Lexical Tone and Pitch Perception in Tone and Non-Tone Language Speakers

Barbara Schwanhäußer & Denis Burnham

MARCS Auditory Laboratories, University of Western Sydney, Australia  
b.schwanhaeusser@uws.edu.au

Abstract

Past research on categorical perception of lexical tone has produced contradictory results. In Experiment 1 tonal (Mandarin, Vietnamese) and non-tonal (Australian) adults were tested for identification and discrimination on speech and non-speech (sine-wave) tone continua. Tonal language speakers’ category boundaries and discrimination peaks were near the middle of the asymmetric continuum, whereas non-tonal speakers used an acoustically flat stimulus as a reference, indicating that tone space is linguistically oriented in tonal, and acoustically oriented in non-tonal language speakers. In Experiment 2, three tonal-language (Thai) groups (musicians, perfect pitch musicians, and non-musicians) were tested on two new continua represented as speech or sine-wave tones. Identification boundaries were in the middle of the continuum for most participants. In discrimination, the flat stimulus was used as a perceptual anchor, and this was independent of musical background, indicating that the musical Thai participants use the same mid-continuum strategy as the Mandarin and Vietnamese speakers in identification, but the flat no-contour strategy in discrimination. Hence, perception depends on the type of task in Thai speakers: it is linguistic in identification, but acoustic in discrimination.

1. Introduction

In categorical perception stimuli equally spaced along a physical continuum spanning two categories are perceived as members of one or the other category with little perceptual ambiguity, and stimuli within categories are difficult to differentiate [1]. Categorical perception is strong for stop consonant contrasts, whereas for vowels perception is continuous: There are no sharp identification functions and discrimination is accurate over the entire continuum [1]. In tone languages pitch differences can be sufficient to distinguish lexical meaning. As tone is carried on vowels, it might be expected to be perceived continuously. Perception of Mandarin tones by native speakers has been found to be categorical, but noncategorical in non-tone (English) speakers [2]. However, noncategorical perception of tones has also been reported for Thai [see 2] and Cantonese [3] tone speakers under certain conditions. Continua used in previous studies were symmetrical, and the focus was on the lexical status of the word rather than categorical tone perception. In the experiments here asymmetrical tone continua were used to examine the influence of continuum shape on strategies in tone perception. Following a recent study [4], both speech and sine-wave tone continua were presented to investigate differences in tone processing in speech vs. non-speech modes. Artificial tone continua were used in order to concentrate on strategies that speakers bring to the task.

2. Experiment 1: Tone & Pitch: Tone & Non-Tone Language Speakers

Experiment 1 is based on a study in which categorical identification was found to be found greater for tone language (Thai) than non-tone language (Australian English) speakers; and for tones in speech than in any of the three non-speech variants (equivalent F0 variations in filtered speech, violin notes, or sine-waves) [4]. As there was no difference between perception of these variants, here only speech and sine-wave stimuli were used. These were tested in separate blocks in that study, but here, to test whether the categorical speech perception carries over to non-speech, (i.e., whether speech and non-speech perception are modular), speech and non-speech stimuli were presented either in separate blocks, or in speech/non-speech mixed blocks. Finally, both identification and discrimination tests were included [4], because both are required to test categorical perception adequately.

2.1. Subjects

47 participants with normal hearing were tested (16 Mandarin; 16 Vietnamese; 15 Australian English speakers), all students at the University of NSW (mean age 25 years, range 18-31), who were paid for participation.

2.2. Design, Stimuli, and Procedure

Design: 3 x 2 x (2) native language (Vietnamese, Mandarin, Australian English) x presentation (mixed, blocked) x within-group tone type (speech, sine-wave).

