SUPPLEMENTARY METHODS

Subjects

Twenty subjects (11 females) participated in this study. None of the subjects had previous exposure to a tone language. Subjects were divided into two groups based on musical training. Amateur musicians were defined as instrumentalists having at least six years of continuous musical training (mean = 10.7 years) starting at or before the age of 12, in addition to currently playing their instrument. Nonmusicians were defined as having no more than three years of musical training (mean = 1.2 years) at any time in their life. Subjects’ musical history information is summarized in Table S1. All subjects were right handed and reported no audiologic or neurologic deficits. All subjects had normal click-evoked auditory brainstem response latencies and normal hearing thresholds at or below 20 dB HL for octaves from 125 to 4000 Hz. The two subject groups did not differ in age or handedness scores.

Stimuli

A native speaker of Mandarin Chinese was asked to produce /mi/ with three Mandarin tones: /mi1/ ‘to squint,’ /mi2/ ‘bewilder,’ and /mi3/ ‘rice’ (by convention, the number indicates tone or lexically meaningful pitch contour: Tone 1 = level tone, Tone 2 = rising tone, and Tone 3 = dipping tone). Recording took place in a sound attenuated chamber using a SHURE SM58 microphone recorded at 44.1 kHz onto a Pentium IV PC. These original productions were then duration-normalized to 278.5 milliseconds (ms) using Praat\(^1\). Using Praat, the pitch (\(f_0\)) contours of each of the original production were extracted and then superimposed onto the original Tone 1 (/mi1/) production using the
Pitch-Synchronous Overlap and Add (PSOLA) method, which resulted in perceptually natural stimuli as judged by four native speakers of Mandarin. The stimuli, therefore, consisted of three instances of /mi/ (in three Mandarin tones) differing only in $f_0$. These stimuli were RMS amplitude normalized using the software Level 16. To accommodate the capabilities of our stimulus presentation software, the stimuli were resampled to 22.05 kHz. **Fig. S1** shows the $f_0$ contours of the three stimuli ($f_0$ ranges: 140-172 Hz, 110-163 Hz, and 89-110 Hz, respectively). It is worth pointing out that we use the term “linguistic pitch” to describe these $f_0$ contours because they were embedded in speech, not music. We realize that none of our subjects spoke a tone language and thus these $f_0$ contours were not lexicalized. It is, therefore, likely that these $f_0$ contours were interpreted as intonational tones, which also carry linguistic functions.

**Physiologic (ERP) Recording Procedures**

Physiologic recording procedures were similar to our published studies (e.g., Russo et al.). During testing, subjects watched a videotape with the sound level set at $< 40$ dB SPL to facilitate a quiet yet wakeful state. Subjects listened to the video soundtrack (presented in free field) with the left ear unoccluded, while the stimuli were presented to the right ear through ER-3 ear inserts (Etymotic Research, Elk Grove Village, IL) at about 70 dB SPL (Stim, AUDCPT, Compumedics, El Paso, TX). The order of the three stimuli was randomized across subjects with a variable inter-stimulus interval between 71.50 and 104.84 ms. Responses were collected using Scan 4.3 (Compumedics, El Paso, TX) with Ag–AgCl scalp electrodes, differentially recorded from Cz (active) to ipsilateral earlobe (reference), with the forehead as ground. Two blocks of 1200 sweeps per block
were collected at each polarity with a sampling rate of 20 kHz. Filtering, artifact rejection and averaging were performed offline using Scan 4.3. Responses were bandpass filtered from 80-1000 Hz, 12 dB/octave, and trials with artifacts greater than 35 µV were rejected. Waveforms were averaged with a time window spanning 45 ms prior to the onset and 16.5 ms after the offset of the stimulus. Responses of alternating polarity were then added together to isolate the neural response by minimizing stimulus artifact and cochlear microphonic. For the purpose of calculating signal-to-noise ratios, a single waveform representing non-stimulus-evoked neural activity was created by averaging the neural activity 45 ms prior to stimulus onset.

