ABSOLUTE PITCH AND LEXICAL TONES:
TONE PERCEPTION BY NON-MUSICIAN, MUSICIAN, AND
ABSOLUTE PITCH NON-TONAL LANGUAGE SPEAKERS

Denis Burnham and Ron Brooker

MARCS Auditory Laboratories, University of Western Sydney and
Sydney Conservatorium of Music, University of Sydney
d.burnham@uws.edu.au

ABSTRACT

In this paper we investigate whether musically trained non-tonal language speakers perceive lexical tone better than their non-musician counterparts. Three groups of English language speakers, non-musicians, musicians, and musicians with absolute pitch (n=24, N=72), were tested for same/different discrimination of Central Thai tone pairs. These were presented in three separate conditions: as speech (on the syllable [ba]), as filtered speech, or as violin sounds. Non-musicians discriminated tones better in music than in filtered speech, and better in each of these than in speech. Musicians without absolute pitch showed the same pattern of results but were better in all three contexts compared with the non-musicians. On the other hand absolute pitch musicians were equally good in all three contexts, and better overall than the other musicians and the non-musicians. It is concluded that speech and music perception are not independent: musical training and absolute pitch ability may affect speech perception.

1. INTRODUCTION

In a series of experiments Burnham and colleagues [1] showed that there is attenuation of lexical tone perception as a function of both language background and stimulus context. More specifically they found that (i) non-tonal language (English) speakers do not discriminate Thai lexical tones as well as either native Thai speakers, or Cantonese speakers for whom the particular tones are unfamiliar; (ii) non-tonal language speakers show better discrimination of lexical tones presented as music (violin sounds), than as filtered speech, and in turn better than for tones presented in speech; and (iii) tonal language speakers perceive tone contrasts equally well in all three stimulus contexts. Attenuation of tone discrimination for non-tonal speakers is thought to result from the developmental absence of exposure to lexical tone. Just as infants start to become less sensitive to non-native consonant contrasts [2] and more tuned to the consonants and vowels in their native language [3] around 6 months, the presence or absence of lexical tone in the ambient language should also modify speech perception. Indeed, preliminary studies of English-language infants’ lexical tone perception in our laboratory suggest attenuation of tone perception ability between 6 and 9 months. Thus, it may be the case that the onset of language specific speech perception around 6 months results in attenuation of non-tonal language infants’ perception of lexical tone, while tonal language infants continue their lexical tone perception development unabated.

Deutsch and her co-workers [4] have reported that tonal language speakers produce “remarkably precise absolute pitch in reading out lists of words”. While further comparative studies are required in this area, this points to a possible parallel between acquiring a tonal language, and musicians with absolute pitch (AP). Absolute pitch possessors (APPs) have the ability to both produce and identify the pitch of a tone without reference to an external standard, which may be similar to the abilities of tonal language speakers. One speculation for the genesis of AP which is gaining “consensus among psychologists is that (this) ability requires activation and training during a critical period (analogous to the critical period for language acquisition)” [5]. This critical period has been suggested to occur roughly between birth and 6 years of age [6]. Indeed there is now evidence that in a pitch tracking task infants track absolute pitch, while adults track relative pitch [8].

The vast majority of AP possessors report the early onset of musical training [8], usually during that period of life when people are developing their native language. Similarly, there are reports that children (even without AP) given musical training outperform their untrained peers on pitch discrimination tasks [9]. Morrongiello speculates that “(t)he benefits of musical training may possibly generalize to enhance children’s perception and processing of other complex auditory stimuli, such as speech…” [10]. Chen-Hafteck [11] suggests that the ability to sing accurately is the same fine pitch discrimination ability required to speak a tonal language, and that appropriate environmental pitch stimulation at an early age will provide non-tonal speakers with the “same advantages as those acquired by children speaking tonal languages”.

These speculations are yet to be tested empirically. The study reported here provides a first step in such tests. Lexical tone perception is tested in three contexts, speech, filtered speech, and music, with non-tonal language speakers who are (a) non-musicians (b) musicians without AP, and (c) musicians with AP. If musicians with and without AP perform the same as non-musicians, then it might be suggested that speech and music perception development are independent. If however, musicians without...
musicians, then it may be concluded that speech and music perception development are not independent.

2. METHOD

2.1. Subjects

A total of 72 native English-speaking subjects were tested, each with little or no experience of other languages (1 or 2 years of high school maximum) and no experience with Thai or any tonal language. 48 of the 72 subjects were experienced musicians, selected from two music-training institutions in Sydney, Australia. 24 were practicing musicians (Musicians) and 24 were practicing musicians with absolute pitch (APPs). The APPs were selected by testing absolute pitch perception on measures devised by Watson [11]. All APPs included in the study achieved above 90% accuracy on three absolute pitch tasks - pitch naming, pitch producing, and musical key identification. The remaining 24 subjects were Introductory Psychology students with no musical training (Non-Musicians).

