BOUNDARIES OF PERCEPTION OF LONG TONES IN TAIWANESE SPEECH

Fran H. L. Jian

Dept. Linguistics, University of Reading, England

ABSTRACT

In this work we set out to investigate the fundamental frequency boundaries of perception of the Taiwanese long tones. We are interested in how the variations in fundamental frequency affect the perception of linguistic tones in Taiwanese speech. Our investigation is adopted from similar studies of tones in Mandarin speech. As opposed to Mandarin tones that can be perceived with little difficulty the seven Taiwanese tones have a more subtle structure and are consequently harder to perceive successfully. The experimental results in this paper allow us to quantify these perceptual boundaries. The experiments consisted of a perception test involving over 150 Taiwanese subjects where the task involved identifying the tone of the words played back in a random sequence. The stimuli consisted of a set of tone pairs and a selection of intermediate tone words obtained by linearly interpolating between the words of the tone pairs.

1. INTRODUCTION

Taiwanese (or Hokkien and Fukienese) is an ancient Chinese dialect originating from the Fujian province in the southeast of China. It is widely spoken in Taiwan although for political reasons Mandarin is the official language of Taiwan. Taiwanese is mutually unintelligible from Mandarin and consequently there are linguistic barriers within the Taiwanese society.

1.1. Tone languages

In common with many ancient Asian tongues such as Thai, Cantonese and Mandarin, Taiwanese is a tone language, where the interpretation of an utterance depends on the tone or fundamental frequency contours applied to the utterance. Taiwanese has seven tones, where there are two level tones, four falling tones and one falling-rising tone. The tones are classified into two main groups according to the syllable endings. Syllables ending in voiceless stops [p, t, k, ?] (i.e. ‘checked’ syllables) are known as 'entering’ or ‘short’ tones. All others (i.e. ‘free’ syllables) are referred to as ‘non-entering’ or ‘long’ tones [1], [2], [3], [4], [5]. We adopt the latter terms to refer to the Taiwanese tonal inventories, namely the ‘long’ and ‘short’ tones. In this paper our primary interest is in the perception of fundamental frequency, so the two short tones will not be included, as it has been found that fundamental frequency is not the primary cue in perceiving the short tones ([9]). We will therefore only consider the five long tones: the high level tone t1, high falling tone t2, mid falling tone t3, falling-rising tone t5 and mid level tone t7. The understanding of the linguistic tones is an essential pre-requisite for the successful implementation of automatic Taiwanese speech recognition systems.

1.2. Perception of Taiwanese tones

Most documented work into Chinese languages focuses on standard Mandarin Chinese (e.g. [6]). Although, some work has been reported on the study of Taiwanese tones, most of this research is based on impressionistic evaluations of the tones. Only a few papers report results based on actual acoustic measurements. Experiments investigating the perception of Taiwanese tones have been limited to the verification of the fundamental frequency as the main cue for long tones [7], and the effects of coarticulation on tones [8].

In this work we set out to investigate the frequency boundaries of perception of the Taiwanese long tones and how the variations in fundamental frequency affect the perception of linguistic tones in Taiwanese speech. Our experiment is adapted from similar work carried out on tones in Mandarin speech [10].

The remainder of this paper is organised as follows: In section 2 the perceptual experiment is described, followed by the results in section 3. The results are discussed in section 4 and the paper is summarised in section 5.

