The role of spectro-temporal fine structure cues in lexical-tone discrimination for French and Mandarin listeners

Laurianne Cabrera, Feng-Ming Tsao, Dan Gnansia, Josiane Bertoncini, Christian Lorenzi

▶ **To cite this version:**

Laurianne Cabrera, Feng-Ming Tsao, Dan Gnansia, Josiane Bertoncini, Christian Lorenzi. The role of spectro-temporal fine structure cues in lexical-tone discrimination for French and Mandarin listeners. Journal of the Acoustical Society of America, 2014, 136 (2), pp.877-882. hal-01968845

HAL Id: hal-01968845
https://hal.science/hal-01968845v1

Submitted on 3 Jan 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The role of spectro-temporal fine structure cues in lexical-tone discrimination for French and Mandarin adult listeners

Laurianne Cabrera *

Laboratoire de Psychologie de la Perception, CNRS, Université Paris Descartes

45 rue des saints Pères, 75006 Paris, France

Feng-Ming Tsao

Department of Psychology, National Taiwan University

No.1, Sec. 4, Roosevelt Road, Taipei, 106, Taiwan

Dan Gnansia

Neurelec, 2720 Chemin de Saint-Bernard Porte, 06224, Vallauris, France

Josiane Bertoncini

Laboratoire de Psychologie de la Perception, CNRS, Université Paris Descartes

45 rue des saints Pères, 75006 Paris, France

Christian Lorenzi

Laboratoire des systèmes perceptifs, CNRS,

Institut d'Etude de la Cognition, Ecole normale supérieure, Paris Sciences et Lettres

29 rue d'Ulm, 75005 Paris, France

* Corresponding author: Laboratoire de Psychologie de la Perception, CNRS-UMR 8158, Université Paris Descartes, 45 rue des saints pères, 75006, Paris, France.
Tel: +33 1 42 86 43 20
E-mail address : laurianne.cabrera@gmail.com

Submitted on June 2014, JASA

Running title: weight of modulation cues in lexical tones

Abstract

The importance of spectro-temporal modulation cues in conveying tonal information for lexical tones was assessed in native-Mandarin and native-French adult listeners using a lexical-tone discrimination task. The fundamental frequency (F0) of Thai tones was either degraded using an 8-band vocoder that reduced fine spectral details and frequency-modulation cues, or extracted and used to modulate the F0 of click trains. Mandarin listeners scored lower than French listeners in the discrimination of vocoded lexical tones. For click trains, Mandarin listeners outperformed French listeners. These preliminary results suggest that the perceptual weight of the fine spectro-temporal modulation cues conveying F0 information is enhanced for adults speaking a tonal language. PACS numbers: 43.71 Rt, 43.71 Hw, 43.66 Mk

I. INTRODUCTION

Tonal variations at the syllable level distinguish word meaning in tonal languages (*e.g.*, Liang, 1963). Native listeners rely mainly on fundamental-frequency (F0)–and thus, voice-pitch cues–to discriminate lexical tones. However, other acoustic cues such as duration, amplitude or voice quality may also play a secondary role (*e.g.*, Kuo *et al*., 2008; Whalen and Xu, 1992).

Over the last decades, psycholinguistic studies have investigated whether being native (*i.e.*, expertise) in tonal language influences the relative weight of these acoustic cues in lexical-tone perception (Gandour and Harshman, 1978). Burnham and Francis (1997) showed that non-native (English-speaking) listeners are less accurate in discriminating lexical tones than native (Thai-speaking) listeners (see also Burnham and Mattock, 2007). Their results also suggested that non-native listeners rely more on mean F0 to perceive tones compared to native listeners who are able to categorize F0 patterns in spite of phonetic and tonal variability (*i.e.*, small variations in pitch within the same category). Lee *et al.* (2008) explored further the

influence of the native language on the identification of lexical tones by using degraded speech sounds (*i.e.*, "fragmented" tones obtained by removing a variable number of F0 periods at the onset, center or final part of the syllables). Lexical tone identification performance was affected by these degradations of F0 information for the non-native listeners (English speakers learning Mandarin) only. These results confirm that non-native listeners rely heavily on mean F0 whereas native listeners (Mandarin speakers) rely more on F0 direction (see also Huang and Johnson, 2010). Altogether, these studies are compatible with the notion that expertise in tonal language influences the relative importance of the acoustic cues involved in lexical-tone perception. From a wider perspective, they are consistent with the idea that linguistic experience shapes the weight of acoustic cues in speech perception (*e.g.*, Burnham and Mattock, 2007).

The search for the acoustic cues used in discriminating or identifying lexical tones has been recently renewed by the use of "vocoders" that manipulate the spectral and temporal modulation components of speech signals (see Shamma and Lorenzi, 2013 for a review). The slowest amplitude-modulation cues (AM, corresponding to the relatively slow variations in amplitude over time) play a major role in consonant recognition, whereas the fast AM cues, frequency-modulation cues (FM, corresponding to the oscillations in instantaneous frequency close to the center frequency of the frequency band), and fine spectral details are more important in lexical-tone recognition. Chinese-speaking listeners have been shown to rely more on the FM cues (*e.g.*, Wang *et al.*, 2011; Xu and Pfingst, 2003) cues compared to English- or French-speaking listeners who hinge on AM cues to identify native speech sounds (*e.g.*, Shannon *et al.*, 1995; Smith *et al.*, 2002). Fu *et al.* (1998) showed that for native Mandarin speakers, lexical-tone recognition was more affected by a reduction of temporal resolution (that is by the selective attenuation of the fast [F0-related] AM cues above 50 Hz) than by a reduction of spectral resolution (tones were vocoded using a decreasing number [4,

3, 2, and 1] of broad frequency bands). In contrast, consonant and vowel recognition was found to be mostly affected by a reduction of spectral resolution. In addition, Xu *et al*. (2002) and Kong and Zeng (2006) showed a clear dependence of lexical tone recognition on spectral resolution when the number of frequency bands varied from 1 to 32.