Stimuli: Speech and sine-wave continua were created with a fixed onset (200Hz) and offset from 160Hz (falling) to 220Hz (rising) in 10 Hz intervals. Thus continua were asymmetrical and consisted of 7 steps, with the 190Hz offset stimulus at middle of the continuum, and the 200Hz offset stimulus as a flat no-contour stimulus. F0 movement was linear to avoid resemblance to tones of specific languages. The speech sound was /wa/, as it was unlikely to be heard as an independent word in any of the languages. The MARCS Auditory Perceptual Toolbox (APT) [5] was used for sine-wave resynthesis, and tone and F0 were matched by resynthesizing speech and sine-waves to the same F0 and duration (495ms) with STRAIGHTv30kr16 [6].

Procedure: Participants were tested individually in a single session. Stimuli were presented on a laptop computer (Compaq Evo N1000c) over headphones (KOSS UR20) at a comfortable listening level in the DMDX [7] experimental environment.

Identification Procedure: Participants were instructed to “press the RIGHT (LEFT) key for one kind of sound and the LEFT (RIGHT) key for the other kind”. Responses timed out after 4 secs, and such trials were not replaced. There were 2
trial blocks, each with practice (8 items), and training phases. After reaching a criterion of 8 consecutive correct responses, the test phase began – 8 repetitions of each item, presented in random order. Given 7 steps on each continuum, there were 112 items. In each group, half the participants had separate speech and sine-wave blocks and the other half had speech and sine-wave mixed within blocks. The task took 15-20 min. For each listener, two crossover (in Hz) and two $d'$ values were computed, one for each tone type. Crossovers were computed by running a logistic regression for each listener and dividing the constant by the slope at the 50% boundary (because log (0.5/0.5) = 0 = constant + slope (crossover)), and converting to Hz. For $d'$ (to measure category boundary sharpness) values were computed for the perceptual distance between the stimuli spanning the 50% crossover. The proportion of responses for one of the two categories for each stimulus was converted to a z-score, then for each stimulus pair, the z-score for the smaller proportion was subtracted from that for the larger proportion to derive the $d'$. To avoid inflated $d'$ estimates, response category proportions of 0 and 1 were converted to 0.005 and 0.995 (1/(2N) and 1-1/(2N)) [8].

**Discrimination Procedure:** Participants listened to stimulus pairs and were instructed to “press the LEFT (RIGHT) key if they are the same sound, and press the RIGHT (LEFT) key if they are different sounds”. Responses timed out after 15000 ms. Omitted trials were not replaced. The task was separated into two parts. Both parts started with 8 practice items, but there were no training trials. Again, there were blocked and mixed manners of presentation. In the former, participants listened to a block of speech stimuli followed by a sine-wave block (or vice versa), in the mixed mode speech and sine-wave stimuli were presented randomly in the same blocks. A roving AX paradigm was used, measuring discrimination accuracy along the whole continuum. Neighbouring stimuli were presented pairwise. The 8 practice trials (4 different, 4 same pairs) were presented with feedback. In test trials there were 4 repetitions (AA, BB, AB, BA) of each possible stimulus-pair for each tone type ($\Sigma=192$). The task took 20-25 minutes. Performance was measured by $d'$, calculated according to models for separate discrimination tasks [9]. The p(Hit) and p(False Alarm) values were forcibly limited to the 0.99 to 0.01 range, so that the maximum possible $d'$ value was 8.715.

### 2.3. Hypotheses

It was hypothesized that perception should be more categorical (i) by tone than non-tonal language speakers [see 4]; (ii) for speech than sine-waves [4]; and (iii) in blocked than mixed presentations, because speech/non-speech mixing would cause interaction of processing modes. In addition, two alternative response strategies were deemed possible. These are schematically presented in Figure 1, and described below.

**Mid-Continuum Response Strategy:** It might be expected that the category boundary and discrimination peak for this synthetic asymmetric continuum would be midway between the extreme continuum values. If so then (i) the identification boundary should lie near the flat 200 Hz stimulus, and (ii) the stimuli around the 200 Hz stimulus should be more discriminable than those at the ends of the continuum.

