別刷

音楽知覚認知研究
Journal of Music Perception and Cognition
Vol. 7, No. 2, 2001

Detecting contour in short spoken and musical items: A comparison of speakers from tonal and non-tonal languages

話し言葉および音楽の部分となりうるような断片における音調曲線の検出：声調言語話者と非声調言語話者との比較

Catherine STEVENS　　Peter KELLER
University of Western Sydney

2001年12月

日本音楽知覚認知学会
The Japanese Society for Music Perception and Cognition

音楽知覚認知研究
Journal of Music Perception and Cognition
2001, Vol. 7, No. 2, 133-143

資料論文
Short report

Detecting contour in short spoken and musical items: A comparison of speakers from tonal and non-tonal languages

Catherine STEVENS and Peter KELLER
kj.stevens@uws.edu.au keller@mpipf-muenchen.mpg.de

MARCS Auditory Laboratories, University of Western Sydney
Locked Bag 1797, South Penrith DC NSW 1797, Australia

Abstract

Contour is characteristic of both spoken and musical events, and extraction of contour enables recognition and discrimination of short musical and spoken items. Based on the rich musicality of the early linguistic environment, we test the hypothesis that perceptual biases such as heightened sensitivity to contour in a tonal language persist into later auditory processing. Our experiments investigate the effect of tonal language background (Thai versus English speakers) on discrimination of contour in the context of tonal (Thai) and non-tonal (English) speech stimuli, musical intervals, and frequency discrimination. As hypothesized, adult participants from a tonal language background were more accurate than non-tonal language participants at discriminating contour in intact words and filtered speech. The effect of language background on discrimination accuracy was upheld for both Thai and English items. However, there was no effect of language background on discrimination of musical intervals, nor was there any evidence of variation in psychophysical thresholds for frequency discrimination. We speculate that the results reflect a form of expertise.

Keywords: tonal language, speech perception, music perception, auditory discrimination, expertise, contour

話し言葉および音楽の部分となりうるような断片における音調曲線の検出：声調言語話者と非声調言語話者との比較

Catherine STEVENS ・ Peter KELLER

MARCS Auditory Laboratories, University of Western Sydney
Locked Bag 1797, South Penrith DC NSW 1797, Australia

概　要

音調曲線は，話し言葉および音楽における諸事象を特徴付けるものであり，音調曲線を抽出することによって，話し言葉や音楽の部分となりうるような断片の，再認および弁別が可能になる。幼児期の言語環境が音楽的に豊かであることに基づいて，声調言語においては音調曲線に対する感度を高めるような知覚上の傾向が生じ，その後の聴覚処理にも残るのではないかとの仮説を立て，検証を試みた。本論文の実験においては，声調言語の環境で育つことが，音調曲線の弁別にどのような効果を及ぼすかを（タイ語の話者と英語の話者とを比較することによって）調べた。この際，

声調言語（タイ語）および非声調言語（英語）から取った音声材料を用いた知覚実験のほか，音程判断，周波数弁別判断の実験を行った。仮説に従えば，声調言語の環境で育った成人の実験参加者は，非声調言語の環境で育った実験参加者に比べて，単語をそのまま呈示したときにも，音声をフィルターに通したときにも，音調曲線の弁別に関して，より正確な判断を下す傾向があると考えられる。この予測が検証された。言語歴が，弁別の正確さに影響を及ぼすということは，タイ語，英語のいずれの材料においても認められた。ところが，音程の弁別に関して言語歴の効果は認められず，周波数弁別実験において得られた閾値にも違いは認められなかった。このような結果は，ある種の熟練技術を反映しているのではないかと考えている。

キーワード：声調言語，音声知覚，音楽知覚，聴覚的弁別，熟練技術，音調曲線

Introduction

Fundamental issues in the study of auditory perception include the effect of early experience, and the search for general mechanisms that underpin a range of behavioural abilities and phenomena. The research reported here is motivated by both of these concerns. We compare the perception of two classes of auditory patterns, namely spoken word and short musical items, and investigate the effect of native language background on sensitivity to pitch change in speech and music. We ask: what links can be drawn between music and speech recognition? What stimulus features and perceptual mechanisms might underpin the two domains? What is the contribution of the early linguistic environment to the development of general auditory feature extraction and weighting processes?