However, the vocoder-based studies cited above did not compare *directly* lexical-tone recognition across listeners from different linguistic backgrounds using the same material and procedure. Thus, it is still unclear whether the native language of adult speakers influences the weight of the spectro-temporal modulation cues conveying F0 information in speech perception. The goal of the present pilot study was to assess the effect of two different native languages on the ability to use fine spectral details and FM cues in lexical-tone discrimination. This comparison should reveal to what extent native language shapes the perception of speech modulation cues. Three conditions were designed to: 1) preserve the original speech modulation cues of lexical tones, 2) degrade the fine spectral details and FM speech cues of lexical tones and thus the F0-related information and 3) generate sounds (click trains) preserving only the F0-related FM cues of the original lexical tones. First, it was hypothesized that native French-speaking listeners should be less accurate in lexical-tone discrimination than native Mandarin-speaking listeners. Second, if Mandarin listeners rely more on F0 direction than French listeners, they should be more affected by the degradation of fine spectral and FM cues than the French ones. Third, if exposure to a tonal language results in a higher perceptual weight of the fine spectral than temporal cues in perceiving F0 information for both speech and click-train sounds, Mandarin-speaking listeners should be more accurate in the discrimination of F0 contours in the click-train condition than French-speaking listeners (see Xu *et al*., 2002). The present study used a same/different task to assess the discrimination of lexical tones. Three Thai lexical tones were used in order to present non-native stimuli for both French and Mandarin-speaking listeners and use similar stimuli than

those used in Burnham and Francis (1997): low, rising, and falling F0 patterns. We assumed that the discrimination abilities of native Mandarin-speaking listeners should be facilitated by their experience in lexical tone processing. Participants had to discriminate the following three contrasts: rising/low, rising/falling, and low/falling using a same/different task. The rising/low contrast would be the most difficult to discriminate for non-lexical users because of the highly similar F0 trajectories until the mid-point of the tone. It is not the case for the rising and falling tones that have totally different F0 trajectories and differ on other cues such as duration, making them easier to distinguish (see Abramson, 1978). In addition, two different inter-stimulus intervals (ISI) were used (500 and 1500 ms) in the present same/different task. As suggested in previous studies (see Burnham and Francis, 1997; Clément *et al*., 1999; Werker and Tees, 1984) long ISI may affect information loss in the short-term memory representation of the cues used to discriminate two sounds. We assumed that a long ISI (1500 ms) may affect the rate of information loss in the short-term memory representation of the voice-pitch cues used to discriminate lexical tones compared to a short ISI (500 ms). This effect may be greater for the French-speaking listeners than Mandarin-speaking listeners.

II. METHOD

A. Participants

One hundred and twenty young adult subjects, were split into 12 groups of 10 subjects. No participant reported any history of speaking or hearing disorders. Sixty native French speakers, born and raised in France, were tested in Paris, France (age range = [21-34 years]; mean = 24; standard deviation (SD) = 2.5 years; 24 females). They did not learn any tonal language. The other 60 participants were native Mandarin speakers, born and raised in Taiwan, and tested in Taipei, Taiwan (age range = [19-29]; mean = 23 years; SD = 2.5 years; 33 females). For each participant the musical experience was assessed and 52 native French-

speaking and 44 native Mandarin-speaking listeners reported to practice music. Table 1 summarizes the participants' characteristics.

-Table 1 about here-

B. Stimuli

Three Thai tones (rising, falling and low) were pronounced by a native female speaker asked to speak clearly (F0 range=100-350 Hz, as estimated by YIN algorithm; de Cheveigné and Kawahara, 2002) with the syllable /ba/ (see the Figure 1 for a representation of the F0 trajectories, the sound files also available online). All stimuli were recorded digitally via a 16-bit A/D converter at a 44.1-kHz sampling frequency. In each lexical tone, eight different tokens were chosen because of their higher clarity. The mean duration of the stimuli was 661.6 ms (SD=32.3 ms) for the rising tones, 509.9 ms (SD=36.8) for the falling tones, and 636 ms (SD=31.2) for the low tones.