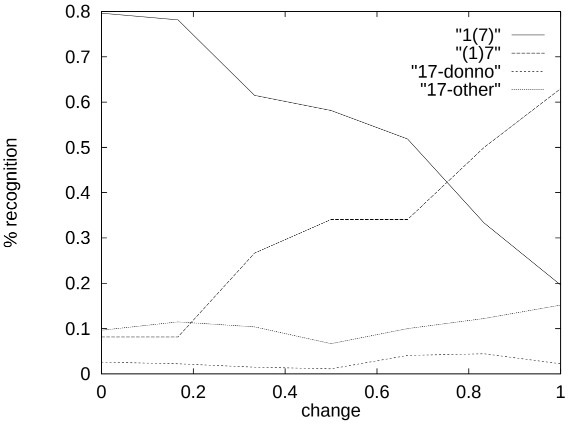
2. METHOD

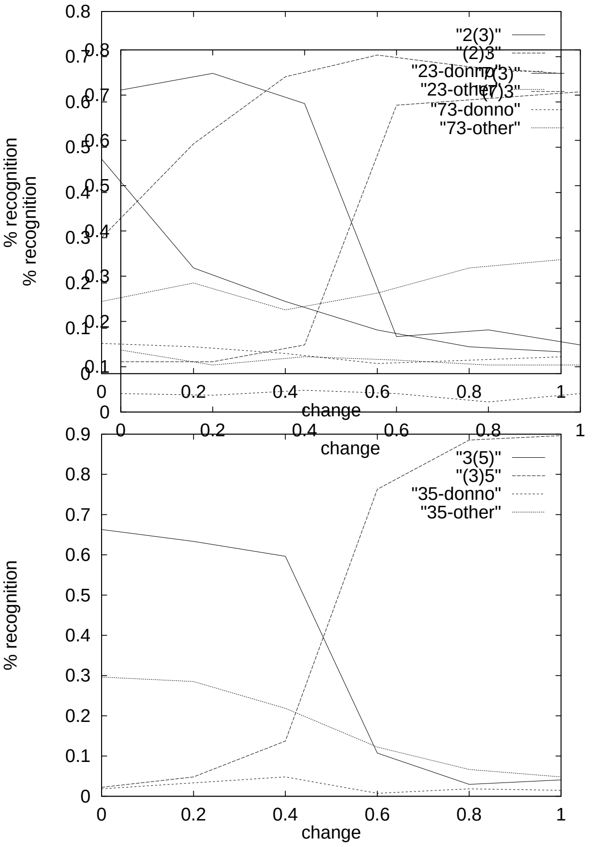
2.1. Subjects

Over 150 subjects participated in the perception tests. About 30 of the subjects were Taiwanese students studying at the University of Reading with good English skills, 30 subjects were taken from the general population of Kaoshiung (a harbour city in the southwest of Taiwan) and the remaining subjects were 13 and 14 years old school-children from Kaoshiung and Pin-dong respectively. These subjects did not speak English regularly.

2.2. Stimuli

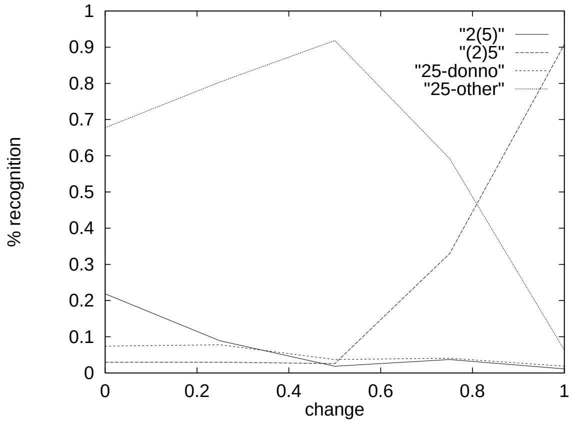
In the experiment we generated a set of stimuli by altering the fundamental frequency of a base word into the five long tones and intermediate tones between tone pairs. There are N(N- 1)/2 = 5*4/2 = 10 possible long tone pairs. However, we only considered pairs that were similar in frequency contour, namely the tone pairs t1/t2, t1/t7, t2/t3, t2/t5, t3/t5, t3/t7 and t5/t7. The pairs t1/t3, t1/t5 and t2/t7 were not considered as they are unlikely to be confused. Further, this reduces the total number of words in our perception test by 30%. A t1 word was chosen

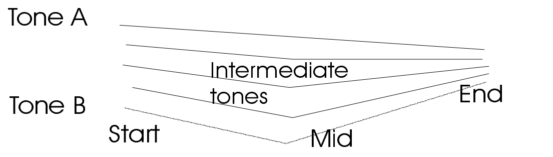




as the base word since it is level in frequency and introduces a minimum of acoustical artifacts during re-synthesis. The fundamental frequency contours of the words were represented using two line segments connected by the start mid and end frequencies.

