Taiwanese Tone Recognition Using Fractionalized Curve-fitting of Prosodic Features

Yu-lun Hsieh[1], *Ching-ting Chuang*[2], *Feng-fan Hsieh*[2], *Yueh-chin Chang*[2], *Wen-lian Hsu*[1]

[1]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[2]Graduate Institute of Linguistics, National Tsing Hua University, Hsinchu, Taiwan
morphe@iis.sinica.edu.tw, d9644510@oz.nthu.edu.tw, ffhsieh@mx.nthu.edu.tw,
ycchang@mx.nthu.edu.tw, hsu@iis.sinica.edu.tw

Abstract

In this paper, we examined different methods of modeling prosodic features of tones, and their effects on a speaker-independent Taiwanese tone recognition system. Tones can be modeled either by plain or curve-fitted features. Plain features represent the original curve faithfully using pitch values, while curve-fitted features can be thought of as an approximation to the values using mathematical functions, such as a Legendre polynomial. In addition, durational information of tones was also proven effective in previous researches. Thus, we proposed a new approach of modeling Taiwanese tones using curve-fitted features extracted from fractions of the pitch curve, along with duration as an additional prosodic feature. Our experimental results showed that using these features in an SVM classifier could substantially improve the accuracy of tone recognition in Taiwanese. Besides, we provided an empirical perspective for theoretic studies on tonal neutralization.

Index Terms: Taiwanese, tone recognition, prosodic feature

1. Introduction

Sinitic languages such as Mandarin and Taiwanese are famous for their syllabic and tonal characteristic, which is different from western languages, such as English. The same syllable structure can carry different lexical tones to indicate different meanings. Tones provide critical information in speech recognition. It was argued that articulatory features such as segmental information, syllable structures and prosodic features may play an important role in tonal recognition in Mandarin [1].

Taiwanese, a relatively understudied language with fewer resources, is a dialect of the Southern Min languages widely spoken in Taiwan. Generally speaking, the structure of Taiwanese syllables is of the form 'CGVC', where 'C' stands for consonants, 'G' for glides, and 'V' for vowels [2]. Compared to Mandarin, Taiwanese has a more complicated tonal system. It has three tonal height contrasts, while Mandarin, on the other hand, only has two. There are seven tones in Taiwanese. Note that Tone 4 (or the *Yangshang* tone in traditional Chinese phonology terms) is missing because it was diachronically merged with Tones 3 or 6. The corresponding F0 curves of Tone 1 through 3 and 5 through 8, after normalizing the duration of the syllable, are shown in Figure 1.

Among them, Tone 7 and 8 are comprised of a stop consonant coda /p, t, k, ʔ/, and are called 'checked tones' or 'entering tones'. Durational differences between checked tones and other tones are significant, with checked tones being shorter. Furthermore, it was observed that the duration of a syllable might change drastically depending on different syllable structures. Duration is argued to be an important phonetic cue even in discourse [3]. Therefore, in order to characterize the tonal structure of Taiwanese, one has to consider not only the pitch curve but also other prosodic qualities such as duration. Moreover, the pronunciations of tones in Taiwanese were under the influence of tone sandhi, in which a citation tone changes to a sandhi tone according to its syntactic position. The details of this phenomenon were described in [4], but they are beyond the scope of this research.

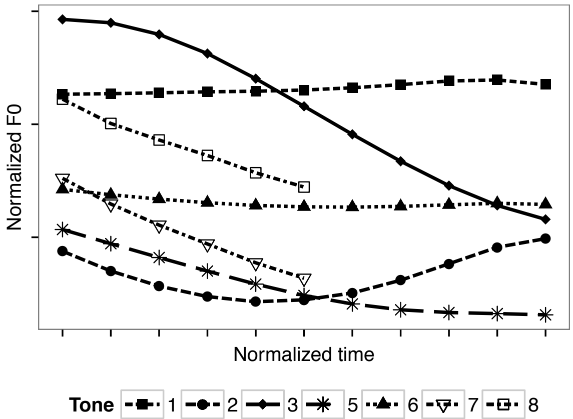

Figure 1: *Mean normalized F0 contour of seven Taiwanese tones.*

In the interest of comprehensive studies of the phonetics and phonology of Taiwanese language, a large speech corpus is required. However, a sizable Taiwanese speech database with accurate labeling is hard to come by. During the collection of such data by hand, one can find it to be time-consuming and error-prone. Thus, an automatic recognition system is crucial for building a large corpus for further studies.