-Figure 1 about here-

In the first condition ("Intact"), each digitized signal was passed through a bank of 32, fourth-order gammatone filters (Gnansia *et al.*, 2009; Patterson, 1987) -also called "analysis filters" hereafter, each 1-ERB_N-wide (average equivalent-rectangular-bandwidth of the auditory filter as determined using young normally hearing listeners tested at moderate sound levels; Moore, 2007) with CFs uniformly spaced along an ERB_N-number scale (one filter per ERB_N) ranging from 80 to 8020 Hz. The Hilbert transform was then applied to each bandpass filtered speech signal to extract the AM and FM components. The temporal envelope (AM component) was lowpass filtered using a zero-phase Butterworth filter (36 dB/octave rolloff). Cochlear filtering (modelled here by gammatone filtering) imposes limitations on the maximum AM rate. As in Gnansia et al. (2009), we adjusted the cutoff frequency of the lowpass filter (following gammatone filtering) to $ERB_N/2$; the ERB_N corresponded to that of the (normal) cochlear filter tuned to the geometric center of the 1-ERB_N wide gammatone

filter. In each band, the FM carrier was then multiplied by the filtered AM function. Finally, the narrowband speech signals were added up and the level of the resulting speech signal was adjusted to have the same root-mean square (RMS) value as the input signal. In this condition, speech processing resulted in near-perfect stimulus reconstruction.

In the second condition ("Vocoded"), the same signal processing scheme was used, except that each digitized signal was passed through a bank of 8, fourth-order gammatone filters, each 4-ERB_N-wide with CFs uniformly spaced along an ERB_N-number scale (one filter per ERB_N) ranging from 80 to 8020 Hz. In this condition, the cut-off frequency of the low-pass filter used to extract temporal envelopes was set to $ERB_N/2$; the ERB_N corresponded to that of the (normal) cochlear filter tuned to the geometric center of the 4-ERB_N wide gammatone filter. The original FM carriers were replaced by sine wave carriers with frequencies at the center frequency of the gammatone filters, and with random starting phase in each analysis band.

In a third condition ("F0-modulation"), the stimuli were generated by first extracting the F0 trajectory of each original lexical tone using the YIN algorithm (de Cheveigné and Kawahara, 2002; implemented in MATLAB, MathWorks, Natick, MA). Then, this F0 trajectory was used to modulate the F0 of a periodic click train (more precisely, the signal was a periodic click train of 88- µs square (*i.e.*, monophasic) pulses, which were repeated at a rate equal to 1/F0). A first order (butterworth) lowpass filter was used to limit the frequency range below 22050 Hz. The "F0-modulation" signals have the same duration as the original lexical tones (sound samples are provided on-line in the JASA supplementary material).

This study intended to measure lexical-tone discrimination on the sole basis on temporal-envelope (AM) cues for lexical-tone users and non-users. Note that the stimuli were not normalized in duration, because duration cues are considered to be an element of

temporal-envelope cues (Rosen, 1992). Thus, normalization in duration would degrade envelope cues.

C. Procedure

A same/different task like Burnham and Francis (1997) was used to assess discrimination using E-Prime 2.0 software (Psychology Software Tools). Two different ISIs (500 and 1500 ms) were used. In each vocoder condition (Intact, Vocoded or F0-modulations), eight practice trials, with feedback, were first presented with unrelated sounds (unprocessed syllables /ko/ and /mi/) to train subjects with the task. This was followed by a test phase composed of 48 test trials, without feedback. Half of the trials consisted of the presentation of two stimuli of the same category, and the other half in the presentation of two stimuli belonging to two different categories (8 trials per contrast: rising/low, rising/falling, and low/falling). "Same" and "different" trials were presented in random order within two blocks (of 24 trials). Each subject was randomly assigned to a given experimental condition (Intact, Vocoded or F0-modulations) and a given ISI duration (500 or 1500 ms) in order to reduce training effect between conditions and prevent recognition of the sounds in the degraded conditions Thus, six independent groups of ten adults from each language background were tested in a soundproof booth in Paris or in Taipei using the same material and by the same experimenter. All stimuli were presented in free field using a Fostex (model PM0.5) speaker at a 70 dB SPL. Subjects sat in front of a computer controlling the experiment and 50 cm from the speaker located on their right side (*i.e.*, at 40 deg azimuth and 0 deg elevation). Subjects were instructed to listen carefully to the pairs of sounds and to respond as fast and as accurately as possible by pressing two different keys (indicating same or different). The subject's accuracy was estimated by a *d' score* where $d' = Z[p(\text{hit}) - Z[p(\text{false alarms})]$ (see Macmillan and Creelman, 1991). If the hit rate was 1.0 or the false alarm rate was 0, d'

was calculated after adjusting the hit rate or false alarm rate by the reciprocal of the number of trials.

III. RESULTS

The *d'* scores of the native-Mandarin and native-French participants for each tone contrast are represented in Figure 2 for the "Intact", "Vocoded" and "F0-modulations" conditions. A *d'* of 0 corresponds to chance, and a *d'* of 2.68 corresponds to perfect discrimination. French and Mandarin speakers showed similar discrimination performance in the "Intact" speech condition. In the "Vocoded" speech condition, the performance of both groups decreased significantly but remained above chance level. Moreover, Mandarin speakers scored lower than French speakers in this condition (when the fine spectro-temporal modulation cues conveying voice-pitch information were severely degraded). In the "F0-modulations" condition, Mandarin speakers showed slightly better discrimination of (F0) pitch contours compared to French speakers.