The re-synthesis was performed by superimposing the five tones on the base word using PSOLA (Pitch Synchronous OverLap and Add) a re- synthesis/analysis extension to xwaves [11]. Eight intermediate tones were generated for each tone pair by linear interpolaption. These eight intermediate tones represent a "morph" between the two tones in the tone pair. Figure 1 shows how these intermediate tones where obtained.





Inte

. **Figure 1: Interpolation between tone pairs.**

The fundamental frequency contours used in the synthesis were acquired from a set of acoustic measurements.

2.3. Procedure

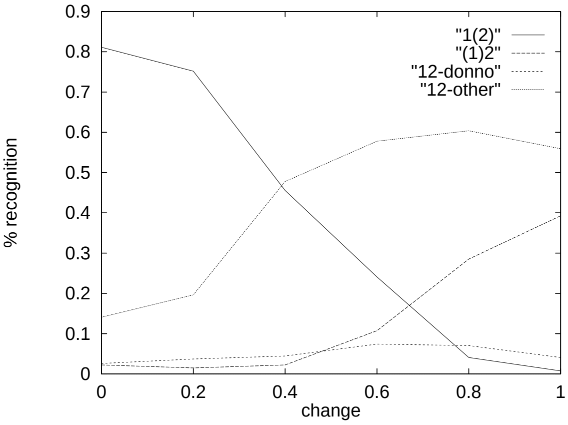
The stimuli consisted of 70 words, which were shuffled into random order twice to form two sets presented to the subjects separated by an intermission. Each word was prefixed by a reference frame (in Taiwanese) saying: "the next word is ...". The questionnaire issued to the subjects consisted of six boxes for each word aligned on individually numbered lines. Each box was associated with a character representing the word for each of the five long tones and one box was provided for subjects to indicate "don't know".

3. RESULTS

The results consist of two sets of graphs. The first set shows the recognition rate for the tone pairs, rate of missing replies and rate of incorrect replies. Incorrect replies are those not matching the tones in the tone pair.

The horizontal axis shows the interpolation step in percentage between the two tones. A percentage of 0 means that the words are completely tone A, while 100% means the work is totally tone B, and 50% indicate that the frequency contour of the word lie halfway between the frequency contours of the words A and B. The vertical axis shows the percentage of replies for the particular condition.

The second group of plots shows the breakdown of the incorrect replies, i.e. which tones were actually perceived by the subjects. Plots are only given for the tone pairs that had a high degree of incorrect replies.



4. DISCUSSION

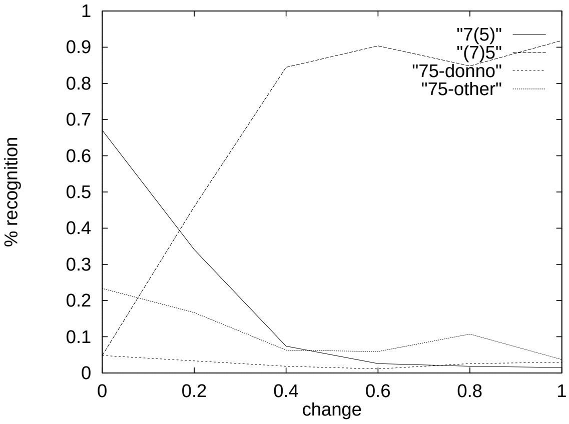
t1/t2. As expected at the 0% interpolation point most subjects recognise the words as t1. As the words are interpolated towards t2 the number of subject recognising the words as t1 decreases. Overall, less than 10% of the subjects reply that they cannot identify the tone. This suggests that the subjects are relatively confident in their judgement. Surprisingly, most subjects recognise t2 incorrectly as other tones. At the 100% interpolation point only 40% of the subjects recognise the words as t2 while the majority of 58% subjects recognise the words incorrectly as different tones.

t1/t7 Less than 10% of the subjects reply that they cannot identify the tone or supply incorrect replies. Moreover, at the 0% interpolation point most subjects recognise the words as t1. Similarly, at the 100% point most subjects recognise the words as t7. The crossing of the recognition lines for t1 and t7 occur at the 75% interpolation point. This indicates that the words must be drastically interpolated towards t7 before words are consistently recognised as t7. Thus, the recognition of t1 is more dominant than t7.