2.2. Stimulus Materials and Apparatus

Three stimulus sets were created, Speech, Filtered Speech, and Music. A female Thai native speaker recorded the original Speech stimuli. The syllable [ba] was used to carry the five tones: the three so-called static tones, mid [ba], low [bæ], and high [bá], and the two so-called dynamic tones, falling [ba̰]; high [bá], rising [bát]. Fifteen exemplars of each were recorded and of these three of each were chosen as stimuli on the basis of their similarity and consistency in terms of duration and F0. These fifteen speech stimuli were then manipulated to create the Filtered Speech and the Music stimuli. Filtered speech was created by repeat filtering the speech sounds with a low pass digital filter. This removed the upper formants from the speech whilst leaving fundamental frequency information intact. The Music stimuli were created on a violin, because the violin can both maintain a continuous sound and reproduce rapid pitch changes, e.g., the pitch dynamics of the Thai falling tone which covers approximately one and half octaves in a short space of time. A professional violinist listened extensively to the speech recordings and then reproduced the sounds. A minimum of 25 exemplars of each tone were recorded and digitized. This removed the upper formants from the speech whilst leaving the similarity to the original speech sounds. All the sounds were digitized and stored on disk.

For each stimulus type, all possible pairings of the five tones were used, a total of 10 tone contrasts. One of the three exemplars of each sound was chosen at random for presentation in each trial. This exemplar variation over trials was employed to encourage linguistic processing, and to discourage acoustic processing, in which idiosyncratic features of particular tokens may become important cues for discrimination.

The experiment was conducted at the University of Western Sydney, the Sydney Conservatorium of Music, University of Sydney, and the Australian Institute of Music, Surrey Hills, using a portable Toshiba 3100e laptop computer with D-A, digital I/O, and filter boards. An in-house program, MAKEDIS, was used to control presentation and timing of the sounds, and record subjects' responses and reaction times for each trial. A response panel attached to the computer contained a "same" and a "different" key for subjects' responses, and a set of colored lights that flashed when subjects made a correct response (only in the initial task competence, not the test phases). The feedback lights were only used in training, not in test phases.

2.3. Procedure

Each subject completed three tasks, in counterbalanced order. The procedure differed only in the stimulus type employed, Speech, Filtered Speech, or Music. Each task comprised four distinct phases. The first phase simply required the subject to listen to a "context-setting" tape recording for each of the three stimulus types, a short tape of Thai speech, filtered speech, or violin music. The second was a task competence phase in which subjects were required to respond correctly on four simple auditory distinctions, two "same" and two "different" presentations of [ræg vs ræg]. This phase was included to ensure that each subject could make simple auditory discriminations, and was acquainted with the task. The two final phases were test blocks, with 40 trials in each block. Five of the possible 10 contrasts were presented in the first block, and the other five in the second block. The order of presentation of blocks was counterbalanced between subjects. These test blocks are the main focus here. Equal numbers of same (AA or BB) and different pairs (AB or BA) were randomly presented to the listener, whose task it was to respond as quickly as possible by pressing either the "same" or "different" key. The actual exemplars presented on any particular trial were selected randomly by the computer from the pool of three possible exemplars for each phone. The 10 tone contrasts comprised of 3 static-static comparisons (mid-low, mid-high, low-high), 6 static-dynamic comparisons (mid-rising, mid-falling, low-rising, low-falling, high-rising, high-falling), and 1 dynamic-dynamic comparison (rising-falling). The program kept track of these so that appropriate statistical comparisons could be made. For each contrast pair, the four possible same-different combinations (AA, AB, BA, BB) were presented twice. Thus in each block of 40, there were 20 same and 20 different trials.

There is evidence that in same-different tasks both native, phonemic differences and non-native phonetic speech differences can be discriminated with an interstimulus interval (ISI) of 500 msec, but that only native differences can be perceived at an ISI of 1500 msec [12]. As the tone contrasts are non-native to English perceivers, and as we were unsure how musicians would process stimuli, ISI was varied between subjects here - in each of the three groups, non-musicians, musicians, and musicians with perfect pitch, half of the subjects had an ISI of 500 msec throughout, and the other half had an ISI of 1500 msec.

2.4. Dependent Variables

The number of correct and incorrect responses and the accompanying reaction times were recorded on disk. The correct and incorrect responses on different trials were converted to a discrimination index (DI) given by [number of correct responses on different trials minus number of incorrect responses on same trials]/number of trials. The resulting score is a measure of how well subjects were able to discriminate speech sounds on AB and BA trials. A score of 1 indicates that
all AB and BA trials were responded to as “different” and all AA and BB trials as “same”. A score of zero indicates chance responding. Reaction times (RTs) were recorded from the onset of the second sound in each tone pair.

3. RESULTS

The DI and the RT data were analyzed in separate analyses of variance (ANOVAs) with the design: 3 Groups (Non-Musicians, Musicians, and Musician APPs) x 2 ISIs (500 msec ISI, 1500 msec ISI) x 3 Stimulus Contexts (Speech, Filtered Speech, Music) x 3 Contrast Types (Static-Static (S-S), Static-Dynamic (S-D), and Dynamic-Dynamic (D-D)) with repeated measures on the last two factors. Planned orthogonal comparisons were conducted within each ANOVA between all Musicians (APPs or not) vs Non-Musicians, APPs vs Musicians; 500 vs 1500 msec ISI; Speech vs Non-Speech (Filtered Speech & Music); Speech vs Non-Speech (Filtered Speech & Music), Filtered Speech vs Music; D-D vs S-D and S-S, and S-D vs S-S. The alpha-level was set at 0.05, and with df = 1, 66 for each contrast, \( F_{\text{critical}} = 3.99 \).

