Automatic Tonal and Non-Tonal Language Classification and Language Identification Using Prosodic Information

Liang Wang¹, Eliathamby Ambikairajah², and Eric H.C. Choi³

¹,² School of Electrical Engineering and Telecommunications, Faculty of Engineering,
The University of New South Wales, Sydney, Australia
³ ATP Research Laboratory, National ICT Australia,
Sydney, Australia
¹l.wang@student.unsw.edu.au
²ambi@ee.unsw.edu.au
³eric.choi@nicta.com.au

Abstract. In this paper, an innovative method for tonal and non-tonal language classification using prosodic information is reported. The normalized feature parameters that measure pitch changing speed and pitch changing level are used to train a 3-layer feedforward neural network for the classification. To demonstrate the effectiveness of the proposed method, the recognition rate and the processing time of the novel system are compared with a PPRLM system on a 2-language identification task. For the evaluation results of identifying English/Mandarin, the novel system can achieve a recognition rate of 83.3\%, compared with 91.7\% of the PPRLM system. However, the processing time of the novel system is only half of that of the PPRLM system. In another extended tonal and non-tonal language classification task with 6 languages, the novel system can achieve a classification rate of 80.6\%. Possible applications of the new method to perform pre-classification in language identification are also discussed.

Keywords: Tonal language, non-tonal language, pitch changing speed, pitch changing level, language identification.

1 Introduction

The goal of language identification (LID) is to identify the language spoken in a particular utterance. Previous research on language identification relies on Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) to model the characteristics of a language described by certain features [1]. In language identification, GMM systems are mostly used to identify the language using acoustic content of the speech signal. The acoustic features are typically modeled as a series of Gaussian mixtures. It is suggested to use as many mixtures as possible, in order to obtain a better LID rate. Thus, more computation time is needed in both the training and testing sessions. Compared with the GMM systems, HMM systems are mostly
used to extract the phonotactic content of a speech signal. The phone-based HMM LID system, such as Parallel Phoneme Recognition followed by Language Modeling (PPRLM), is widely used to perform LID and is indicated to be the most successful approach [2]. In the PPRLM system, the speech utterances are required to be labeled at the phonotactic level. Also, to perform the phoneme recognition is very time consuming.

Recent research has focused on building LID systems based on prosodic information. Prosodic information refers to the duration characteristics of phoneme, intonation (pitch contour variation) and stress patterns [3]. This is the longer term information compared with the acoustic and phonotactic information.

Based on our observations of the tone patterns in Chinese and English, we found that it is possible to use pitch changing speed and pitch changing level to identify Chinese and English, which can also be implemented to perform the tonal and non-tonal language classification.

This research work describes a novel tonal and non-tonal language classification scheme based on pitch information. Unlike the LID systems based on acoustic and phonotactic information, the novel classification system does not need detailed phone transcription and the computation time of the novel system is proved to be much less. The new method is universal and it can be applied to all tonal and non-tonal languages. Moreover this method can be combined with other language identification systems to improve the accuracy and shorten the computation time.

2 Characteristics of Pitch Information

In human languages, pitch is regarded as one of the important prosodic features that relate to phonation. It is obvious that the vocal folds can vibrate at different frequencies, and thus that voice can be produced at different pitches. The change of pitch is produced in two ways, that is, by stretching and tensing the vocal folds—the tenser the folds the higher the pitch, or by changing the pressure of the subglottal—the higher the subglottal pressure the higher the pitch [4, 5].

Pitch and pitch changes are utilized in languages in two distinct ways. On the one hand, variations of pitch may be related to relatively long stretches of speech which are many syllables in length and correspond to relatively large grammatical units such as the sentence. Pitch variation used in this way is called intonation. All languages use intonation to express emphasis, contrast, and emotion. On the other hand, the pitch variation is used in short stretches of syllable length, such as in small grammatical units like words and morphemes. Pitch variation used in this way is called tone. Tonal languages are the languages use tone to distinguish lexical meaning.

A slight majority of the languages in the world are tonal. However, most Indo-European languages, which include the majority of the most widely-spoken languages in the world today, are not tonal.

In the case of Mandarin, which is monosyllabic and tonal, for each syllable there are four lexical tones and one neutral tone. Each tone has a particular pattern of pitch trajectory. For example, the phone /ma/ may have four lexical tones, like {/ma1/ /ma2/ /ma3/ /ma4/}. Although the phones are the same, the real acoustic realization
are different because of the tone types are different, and hence the actual meanings are different (see Table 1).

Table 1. Four Mandarin tones in the monosyllable /ma/.