-Figure 2 about here-

Even if native French-speaking listeners had more practice in music than native Mandarin-speaking listeners [mean=6.30, sd=6.14, vs. mean=3.61, sd=3.8 for years of musical practice], a preliminary analysis of variance (ANOVA) on the total *d'* score showed no main effect of the factor Musician or interaction with Language. Thus, data were collapsed across this variable in the main analyses. A repeated-measures ANOVA was performed with Condition (Intact vs. Vocoded vs. F0-modulations), ISI (500 ms vs. 1500 ms) and Language (French vs. Mandarin) as between-subject factors and Contrast (rising/low vs. rising/falling vs. low/falling) as within-subject factor, to assess the role of Language and ISI in the three experimental conditions on the *d'* scores. This analysis revealed a main effect of Condition [$F(2,108)=125.54$, $p < .001$] and a *post-hoc* Tukey's HSD test showed that the "Vocoded" speech condition led to lower discrimination scores compared to "Intact" and "F0-

modulations" conditions (see Figure 2). No main effect of ISI or Language was found. However, a main effect of Contrast was found [F(2,216)=5.4, p=.005]. *Post-hoc* comparisons (Tukey' HSD test) showed that as expected, the "rising-falling" contrast was easier to discriminate than "rising-low"; no difference was observed between the other two contrasts. Moreover, a significant Condition x Contrast interaction was observed [F(4,216)=4.1, p=.003] indicating that the "rising-falling" contrast was easier to discriminate than the "rising-low" contrast in the "Vocoded" speech condition. Furthermore, the significant interaction of Condition x Contrast x Language [F(4,216)=3.51, p=.008] revealed that the higher scores obtained for the "rising-falling" contrast were mainly obtained by Mandarin-speaking listeners.

The significant interaction Contrast x ISI x Language [F(2,216)=3.28, p = .04] indicated that the better *d'* scores for the "rising-falling" contrast were exhibited by the French-speaking listeners with an ISI of 500 ms and by the Mandarin ones with an ISI of 1500 ms. Finally, a significant interaction between Condition x ISI x Language [F(2,108)=4.12, p = .02] showed that French-speaking participants were better than Mandarin ones with a short ISI in the "Vocoded" condition.

A significant interaction Condition x Language was also found [F(2,108)=5.79, p =.004]. To explore further this interaction and to compare the discrimination performance of Mandarin and French-speaking listeners in each condition, separated repeated-measures ANOVAs were run with Language (French vs. Mandarin) and ISI (500 ms vs. 1500 ms) as between-subject factors and Contrast (rising/low vs. rising/falling vs. low/falling) as within-subject. In the "Intact" condition, no main effect or interaction with Language was observed. A main effect of Contrast was observed [F(2;72) = 3.75, p = .03] indicating that the "falling-low" contrast was more difficult to discriminate than the two other contrasts. In the "Vocoded" speech condition, a main effect of Contrast was observed [F(2;72) = 4.69, p = .01]

but indicated that the "rising-falling" contrast was easier to discriminate than the "rising-low". A marginal effect of Language was observed [$F(1;36) = 3.83$; $p = .058$], showing that French participants tended to be slightly better at discriminating the vocoded lexical tones than Mandarin participants. Although the *d'* scores of both groups remained above chance level (Student *t* test; all $p<.001$), a significant interaction between Contrast and Language [$F(2,72)=3.4$, $p = .04$] was observed. *Post-hoc* comparisons (Tukey's HSD test) indicated that the *d'* scores varied between contrasts only for Mandarin-speaking listeners and that they better discriminated the "rising-falling" than the "falling-low" contrast. A significant interaction between ISI and Language was also found [$F(2,72)=4.39$, $p = .04$] indicating that French speaking-listeners were better than the Mandarin ones with a ISI of 500 ms. In the "F0-modulations" condition, a main effect of Language was observed [$F(1,36)=7.31$, $p = .01$] and *post-hoc* comparisons revealed that Mandarin-speaking listeners obtained (slightly) higher *d'* scores compared to French-speaking listeners.

IV. DISCUSSION

The present study aimed to investigate the role of native language on the processing of spectro-temporal cues in lexical-tone discrimination. In apparent contrast with previously published investigations on lexical-tone discrimination (*e.g.*, Burnham and Francis, 1997; Hallé *et al.*, 2004; Sun and Huang, 2012), both French- and Mandarin-speaking listeners were able to correctly perceive differences in pitch contours with the present lexical tones. The absence of difference between language groups results from a ceiling effect (that is from the high performance of both groups for the current discrimination task and for the present speech stimuli that were carefully selected from a large set of clearly articulated utterances). As expected in the "Vocoded" speech condition, results indicate an effect of native language on the perception of F0 variations and suggest that lexical-tone users are more dependent on fine spectral and FM cues than non-users when perceiving lexical tones. Moreover, French-

speaking listeners are better able to make use of the remaining (AM) information (conveying duration and loudness information; *cf*. Rosen, 1992 (p. 74, lines 13-14)) than Mandarin-speaking listeners in this condition. One possibility is that the experience with French prosody and phonological categories makes French-speaking listeners more likely to focus on the AM cues available (related to rhythm information) in the Vocoded stimuli and potentially relevant for French language. Moreover, these results are consistent with previous studies showing that native lexical-tone users rely more on the F0 direction than non-users (*e.g.,* Lee *et al.,* 2008, 2010). Furthermore, the duration of ISI influenced French and Mandarin participants' performance differently in that Vocoded condition. As expected, better performance for the "rising-falling" contrast was observed with a short ISI for French speakers and with a long ISI for Mandarin speakers. This may reveal that native language affects the rate of information decay in the short-term memory representation (and/or its use for further perceptual processes) of the voice-pitch cues used to discriminate lexical tones (Mandarin speakers showing less information decay than French speakers).