Similarities Between Speech and Music Stimuli

It is possible to identify potent similarities between music and speech. For example, at the very least speech and music are conveyed by sound patterns so that general purpose perceptual mechanisms will be engaged by spoken language and music (Repp, 1990). Speech and music consist of structured patterns of pitch, duration and intensity, and are temporal, hierarchical, communicative and expressive (Handel, 1989; Tenney & Polansky, 1980; Wallin, Merker & Brown, 2000). With exposure to spoken language and to music, listeners learn to segment the signal and develop expectations based on statistical regularities and cultural conventions (Dolson, 1994; Jusczyk, 1997; Trainor & Trehub, 1992). Finally, speech and music contain similar prosodic features such as pitch, intonation, stress, loudness and duration (Bollinger, 1986; Fernald, 1996; Halliday, 1970; Noteboom, 1997).

Comparing Speech and Music Perception

An obvious basis for comparing music and speech perception is to consider the

prosodic structure of the two kinds of event. Prosody refers to non-segmental or suprasegmental aspects of speech and arises from vocal tone or pitch (fundamental frequency), stress (a combination of fundamental frequency, amplitude and relative duration), and timing (syllable duration and pause) (Bollinger, 1986; Cutler, Dahan & van Donselaar, 1997; Handel, 1989). Prosody is the melody line or contour of speech. In speech, contour refers to the trajectory of the fundamental frequency over time; the intonation. Contour in speech marks boundaries, distinguishes the categories of utterances (e.g., distinguishing a statement from a question), signals focus, and communicates affect (Cutler et al., 1997). A feature that is comparable across music and speech is contour or intonation. Tonal languages, such as Thai or Vietnamese where phonemes pronounced with varying pitch can convey differential meaning, provide a tool to closely analyse the extraction of contour and intonation features in perception of musical and spoken items.

Another way to compare musical and speech prosody is to consider the musical dimensions of the environment in which infants acquire their first language. Trehub (Trehub & Trainor, 1993; Trehub, Schellenberg & Hill, 1996) has identified rich musical features present in infant-directed speech. There are observable parallels between speech and song directed to infants: i) similar adjustments are made to both to make simple contours, elongated vowels, and emotional expressiveness; ii) the use of speech to capture and hold attention has a counterpart in play songs; and iii) soothing speech has a counterpart in lullabies. In addition to the importance of prosodic features for expression of affect, augmented pitch, rhythm and stress features are thought to facilitate segmentation, discrimination and classification of speech sounds. For example, Jusczyk, Houston, and Newsome (1999) demonstrated sensitivity of 7.5-month-old infants to strong/weak stress patterns in fluent speech. Strong syllables were treated as markers of word onset and Jusczyk et al. concluded that English learners may rely heavily on stress cues when they begin to segment words from fluent speech. We investigate the hypothesis that perceptual biases acquired from the early auditory environment such as intonation features of a tonal language persist into later auditory perception.

We assume that there are common pitch features in musical and speech patterns and that perception of aspects of music and speech events is subserved by similar mechanisms. If contour is a feature common to perception of music and speech, then discrimination accuracy and reaction time are comparable across stimulus type. If the early tone language environment develops sensitivity to contour, and contour components of speech and music are processed similarly, then speakers with a tonal language background are more accurate in response to pitch manipulations in speech and music relative to speakers with a non-tonal language background. We test these

hypotheses in the context of tonal and non-tonal speech stimuli, successive complex tones indicating musical intervals, and sine tones to be discriminated.

The aim of Experiment 1 was to examine the effect of tonal language background on accuracy in discriminating short Thai and English words. The 2×2×2 factorial design consisted of language background (tonal, non-tonal), sound type (intact word, filtered word), and language type (Thai, English) with repeated measures on the latter two factors. The dependent variables were discrimination accuracy (hit rate minus false alarm rate) and response time.