Several researches have focused on different aspects of automatic processing of Taiwanese speech data. For example, in [5], a large vocabulary Taiwanese speech recognizer is built using HMM with raw pitch features in addition to a multiple pronunciation lexicon for sandhi tones. Specifically, two pitch smoothing techniques of the unvoiced regions, namely, random padding and exponential function linking between two consecutive pitch values, were compared to examine their abil-

10.21437/SpeechProsody.2014-142

ity to lower the character and utterance errors. Their results showed that using pitch information with exponential function for smoothing, in addition to the multiple pronunciation model, could significantly decrease the error rates in speech recognition. Another research focused on tone labeling of Taiwanese [6], in which both the citation and sandhi tones were jointly represented using statistical pitch contour models in order to eliminate contextual effects. It was proven to outperform the vector quantization method.

In this research, we want to focus on the recognition of Taiwanese tones using plain or curve-fitted features of the pitch contour along with durational feature. Following a similar approach in [7], we adopted the sub-sectioning method of splitting a pitch contour into different numbers of sections, and modeling them separately as our curve-fitted features. Also, the duration of a tone was included as an additional prosodic feature. We want to examine the effect of plain versus curve-fitted modeling of the F0 contour, as well as the effect of the number of sections and durational feature on tone recognition.

This paper is organized as follows. Section 2 describes our method of modeling and recognizing Taiwanese tones. Section 3 presents the experimental results along with some discussions. Finally, Section 4 concludes this paper.

2. Methods

We implemented several methods of extracting features from the F0 contour in order to compare the effect of them. First, raw F0 values were extracted by a Praat [8] script provided by [9]. Then, three types of feature sets, namely, plain, curve-fitted, and duration, were obtained from the raw F0 values to model different aspects of a tone. Plain and curve-fitted features were intended to capture the shape of the pitch curve, while the duration feature was included to describe the time-domain information. The following sections explain the definitions and extraction methods of these feature sets. Both the detailed and curve-fitted features were then paired with the duration features to train SVM classifiers and evaluation their performance on tone recognition.

2.1. Plain features

Plain features were simply the raw pitch values, in Hertz, computed from the audio. They precisely represent the original form of tones. There are 11 pitch values from equally-spaced points for each rhyme part of the syllable. The purpose of using these features is to provide fine-grained information of the pitch curve for tone model training.

2.2. Curve-fitted features

On the other hand, curve-fitted features were statistical pitch contour models used to capture the general characteristic of F0 variation within a tone. They can be further divided into two kinds. One is the method proposed by [10], in which a 3rd order orthogonal polynomial was used to represent the entire F0 contour. The basis polynomials, which are discrete Legendre polynomials, were normalized to the interval between 0 and 1. The detailed formulation is as expressed in (1).

$$
\begin{aligned}
\Phi_0(\frac{i}{N}) =& 1, \\
\Phi_1(\frac{i}{N}) =& \sqrt{\frac{12 \cdot N}{N+2}} \left(\frac{i}{N} - \frac{1}{2} \right), \\
\Phi_2(\frac{i}{N}) =& \sqrt{\frac{180 \cdot N^3}{(N-1)\,(N+2)\,(N+3)}} \cdot \\
& \left[\left(\frac{i}{N} \right)^2 - \frac{i}{N} + \frac{N-1}{6 \cdot N} \right], \\
\Phi_3(\frac{i}{N}) =& \sqrt{\frac{2800 \cdot N^5}{(N-1)\,(N-2)\,(N+2)\,(N+3)\,(N+4)}} \cdot \\
& \left[\left(\frac{i}{N} \right)^3 - \frac{3}{2} \left(\frac{i}{N} \right)^2 \right. \\
& + \frac{6 \cdot N^2 - 3 \cdot N + 2}{10 \cdot N^2} \left(\frac{i}{N} \right) \\
& \left. - \frac{(N-1)(N-2)}{20 \cdot N^2} \right]
\end{aligned}
\tag{1}
$$