<table>
<thead>
<tr>
<th>Chinese Character</th>
<th>妈</th>
<th>麻</th>
<th>马</th>
<th>骂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tone Description</td>
<td>High level</td>
<td>High rising</td>
<td>Low falling rising</td>
<td>High falling</td>
</tr>
<tr>
<td>Tone symbol</td>
<td>–</td>
<td>/</td>
<td>∨</td>
<td>\</td>
</tr>
<tr>
<td>English Gloss</td>
<td>‘mother’</td>
<td>‘hemp’</td>
<td>‘horse’</td>
<td>‘scold’</td>
</tr>
</tbody>
</table>

As discussed before, intonation is related to relatively long stretches of speech, which means the pitch changing may be happened for several syllables, while tone is related to short stretches of syllable length, which means the pitch changing may be happened for each syllable. Also, intonation is used across all different languages while tone is only used in tonal languages. It can be seen that the speakers of a tonal language should have better ability to make local pitch changes, thus the pitch may be changed more frequently for tonal languages. In order to perform the tonal and non-tonal language classification, we build a novel classification system that is capable of measuring the pitch changes. The pitch changes are measured in two different ways: the pitch changing speed and the pitch changing level.

3 Tonal and Non-Tonal Language Classification

The proposed tonal and non-tonal language classification scheme is described in this section. In order to measure the performance of our novel tonal and non-tonal language classification system, we also built a baseline PPRLM LID system for comparison purpose.

3.1 Tonal and Non-Tonal Language Classification Using PPRLM System

The baseline PPRLM system (Fig. 1) can be decomposed as two sub-systems: the phoneme recognizer and the language model. The phoneme recognizer maps a speech utterance into a sequence of phoneme symbols. In this experiment, we use the labeled Mandarin and English speech utterances in the OGI-TS speech corpus [6] to train the phoneme recognizer. Each phoneme symbol is modeled by a 3 state HMM and each state distribution is modeled by 8 Gaussians. Each basic feature vector contains 12 Mel-frequency cepstral coefficients (MFCC) and a log-energy. Additionally, the delta and acceleration coefficients are appended which results in a 39-dimension feature vector.

With a given phoneme recognizer in hand, an N-gram language model is employed to estimate the probability of the occurrence for a particular phoneme sequence. In
this experiment, a tri-gram language model with the Witten-Bell discounting method is implemented in the PPRLM system.

Once the unknown utterance is identified, its tonal characteristic can be obtained from a lookup table and therefore classification can be done.

Fig. 1. The block diagram of the PPRLM tonal and non-tonal language classification system.

3.2 Novel Tonal and Non-Tonal Language Classification System

As is discussed in Section 2, the novel tonal and non-tonal language classification system (Fig. 2) would perform the classification based on the analysis of the pitch changing speed and the pitch changing level. Each component of the system will be described in detail below.

Fig. 2. The block diagram of the novel tonal and non-tonal language classification system.
In the pitch extraction module, the raw F0 contour is automatically extracted from the input speech signal. In this experiment, we use the pitch analysis based on an autocorrelation method. The pitch values are extracted every 10 ms with a frame size of 40ms.

Since pitch extraction from the speech signal is a difficult task to perform automatically, the resulting F0 contours often contain “drop-out” during voiced speech segments or “spurious” pitch values in regions of unvoiced speech segments. Lengthy stretches of aperiodicity due to creakiness also contribute to a phenomenon in pitch extraction called “double pulsing” in which the extracted pitch values are twice the actual value [7, 8]. In order to get a preferable accurate pitch contour, the pitch smoothing and trimming module is used. At first, all the voiced segments of which the lengths are less than 100ms are removed. It is because presumably these voice segments may be produced by the “spurious” pitch values in regions of unvoiced speech segments, or these voiced segments do not carry any actual meaning (these voiced segments may be caused by coughing, laughing, etc.). Then, a trimming algorithm is used to smooth the F0 curves. The trimming algorithm compares the average pitch values ($f_i$) of a certain voiced segment against the average pitch value ($\bar{F}$) of the whole utterance. The voiced segments will be kept for further processing only if:

$$C_1 \ast F < f_i < C_2 \ast F.$$  

where $C_1$ and $C_2$ two thresholds, determined by experiments. The trimming algorithm effectively eliminates sharp “spikes” in the pitch tracing often seen around nasal-vowel junctions. Finally, a 5-order median filter is used to smooth the pitch again, and also compensate the “drop-out” during voiced speech segments.