t2/t3 Few subjects reply that they cannot identify the tone. However, about 20% of the subjects supply incorrect replies for the entire range of interpolation points. Further, at the 0% interpolation point, the small majority of 48% subjects respond that they hear t2. The number of subjects that answer t2 quickly decreases as the words are interpolated towards t3. Further, at the 0% interpolation point a massive 30% reply that they hear t3. This percentage steadily increase with the interpolation points. The t2 and t3 curves cross at the 10% interpolation point, suggesting that t3 strongly dominate t2.

t2/t5 Less than 10% of the subjects reply that they cannot identify the tone. For most of the interpolation range (0-80%) most subjects recognise the words incorrectly. At the 100% interpolation point most subjects successfully recognise the words as t5. However, the recognition of t5 only excels after the 60% interpolation point.

t3/t5 Most subjects successfully recognise t3 words at the 0% interpolation point. At the 100% interpolation point most subjects recognise the words as t5. The crossing of the two curves occurs at the 50% interpolation point, i.e. midway between the tone-pair. At the 0% interpolation point about 30% of the subjects misidentify the tone. This percentage decrease as the interpolation percentage is increased. This implies that t3 is easily confused with other tones, but as we approach t5 the ambiguity diminishes gradually.



t7/t3 Most subjects recognise t7 and t3 successfully. The crossing of the two curves occurs at the 50% interpolation point. About 2.5 % of the subjects reply that they do not know and about 10% of the subjects provide incorrect answers.

t7/t5 Both t5 and t7 are recognised with great accuracy at the respective start and end positions on the interpolation scale. However, the crossing of the two curves representing t7 and t5 occurs at the 20% point on the interpolation scale. Thus, the recognition of t7 decreases quickly while the recognition of t5 dominates the interpolation range. Further, the curves indicate that t7 words are misidentified at a rate of 22% at the 0% interpolation point.

The other set of plots shows how the incorrect replies were distributed.

t1/t2-other This plot shows that 45% of the subjects recognise the words as t7. Further, as we interpolate towards t2 more subjects recognise the words as t3. Thus, t2 and t3 may be easily confused. Less than 5% of the subjects recognised words as t5.

t2/t3-other This plot shows that 14% of the subjects recognised t2 and t3 as t7.

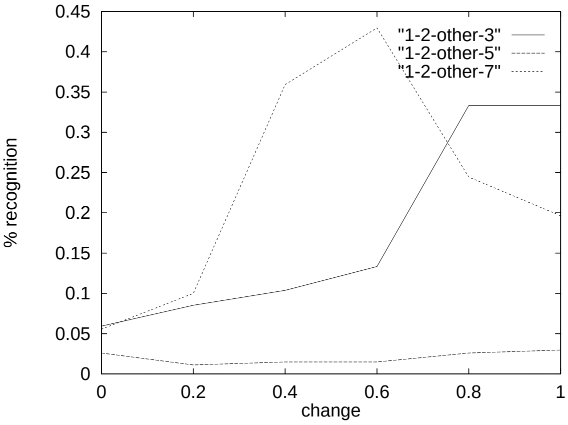
t2/t5-other This plot shows that 80% of the subjects recognise the words incorrectly as t7. The remaining subjects misidentify the tones as t3 (5-15% of the subject) and t1 (1-7% of the subjects)

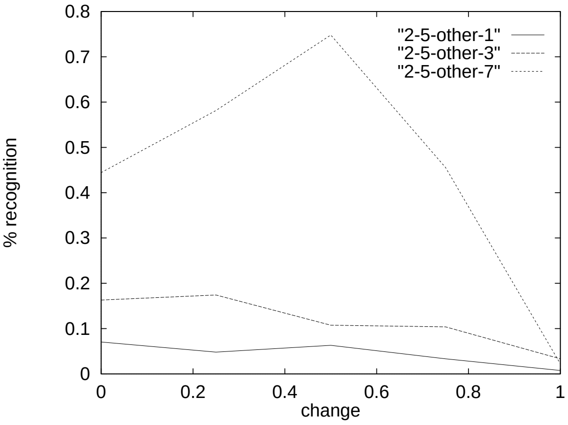
t3/t5-other At the 0% interpolation point about 20% of the subjects recognise the words as t7.