Finally, when the fine spectro-temporal modulations conveying the voice-pitch information were presented for discrimination using "F0-modulations" sounds, Mandarin-speaking listeners showed better discrimination of (F0) pitch contours compared to French-speaking listeners. These results are in line with several studies showing an effect of native language on the identification of pitch contours in non-linguistic signals such as sine waves, harmonic complex tones, or iterated rippled noises (*e.g.,* Bent *et al*., 2006; Swaminathan *et al.*, 2008; Xu *et al.*, 2006).

Overall, the present results suggest that Mandarin-speaking listeners are more dependent on F0 variations–and thus on FM and fine spectral cues–than French-speaking listeners when discriminating lexical tones. The results of this pilot study are consistent with the notion that native language shapes the weight of spectro-temporal fine structure cues in

processing speech sounds. Results obtained with click trains suggest that this influence of native language could extend to non-linguistic sounds. Thus, temporal modulation processing in the auditory system may be influenced by higher-level mechanisms fine-tuned by language expertise. This influence is not incorporated in current models of modulation processing (Jørgensen and Dau, 2013). However, our conclusions should be taken with caution because several factors such as genetic and socio-cultural backgrounds and environmental factors were not controlled in the present study. Further work investigating the development of lexical-tone processing for infants learning tonal and non-tonal languages is required to address these issues. Moreover, the perceptual strategies (*i.e.*, Bent *et al.*, 2006) used by native and non-native lexical-tone listeners when listening to degraded lexical tones may be further explored using a more cognitively demanding task such as an identification task.

Acknowledgments

The authors wish to thank all the participants of this study. C. Lorenzi was supported by a grant from Agence Nationale de la Recherche (ANR; HEARFIN project). This work was also supported by ANR-11-0001-02 PSL* and ANR-10-LABX-0087. J. Bertoncini was also supported by ANR-12-ISH2-0001-01 in France and F. M. Tsao was supported by NSC-102-2923-H-002-001-MY3 in Taiwan.

REFERENCES

Abramson, A. S. (**1978**). "Static and dynamic acoustic cues in distinctive tones," Lang. Speech, **21**, 319–325.

Bent, T., Bradlow, A. R., and Wright, B. A. (**2006**). "The influence of linguistic experience on the cognitive processing of pitch in speech and nonspeech sounds," J. Exp. Psychol. Hum. Percept. Perform., **32**, 97–103.

Burnham, D., and Francis, E. (**1997**). "The role of linguistic experience in the perception of Thai tones," Southeast Asian linguistic studies in honour of Viehin Panupong (Vol. 8, Chulalongkorn University press, Bangkok), pp 29-47.

Burnham, D., and Mattock, K. (**2007**). "The perception of tones and phones," Language Experience in Second Language Speech Learning: In honor of James Emil Flege, pp 259-280.

De Cheveigné, A., and Kawahara, H. (**2002**). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., **111**, 1917–1930.

Clément, S., Demany, L., and Semal, C. (**1999**). "Memory for pitch versus memory for loudness," J. Acoust. Soc. Am., **106**, 2805–2811.

Fu, Q.-J., Zeng, F.-G., Shannon, R. V., and Soli, S. D. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am., **104**, 505–515.

Gandour, J. T., and Harshman, R. A. (**1978**). "Crosslanguage differences in tone perception: A multidimensional scaling investigation," Lang. Speech, **21**, 1–33.

Gnansia, D., Péan, V., Meyer, B., and Lorenzi, C. (**2009**). "Effects of spectral smearing and temporal fine structure degradation on speech masking release," J. Acoust. Soc. Am., **125**, 4023–4033.

Hallé, P. A., Chang, Y.-C., and Best, C. T. (**2004**). "Identification and discrimination of Mandarin Chinese tones by Mandarin Chinese vs French listeners," J. Phon., **32**, 395–421.

Huang, T., and Johnson, K. (**2010**). "Language specificity in speech perception: perception of Mandarin tones by native and nonnative listeners," Phonetica, **67**, 243–267.

Jørgensen, S., and Dau, T. (**2013**). "Modelling speech intelligibility in adverse conditions," Adv. Exp. Med. Biol., **787**, 343–351.

Kong, Y.-Y., and Zeng, F.-G. (**2006**). "Temporal and spectral cues in Mandarin tone recognition," J. Acoust. Soc. Am., **120**, 2830–2840.

Kuo, Y.-C., Rosen, S., and Faulkner, A. (**2008**). "Acoustic cues to tonal contrasts in Mandarin: Implications for cochlear implants," J. Acoust. Soc. Am., **123**, 2815–2864.

Lee, C.-Y., Tao, L., and Bond, Z. S. (**2008**). "Identification of acoustically modified Mandarin tones by native listeners," J. Phon., **36**, 537–563.

Liang, Z. A. (**1963**). "The auditory perception of Mandarin tones," Acta Physiol. Sinica, **26**, 85–91.

Patterson, R. D. (**1987**). "A pulse ribbon model of monaural phase perception," J. Acoust. Soc. Am., **82**, 1560–1586.

