---|---|---|---|---|---|---|---|---|---|
| | | Farsi | Hindi | Japanese | Korean | Mandarin | Tamil | Vietnam. | |
| F0 | span (UB-file) | 2.58 | 6.53 | 8.10 | 0.83 | 9.09 | 0.04 | 8.38 | 5.08 |
| | difference (UB-file) | 2.57 | 8.69 | 4.69 | 1.72 | 9.40 | 1.28 | 9.30 | 5.38 |
| EN | span (UB-file) | 0.56 | 0.30 | 1.89 | -0.23 | -0.17 | -0.01 | 1.20 | 0.51 |
| | difference (UB-file) | -0.09 | 0.72 | 3.02 | 4.35 | 3.45 | 0.13 | 3.96 | 2.22 |
| DUR | nuclei separation (UB-file) | 0.63 | 8.97 | 2.88 | 5.97 | 1.92 | 6.89 | 0.62 | 3.98 |
| | Syllable width (UB-file) | 2.76 | 7.35 | 5.40 | 8.27 | 0.27 | 0.12 | 1.41 | 3.65 |

$^*$ Attributes selected to train language recognizer have their *z*-scores underlined

concave curve as opposed to the rather linear F0 segment contour (Figure 1). Residual attributes over triplet is preferred to that over utterance.

After the above analysis, eight non-syllable-based attributes are selected to be analyzed in trigram forms. They are F0-nucleus (UB-triplet), EN-nucleus (UB-triplet), first-order regressions on F0 segment contour, on F0 pair contour and on EN pair contour, second-order regression on EN pair contour, residuals over F0 and EN triplet. For syllable-based attributes, all the six attributes in Table 4, with UB-file normalization, will be considered in trigram forms.

### 4.2 Trigram Attributes

There is a trade-off between high quantization resolutions and avoiding data sparseness in a fine quantization scheme. In this experiment, trigram attributes are constructed from 6-bin quantized unigram attributes. The *mixture-of-bigrams* approach is adopted to model a trigram with the combination of ($N$,$N$–1),($N$,$N$–2) and ($N$–1,$N$–2) bigrams (Lin and Wang 2006). The *z*-scores for three mixtures of bigrams are averaged as an evaluation metric to the trigram.

The *z*-scores of the non-syllable-based attributes in trigram forms are given in Table 6. Generally, F0 related attributes are more reliable than EN related ones. Almost all F0 attributes have *z*-scores greater than 8 when the hypothesis language is Mandarin, Hindi, Japanese or Vietnamese. Vietnamese even gives *z*-scores greater than 8 for all non-syllable-based attributes. Given the reliability reflected in the first-order F0-regression on segment contour ($z$ =7.16), its second-order counterpart (F0-regression, 2nd-order on segment, $z$=5.13) is also chosen for training language recognizer. Among the non-syllable-based EN attributes, less reliable attributes with *z*-scores smaller than 4 can be

Table 8. *List of unigram / trigram for language recognition*

| Type | Measure | Intuitive normalization/ regression | Unigram | Trigram | Optimal normalization/ regression | Unigram | Trigram |
|---|---|---|---|---|---|---|---|
| F0 | nucleus | Z-utterance | √ | √ | UB-triplet | √ | √ |
| | maximum | Z-utterance | √ | | UB-triplet | √ | |
| | minimum | Z-utterance | √ | | UB-triplet | √ | |
| | span | Z-utterance | √ | | UB-file | √ | |
| | difference | Z-utterance | √ | √ | UB-file | √ | √ |
| | regression on segment | 1st-order | √ | √ | 1st-order | √ | √ |
| | regression on segment | 2nd-order | √ | √ | 2nd-order | √ | √ |
| | regression on pair | 1st-order | √ | √ | 1st-order | √ | √ |
| | regression on pair | 2nd-order | √ | | 2nd-order | √ | |
| | residue | over utterance | √ | √ | over triplet | √ | √ |
| EN | nucleus | Z-utterance | √ | √ | UB-triplet | √ | √ |
| | maximum | Z-utterance | √ | | UB-triplet | √ | |
| | difference | Z-utterance | √ | √ | UB-file | √ | √ |
| | regression on segment | 2nd-order | √ | √ | 2nd-order | √ | √ |
| | regression on pair | 1st-order | √ | √ | 1st-order | √ | √ |
| | residue | over utterance | √ | √ | over triplet | √ | √ |
| DUR | nuclei separation | Z-utterance | √ | √ | UB-file | √ | √ |