for $0 \leq i \leq N$ where $N+1$ is the number of samples in the pitch contour. In this way, the original pitch values $\hat{f}\left(\frac{i}{N}\right)$ can be approximated by (2)

$$
\hat{f} = \left(\frac{i}{N} \right) \sum_{j=0}^{3} \alpha_j \cdot \Phi_j \left(\frac{i}{N} \right), 0 \leq i \leq N \tag{2}
$$

where

$$
\alpha_j = \frac{1}{N+1} \sum_{i=0}^{N} f\left(\frac{i}{N} \right) \cdot \Phi_j \frac{i}{N} \tag{3}
$$

Afterwards, the four coefficients $[\alpha_0, \alpha_1, \alpha_2, \alpha_3]$ from (3) were kept as one type of the curve-fitted features.

The other kind is a fractionalized fitting method similar to [7], in which the F0 curve was divided into four sections, and each section was represented by various parameters. Following this angle of approach, we used the 2nd order polynomial with the first three of the coefficients in (3). The F0 curve was first split into different numbers of equal-length sections, and each section was fitted separately. Figure 2 illustrates the difference between using a 3rd order polynomial to fit the entire curve, and fitting four sections separately using a 2nd order polynomial. We can see that a fractionalized fitting is more faithful to the original curve, while the higher order polynomial can capture the general shape of the curve. The resulting coefficients of the fitted functions were used as another set of curve-fitted features. The number of sections in a tone is another variable that we want to examine in this research.

2.3. Duration features

Lastly, the durations of each rhyme was used as another prosodic feature. As mentioned in Section 1, the duration information may be useful in distinguishing tones, especially the checked ones. We want to examine its effectiveness in distinguishing checked and non-checked tones, as well as other tones that were reported to have durational differences.

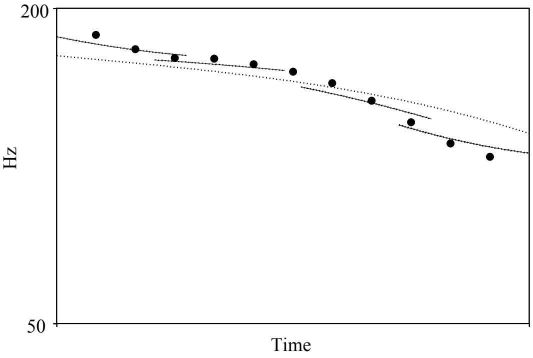

Figure 2: *Illustration of different curve-fitting methods on a series of samples of Tone 3. The dotted line represents using the whole series for fitting, and the solid lines represent fitting four sections separately.*

3. Experiments and Analysis

3.1. Experimental setup

We compiled a Taiwanese read speech corpus containing a total of 11,352 syllables from 10 different native speakers of Taiwanese. The recordings were conducted in a quiet room with a sampling rate of 44.1K and a background noise level lower than 30 dB. The wordlist consists of various categories of CVC combinations as described in Table 1, in which syllables with a stop coda were all carrying checked tones. The syllables represent real Taiwanese monosyllabic words, and were embedded in a carrier sentence to prevent influence from syntactic structures. Note that not all consonants and vowels in Taiwanese were included, as we are still in the process of building a comprehensive inventory now.

Table 1: *Inventory of our Taiwanese speech corpus.*

Position	Category	Inventory
Onset	Stop	p, t, k, ʔ, g
	Fricative	s
	Affricate	ts
	Liquid	l
	Nasal	m, n, ŋ
Vowel	Oral	i, e, a, ə, u, o, ɔ, ɤ
	Nasal	ã, ĩ
Coda	Stop	p, t, k, ʔ
	Nasal	m, n, ŋ

The total duration is about 11 hours. Since durational information may be used in training our tone models, the average duration is also reported in Table 2. The wave files were labeled by trained phoneticians using IPA symbols and numbers that denote surface tones. Pitch values from the rhyme part of the syllable were extracted using a Praat script "TimeNormalizedF0.praat" [9] from 11 equally spaced points. Then, different features described in Section 2 were obtained and used to train speaker-independent SVM classifiers using LibSVM [11], for comparing the effectiveness of our modeling methods.

