CLASSIFICATION OF TAIWANESE TONES BASED ON PITCH AND ENERGY MOVEMENTS

F .H.L. Jian

Dept. Linguistics, University of Reading, England

ABSTRACT

This paper addresses the difficulties associated with automatically distinguishing the seven Taiwanese tones. The tone recogniser is an essential component of any automatic speech recognition system customised for tone languages such as Taiwanese. We show that it is difficult to distinguish between the Taiwanese tones simply employing the fundamental frequency contours and that the task is simplified by employing energy contour features besides the fundamental frequency features. To allow energy to be accommodated into the classification model an energy-contour feature extraction approach is presented. The proposed approach is inspired by the ADSR model employed in musical instrument synthesis where the envelopes of complex sounds are modeled employing only a few parameters. Our experiments demonstrate that the inclusion of energy into the recognition model allows the seven Taiwanese tones to be discriminated successfully. The paper also presents acoustical measurements of the fundamental frequency and energy features described.

1. INTRODUCTION

Taiwanese (also known as Hokkien or Fukienese) is an ancient Chinese dialect originating from the Fujian province of southeast China. Although Mandarin Chinese is the official language in Taiwan, Taiwanese is widely used in conversation. It is the main language used by older generations and it is mutually unintelligible from Mandarin Chinese. Computer aided Taiwanese speech recognition appears to be a promising technology for bridging language gaps in the current Taiwanese society. Thus, the objective of this work is to make a contribution towards this common goal.

1.1. Tone languages

In common with other oriental languages such as Mandarin Chinese [2], [9], Thai [10], [19] and Cantonese [8], Taiwanese is a tone language [1], [2]. In a tone language the lexical meaning of an utterance is dependent on the tone or fundamental frequency contour applied to the utterance. The term "tone" refers here to the fundamental frequency contour of the word.

1.2. Tone classification

A tone classifier is an essential component of a speech recognition system for tone languages. Much research has been conducted into the automatic recognition of Mandarin Chinese [3], [4], [5], [13], [15], [16], [17], [11], [12], [13], [17], [18], [7]. Mandarin Chinese has four distinct tones that many non-Chinese speakers can distinguish with little or no training. Taiwanese however have seven tones with a subtle structure that makes them much harder to distinguish. Some Taiwanese tones may even appear undistinguishable for the un-trained ear. Consequently, the tone recognition methodologies developed for Mandarin tones cannot readily be applied to Taiwanese tones.

1.3. Fundamental frequency

It is relatively easy to distinguish the four Mandarin tones because their fundamental frequency contours are distinct. There is a level (high) tone, rising tone, falling- rising tone and a falling tone. Simply by analysing the fundamental contour of a word its tone is easily determined. However, in Taiwanese there are two level tones, four falling tones and one rising tone. Some tones are overlapping in range. Simply using the fundamental frequency contour to distinguish the tones is difficult.

1.4. Long and short tones

Besides fundamental frequency movement, ancient tone languages including Cantonese and Taiwanese have the notion of “free” and "checked" syllables, which are also referred to as non-entering or long tones and entering or short tones respectively. Although the terms “long” and “short” often imply that long tones are longer and short tones are shorter in duration, in this paper we will show that the duration of so called "long" and "short" tones are overlapping. Therefore, duration is not a suitable cue for distinguishing long and short tones. This confirms the work of Jian [1] in a perceptual study on Taiwanese long and short tones that duration is not a main perceptual cue. Instead, the rate of energy decay of a word was found to be a major cue for perceptually distinguishing the tones. The explanation for the differences in energy contour is that short tone words are accommodated by unreleased voiceless stops [p, t, k, ?]. The sudden closure of the glottis that results in utterances with glottal-stop is visible as steep energy decays when plotting energy against time. The closure "removes" the energy from the vibrating vocal chords, so, instead of a naturally decaying energy contour, the result is a very sudden and violent energy decay. Based on these findings we present a method for extracting energy features from utterances in isolation, and use these to aid the classification of the seven Taiwanese tones. Extracting energy features is more complex than extracting fundamental frequency features,

because the energy location depends on the intensity of the utterance and distance from the speaker. Further, the energy movement is the sum of many factors. In this paper we adapt a method inspired by the ADSR (attack, decay, sustain and release) representation commonly used musical instrument sound synthesis. This allows the envelope of a complex sound to be represented using only a small set of parameters. By combining fundamental frequency and energy movements it is possible to fully distinguish the seven Taiwanese tones.