The pitch changing speed analysis is firstly performed on each voiced segments. Assume $s_1$, $s_2$, $s_j$, ..., $s_M$ stand for the voiced segments in a particular utterance and $f_1$, $f_2$, $f_j$, ..., $f_N$ stand for the pitch values within a certain voiced segment in that speech utterance. Then the absolute value of the pitch varies within the voiced segment $j$ is firstly calculated:

$$p_{v_j} = \sum_{i=1}^{N-1} |f_i - f_{i+1}|.$$  

thus the total pitch changing can be calculated by:
Similar to the pitch changing speed analysis, the pitch changing level analysis is also performed for each voiced segment first. Let $\sigma p_j$ be the standard deviation of the pitch value of the $j$th voiced segment of the speech utterance. Then the pitch changing level may be measured by summing $\sigma p_j$ across the whole utterance:

$$STDP = \sum_{j=1}^{M} \sigma p_j.$$ 

Thus, the pitch changing speed is a measurement of the pitch’s local variation pattern, while the pitch changing level is used to measure the pitch’s global variation pattern.

The voiced segments duration, which is represented by $VD$, is estimated by counting the total number of voiced segments in the speech, e.g., the total voiced sound duration of a speech utterance will be $VD \times 10\, \text{ms}$ since the pitch values are extracted every 10ms in this experiment. The voiced duration can be used to compensate the differences of the speed of the different speakers.

The pitch average value ($AVE$) is obtained by averaging the pitch value across the whole utterance. The pitch average value is used to normalize the pitch changing speed and pitch changing level between different genders. Because presumably, females would have higher average pitch values and wider pitch ranges than males, and also have a faster pitch changing speed and bigger pitch changing level.

The voiced phoneme count is represented as $VC$ in this experiment and is defined as the number of voiced phonemes in a speech utterance. Voiced phonemes are identified by using a broad-phone-class recognizer. The voiced phoneme count is used as a normalization factor together with the pitch changing speed analysis and the pitch changing level analysis. It is because we want to measure both the pitch changing speed and the pitch changing level parameters against each phoneme symbol, especially for Mandarin which is a monosyllable language.

Based on these observations, a normalization module is used to generate the feature parameters for the neural network. Normalization of feature parameters is indispensable to reduce the undesirable variation caused by speaker difference and other possible factors [9]. In this study, $PV$ and $STDP$ are normalized as:

$$p\hat{V} = \frac{PV}{AVE \times VD}.$$
\[ S\hat{DP} = \frac{STDP}{AVE \ast VC}. \] (6)

So \( \hat{P} \) and \( \hat{STDP} \), together with \( AVE \), \( VD \) and \( VC \) are the five feature parameters that are fed into the neural network to perform the final classification.

The last module of the novel tonal and non-tonal language classification system is a simple three-layer feedforward neural network. The neural network is trained using the standard error backpropagation algorithm. The input layer consists of five neurons, each accepting one feature parameter. The output layer has only two neurons to represent the tonal and non-tonal languages respectively. The size of hidden layer is determined empirically and task dependent. Presumably if the tonal and non-tonal language classification system is performed for more languages, more hidden neurons would be required. In this experiment, the best recognition performance is obtained by using 10 hidden neurons.

4 Experiments and Results

As virtually no literature can be found on the tonal and non-tonal language classification, there is no way for us to compare our novel classification system with the others. In this experiment, two tonal and non-tonal language classification tasks are carried out. The tonal and non-tonal language classification is firstly performed on Mandarin and English, which are recognized as tonal language and stress language respectively. With another experiments for other four languages, we will show that this novel method is universal, which means that it can be applied to all tonal and non-tonal language classification tasks. Moreover this method can be combined with other LID systems to improve the accuracy and shorten the computation time for a language identification task.

For the 2-language classification task, Mandarin with mainland dialect and English with non-southern dialect are chosen to represent tonal and non-tonal languages respectively. For two languages, the tonal and non-tonal language classification can be viewed as language identification (the language is to be classified as tonal or non-tonal language by our novel system, or to be identified as Mandarin or English in the language identification task). So in this task, the Mandarin and English language identification task is performed on the baseline PPRLM system and our novel system. The results of LID recognition rate of Mandarin vs. English are compared.

For the 6-language classification task, another two tonal languages and two non-tonal languages are added. Again, the PPRLM is used in order to compare the results with the novel system. Firstly we use the PPRLM to perform the language identification, then we map the identified language into a decision of tonal or non-tonal language based on prior knowledge in order to obtain the classification rate.