Rosen, S. (**1992**). "Temporal information in speech: acoustic, auditory and linguistic aspects," Philos. T. R. Soc. B, Biol. Sci., **336**, 367–373.

Shamma, S., and Lorenzi, C. (**2013**). "On the balance of envelope and temporal fine structure in the encoding of speech in the early auditory system," J. Acoust. Soc. Am., **133**, 2818–2833.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science, **270**, 303–304.

Smith, Z. M., Delgutte, B., and Oxenham, A. J. (**2002**). "Chimaeric sounds reveal dichotomies in auditory perception," Nature, **416**, 87–90.

Sun, K.-C., and Huang, T. (**2012**). "A cross-linguistic study of Taiwanese tone perception by Taiwanese and English listeners," J. East Asian Linguist., **21**, 305–327.

Swaminathan, J., Krishnan, A., and Gandour, J. T. (**2008**). "Pitch encoding in speech and nonspeech contexts in the human auditory brainstem," Neuroreport, **19**, 1163–1167.

Wang, S., Xu, L., and Mannell, R. (**2011**). "Relative contributions of temporal envelope and fine structure cues to lexical tone recognition in hearing-impaired listeners," J. Assoc. Res. Otolaryngol., **12**, 783–794.

Werker, J. F., and Tees, R. C. (**1984**). "Phonemic and phonetic factors in adult cross-language speech perception," J. Acoust. Soc. Am., **75**, 1866–1878.

Whalen, D. H., and Xu, Y. (**1992**). "Information for Mandarin tones in the amplitude contour and in brief segments," Phonetica, **49**, 25–47.

Xu, L., and Pfingst, B. E. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Am., **114**, 3024–3027.

Xu, L., Tsai, Y., and Pfingst, B. E. (**2002**). "Features of stimulation affecting tonal-speech perception: implications for cochlear prostheses," J. Acoust. Soc. Am., **112**, 247–258.

Xu, Y., Gandour, J. T., and Francis, A. L. (**2006**). "Effects of language experience and stimulus complexity on the categorical perception of pitch direction," J. Acoust. Soc. Am., **120**, 1063–1074.

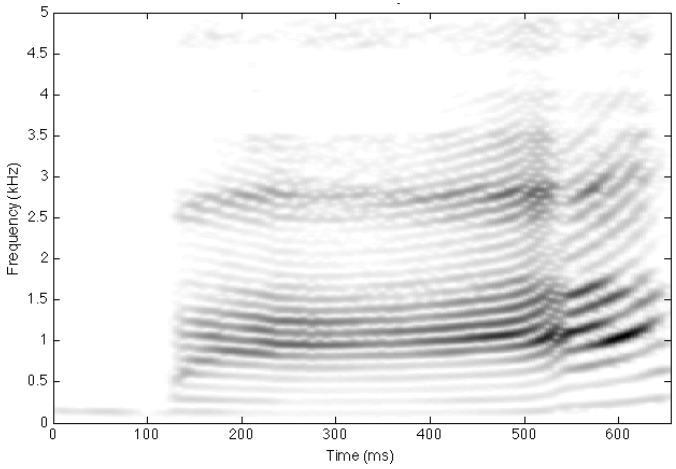
Figure captions

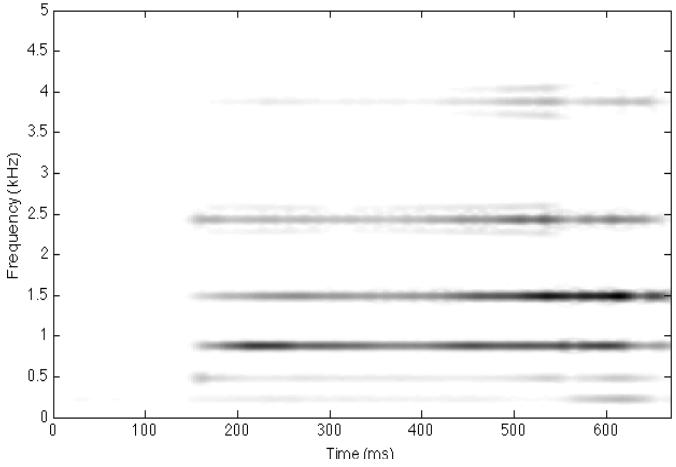
Table 1. Participants' information

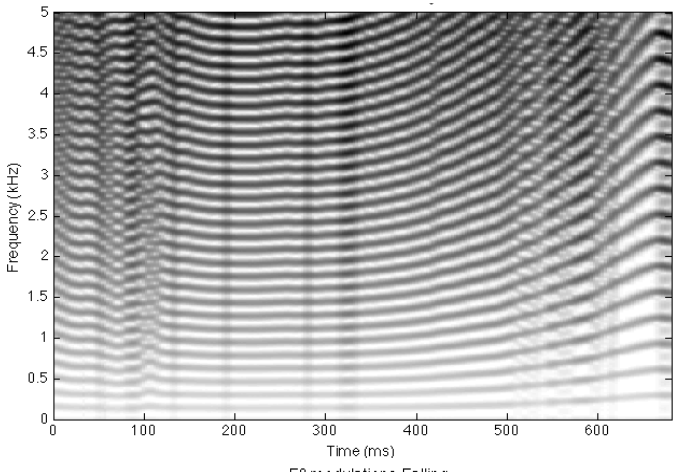
Figure 1. Spectrograms of the Intact (left panels), Vocoded (middle panels) and F0-modulation (right panels) versions of a /ba/ rising (first row), /ba/ falling (second row) and /ba/ low (third row) stimulus.