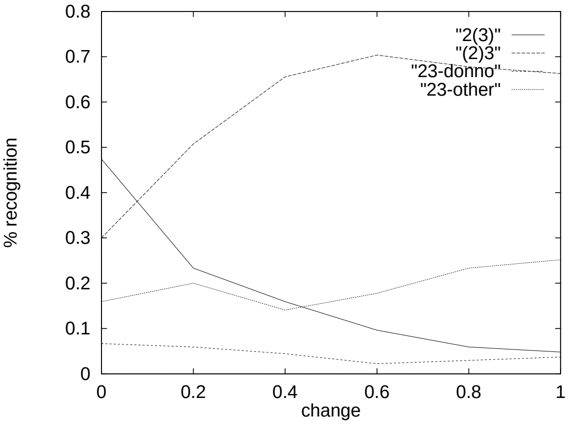
In general, the 50% region in the interpolation range did not appear to confuse subjects as there are few peaks in the curves depicting the "do not know" replies. One could expect to see small peaks in these curves at the boundary between tone-pairs.

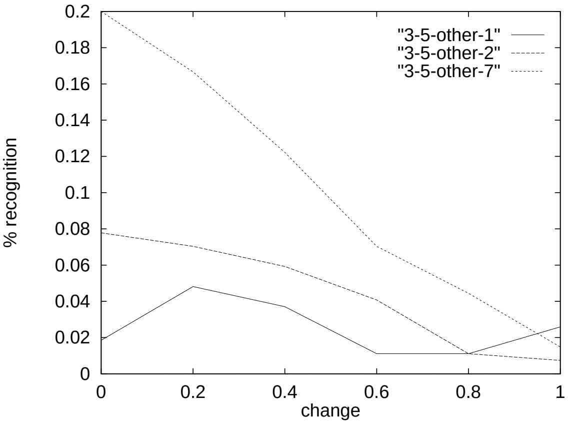
5. SUMMARY

Subjects appear confident in their replies as less than 10% of the subjects reports that they cannot identify the tones. The









experiments revealed that subjects had no difficulty in distinguishing between t1, t3, t5 and t7. However, the identification of t2 was sensitive to noise. By interpolating between t1 and t2, intermediate tones were perceived as t7. Similar patterns could be seen by interpolating between t2 and t3 and t2 and t5. Further, the experiments revealed that some tones were recessive with respect to other tones, i.e. they were more difficult to perceive. Surprisingly, the experiments showed that subjects were confident in their evaluations even half way between the two pairs of a tone pair.

7. REFERENCES

1. Chin-Chu Yang, *Bilingual dictionary of Mandarin and Taiwanese*, Dun Li Publishing Co. Inc., 1996.

2. K.T.Hsu, *A study of the Taiwanese sound pattern representation and orthography*, Wen-He Ltd, 1988.

3. Eric Zee, Duration and intensity as correlates of F0, *Journal of Phonetics*, 6, 213-220, 1978.

4. R.L.Cheng, *Taiwanese and Mandarin structures and their evelopmental trends in Taiwan*, Yuan-Liou Publishing Co., Ltd., 1997.

5. Tsai-Chwun Du, *Tone and stress in Taiwanese,* University of Illinois at Urbana-Champaign, Ph.D. Thesis, 1988.

6. M.C.Lin,The acoustic characteristics and perceptual cues of standard Chinese, *Zhongguo Yuwen* 204:182-193, 1988.

7. H.B.Lin and B.Repp, Cues to the perception of Taiwanese tones, *Language and Speech*, 32 (1), 25-44, 1989.

8. Shu-Hui Peng, Production and perception of Taiwanese tones in different tonal and prosodic contexts, *Journal of Phonetics*, 25, 371-400, 1997.

9. 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.

10. W.S.Chan and C.K.Chuang and W.S-Y.Wang, Crosslanguage study of categorical perception for lexical tone, *Journal of the Acoustical Society of America*, 58, 119 (Abstract) 1975.

11. Gregor Mohler, Rule Based Generation of Fundamental Frequency Contours for German Utterances, *Proceedings of the 2^nd^ "Speak!" Workshop*, Darmstadt, (abstract) 1995.