Table 2: *Mean duration of seven Taiwanese tones in the corpus.*

Tone	Duration (ms)
1	202
2	233
3	183
5	180
6	209
7	122
8	107

3.2. Results and discussion

As shown in Table 3, the five-fold cross-validation accuracies of different feature sets were computed to evaluate the overall performance of our system. A few observations can be made from the results.

First, including the duration feature can indeed assist in identifying tones, with a 3% improvement in accuracy. It indicated that both frequency domain and time domain information are essential in the modeling of tones. Secondly, using curve-fitted features of the F0 curve outperforms using the plain features of the raw F0 values, which conforms to previous researches of tone modeling. The technique of dividing the curve into a number of sections was proven effective as well. By just splitting the curve into two sections, we can achieve a 0.5% increase of accuracy.

Table 3: *A comparison of tone recognition accuracy between different feature sets.*

Type	Feature	Accuracy (%)
Plain Features	Raw F0	77.10
	Raw F0 + duration	80.02
Curve-fitted Features	Entire curve fit + duration	80.08
	2-section fit + duration	80.51
	3-section fit + duration	80.32
	4-section fit + duration	**80.75**
	5-section fit + duration	80.60

The best performance was found in the feature set of 4-section polynomial fitting plus duration, with the accuracy of 80.75%. Notably, a higher partitioning of the F0 contour, i.e. 5-sections, resulted in a lower accuracy. It showed that the number of features is not positively related to accuracy, as the separation of a tone contour into too many sections might cause an overfitting effect that compromised the robustness of a model and its ability to identify the general characteristics of a tone. A 4-section method may be appropriate in that it can capture the left and right contextual variations resulting from neighboring segments using the two boundary sections, while the fluctuations in the center regions of a tone were well-represented by the remaining two sections. It has also been proven successful in [7], in which the language being studied is Mandarin. On the other hand, fitting the whole curve with a higher-order polynomial function is too coarse to be effective in representing a tone, and thus resulting in a lower accuracy.

In order to further analyze the effectiveness of our model on each of the seven tones, we trained the model using the complete set of F0 and duration data with the 4-section plus duration feature set, and then computed the accuracy for each tone. The results were shown in Table 4. As we can see, Tone 7 and 8 have considerably lower accuracies than others. It may simply be due to the fact that the sample size is too small for generating a robust model. However, there are in fact two kinds of checked tones in Taiwanese, one ending with /p, t, k/ and the other with /ʔ/. Previous studies showed that the realizations of them are slightly different [4]. For a deeper understanding and modeling of these two checked tones, additional data as well as research on the modeling techniques are required.

Table 4: *Recognition accuracy grouped by tone types.*

Tone	Number of correct/total syllables	Accuracy (%)
1	2204/2525	87.29
2	805/939	85.73
3	2033/2193	**92.70**
5	1737/2117	82.05
6	2200/2645	83.18
7	441/588	75.00
8	154/344	44.77

Nonetheless, for the non-checked Tone 1 to 6, the best performance was found on Tone 3. It could be attributed to the unique shape and range of the F0 curve, as depicted in Figure 1, along with a shorter duration that gave rise to a more distinctive tone model. Contrastively, the relatively lower accuracy of less than 85% occurred in Tone 5 and 6. It could be accounted for if we look at the wrong predictions of the classifier, as explained separately below.

- For Tone 5, we found that the most common mistakes were Tone 3 and 6, each occurred about 140 times. The similar pitch height of Tone 5 and 6 might have caused a confusion for the recognizer. As for the other two tones with comparable pitch height, namely, Tone 7 and 8, they can be easily distinguished by duration. Meanwhile, Tone 5 and 3 may have been indivisible because they were alike in both shape and durational feature.
- For Tone 6, the most common errors were Tone 1 and 5, with around 260 and 140 occurrences, respectively. The analogous reasoning above could be applied to explain the indistinguishability between Tone 6 and 1, in that their shape, height, and duration were all comparable. The multiple resemblances between them could have contributed to the errors.