Figure 2. *d'* scores of the Mandarin-speaking and French-speaking listeners in the three experimental conditions (Intact, Vocoded and F0-modulations) for each lexical-tone pair (RL: Rising-Low; RF: Rising-Falling; FL: Falling- Low) presented in the two ISI (500 ms; 1500 ms). The bars represent the standard errors.

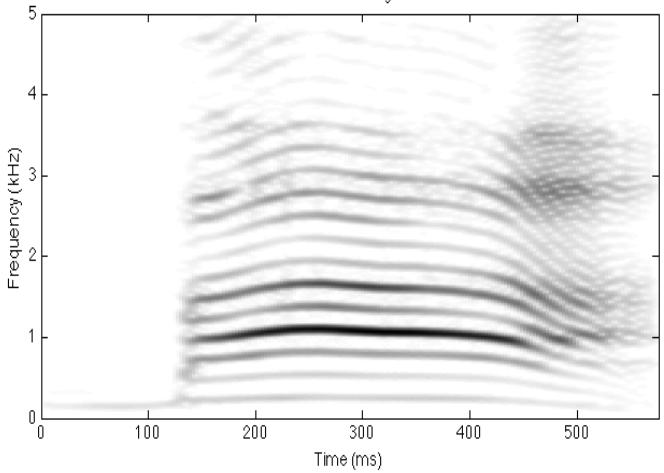
Condition	Isi	Language	mean age (years)	sd age (years)	age range (years)	number of females	number of musicians	mean musical practice (years)	sd of musical pratice	range of musical pratice (years)
Intact	1500	French	25	3.4	22-31	7	6	6.9	7.7	0-23
Intact	500	French	24	1.9	21-27	6	6	5.9	6	0-15
Intact	1500	Mandarin	24	2.4	22-29	7	5	3.4	4.1	0-14
Intact	500	Mandarin	23.6	2.5	20-27	6	6	4.9	2.9	0-10
Vocoded	1500	French	23.8	1.4	22-26	1	6	5	5.3	0-15
Vocoded	500	French	25.3	3.7	22-34	0	7	7.3	6.3	0-20
Vocoded	1500	Mandarin	22.1	1.7	19-24	4	4	4.5	4	0-12
Vocoded	500	Mandarin	23.2	2.9	20-29	6	6	3.8	4.1	0-13
F0-modulations	1500	French	24.9	1.97	22-29	7	6	5.4	6	0-16
F0-modulations	500	French	24.6	1.9	22-28	3	7	7.3	6.5	0-18
F0-modulations	1500	Mandarin	23.9	3.1	20-28	5	4	2	2.2	0-7
F0-modulations	500	Mandarin	21.7	1.9	20-26	6	5	3.4	4.7	0-15