Experiment 1

Method

Participants (Experiments 1-3). The sample consisted of 47 female and male adult participants. Twenty-four were native speakers of Thai studying at Chulalongkorn University in Bangkok (11 males and 13 females; $M=23.8$ years, $SD=3.86$). The remaining 23 participants had English as a first language and were recruited from Psychology I at the University of Western Sydney (1 male; 22 females; $M=23.8$ years; $SD=9.05$). All participants had self-reported normal hearing. Experiments were conducted in one session. Experiment 1 (speech stimuli) was always completed first. The order of Experiments 2 and 3 was counterbalanced across participants.

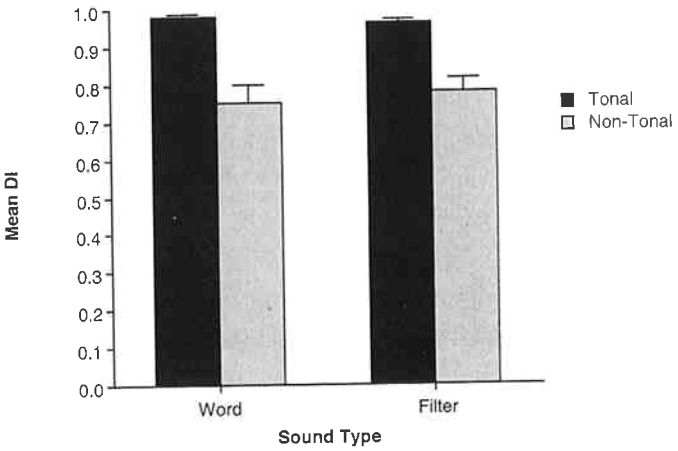
Stimuli and equipment. Spoken word stimuli consisted of two-syllable items with a weak-strong stress pattern spoken in Thai and English. The items were recorded by two female speakers: one a native Thai speaker and the other a native English speaker. Twelve English and 12 Thai words were recorded with a rising and then a falling intonation. The Thai items were matched as closely as possible with the structure of the English words. All items were comparable in length (Thai words mean duration=558 ms, $SD=121$ ms, range 362-735 ms; English words mean duration=639 ms, $SD=83$ ms, range 453-778 ms), fundamental frequency (range: 380-390 Hz), and frequency of occurrence, and normalised for loudness. To eliminate the semantic content of the words but retain their speech-like quality, items were low-pass filtered at 400 Hz. Speech stimuli were recorded digitally using a DAT recorder and presented in all experiments on a Macintosh G3 notebook through AKG stereo headphones.

Procedure. Stimuli were presented in four blocks of 96 trials each: Thai-intact, Thai-filtered, English-intact, and English-filtered. The presentation order of blocks was counterbalanced across participants. Participants were instructed to ignore the meaning of the items, attend to their acoustic properties and judge whether items in a pair were exactly the same or different. Experiment 1 took 30 minutes.

Results

Accuracy was calculated as a discrimination index consisting of hit rate minus false alarm rate for pairs of items. Response time analyses were based on times recorded for correct responses only. It was hypothesised that participants with a tonal language background would show heightened sensitivity to contour in speech

a) Thai Items



b) English Items

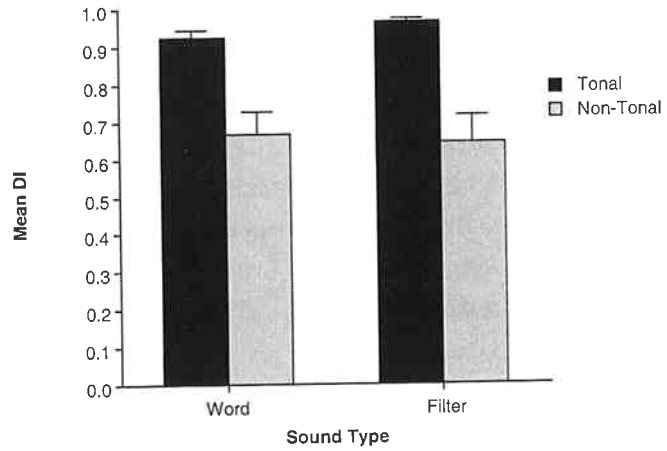


Figure 1. a) Mean discrimination index recorded from participants with tonal versus non-tonal language backgrounds in response to spoken Thai items. b) Mean discrimination index recorded from participants with tonal versus non-tonal language backgrounds in response to spoken English items.

