ENTROPY-BASED ANALYSIS OF THE PROSODIC FEATURES OF CHINESE DIALECTS

Raymond W. M. Ng and Tan Lee

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong

{wmng, tanlee}@ee.cuhk.edu.hk

ABSTRACT
In this paper, a novel approach is proposed to analyze prosodic features of four Chinese dialects: Wu, Cantonese, Min and Mandarin. The ultimate goal is to exploit these features in the task of automatic spoken language identification. Two entropy-based evaluation metrics are formulated to address the problems of data sparseness and lack of speakers. Different prosody-related acoustic features and their combinations are evaluated. F0, F0 gradient and intensity are found to contain the most language-related information. Maximum language-related information are observed in multi-dimensional N-gram features with F0, F0 gradient and syllable position in sentence. There are also some uncertain results that reveal the limitations of the proposed metrics.

Index Terms— Chinese dialect identification, prosody, Cantonese, Mandarin, Min, Wu

1. INTRODUCTION
Automatic spoken language identification (LID) is the process of automatically determining the language of spoken utterances. Different languages have different phonetic features, such as distribution of vowels. Languages also differ in terms of their phonotactics, which are rules governing the sequence of allowable speech units [1]. An LID system exploits these features in distinguishing languages. Chinese dialect recognition has not received much attention until recently. In NIST LRE-2007, the language detection tasks were expanded. Four Chinese dialects, Wu (吳), Cantonese (粵), Min (閩), and Mandarin (官話), were targeted in a Chinese Language Recognition task.

Chinese being a tonal language, its dialects differ not only in terms of phonetic properties, but also in prosodic properties. Prosodic features have been less explored in LID. They include the rhythmic and intonational properties in speech. Voice fundamental frequency (F0), intensity and duration are the basic acoustic features that determine speech prosody. The use of such features in LID of six languages (including Mandarin) was reported in a study by Rouas [2]. Prosodic features were also tested and confirmed in a Chinese dialect identification task in [3]. While there are sufficient reasons of using prosodic features in LID, there were few in-depth analyses comparing different prosodic features in the task.

In this paper, the language-related information of different prosodic feature parameters and their combinations will be studied. To fit into the general notation of LID, ‘languages’ and ‘dialects’ are used interchangeably to refer to the four target dialects in this paper.

2. SPEECH DATA AND PROSODIC FEATURES

2.1. Speech Corpus
Xiandai Hanyu fanyan yinku (現代漢語方言音庫) [4] is used. The corpus includes speech data of 46 speakers from 40 Chinese cities. There are four types of speech for each city, including about 900 isolated phones, 180 words, 55 sentences, and a discourse of one to two minutes. In this study, nine speakers from eight cities in four dialects - Wu (吳), Cantonese (粵), Min (閩), and Mandarin (官話) are chosen for analysis. For each chosen city, the 55 sentences are analyzed. Each sentence is approximately 3 seconds long. Details of the experimental data can be referred to in Table 1.

In the experiment, segmentation is conducted to give syllable-like discrete units called pseudosyllables. The temporal locations of pseudosyllabic nuclei are also determined. Interested reader is referred to [5] for syllabification methods and [2] for the concept of pseudosyllables. Pseudosyllables are the basic units from which prosodic feature parameters are extracted and evaluated in this study.

2.2. Feature extraction
Table 2 gives a summary of the 27 features being investigated. They include four types of features, prefixed A-D. For each type A feature, which is one-dimensional, scalar quantization is performed on the corresponding prosodic parameter. Type B-D features are derived by combining two or more independently quantized parameters. For example, the three quintile-quantized features A2, A3 and A5 are combined to give C1 which consists of 5 × 5 × 5 = 125 feature classes.

A1-A8 are one-dimensional prosodic features over a single pseudosyllable. F0 (A1) and intensity (A4) are obtained from frame based measurements at detected pseudosyllabic nuclei. F0 gradient (A3) is calculated from an automatically extracted syllable-wise F0 contour. Inter-pseudosyllabic duration (A6) is the temporal interval length between two consecutive pseudosyllabic nuclei. Exceptionally long durations
due to utterance breaks are detected and removed by an outlier detection algorithm. Voiced ratio (A8) is the ratio of the temporal length of syllable-wise F0 contour to the inter-pseudosyllabic duration (A6). F0, intensity and duration features are normalized over all measurements from each speaker to give A2, A5 and A7 respectively. A particular feature \( f \) is normalized by first subtracting the speaker mean \( \bar{f} \) from the measured \( f \), then dividing it by the difference of the upper and lower quartiles of that feature. Features A9-A13 incorporate some sentence-level measurements. Sentence-level F0 gradient (A9) is obtained from a phrase F0 curve generated by linear regression. Point residual F0 (A10) is calculated by subtracting the phrase F0 curve at nucleus from the F0 measurement at the same position. Average residual F0 (A11) is similar to A10, but the difference over the whole pseudosyllable span is considered. Sentence length (A12) is the number of detected pseudosyllabic nuclei in a sentence. Pseudosyllable position (A13) ranges from 0 to 1 and denotes a pseudosyllable’s relative position in a sentence. A9 and A12 are sentence-level measurements and the values are duplicated at all pseudosyllables within the sentence.