The remainder of this paper is organised as follows: In section 2 we investigate why the fundamental frequency contour does not encompass sufficient information to completely classify the seven Taiwanese tones. In section 3 we show how this is solved by extending the recognition model by introducing energy-contour features. The paper is closed in section 4 by a set of concluding remarks.

2. FUNDAMENTAL FREQUENCY MOVEMENTS

The data used in this study was acquired by recording the voice of a native female Taiwanese speaker studying at the University of Reading. The recorded material consisted of 20 words from each of the seven tone groups with each word repeated three times. The recorded words were imported into the Xwaves digital signal processing and analysis environment. Xwaves facilitates the extraction of fundamental frequency (f0) and the root-mean-square (RMS) energy contours.

type	tone	Sf	mf	Ef	dur
long	t1	246	239	241	33
	t2	279	248	190	19
	t3	225	198	171	20
	t5	202	193	230	50
	t7	221	216	213	32
short	t4	239	225	195	9
	t8	216	216	199	14

Table 1: Mean start, mid and end frequencies (in Hz) for the seven Taiwanese tones, and the mean word duration (in cs) for words spoken by one female Taiwanese speaker.

2.1. Feature extraction

To allow the different words to be compared, a common set of word features were extracted from the Xwaves output. These features include the word duration and the start, mid and end frequencies. These frequency triplets represent the fundamental frequency contour approximated by two line segments connecting the three equally spaced points. Table 1 summarises these measurements where the values represent the mean duration, start, mid and end frequency for each tone respectively.

2.2. Frequency and classification

Note that the values in Table 1 only represent the mean and does not reveal the spread of the measurements. The data shows that tone 1 is unique being the highest level tone. Further, tone 5 is the longest tone and the only tone that is falling-rising. Tone 7 is also characteristic as it is the only tone at medium height that is level, and distinctly different to tone 1. The difficult tones are tone 2, 3, 4 and 8. Clearly, tone 2 is distinct from tone 3 in height and tone 4 is distinct from tone 8 in height. However, tone 2 is not distinct with respect to either tone 4 or tone 8 as the frequency ranges are often overlapping. Similarly, tone 3 is not distinct with respect to tone 4 and tone 8.

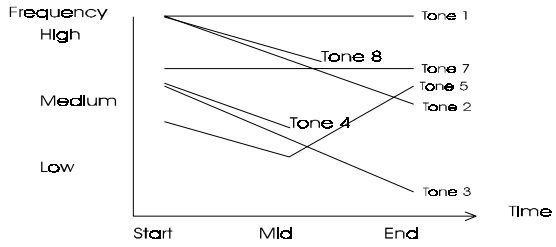

Figure 1: The seven Taiwanese tones (mean).

Figure 1 depicts the seven Taiwanese tones. It is clear from Figure 1 and Table 1 that although it is possible to classify tones using the fundamental frequencies and durations, it is harder to separate the long falling from the short falling tones. To enable the separation of long and short tones it is necessary to investigate the energy contour of the word.

3. ENERGY MOVEMENTS

The previous section revealed that neither frequency nor duration can be used to completely separate the "long" tones from the "short" tones. As mentioned, Taiwanese short tones always end with unreleased voiceless stops [1], [2]. In other words the short tone words can be viewed as "bursts" of energy, starting and ending suddenly and harshly, while the long tone words have more "soft" or smooth contours. Figure 2 illustrates this phenomenon. It shows the waveform and the RMS energy contour for a long and a short tone word. Although, both words have similar duration, the short word has a more rapid energy decay than the long word.

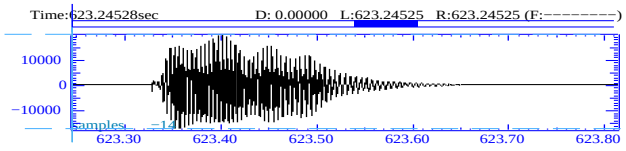

Figure 2: The Xwaves waveform plot for a short and long tone word.

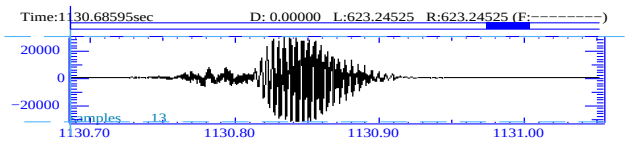

Figure 3: The Xwaves waveform plot for a short and long tone word.