items. As hypothesised, the accuracy (discrimination index) main effect was significantly higher for Thai participants than for Australian participants, evident in response to both Thai items, F (1, 45)=28.05, p=.000003 (see Figure 1a) and English items, F (1, 45)=25.09, p=.000008 (Figure 1b). The effect of language background was evident for intact spoken items and low-pass filtered items. Analysis of response times revealed no main effects of, or interaction between, language background and sound type.

Experiment 2

Experiment 2 investigated whether tonal language background influenced sensitivity to pitch change in short musical items. The 2×3 factorial design consisted of the variables language background (tonal, non-tonal) and pitch change (contour only, interval only, contour + interval). The dependent variables were speed and accuracy.

Method

Stimuli. Stimuli consisted of descending and ascending major and minor third musical intervals. An interval of a third was chosen as it was similar in magnitude to the variation in pitch in the rising and falling contours of the spoken items used as stimuli in Experiment 1. (Finer frequency discrimination (< 3rd) would be made in Experiment 3.) The musical intervals were also played with a rhythm pattern with the first note of the interval lasting for 200 ms and the second note lasting for 400 ms (crotchet-minim pattern). Like the spoken items, the musical intervals consisted of a complex waveform, in this case a sawtooth. Ascending intervals began on G_4 (392 Hz) and descending intervals ended on G_4. There were 12 same and 12 different interval combinations. The 12 "different" trials consisted of four trials wherein the contour or direction of the interval differed (e.g., ascending major 3rd paired with descending major 3rd); four trials where both contour and interval differed (e.g., ascending major 3rd paired with descending minor 3rd); and four trials where only the interval differed (e.g., ascending major 3rd paired with ascending minor 3rd). Four practice trials consisting of major and minor 3rds from D_4 (293.66 Hz) preceded the experimental trials.

Procedure. Pairs of musical intervals were presented in random order. Participants were asked to listen to each pair and decide whether the second interval was exactly the same or different from the first. There were 24 trials, 12 in which "same" was the correct response and 12 in which "different" was the correct response. The experiment took eight minutes.

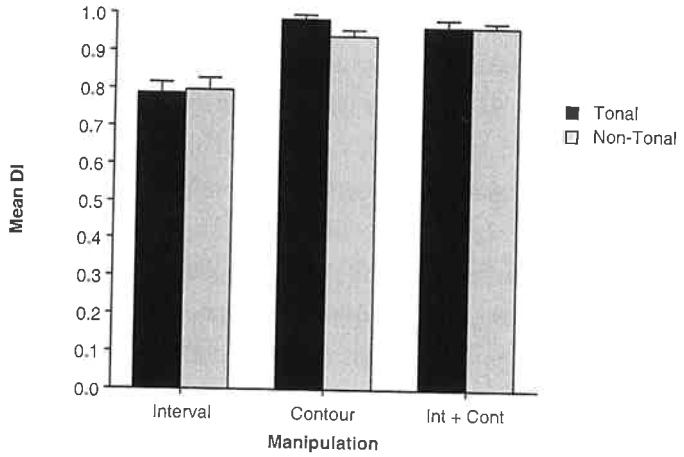


Figure 2. Mean discrimination index recorded from participants with tonal versus non-tonal language backgrounds in response to musical intervals — ascending and descdending major and minor 3rds.

Results

It was hypothesised that if prosodic components of spoken and musical items are similar and the native language environment develops sensitivity to distinctive features of the language then listeners with a tonal language background reflect sensitivity to pitch change in musical items. The discrimination index revealed no significant effects of language background on discrimination of musical intervals. Pitch manipulations within interval stimuli involved a change to interval (major 3rd to minor 3rd or minor 3rd to major 3rd) and/or to contour (ascending to descending or descending to ascending). Significant linear and quadratic trends were evident across these manipulations indicating that, across participants, discrimination was easier for contour than for interval manipulations, $F(1, 45)=61.68$, $p=.0000000005$. As Figure 2 shows, it was not necessarily the case that a contour plus interval manipulation was more discernible than a contour manipulation alone.