In these experiments, a desktop computer with a 3.2GHz single-core CPU and 1GBytes of RAM is used.
4.1 Speech Corpora

In the experiments, the OGI-TS speech corpus and the CALLFRIEND speech corpus [6] are used. In the OGI-TS speech corpus, each speech utterance was spoken by a unique speaker over a telephone channel and was sampled at 8000Hz. The average duration of an utterance is about 45s. The CALLFRIEND speech corpus is similar to the OGI-TS, except that the average duration of an utterance is about 30 minutes. It should be noted that in the experiments, there is no overlapping in the training and testing data for all systems.

In the 2-language task, for the PPRLM, all labeled speech of Mandarin and English in the OGI-TS is used to train and test the phoneme recognizers. The language models are trained and tested by the data from CALLFRIEND Mandarin Chinese-Mainland Dialect (108 utterances for training and 12 utterances for testing) and CALLFRIEND American English-Non-Southern Dialect (108 utterances for training and 12 utterances for testing). For the novel system, the phone-class recognizers are trained and tested by the same data sets used to train and test the phoneme recognizer in PPRLM. With the five pitch feature parameters as the input, the neural network is trained and tested by the same data used to train and test the language model in PPRLM.

In the 6-language task, the phoneme recognizers and the phone-class recognizers are trained and tested by the same data used in the 2-language task. For the PPRLM, the language models are trained and tested by the data from CALLFRIEND Mandarin Chinese-Mainland Dialect (114 utterances for training and 6 utterances for testing), CALLFRIEND American English-Non-Southern Dialect (114 utterances for training and 6 utterances for testing), CALLFRIEND Mandarin Chinese-Taiwan Dialect (6 utterances for training and 6 utterances for testing), CALLFRIEND American English-Southern Dialect (6 utterances for training and 6 utterances for testing), CALLFRIEND Vietnamese (6 utterances for training and 6 utterances for testing) and CALLFRIEND German (6 utterances for training and 6 utterances for testing). For the novel system, again the neural network is trained and tested by the same data used to train and test the language model in PPRLM.

4.2 Experiment Results

To perform the voiced phoneme counting, we implement a phone-class recognition system. As the OGI-TS speech corpus has already labeled Mandarin and English at the phonotactic level, we first translate the phonotactic level labeling into phone-class level labeling. VOI, UNV and SIL are used to represent the voiced segment, unvoiced segment and the silent segment. Then we build a HMM for each of the phone-class. Each phone-class is modeled by a 3-state HMM with 8 Gaussian mixtures. The OGI-TS speech corpus is used to train and test the phone-class recognizer. For testing the recognizer, the 45s “story-bt” utterances (about 12 for each language) are used. None of the testing utterance is seen from the training set. In the experiment, the processing time for the phone-class recognition is much less than the processing time of the phoneme recognition. The accuracy rate of the phone-class recognition can be
achieved at 84.17%. The resulting confusion matrix of the phone-class recognition system and the corresponding accuracy rates are shown in Table 2.

Table 2. Confusion matrix of the phone-class recognition system on OGI-TS test set.

<table>
<thead>
<tr>
<th></th>
<th>VOI</th>
<th>UNV</th>
<th>SIL</th>
<th>Total</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>VOI</td>
<td>46694</td>
<td>6336</td>
<td>1981</td>
<td>55011</td>
<td>84%</td>
</tr>
<tr>
<td>UNV</td>
<td>2547</td>
<td>13692</td>
<td>1348</td>
<td>17587</td>
<td>77%</td>
</tr>
<tr>
<td>SIL</td>
<td>1380</td>
<td>2158</td>
<td>23350</td>
<td>26888</td>
<td>86%</td>
</tr>
</tbody>
</table>

The comparison in the 2-language task of the recognition rate and the processing time of the PPRLM and the novel system is shown in Table 3. The processing time is measured by taking the ratio of the whole system’s running time and the duration of the utterance (i.e., the 2 * real-time means that for a 10s utterance, it would take about 20s to process on the desktop computer used in this experiment).

Table 3. Language identification results comparing the novel system and the PPRLM system on 2-language task

<table>
<thead>
<tr>
<th></th>
<th>PPRLM</th>
<th>Novel System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recognition rate</td>
<td>91.7%</td>
<td>83.3%</td>
</tr>
<tr>
<td>Processing time</td>
<td>2 * real-time</td>
<td>1 * real-time</td>
</tr>
</tbody>
</table>

Compared with the PPRLM system, which is known as the most successful LID system without using any fusion technique, our novel system shows a promising performance which is only 8.4% less in the recognition rate. It should be noted that, the processing time of the novel system is only the half of that of the PPRLM system.