Previous attempts have been made at extracting energy features from words. Lee [6] introduced an energy factor that is the slope from the maximum to the 10% threshold of the maximum. The problem with such simple factors is that the drop in energy may occur just at the end of the word and will in such cases provide a similar result as for an ordinary long word.

3.1 Energy features

Our contribution is to introduce a model of energy that more generally describes the shape of the energy contour irrespective of local variations, absolute word intensity and word duration. Such a model has existed for some time, namely the ADSR model (attack, decay, sustain and release). This model is used to model the envelope contour of complex musical instruments . The model approximates the energy contour using four line segments, namely, the attack going from the start to the maximum peak, the decay going from the maximum down to the sustain level which is held for the sustain period, followed by the release that is the slope from the end of the sustain down to the end of the word. Figure 4 illustrates the ADSR parameters. In practice, it is hard to extract or approximate a real word envelope using the ADSR model as the end changes are subtle. Therefore, we use a specialised adaptation.

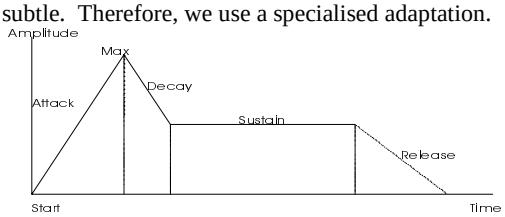

Figure 4: The ADSR showing how the energy of a sound changes over time.

3.2. Feature extraction algorithm

The first thing to note is that the start segments of the words are similar, thus we ignore the attack (the attack is the segment from the start of the word to the maximum RMS amplitude). The real question is how the slope of the energy contour changes with time. The following approach was adopted: The segment of the word starting at the maximum energy peak to the end of the word is split into three parts, the start part, middle part and end part. For each part a histogram of slopes is generated. The histograms summarise the distribution of the energy contour, i.e. the percentage of the word with level slope, the percentage of the word with medium slope and the percentage of the word with a steep slope. Only negative slopes are considered since we are interested in the nature of the word intensity decay. Figure 5 illustrates the steps involved.

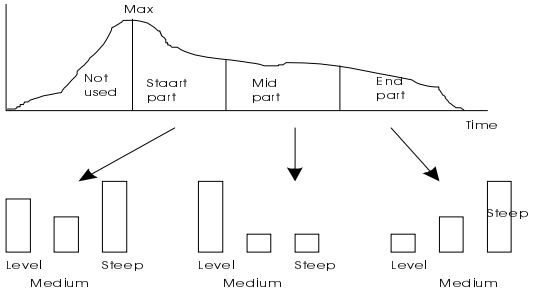

Figure 5: Extracting energy contour features from a Taiwanese word.

One advantage of this representation is that a complex energy contour is modelled using a vector of 9 scalars. Further, it encompasses the change in slope over time and it is independent of absolute word length and intensity.

3.3. Results

Table 2 shows the results of applying this approach to the energy contours of our speech material. The table shows the average of the nine parameters for each of the seven tones and the average parameters for all the long and the short tones. Figure 6 shows the data in Table 2 diagramatically.

		start			medium			end		
type	Tone	level	Med	steep	level	Med	Steep	level	med	Steep
short	t1	26.3	2.6	0.3	31.6	0.1	0.0	35.7	1.5	0.0
	t2	21.0	6.2	1.3	28.2	4.9	0.4	31.0	3.7	1.3
	t3	23.9	4.8	0.2	31.3	2.8	0.1	33.0	1.7	0.2
	t5	33.6	0.8	0.0	21.1	0.7	0.2	37.4	4.2	0.1
	t7	26.7	1.6	0.3	32.2	0.2	0.0	36.7	0.9	0.0
	mean	26.3	3.2	0.4	28.9	1.7	0.1	34.7	2.4	0.3
short	t4	16.4	5.6	1.8	19.9	9.7	3.1	20.4	10.7	6.4
	t8	17.4	3.4	1.2	20.7	7.9	2.6	20.0	9.2	6.7
	mean	16.8	4.5	1.5	20.3	8.8	2.9	20.7	10.0	6.6

Table 2: Average energy features of the seven Taiwanese tones and the average vector for all long and short tones.