Another perspective on the lower accuracy group is that there may be mutual affinity among them. In fact, previous study [12] showed that, in some dialects of Southern Min, Tone 5 and 6 were merged or 'neutralized'. Our findings could lend support to the theory that tonal neutralization is the result of the similarity and difficulty in maintaining contrast between tones.

In sum, a more robust model with the capability to tackle with these problems is required to improve the accuracy of our system. Moreover, a full-fledged Taiwanese tone recognition system must be able to distinguish sandhi tones from surface tones, and our system is yet to achieve this goal. A more sophisticated modeling scheme is necessary to incorporate such complications.

4. Conclusions

In this paper, we examined the effect of fractionalizing prosodic features on Taiwanese tone recognition, and proposed a new approach to modeling Taiwanese tones. Our results showed that using curve-fitting of four fractions of F0 values, and including the duration feature into model training were useful means of improving the accuracy of tone recognition. By further analyzing the outcomes, we provided an empirical point of view for theoretic studies on tonal neutralization. Future work can be done on investigating the effect of other prosodic and articulatory features, such as consonant context or energy. Since previous research suggests that these factors may play a role in tone recognition [1], incorporating them into Taiwanese speech recognition systems could be fruitful. In addition, expanding our corpus to include more tokens of checked tones is necessary for improving the accuracy. Furthermore, the current system only dealt with the surface tones. We will have to derive a more extensive model to resolve the problem of recognizing sandhi tones in Taiwanese.

5. Acknowledgments

The Taiwanese recording corpus was provided by the National Tsing Hua University Phonetics Lab, and was supported in part by the National Science Council grant of Taiwan (NSC 97-2410-H-007-025 and NSC 99-2410-H-007-050). The authors would like to thank Dr. Yu Tsao for insightful remarks and assistance, as well as comments from two anonymous reviewers.

6. References

[1] H. Chao, Z. Yang, and W. Liu, "Improved tone modeling by exploiting articulatory features for Mandarin speech recognition," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2012, pp. 4741–4744.

[2] L. W. R. Cheng and S. J. Cheng Xie, *Phonological Structure and Romanization of Taiwanese Hokkien*. Taipei: Student Book Company, 1977.

[3] S.-F. Wang and J. Fon, "Durational cues at discourse boundaries in Taiwan Southern Min," in *Proc. 6th International Conference on Speech Prosody*, 2012.

[4] R. L. Cheng, "Tone sandhi in Taiwanese," *Linguistics*, vol. 6, no. 41, pp. 19–42, 1968.

[5] D.-C. Lyu, M.-S. Liang, Y.-C. Chiang, C.-N. Hsu, and R.-Y. Lyu, "Large vocabulary Taiwanese (Min-nan) speech recognition using tone features and statistical pronunciation modeling," in *Proc. 8th EuroSpeech*, 2003.

[6] W.-C. Kuo, Y.-R. Wang, and S.-H. Chen, "A model-based tone labeling method for Min-nan/Taiwanese speech," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2004, pp. 505–8.

[7] Y. Tian, J. L. Zhou, M. Chu, and E. Chang, "Tone recognition with fractionized models and outlined features," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, 2004.

[8] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.3.55)," 2013.

[9] Y. Xu, "Timenormalizef0.praat," 2009.

[10] S.-H. Chen and Y.-R. Wang, "Vector quantization of pitch information in Mandarin speech," *IEEE Trans. On Communications*, vol. 38, no. 9, pp. 1317–1320, 1990.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27:27, 2011.

[12] C.-T. Chuang, Y.-C. Chang, and F.-F. Hsieh, "Complete and not-so-complete tonal neutralization in Penang Hokkien," in *Proc. International Conference on Phonetics of the Languages in China*, W.-S. Lee, Ed., 2013, pp. 54–57.