Features B1-B8 are \( N \)-gram prosodic features of F0 gradient and three normalized type A features: F0, intensity and duration. In Table 2, \( A_{2n} \) denotes the feature value for the present pseudosyllable, \( A_{2n-1} \) and \( A_{2n+1} \) are values for the pseudosyllables on the left and right respectively. Features C1-C3 are multi-dimensional prosodic features. From our analysis of uni-gram features (see Section 4.1), F0 gradient (A3) contains the most language-related information. It is chosen to combine with F0 (A2), intensity (A5), sentence gradient (A9), residual F0 (A10) and pseudosyllabic position (A13) to form C1-C3. D1-D3 are multi-dimensional \( N \)-gram prosodic features, which is an extension to type C features where multi-dimensional features up to bi-gram are used.

### 3. CRITERIA OF FEATURE EVALUATION

#### 3.1. Mutual Information

We use information theories to evaluate and compare the effectiveness of the feature sets. Consider two discrete random variables \( F \) and \( L \). \( F \) refers to the feature class identity of a pseudosyllable (or its N-gram), evaluated based on the feature parameters specified in Table 2. \( L \) refers to the language identity of a pseudosyllable (or its N-gram). Mutual information measures the decrease of uncertainty of \( L \) after knowing \( F \) (or that of \( F \) after knowing \( L \)) [6]. Here entropy is used to quantify uncertainty. The initial uncertainty in \( L \) and the mutual information are represented respectively by (1) and (2):

\[
H(L) = - \sum_{l=1}^{K_L} P(l) \log P(l)
\]

\[
I(L; F) = H(L) - H(L|F)
\]

where

\[
H(L|F) = - \sum_{f=1}^{K_F} P(f) \left( \sum_{l=1}^{K_L} P(l|f) \log P(l|f) \right)
\]

\( K_F \) is the total number of feature classes, \( K_L \) is total number of language classes. In this study, \( K_L \) equals four. Theoretically it is possible to enumerate all features \( F \) and look for a desirable \( F^* \) in the sense that \( I(L; F^*) \) is the maximum.

\[
F^* = \arg \max_F I(L; F)
\]

#### 3.2. Low completeness, feature bias and z-score

Given unlimited amount of unbiased data, equation (4) can be applied directly to find the optimal \( F^* \). However, there is

---

**Table 1. Details of the four dialects investigated.**

<table>
<thead>
<tr>
<th>Dialect</th>
<th>City</th>
<th>Speaker ID</th>
<th>Gender of Speaker</th>
<th>Number of Pseudosyllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wu</td>
<td>Shanghai</td>
<td>SH</td>
<td>M</td>
<td>509</td>
</tr>
<tr>
<td></td>
<td>Suzhou</td>
<td>SZ1</td>
<td>M</td>
<td>498</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SZ2</td>
<td>F</td>
<td></td>
</tr>
<tr>
<td>Cantonese</td>
<td>Guangzhou</td>
<td>GZ</td>
<td>M</td>
<td>559</td>
</tr>
<tr>
<td></td>
<td>Hong Kong</td>
<td>HK</td>
<td>F</td>
<td>566</td>
</tr>
<tr>
<td>Min</td>
<td>Xiamen</td>
<td>XM</td>
<td>M</td>
<td>593</td>
</tr>
<tr>
<td></td>
<td>Taipei</td>
<td>TP</td>
<td>F</td>
<td>468</td>
</tr>
<tr>
<td>Mandarin</td>
<td>Beijing</td>
<td>BJ</td>
<td>M</td>
<td>472</td>
</tr>
<tr>
<td></td>
<td>Tianjin</td>
<td>TJ</td>
<td>F</td>
<td>513</td>
</tr>
</tbody>
</table>