Table 9. *List of cross-attribute pairs*

| Type | Measure | Intuitive pairs | | Optimal pairs | |
|---|---|---|---|---|---|
| | | Attribute 1 | Attribute 2 | Attribute 1 | Attribute 2 |
| F0 | nucleus | Z-utterance | Z-triplet | Z-file | UB-file |
| | maximum | Z-utterance | Z-triplet | Z-file | UB-file |
| | minimum | Z-utterance | Z-triplet | Z-file | UB-file |
| | span | Z-file | Z-utterance | UB-file | Difference (UB-file) |
| | difference | Z-file | Z-utterance | Raw | Z-file |
| | regression | 1st-order on segment | 2nd-order on segment | 1st-order on segment | F0-difference (Z-file) |
| | regression | 1st-order on segment | 1st-order on pair | 1st-order on segment | 1st-order on pair |
| | regression & residue | 1st-order regression on triplet | residue over utterance | 1st-order regression on triplet | residue over triplet |
| EN | nucleus | Z-utterance | Z-triplet | Z-file | UB-file |
| | maximum | Z-utterance | Z-triplet | Z-file | UB-file |
| | regression | 2nd-order on segment | 1st-order on pair | 2nd-order on segment | 3rd-order on pair |
| | regression & residue | 1st-order regression on triplet | residue over utterance | 1st-order regression on triplet | residue over triplet |
| DUR | nuclei separation | Z-file | Z-utterance | Raw | Z-file |
| | syllable width | Z-file | Z-utterance | Raw | Z-file |

found sporadically. The detection of Hindi by non-syllable-based EN attributes, on the average, is the least reliable.

Three pairs of unigram syllable-based attributes are compared in Section 4.1. The same comparison is performed for trigrams. Consistent with the unigram analysis results, EN-difference and nuclei separation are preferred to their similar counterparts. On the other hand, F0-difference is preferred to F0-span in trigram forms. It is worth noticing that these pairs of similar attributes have different reliability in different target languages. Using $z$>4 as a threshold, reliable syllable-based trigram attributes include F0-difference for detecting Hindi, Japanese, Mandarin and Vietnamese; EN-difference for detecting Korean; and nuclei separation for detecting Hindi, Korean and Tamil.

Table 10. *z-score calculation and LID test*

| Item | Information |
|---|---|
| Feature used | F0-nucleus (Z-file)<br>EN-nucleus (Z-file)<br>1st-order regression of F0 segment<br>Nuclei separation (Raw)<br>Nuclei separation (Z-file)<br>Voiced ratio<br>F0-residue over utterance line |
| Resolutions | (1-gram) 3, 6, 9<br>(3-gram) 27 [3x3x3][#] , 54 [3x6x3][#], 108 [3x6x6][#], 216 [6x6x6][#], 324 [6x9x6][#], 729 [9x9x9][#] |
| Configurations | (1-gram) 315 (3-gram) 630 |
| | (1-gram) 0.59 (3-gram) 0.66 |

[#] Numbers in square brackets [ ] represents the quantization resolutions of the three segments in a trigram

A list of selected unigram and trigram attributes are summarized in Table 8. Feature attribute selection returns an optimal normalization or regression method for different attributes. UB-triplet and UB-file are the best normalization methods for frame-based and syllable-based attributes respectively. F0-regression should be in low order. Residual attributes should be extracted over triplet contours. This set of optimal attributes will be compared with a intuitive attribute set where attributes are selected with two rules: (I) use Z-utterance normalization whenever possible; (II) use residual attribute over utterance contours. These are the rules we have previously applied without the knowledge brought by feature attribute analysis in this paper.

The optimal regression order for EN pair contours is 1st-order, which is not consistent with that for segment contours (2nd-order). It does not either fit in with the intuition that EN contours have higher order of curvature mentioned in Section 4.1. Regression both in low and high orders for EN pair contours will be tested in the latter experiments.

### 4.3 Cross-attribute Pairs

Apart from the unigram and trigram attributes, cross-attribute pairs are analyzed. $93^2$ prosodic attribute pairs are constructed by the Cartesian products of any two prosodic attributes with 6-bin quantization, and they are evaluated as if bigram attributes with the optimality conditions specified in Eq. (6). Analysis results show that all reliable pairs fall within the same type. Most of the reliable pairs are within the same measure, only differ by normalization or regression methods. 14 cross-attribute pairs are selected to train language recognizer and they are listed in Table 9. A corresponding set of intuitive cross-attribute pairs is also constructed, which are to be used as a comparison baseline.