**Flat-Anchor Response Strategy:** It might alternatively be expected that the category boundary and discrimination peak for this synthetic asymmetric continuum would be near the flat no-contour tone (200 Hz). If so then (i) the identification boundary should lie near the flat 200 Hz stimulus, and (ii) the stimuli around the 200 Hz stimulus should be more discriminable than those at the ends of the continuum.

![Figure 1: Possible Response Flat, Mid-Continuum and Anchor, for the asymmetric continuum in experiment 1.](image)

#### 2.4. Results

**Identification:** Mean crossover values and slopes across listener groups, contexts and presentation are shown in Table 1. Planned comparisons within a 3 x 2 x (2) language x presentation mode x stimulus type analysis of variance (ANOVA), with repeated measures on stimulus type were conducted on crossover boundary values. Crossover values of tonal language speakers (Vietnamese, Mandarin) combined were significantly lower than for the non-tonal language speakers ($M_{tonal}=187$, $SD_{tonal}=194$, $SD_{non-tonal}=5.5$; $F(1,31)=10.00$, indicating that tonal speakers perceptually halve the continuum, whereas the non-tonal speakers divide the continuum with the acoustically salient flat 200 Hz stimulus). Crossovers for speech were significantly lower than for sine-waves, $F(1,31)=6.76$ ($M_{speech}=187$, $SD_{speech}=7.96$; $M_{sine}=190$, $SD_{sine}=6.97$), indicating a difference between speech and non-speech processing.

<table>
<thead>
<tr>
<th>Language Group</th>
<th>Vietnamese</th>
<th>Mandarin</th>
<th>Australian</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Crossover</td>
<td>Slope</td>
<td>Crossover</td>
</tr>
<tr>
<td>Mixed</td>
<td>Sine</td>
<td>191</td>
<td>189</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>187</td>
<td>1.60</td>
</tr>
<tr>
<td></td>
<td><strong>Means</strong></td>
<td><strong>190.4</strong></td>
<td><strong>1.78</strong></td>
</tr>
<tr>
<td>Blocked</td>
<td>Sine</td>
<td>188</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>188</td>
<td>1.25</td>
</tr>
<tr>
<td></td>
<td><strong>Means</strong></td>
<td><strong>188.2</strong></td>
<td><strong>1.83</strong></td>
</tr>
<tr>
<td><strong>Grand Means</strong></td>
<td>189.2</td>
<td>1.85</td>
<td>1.78</td>
</tr>
</tbody>
</table>

Table 1: Mean crossovers (Hz) and slopes ($d'$) across listener groups, contexts, and presentation type

For the slope values, planned contrasts in ANOVA showed that the presentation manner x language background interaction approached significance, $F(1,29)=4.17$; $F_{rural}(1,29)=4.18$. Identification is more categorical in mixed than blocked conditions for non-tonal ($M_{mixed}=1.89$, $SD_{mixed}=5.4$; $M_{blocked}=1.68$, $SD_{blocked}=3.5$) but not tonal speakers. ($M_{tonal}=1.49$, $SD_{tonal}=69$; $M_{non-tonal}=1.54$, $SD_{non-tonal}=56$). So, tonal language speakers’ perception appears resistant to context effects, and non-tonal language speakers show greater categoricality in mixed presentations.
Discrimination: Table 2 shows d’ values with the middle and the flat stimuli as reference points for tone and non-tonal language speakers respectively. Planned contrasts within a 3 x 2 x (2) ANOVA revealed no significant differences between sine-wave and speech perception, and no mixed vs. blocked differences. The results indicate that tone language speakers are better at discriminating stimuli in the middle of the continuum than the other stimuli, (Mandarin: M_mid=2.91, SD=3.34; M_flat=1.81, SD=3.44; Vietnamese: M_mid=2.57, SD=3.36; M_flat=1.47, SD=3.39; F(1,34)=4.70), whereas the Australian listeners are better at discriminating the stimuli around the flat tone than the other tones (M_mid=3.68, SD=46; M_flat=3.21, SD=92 (F(1,10)=5.92)).