Experiment 3

Experiment 3 involved a frequency discrimination task to investigate whether speakers of a tonal language have lower difference thresholds for frequency.

Method

Stimuli. Tones 400 ms in length were generated as sine waves sampled at 22 kHz, 8-bit resolution. The tones ranged from 392 Hz (G_4) to 416 Hz ($\sim G\#_4$) and were separated by steps of 2 Hz yielding 13 different tones. All tones were paired

with the lowest tone in both a rising and falling combination, e.g. 392 Hz-394 Hz; and 394 Hz-392 Hz. A gap of 100 ms separated presentation of the two 400 ms tones. There was a three-second pause between trials in which time participants made their response.

Procedure. Participants were presented with a pair of sine tones and asked to decide as quickly and as accurately as possible whether the tones were the same or different. Trials were presented in a new random order for each participant. There were 30 trials in total, 24 consisting of pairs of frequencies that differed, and six trials where the pairs of frequencies were the same (to maintain vigilance). The experiment took five minutes.

Results

A general index of difference threshold was calculated based on the point at which two correct ("different") responses were given to a frequency difference. Only trials with a difference between frequencies were scored. The minimum possible difference (and therefore lowest threshold) was 2 Hz and the maximum possible difference was 24 Hz. An independent samples t-test revealed no significant difference between the minimum difference discernible by tonal language participants (M = 11.08 Hz, SD = 6.08) and non-tonal language participants (M = 13.4 Hz, SD = 6.96).

General Discussion

The results provide support for the main hypothesis that tonal language background sensitises adult listeners to contour in short spoken items: Thai speakers were more accurate in recognizing contour manipulations to spoken items than English speakers. The effect was upheld not only when the items were intact words but also when semantic content had been removed by low-pass filtering. Interestingly, listeners from a tonal language background were more accurate than their non-tonal language counterparts at discriminating contour in *both* Thai and English items. Semantic content of the items had little effect: contour was semantically meaningful in Thai items for Thai speakers but Thai speakers were equally adept at detecting non-meaningful changes to contour in English items. It appears that the linguistic environment does sensitise an individual to distinctive features of the first language speech signal, such as contour.

Motivational variations could account for the differences between performance of tonal and non-tonal language participants. However, the results of Experiments 2 and 3 eliminate this possibility as performance across groups was comparable. Experiment 2 showed no effect of tonal versus non-tonal language background on discrimination of pitch change in a music task; sensitivity to contour in speech items

does not appear to generalise to pitch change of a similar magnitude in a musical context. This finding corroborates the argument that speech is special (Liberman, 1982; Mattingly, Liberman, Syrdal, and Halwes, 1971) and is in keeping with Van Lancker and Fromkin's (1973, 1978) observations that pitch discrimination is lateralized to the left hemisphere when the pitch differences are linguistically processed but not when presented in a non-linguistic context. Another possibility, however, is that heightened sensitivity to contour in speech among tonal language speakers is an effect of expertise. An appropriate test of an expertise hypothesis would be to investigate the effect of another form of expertise—musical training—on performance in the speech and music discrimination tasks. If expertise underpins enhanced sensitivity—that is, expertise with either speech or music—then musically trained, non-tonal language participants should perform at a comparable level to tonal language speakers in the present speech task, and we would expect them to excel at pitch discrimination in the music task. Both Thai and English participants in Experiment 2 found contour differences more discernible than interval differences. This finding is being further explored in an experiment where the pentatonic scale characteristic of Thai traditional music is included in the experimental trials.