Figure 6: A graphical representation of the average energy features.

The data suggests two distinct trends. The long tones have a higher percentage of level segments especially in the start part, while the short tone words have a higher percentage of steep

and medium slope segments especially in the middle and end parts.

4. SUMMARY

In this paper we showed that classifying Taiwanese using the fundamental frequency is difficult and that the task is simplified combining the use of the fundamental frequency contour with the use of energy contours. A strategy for extracting energy features was presented. Based on actual acoustical measurement we showed that the seven tones can be classified successfully.

5. REFERENCES

1. F.H.Jian, Perception of long and short tones in Taiwanese Speech, *The Journal of the Acoustical Society of America*, vol. 102, No.5, Pt.2 (Abstract), 1997.
2. R.L.Cheng, *Taiwanese and Mandarin structures and their developmental trends in Taiwan*, Yuan-Liou Publishing Co., Ltd., 1997.
3. W.J. Yang, J.C.Lee, Y.C.Chang and H.C.Wang, Hidden Markov model for Mandarin lexical tone recogniton, *IEEE Trans. Acoust., Speech, Signal Processing*, vol.36, pp.988-992, 1988.
4. H. Ma, Chinese four tones recognition based on a simplified vector quantization method, *Proc. ICPR86*, Paris, 1986.
5. X.Hu and K.Hirose, Recognition of Chinese tones in monosyllabic and disyllabic speech using HMM, *Proc. ICSLP 94*, Yokohama, vol.1, pp.203-206, 1994. *
6. Tan Lee et. al., Tone recognition of isolated Cantonese syllables, *IEEE Trans. on speech and audio processing*, 1995.
7. J.K.Chen, F.K.Soong and L.S. Lee, Large vocabulary word recognition based on tree-trellis search, *ICASSP*, pp.II 137-140, 1994. *
8. T.J.Vance, Tonal distinctions in Cantonese, *Phonetica* 34:93-107, 1977.
9. X-N.S.Shen and M.Lin, A perceptual study of Mandarin tones 2 and 3, *Language and speech*, 34(2), 145-156, 1991.
10. J.Gandour, On the interaction between tone and vowel length: evidence from Thai dialects, *Phonetica*, 34:54-65, 1977.
11. T.Lee, P.C.Ching, L.W.Chan, Y.H.Cheng and B.Mak, Tone recognition of isolated Cantonese syllables, *IEEE Transactions on speech and audio processing*, vol.3, no.3, 1995.
12. Fangxin Chen and Anton J. Rozsypal, Computer modelling of lexical tone perception, *Canadian Acoustics*, 19, no.4, 103-104", 1991. *
13. Davies, P., Hidden Markov Modelling of modern standard Chinese tones in connected speech, *Speech Hearing and Language*, vol.3, 1989.
14. Y.P.A Ng, P.C.Ching and L.W.Chan, Automatic recognition of Cantonese lexical tones in connected speech by multi-layer perception, *ESCA. EUROSPEECH 1995 4TH European conference on speech communication and technology*, 1995.
15. Sin-Horng Chen and Yih-Ru Wang, Tone recognition of continuous Mandarin speech based on neural networks, *IEEE Transactions on speech and audio processing*, vol. 3, no.2, 1995.
16. K.Hirose and X.Hu, HMM-based tone recognition of Chinese trisyllables using double codebooks on fundamental frequency and waveform power, *ESCA. EUROSPEECH'95 4TH*, 1995.
17. Jung-Kuei Chen, Lin-Shan Lee and Frank K. Soong, Large vocabulary, word-based Mandarin dictation system, *ESCA.EUROSPEECH'95 4TH*, 1995. *
18. Lin-Shan Lee, Chiu-Yu Tseng and Hung-Yan Gu et. al., Golden Mandarin (I)--a real time Mandarin speech dictation machine for Chinese language with very large vocabulary, *IEEE Transactions on speech and audio processing*, vol.1, no.2, 1993.*
19. Jack Gandour, Tone perception in Far Eastern languages, *Journal of Phonetics*, 11, 231-242, 1983.
20. Yuh-Shiow Lee, Douglas A.Vakoch and Lee H.Wurm, Tone perception in Cantonese and Mandarin:a cross-linguistic comparison, *Journal of Psycholinguistic Research*, vol.25, no.5, 1996.