**Table 2. Prosodic features under investigation**

<table>
<thead>
<tr>
<th>Feature Set</th>
<th>Prosodic parameter(s)</th>
<th>Quantization Levels</th>
<th>( K_F^* )</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>F0</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A3</td>
<td>F0 gradient</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A5</td>
<td>Normalized A4</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A6</td>
<td>Inter-pseudosyllabic duration</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A7</td>
<td>Normalized A6</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A8</td>
<td>Voiced ratio of pseudosyllable</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A9</td>
<td>Sentence-level F0 gradient</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A10</td>
<td>Point residual F0</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A11</td>
<td>Average residual F0</td>
<td>[5]</td>
<td>5</td>
</tr>
<tr>
<td>A12</td>
<td>Sentence length</td>
<td>[5]</td>
<td>5</td>
</tr>
</tbody>
</table>

\( K_F^* \) is the number of feature classes.
limited amount of data and a lack of speakers in this experiment. First, consider a feature $F^*$ with many feature classes and limited data. In an extreme case, every feature class comprises only one data point. This is referred to as low completeness [7]. Second, consider a scenario when each language is represented by very few speakers. Strong individual characteristics tend to show up. This is referred to as feature bias. In these two scenarios, equation (4) is no longer sufficient for feature analysis.

To illustrate the two scenarios, we consider a set of random variables $S$, whose element is any grouping scheme by which eight speakers are divided into four groups of twos (SZ1 and SZ2 are considered as one speaker for convenience). $S$ has 105 elements, as $|S| = (C_4^8 \times C_2^8 \times C_2^8 \times C_2^8)/4! = 105$. Note that $L$ is one of the 105 grouping schemes which groups speakers according to their language identities, i.e. $L \in S$.

Consider $S_{arb}$, an arbitrary grouping of speakers ($S_{arb} \in S - L$). $S_{arb}$ does not reflect the language identities of speakers. Nevertheless, for an $F^*$ with low completeness, $H(S_{arb} | F^*) = H(L | F^*) = 0$ because of zero conditional entropy for every single-member feature class. For an $F^*$ with feature bias, the distribution of this $F^*$ is governed by speakers’ individual characteristics, and there exists some $S_{arb}$ where $H(S_{arb} | F^*) \approx H(L | F^*)$. With the help of equation (2) and the fact $H(L) \approx H(S_{arb})$, we know for $F^*$ with low completeness or feature bias:

$$I(L; F^*) \approx I(S_{arb}; F^*) \ \exists S_{arb} \in S - L$$  \hspace{1cm} (5)

To avoid (5), a new criteria is proposed as:

$$I(L; F^*) \gg I(S_{arb}; F^*) \ \forall S_{arb} \in S - L$$  \hspace{1cm} (6)

$I(S; F^*)$ has different dynamic ranges for different features $F^*$ and different grouping sets $S$. We do not know how large the difference between $I(L; F^*)$ and $I(S_{arb}; F^*)$ should be. Assuming the measures of $I(S; F)$ are independent and normally distributed given any $F$, we can find $F^*$ by maximizing the $z$-score of $I(L; F)$.

$$F^* = \arg \max_{F} \max_{F} \frac{I(L; F) - I(S; F)}{\text{STD}(I(S; F))}$$  \hspace{1cm} (7)

Therefore, an optimal feature $F^*$ containing abundant language-related information can be found by equation (4) and (7), i.e. to evaluate $I(L; F)$ and $z$. In cases where $I(S; F)$ are not normally distributed, assumption needed by equation (7) is violated. The evaluation of $I(S; F)$ for other speaker groupings will be necessary in order to verify the criteria specified in equation (6).

4. Results and Discussion

The 27 chosen features are compared in terms of $I(L; F)$ and $z$ in Table 3. An optimal feature is expected to have large $I(L; F)$ and $z$ values. For each chosen feature $F$, $I(L; F)$, as well as the distribution of 105 $I(S; F)$ values are shown in a boxplot in Figure 1. The value of $z$ for a particular $F$ can be visualized, by noticing the relative position of the grey asterisk ($I(L; F)$) in each column of $I(S; F)$. For example, the asterisk for $A1$ lies far below the median. It probably lies below the mean also and $z$ of $A1$ should be negative.

$z$ value is useful in indicating feature bias. The low $z$ value for $A1$ reflects the strong individual characteristics in raw features. Normalization is performed to align feature’s magnitude and dynamic range across speakers, such that biases can be removed as far as possible. By comparing $A1$ with $A2$, and $A4$ with $A5$, we can observe that $z$ increases, indicating less feature bias when features are normalized.