The results of Experiment 3 are consistent with those of Experiment 2. In the frequency discrimination task there was no threshold difference between speakers of tonal and non-tonal languages. Differences observed in Experiment 1 then are not the result of threshold variations. Rather, the sensitivity to pitch change seems to be evident only in a speech context where items are speech-like - either intact or filtered. An important future study will examine the complex and intriguing relationship between the degree of tonality of a particular language and the tonality of the associated musical environment. For example, Deutsch, Henthorn and Dolson (1999) suggest that, when performing a speech production task, speakers of the tonal languages, Vietnamese and Mandarin, possess absolute pitch. Other investigations will shed light on the relation between general auditory feature discrimination and expertise that may develop in musical and/or linguistic contexts.

Acknowledgements

This research was supported by a grant from the Australian Research Council awarded to the first author. We thank Dr Sudaporn Luksaneeyanawin for her advice and assistance in creating and recording the Thai stimuli and the students of Chulalongkorn University for participating in the experiments. We also thank the reviewers for their helpful comments on an earlier version of the manuscript. The second author assisted in the conduct of this study in his capacity as Research Assistant in MARCS; he now holds a postdoctoral position at Haskins Laboratory,

Connecticut, USA. Correspondence regarding the study may be directed to Catherine Stevens, School of Psychology/MARCS, University of Western Sydney - Bankstown Campus, Locked Bag 1797, South Penrith DC, NSW 1797, Australia. E-mail: kj.stevens@uws.edu.au; Internet: http://www.uws.edu.au/marcs/

References

Bollinger, D. (1986). *Intonation and its parts*. Stanford, CA: Stanford University Press.

Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*, 141-201.

Deutsch, D., Henthorn, T., & Dolson, M. (1999). Absolute pitch is demonstrated in speakers of tone languages. *Journal of the Acoustical Society of America, 106*, 2267 (Abstract).

Dolson, M. (1994). The pitch of speech as a function of linguistic community. *Music Perception, 11*, 321-331.

Fernald, A. (1996). Human maternal vocalizations to infants as biologically relevant signals: An evolutionary perspective. In P. Bloom (Ed.), *Language acquisition: Core readings* (pp.51-94). Cambridge, MA: MIT Press.

Halliday, M. A. K. (1970). *A course in spoken English: Intonation*. London, UK: Oxford University Press.

Handel, S. (1989). *Listening*. Cambridge, MA: MIT Press.

Jusczyk, P. W. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.

Jusczyk, P. W., Houston, D., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology, 39*, 159-207.

Liberman, A. M. (1982). On finding that speech is special. *American Psychologist, 37*, 148-167.

Mattingly, I. G., Liberman, A. M., Syrdal, A. K., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology, 2*, 131-157.

Noteboom, S. (1997). The prosody of speech: Melody and rhythm. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp.640-673). Oxford, UK: Blackwell.

Repp, B. (1990). Some cognitive and perceptual aspects of speech and music. In J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, language, speech and brain*. Macmillan Press.

Tenney, J., & Polansky, L. (1980). Temporal Gestalt perception in music. *Journal of Music Theory, 24*, 205-239.

Trainor, L. J., & Trehub, S. E. (1992). A comparison of infants' and adults' sensitivity

to Western tonal structure. *Journal of Experimental Psychology: Human Perception and Performance, 18*, 394-402.

Trehub, S. E., & Trainor, L. J. (1993). Listening strategies in infancy: The roots of music and language development. In S. McAdams & E. Bigand (Eds.), *Thinking in sound: The cognitive psychology of human audition* (pp.278-327). Oxford: Oxford University Press.

Trehub, S. E., Schellenberg, E. G., & Hill, D. S. (1996). The origins of music perception and cognition: A developmental perspective. In I. Deliege & J. Sloboda (Eds.), *Perception and cognition of music* (pp.103-128). Hove, UK: Psychology Press.

van Lancker, D., & Fromkin, V. A. (1973). Hemispheric specialization for pitch and "tone": Evidence from Thai. *Journal of Phonetics, 1*, 101-109.

van Lancker, D., & Fromkin, V. A. (1978). Cerebral dominance for pitch contrasts in tone language speakers and in musically untrained and trained English speakers. *Journal of Phonetics, 6*, 19-23.

Wallin, N., Merker, B., & Brown, S. (Eds.) (2000). *The origins of music*. Cambridge, MA: MIT Press.

Received September 10, 2001
Accepted November 8, 2001