3.1. Discrimination Index Data

Mean DI scores for the three groups on each of the three stimuli types are shown in Figure 1. As can be seen, the two Musician groups performed better than the Non-Musicians, \( F(1,66) = 18.85 \). Scores were also generally significantly higher for Speech than Non-Speech, \( F(1,66) = 21.15 \), and also for Filtered Speech than Music, \( F(1,66) = 10.78 \). Most interesting are the interactions between subject groups and stimulus type. There was a significant interaction of all Musicians vs Non-Musicians with Speech vs Non-Speech, \( F(1,66) = 12.33 \), and also between all Musicians vs Non-Musicians, and Filtered Speech vs Music, \( F(1,66) = 9.34 \). As can be seen in Figure 1, Non-Musicians’ performance improved as the stimuli became less speech-like, from Speech to Filtered Speech to Music; while for both the Musician groups the functions were relatively flat. A significant interaction between APPs vs Musicians and Filtered Speech vs Music, \( F(1,66) = 6.32 \), shows that non-APP musicians performed better with Music than Filtered Speech, while APPs performed better with Filtered Speech than Music.

There were a number of effects for the type of speech contrast, S-S, S-D, D-D. These are illustrated in Figure 2. There were significant interactions between all Musicians vs Non-Musicians, and both D-D vs S-D and S-S combined, \( F(1,66) = 11.98 \); and S-S vs S-D contrasts, \( F(1,66) = 4.50 \). Non-Musicians’ performance was linked to the degree of dynamism in the contrast; their performance was best for D-D, then S-D, and then S-S contrasts. On the other hand, all Musicians performed better the more static the tonal contrast, best on S-S,
3.2. Reaction Time Data

Figure 3 shows mean RTs on AB (Different) trials. These are shown for 500 and 1500 msec ISI as a number of ISI effects were significant. In particular, while Musicians and APP Musicians were both quicker to respond than Non-Musicians, F(1,66)= 26.39, there was an interaction of Musicians with and without APP with ISI, F(1,66)= 12.54. This indicates that non-APP musicians were faster to respond than APPs at the shorter ISI, while APPs were quicker to respond for the longer 1500 msec ISI. Finally, while responding was generally faster for speech than non-speech, F(1,66)= 40.89, and for Filtered Speech than Music, F(1,66)= 10.22, an interaction of Speech vs Non-Speech with ISI, F(1,66)= 7.18, 20.83, indicates that responses to filtered speech were generally faster than to music at 500 msec, but not at 1500 msec.

4. DISCUSSION AND CONCLUSIONS

Musical training and absolute pitch has a positive effect upon the discrimination of lexical tone, both in terms of accuracy and reaction time. Plainly, the effects of musical training found in children by Morrongiello [9], have now also been found in adults. The superior performance by musicians both with and without absolute pitch suggests that that they have either avoided or reversed the attenuation of lexical pitch discrimination normally suffered by non-tonal language speakers. This transfer of perceptual skill from music to speech represents an apparent reversal of the effect in which speaking a tonal language appears to increase Cantonese children's singing accuracy [10]. It appears that musical training, commenced early in life, has a similar effect on tone perception to that of language acquisition by tonal language speakers.

Within the musician groups, the APPs showed greater accuracy than musicians without AP on each measure and stimulus style. In addition to this quantitative superiority, there was a notable qualitative difference between discrimination by the APPs and the Musicians and Non-Musicians alike. While performance was best for Music, then Filtered Speech then Speech for both Non-Musicians and non-APP Musicians, for those with Absolute Pitch, performance was similar on all three stimulus styles, with a slight superiority for filtered speech over music. This qualitative difference is also seen for the types of tone contrasts: whereas for Non-Musicians and non-APP musicians the D-D contrasts were easiest, followed by S-D, and then S-S, for APPs the order was reversed. These differences deserve further investigation, however, one possible reason may be posited here. As exemplars varied randomly over trials, if these slight differences were differentially noticed by the APPs in the more music-like stimuli then their relative performance for these items may have been compromised. While this does not explain the whole complex of results found here, this should be borne in mind in future studies. It is possible that the generally greater accuracy for Non-Speech stimuli may have been understated here for the APPs.

The RT data on different ISIs across the three stimulus styles shows that while the musicians without AP performed better at 500 msec ISI, the APPs performed better in the 1500 msec condition. This is consistent with the view that AP involves long-term pitch memory that draws on internal pitch standards (a template) to identify tones [13]. Memory decay experiments indicate APPs do not have better echoic memory [8]. Here it is suggested that APPs use efficient abstract processing of internal pitch sense. Speech perception and music perception are not independent. Music training and absolute pitch ability affect speech perception ability. Just when and how this occurs in development awaits further investigation.

5. REFERENCES