4.1. One-dimensional prosodic features

Because $A1$-$A13$ have the same number of classes ($K_F$), $I(L; F)$ of these 13 features are directly comparable. Raw features $A1$, $A4$ and $A6$ are included only to illustrate the effect of normalization. Let alone these raw features and sentence gradient ($A9$), in the remaining nine type-$A$ features F0 gradient ($A3$) stands out with high $I(L; F)$. The other eight features have similar $I(L; F)$ around 0.01. Their $z$ values are compared. Normalized F0 ($A2$) and intensity ($A5$) are found to have large $z$ values. These three features are considered to be reliable. Voiced ratio ($A8$) and residual F0 ($A10, A11$) are considered to be auxiliary as reflected from their smaller $z$ values. Sentence length ($A12$) and syllable position ($A13$) are unrelated features, as shown by negative $z$.

$I(L; F)$ and $z$ of duration ($A6$) and sentence gradient ($A9$) are high and this is counter-intuitive. The distributions of these features are studied. For $A6$, pseudosyllabic durations tend to be short in Mandarin and long in Min. For $A9$, rising phrase F0 curves dominate only in Cantonese. Because there are only two speakers in one dialect, it could be possible that the small number of speakers just happen to have large $z$ values compared with their other features. The other hand, the distributions of $A6$ and $A9$ may be affected by dialect-dependent characteristics of agglutination manners and lexical tone profiles. With the lack of domain knowledge and the lack of speakers, no conclusion can be drawn here.

4.2. Multi-dimensional and N-gram features

Types $A$ and $B$ features are compared in Figure 1(b). $I(L; F)$ increase with the temporal span of the unit. However, F0 gradient bi-gram ($B3$), intensity tri-gram ($B6$) and duration tri-gram ($B8$) have smaller $z$ values compared with their corresponding uni-grams. That means an increase of the unit’s temporal span doesn’t necessarily mean an increase of language-related information in $F$. F0 gradient tri-gram ($B4$) is the most reliable type $B$ feature with the highest $I(L; F)$ (0.043). F0 bi-gram ($B1$) and tri-gram ($B2$) are also preferred features with large $z$. Intensity uni-gram ($A5$) is marginally better than F0 uni-gram ($A2$), but its N-gram counterparts $B5, B6$ have smaller $z$ values and are inferior to $B1, B2$.

For type $C$ and $D$ features, $K_F$ increases significantly. Data sparseness may lead to low completeness. This fear can
be ruled out by the high $z$ values for these features. $C3$ and $D3$ have a component parameter $A9$ whose reliability is in doubt, although both $C3$, $D3$ have quite large $I(L;F)$ and $z$. $C1$ and $C2$ have similar values of $I(L;F)$. With larger $z$, $C2$ is considered a better feature. $D2$ with $F0$ bi-gram, $F0$ gradient bi-gram and syllable position have the largest $z$ (and the most language-related information) among 27 features, although syllable position ($A13$) alone conveys no language information.

Finally, the optimal condition specified by equation (6) is verified. There are only four features for which equation (6) could be fulfilled. These four features are $B1$, $B2$, $D2$ and $D3$. As visualized in Figure 1, the grey asterisk ($I(L;F)$) of these four features are located above the top whiskers. Furthermore, it is found in the data set that $S$ with subgroups of HK-GZ and XM-TP tends to give high $I(S;F)$. This indicates the higher resemblance of the Cantonese pair and Min pair of speakers in this database. If the number of speaker used had been greater, this observation could have been generalized to a conclusion of the homogeneity of a dialect.

5. CONCLUSIONS

In this paper, two entropy-based evaluation metrics are formulated to evaluate the effectiveness of 27 prosodic features in LID. These two metrics can be applied for analysis in cases of data sparseness and feature bias which are common in practical applications. $F0$ gradient ($A3$), $F0$ ($A2$), intensity ($A5$) are one-dimensional prosodic features which contain abundant language-related information. Multi-dimensional N-gram features with $F0$, $F0$ gradient and syllable position ($D2$) is shown to be the most robust. These statistics can be readily used to select optimal prosodic features in an LID system. On the other hand, there are still cases (e.g. duration and sentence gradient) in which the evaluation metrics fail to tell. Domain knowledge to different prosodic features are yet to extensively apply. In view of the LID tasks, the comparison between the conventionally used phonetic features and the prosodic features are necessary before prosodic features can be applied to enhance LID performance.

6. REFERENCES