Table 4 shows the classification results for the 6-language task.

Table 4. Tonal/non-tonal language classification results comparing the novel system and the PPRLM system on 6-language task

<table>
<thead>
<tr>
<th></th>
<th>PPRLM</th>
<th>Novel System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification rate</td>
<td>86.1%</td>
<td>80.6%</td>
</tr>
<tr>
<td>Processing time</td>
<td>2.4 * real-time</td>
<td>1 * real-time</td>
</tr>
</tbody>
</table>

It shows that the novel system is also capable of performing the tonal and non-tonal language classification in this task. The processing time is also much shorter than the PPRLM system. Although the novel system is only tested on a limited number of speech utterances, we will discuss the possibility of its application for all the tonal and non-tonal language classification tasks in detail.
5 Other Tonal Languages

As mentioned earlier, the novel method described here is universal, which can be applied to any tonal and non-tonal language classification. As examples, we discuss the applications in other tonal languages and the pitch accent languages.

5.1 Cantonese, Thai and Vietnamese

Tonal languages fall into two broad categories: contour tone systems and register tone systems. Mandarin has a contour tone system, in which the distinguishing feature of the tones are their shifts in pitch rather than the pitch relative to each other as in a register tone system. Similarly, Cantonese, Thai and Vietnamese are monosyllable languages and have the contour tone system. Cantonese has nine lexical tones, Thai has five tones and Vietnamese has six tones. So the novel method can still be used to perform tonal and non-tonal language classification for these three languages. It should be noted that, with different number of tones, the pitch changing pattern may be different as well. So the novel method may also be able to perform the language identification across the tonal languages.

5.2 Swedish and Norwegian

Swedish and Norwegian are also tonal languages. But pitch is also used to differentiate two-syllable words depending on their morphological structure. For many multi-syllable words in those two languages, there are two pitch contours, which are called Tone 1 and Tone 2 respectively. Words with the same vowels and consonants but different pitch contours often have different meanings. So these two languages could be treated same as Mandarin by the novel method, as the pitch changing is against each word, rather than the whole sentence.

5.3 African Languages

Most of the African languages are tonal languages, but they have register tone system. For those tonal languages that have register tone system, all the tones have a “flat” shape, which means, pitches are only differentiated by their relative values. As we can see, the pitch changing is still against each word rather than the whole sentence. So the tonal and non-tonal language classification can still be used in tonal language with the register tone system.

There are also some African languages that combine register and contour tones, such languages are also suitable to be processed by the novel system.
5.4 Japanese

The role of pitch in Japanese is quite different from other Asia languages such as Chinese, Thai and Vietnamese [10]. The Japanese language does not have tone, but does have pitch accent. In Japanese, the pitch of each syllable could be either high or low. In a pitch accent language, there is one accented syllable in a word, and the position of that accented syllable determines the tonal pattern of the whole word. This is unlike the situation in tonal languages, where the tones of each syllable can be independent of the other syllables in the word. For example, in the tonal language and the pitch accent language both only have two distinguishing tones, the two-syllable words /ada/ would have four possible patterns (low-low /ada/, high-high /ada/, low-high /ada/ and high-low /ada/) in the tonal language while only has two possible patterns (high-low /ada/ and low-high /ada/) in the pitch accent language.

With the discussion above, it can be found that though for the pitch accent languages, the pitch changing pattern is different from that of the tonal language, for the same phones, the real acoustic realization might be different because of the pitch patterns are different, and hence the actual meanings are different. So in the tonal and non-tonal language classification task, the pitch accent language can still be treated as the tonal language.

6 Applications in LID Task

Rather than performing the tonal and non-tonal language classification task, this novel system is also possible to perform the language identification task, by analyzing the pitch changing speed and the pitch changing level. Our novel system can be used as a pre-classification for the PPRLM system to improve the accuracy and efficiency. In the pre-classification stage, all the languages to be identified will be classified as tonal language or non-tonal language, and then the PPRLM system with the corresponding subsets of models can be applied to the classified tonal and non-tonal languages. Presumably, the PPRLM system would perform more accurately and take less processing time since fewer number of languages need to be identified.

7 Conclusion

By measuring the pitch changing speed and pitch changing level, we show that the novel system can be used to perform tonal and non-tonal language classification. Compared with the PPRLM system, the resulting LID recognition rate of our novel system is promising, while its processing time is only half of that of the PPRLM system. We also show that, this novel method can be applied to all tonal languages and even pitch accent languages to perform tonal and non-tonal language classification. Also this method can be used as a pre-classification stage for the PPRLM system in language identification.
References