## 5 Correlations to Language Identification

The motivation of the mutual information analysis is to make feature selection possible as a frontend process. To justify this, a separate experiment will be performed to obtain the correlation statistics between the $z$ evaluation metric and the LID performance. 15 language pairs among English, French, German, Japanese, Mandarin and Spanish are considered. NIST Language Recognition Evaluation (LRE) 1996 and 2003 corpora are used (NIST 1996; NIST 2003). As shown in Table 10, 315 unigram and 630 trigram prosodic features are

Table 11. *Language recognition EER with NIST LRE 2003 (30-seconds)*

| Target language | (Mary et al. 2008) | ITU-1 | OPT-1 | OPT-1EN | ITU-2 | OPT-2 |
|---|---|---|---|---|---|---|
| Farsi | n/a | 27.5% | 33.8% | 30.0% | 30.0% | 26.5% |
| Hindi | n/a | 28.8% | 26.3% | 25.1% | 26.3% | 21.3% |
| Japanese | n/a | 26.9% | 21.3% | 20.1% | 22.5% | 16.9% |
| Korean | n/a | 26.2% | 26.5% | 26.3% | 24.9% | 21.0% |
| Mandarin | n/a | 17.5% | 14.8% | 13.8% | 16.3% | 14.9% |
| Tamil | n/a | 25.0% | 22.5% | 23.8% | 22.6% | 21.6% |
| Vietnamese | n/a | 11.3% | 8.8% | 10.0% | 11.3% | 7.5% |
| 7 Asian Languages | n/a | 23.3% | 22.0% | 21.3% | 22.0% | 18.5% |
| Full Set* | 32.0% | 24.1% | 23.0% | 22.8% | 22.6% | 20.1% |

*Full Set EER is the EER average of 12 language hypotheses in NIST LRE 2003

Table 12. *Pairwise LID with OGI-TS: identification rates*

| | (Rouas et al. 2003) | (Lin and Wang 2006) | Our system with feature selection |
|---|---|---|---|
| Identification | 65.01% | 80.13% | 85.45% |

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 1996 data set only, while LID test data comes solely from NIST LRE 2003 evaluation data set. Table 10 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.

## 6 Experiments

### 6.1 Pairwise Language Identification

From this section, we present the results of language identification (LID) experiments. In all LID tests described below, the *bag-of-sounds* paradigm is applied for language classifier training (Li et al. 2007). 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 (Muthusamy et al. 1992). 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 12, the LID results from our prosodic LID system are compared with the studies of Rouas (Rouas et al. 2003) and Lin (Lin and Wang 2006) with an identical task. Features after selection are used (Optimal features in Table 8 and Table 9). Both of the quoted researches focus on the explicit use of prosodic features in language identification.

### 6.2 NIST Language Recognition

In the second experiment, we extend from the rather small prosodic corpus to a typical language detection task with NIST LRE corpora. NIST LRE 1996 development and evaluation sets are used for training and NIST LRE 2003 data set is used for testing (NIST 1996; NIST 2003). To illustrate the significance of feature analysis, we compare across five testing conditions:

- **ITU-1(Intuitive parallel set)**: Intuitive unigram and trigram attributes from Table 8;

- **ITU-2(Intuitive cross-attribute set)**: ITU-1 plus the intuitive cross-attribute pairs from Table 9;
- **OPT-1(Optimal parallel set)**: Optimal unigram and trigram attributes from Table 8;
- **OPT-1EN(Optimal parallel set, high order EN-regression)**: OPT-1 with EN-regression on pair contour changed to 3rd-order regression;
- **OPT-2(Optimal cross-attribute set)**: OPT-1EN plus the optimal cross-attriubte pairs from Table 8.

Using the *bag-of-sound* paradigm in LID (Li et al. 2007), the five conditions give rise to vector space models of the similar dimensions (1449 terms for the conditions ITU-1, OPT-1, OPT-1EN and 1953 terms for condition ITU-2 and OPT-2). The results are compared with Mary (Mary 2008) with an identical task in Table 11.

It can be seen that all the five conditions give a lower equal error rate (EER) compared with the previously reported result (Mary 2008). Selection in unigram and trigram attributes (OPT-1) brings 5.6% relative EER reduction over the intuitive set (ITU-1) from 23.3% to 22.0%. Section 4.2 talks about the unclear reliable patterns of EN-regression. OPT-1EN uses 3rd-order regression on EN pair contours as opposed to 1st-order suggested by the *z*-scores. A further 3% relative EER reduction is obtained. Finally, when we consider cross-attribute pairs as well, from ITU-2 to OPT-2 we can see a 16.0% relative EER reduction brought by feature attribute selection.