2.5. Discussion
The results show that tonal language speakers perceptually bisect the available tone space (suggesting perception in a linguistic manner); whereas non-tonal language speakers tend to divide the tone space into stimuli above and below the acoustically salient flat no-contour stimulus (suggesting an acoustic approach). There also appears to be a processing difference for identification of speech vs. sine-wave stimuli, but in discrimination, sine-wave and speech are treated similarly. It is possible that the cognitive load is greater in discrimination (two stimuli compared) than in identification (of one stimulus at a time).

3. Experiment 2: Tone & Pitch: Musical and Non-musical Tone Language Speakers
On the basis of the strategy differences found in Experiment 1 for tone and non-tone language speakers, Experiment 2 investigated differences in categorical tone perception for two differently shaped asymmetric continua to accentuate any difference between the mid-continuum, and flat-anchor strategies. Three Thai (tone language) groups participated: non-musicians, musicians, and perfect pitch musicians. As in Experiment 1 both identification and discrimination and speech and sine-wave continua were included (presented in separate blocks or mixed within blocks).

3.1. Subjects
48 Thai-speaking adults were tested (16 non-musicians, 16 musicians and 16 perfect pitch musicians - determined by a perfect pitch perception test). All were students at Chula- longkorn University, Bangkok (mean age 21 yrs, range 18-27 yrs), and paid for participation.

3.2. Design, Stimuli, and Procedure
Design: 2 x 2 x (2) presentation manner (mixed or blocked) x continuum shape (falling or rising) x a within-group factor tone type (speech and sine-wave).

Stimuli: Two continua, falling and rising, were created for each tone type, with a fixed onset (220Hz) and offsets from 182.5 to 235Hz in the falling continuum; and 205 to 257.5Hz in the rising continuum (step size: 7.5Hz). So each continuum consisted of 8 steps. F0 movement in all tones was linear to avoid resemblance to independent Thai words.

Identification Procedure: In test trials, there were 8 repetitions of each of the 8 steps on each of two continua (Σ=128).

Discrimination Procedure: In test trials neighbouring stimuli were presented pairwise with 4 repetitions (AA, BB, AB, BA) of each possible stimulus-pair in each tone type (Σ=224).

3.3. Hypotheses
It was hypothesized that musicians (perfect pitch and other) would show better categorical perception than non-musicians. Additionally, it was expected that non-musicians would favour the flat (acoustic) strategy for non-speech stimuli and the mid strategy (linguistic) for speech sounds (see 2.5), as they are familiar with tones only in the linguistic context; whereas musicians should favour the mid-continuum strategy for both speech and non-speech tones, as they would be familiar with pitch variations in both contexts. Over and above this, absolute pitch musicians may show less categorical perception than either the musicians or the non-musicians because their skill in identifying pitch absolutely would mitigate against their clumping together into categories pitches perceived to be different.

3.4. Results
Identification: There is no significant difference between the crossovers for speech and sine-wave in the falling continuum, but in the rising continuum, the crossover for speech lies below the middle, whereas the crossover for sine sounds lies above the middle of the continuum (M_falling=231, SD=9.71; M_sine=212, SD=7.55; M_perfect pitch=214, SD=11.71; rising: M_falling=232, SD=4.09, M_sine=225, SD=6.29; M_perfect pitch=230, SD=4.63).

Table 2: Mean d’ discrimination scores across listener groups, contexts, and presentations for reference and other stimuli

Table 3: Means and standard deviations for identification and discrimination scores across listener groups, contexts, and presentations for reference and other stimuli.
Planned contrasts in a 3 x 2 x (2) ANOVA showed blocked stimuli were perceived more categorically than mixed stimuli across listener groups, continua and stimulus types (F(1,31) = 4.28), M\text{Musicians}=1.66, SD=0.63; M\text{Mixed}=1.34, SD=0.57). As hypothesised musicians perceive tones more categorically than do the perfect pitch musicians (F(1,14)=4.60), (M\text{Musicians}=1.71, SD=0.60; M\text{Perfect Pitch}=1.32, SD=0.61). Also interesting to note the relationship between speech and sine-wave perception for the three groups: While there was little difference in slope for speech and sine-wave stimuli for the absolute pitch musicians, musicians showed greater slopes for sine-waves than speech, and the non-musicians showed greater slopes for speech than sine-waves (see Figure 2).