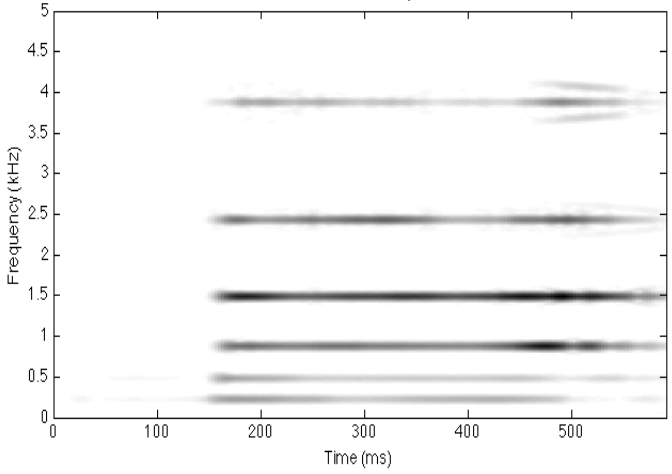




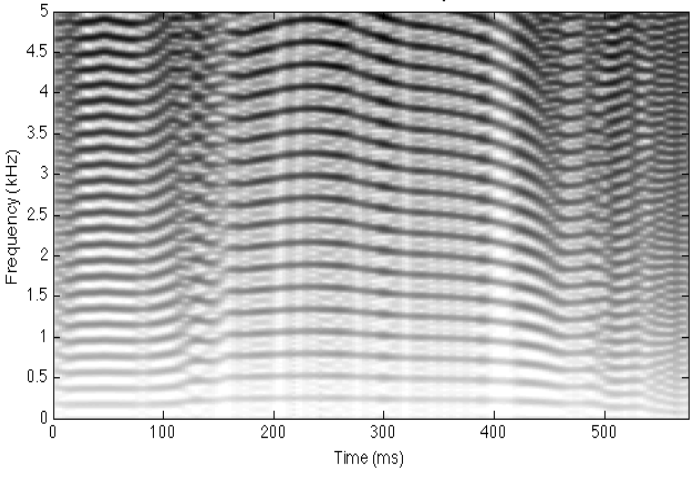
Intact-Falling



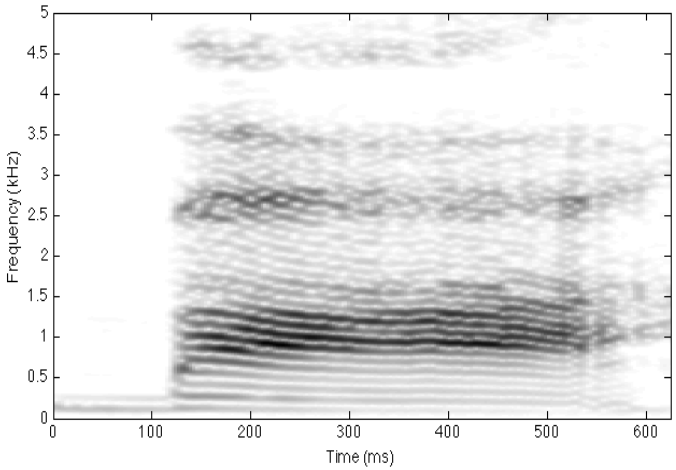
Vocoded-Falling



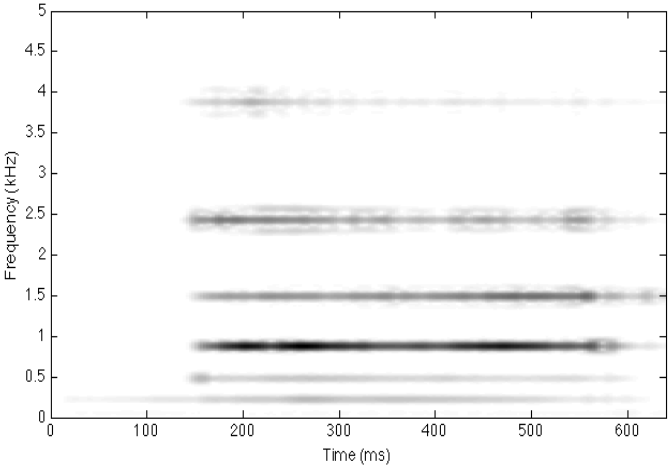
F0 modulations-Falling



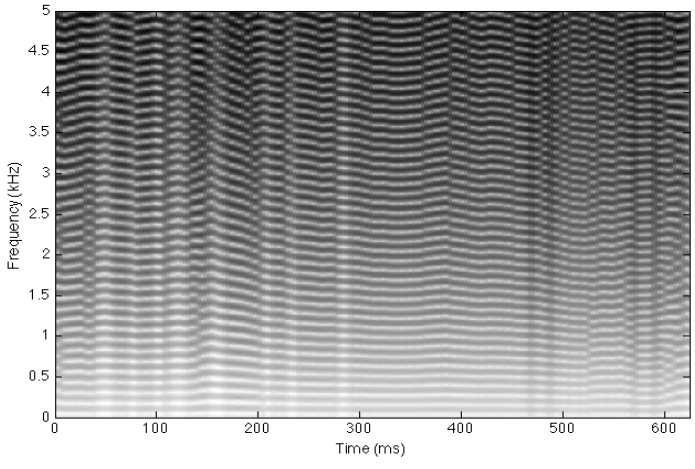
Intact-Low

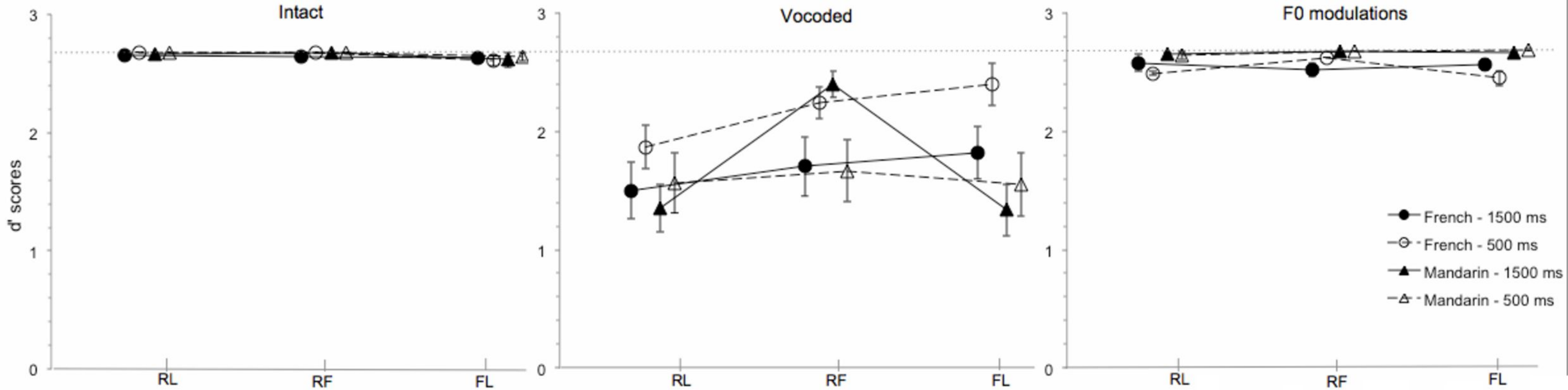


Vocoded-Low



F0 modulations-Low




Intact
Vocoded
F0 modulations
d' scores
0
1
2
3
RL
RF
FL
French - 1500 ms
French - 500 ms
Mandarin - 1500 ms
Mandarin - 500 ms