Looking into particular language hypotheses in Table 6 and Table 11, we can see the correspondence between the high *z* and low EER in Vietnamese. The attribute set OPT-1, OPT-1EN and OPT-2 are not language-specific, thus the OPT sets may not be truly optimal for the detection of every target language. Detection of Farsi does not benefit much after feature attribute selection.

### 6.3 Fusion with a Phonotactic System

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 (Li et al. 2009) (weight=0.9) is done. The prosodic system uses the feature set after feature selection (Optimal features in Table 8 and Table 9). A 13% relative EER reduction from 7.14% to 6.21% is achieved.

Table 13. *Phonotactic and prosodic LID system fusion*

| | Phonotactic system (Li et al. 2009) | Prosodic system | System fusion |
|---|---|---|---|
| EER | 7.14% | 22.71% | 6.21% |

## 7 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 to be successful. Some language dependent analysis shed lights on further studies on the prosodic characteristics in languages. While the analysis is specific to prosodic features in LID, the paradigm of analysis is general and can be replicated in other classification tasks.

## 8 Acknowledgement

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

## 9 References

The 1996 NIST Language Recognition Evaluation Plan. [Online]. Available: http://www.itl.nist.gov/iad/mig/tests/lre/1996/LRE96EvalPlan.pdf

The 2003 NIST Language Recognition Evaluation Plan. [Online]. Available : http://www.itl nist.gov/iad/mig/tests/lre/2003/LRE03EvalPlan-v1.pdf

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

Fujisaki, H., Information, Prosody and Modeling – with Emphasis on Tonal Features of Speech. In: Proc. Speech Prosody, 2004, pp. 1-4.

Li, H. et al., IIR System Description for the 2009 NIST Language Recognition Evaluation. Materials from NIST LRE 2009 workshop.

Li, H., Ma, B. and Lee C.-H., 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.

Lin, C.-Y. and Wang, H.-C., Language Identification Using Pitch Contour Information in the Ergodic Markov Model. In: Proc. ICASSP, 2006, pp. 193-196.

Lin, C.-Y. and Wang, H-C., Language Identification Using Pitch Contour Information. In: Proc. ICASSP, 2005, pp. 601-604.

Lee, T. and Qian, Y., Tone Modeling for Speech Recognition. In: Advances in Chinese Spoken Language Processing, Eds.: Lee, C.-H. et al., World Scientific Publishing, Singapore, 2007.

Mary, L. and Yegnanarayana, B., Extraction and Representation of Prosodic Features for Language and Speaker Recognition. Speech Commun., vol. 50, no. 10, pp. 782-796, 2008.

Muthusamy, Y.K., Barnard, E. and Cole, R.A., Reviewing Automatic Language Identification. IEEE Signal Process. Mag., vol. 11, no. 4, pp. 31-41, Oct. 1994.

Muthusamy, Y.K., Cole, R.A. and Oshika B.T., The OGI Multilanguage Telephone Speech Corpus. In: Proc. ICSLP, 1992, pp. 895-898.

Ng, R.W.M., Lee, T., Leung, C.-C., Ma, B. and Li, H., Analysis and Selection of Prosodic Features for Language Identification. In: Proc. IALP, 2009, pp. 123-128.

Ng, R.W.M. and Lee, T., Entropy-based Analysis of the Prosodic Features of Chinese Dialects. In: Proc. ISCSLP, 2008, pp. 65-68.

Pfitzinger, H.R., Burger, S. and Heid S., Syllable Detection in Read and Spontaneous Speech. In: Proc. ICSLP, 1996, pp. 1261-1264.

Ramus, F. and Mehler, J., Language Identification with Suprasegmental Cues: A Study Based on Speech Resynthesis. J. Acoust. Soc. Am., vol. 105, no. 1, 512-521, 1999.

Rouas, J.-L., Automatic Prosodic Variations Modeling for Language and Dialect Discrimination. IEEE Trans. Audio, Speech, and Language Processing, vol. 15, no. 6, pp. 1904-1911, 2007.

Rouas, J.-L. et al., Modeling Prosody for Language Identificatioin on Read and Spontaneous Speech. In: Proc. ICASSP, 2003, pp. 40-43.

Shriberg, E. et al., Modeling Prosodic Feature Sequences for Speaker Recognition. Speech Commun., vol. 46, no. 3-4, pp. 455-472, 2005.