### Table 4: Mean crossovers (Hz) and slopes (d’) for the rising continuum over listener groups, contexts, & presentation type

<table>
<thead>
<tr>
<th>Musical Background</th>
<th>Non-Musicians</th>
<th>Musicians</th>
<th>Perfect Pitch Musicians</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Slope</td>
<td>Slope</td>
<td>Slope</td>
</tr>
<tr>
<td>Rising Continuum</td>
<td>Cross-over</td>
<td>Cross-over</td>
<td>Cross-over</td>
</tr>
<tr>
<td>Mixed</td>
<td>Sine</td>
<td>235</td>
<td>1.44</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>229</td>
<td>1.34</td>
</tr>
<tr>
<td>Means</td>
<td>Mixed</td>
<td>232</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>231</td>
<td>1.80</td>
</tr>
<tr>
<td>Blocked</td>
<td>Sine</td>
<td>233</td>
<td>1.33</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>232</td>
<td>1.57</td>
</tr>
<tr>
<td>Means</td>
<td>Blocked</td>
<td>234</td>
<td>1.39</td>
</tr>
<tr>
<td></td>
<td>Speech</td>
<td>230</td>
<td>1.57</td>
</tr>
<tr>
<td>Grand Means</td>
<td>232</td>
<td>1.48</td>
<td>235</td>
</tr>
</tbody>
</table>

**Discrimination**: Discrimination d’ values show that stimuli appear to be better discriminated around the flat tone than the other stimulus pairs. This is significant across participant groups for the falling continuum, but in the rising continuum the difference between d’ values for flat vs. other pairs is greater in non-musicians (F(1,18)=7.23), (M\text{Musicians}=1.16, SD=3.94; M\text{Non-Musicians}=0.90, SD=3.25) than musicians (M\text{Musicians}=1.84, SD=4.06; M\text{Non-Musicians}=1.67, SD=3.13).

### 3.5. Discussion

In identification, the mid-continuum strategy is generally used more than the flat-anchor strategy, confirming the results for the tone speakers in Experiment 1. As expected, perception is more categorical in musicians than perfect pitch listeners, indicating that perfect pitch may immunise listeners against categorizing perceptually similar stimuli. In addition, perception is more categorical in musicians for sine-wave than for speech, and the opposite is the case for non-musicians, suggesting that musical training may enhance categorical perception in non-speech but not speech contexts, a finding consistent with the modularity of speech processing. In discrimination, the flat-anchor strategy was used, suggesting that different perceptual strategies are used in different tasks. It could be that comparing two stimuli is more demanding than identifying just one, such that a simpler, less refined perceptual strategy is used in discrimination.

### 4. Conclusions

Experiment 1 results suggest that tonal language speakers use a linguistic approach to identify and discriminate tones - they perceptually bisect the continuum, whereas non-tonal language speakers rather use the acoustically salient (flat) stimulus to divide the continuum. These tone-language perceptual strategies were confirmed in Experiment 2 with Thai listeners for identification, but not discrimination. Thus perceptual strategies depend not only on language background, but rather an interaction between language background, musical background, and type of task. Perfect pitch results in less categorical perception overall, and no difference in the perception of speech and non-speech tone. On the other hand, musical training specifically enhances categorical perception of sine-waves (but not speech), indicating modularity of speech and non-speech processing, and reverses the usual superior categorical perception of tones in speech than sine-waves.

### 5. Acknowledgements

The authors thank Prof Dr. Marcus Taft (UNSW), and Dr. Sudaporn Lukanseeyanawin, Benjawan Kasisopa, Sarinya Chompoobutr and Noppawan Nomura (Chulalongkorn U.).

### 6. References