*Analysis Procedures*

For each subject, we calculated two primary measures of FFR pitch-tracking: *stimulus-to-response correlation* and *autocorrelation*. These measures were derived using a sliding-window analysis procedure in which 40-ms bins of the FFR were analyzed in the frequency and lag (autocorrelation) domains. The FFR was assumed to encompass the entire response beginning at time 1.1 ms, the transmission delay between the ER-3 transducer and ear insert. The 40-ms sliding window was shifted in 1 ms steps, to produce a total of 238 overlapping bins. A narrow-band spectrogram was calculated for each FFR bin by applying the Fast Fourier Transform (FFT) to windowed bins (Hanning window) of the signal. To increase spectral resolution, each time bin was zero-padded to 1 second before performing the FFT. The spectrogram gave an estimate of spectral energy over time and the $f_0$ (pitch) contour was extracted from the spectrogram by finding the frequency with the largest spectral magnitude for each time bin. Spectral peaks that
Supplementary Methods: Wong, Skoe, Russo, Dees, Kraus

did not fall above the noise-floor were excluded as possible $f_0$ candidates. Both $f_0$ frequency and magnitude were recorded for each time bin, and the $f_0$ amplitude measure was calculated as the average magnitude across bins. The same short-term spectral analysis procedure was applied to the stimulus waveforms to calculate the degree of similarity (stimulus-to-response correlation) between the stimulus and response $f_0$ contours, defined as the Pearson’s correlation coefficient ($r$) between the stimulus and response $f_0$ contours. This measure represents both the strength and direction of the linear relationship between to two signals.

The second measure of pitch-tracking, autocorrelation, was derived using a pitch-detection short-term autocorrelation method\(^6\). Each of the 238 time bins was cross-correlated with itself to determine how well the bin matched a time-shifted version of itself. The maximum (peak) autocorrelation value (expressed as a value between 0 and 1) was recorded for each bin, with higher values indicating more periodic time frames. The autocorrelation pitch tracking measure was calculated by averaging the autocorrelation peaks ($r$-values) from the 238 bins for each tone for each subject. Running-autocorrelograms (lag versus time) (see Krishnan et al.\(^7\)) were calculated as a means of visualizing and quantifying periodicity and pitch strength variation over the course of the response. In the pitch-tracking and autocorrelation plots (Fig. 1, middle and bottom panels), the time indicated on the x-axis refers to the midpoint of each 40-ms time bin analyzed. For example, the $f_0$ extracted from the first FFR time bin (1.1 ms - 40.1 ms) is plotted at time 21.1 ms.

We also measured the RMS (Root-Mean-Square) amplitude of the FFR waveform, which is the magnitude of neural activation over the entire FFR period (1.1 –
295 ms). This measure takes both negative and positive peaks into consideration. This *FFR RMS amplitude* is driven largely by the amplitude of the $f_0$ (a description of the $f_0$ amplitude calculation is provided above). If a subject has robust pitch-tracking, the largest peaks in the response waveform will fall at the period of the $f_0$. In addition, to quantitatively consider the proportion of the $f_0$ amplitude relative to the overall *FFR RMS amplitude*, we calculated $f_0$-*FFR proportion*, which is the average $f_0$ amplitude divided by the total RMS amplitude.

The use of multiple pitch-tracking measures allows us to more comprehensively observe and quantify pitch encoding differences between the two groups. All pitch-tracking analyses were performed using routines coded in Matlab 7.4.1 (Mathworks, Natick, MA, 2005).

**Behavioral Testing (Tone Identification and Discrimination)**

Subjects also participated in two behavioral experiments designed to test their ability to identify and discriminate Mandarin tones. The stimuli and procedures, summarized briefly here, were essentially identical to Alexander, Wong, and Bradlow. Stimuli consisted of twenty monosyllabic Mandarin Chinese words. The five syllables /bu/, /di/, /lu/, /ma/, /mi/ were each produced in citation form with the four tones (level, rising, dipping, and falling) of Mandarin. Talkers consisted of two male and two female native speakers of Mandarin Chinese. Subjects participated in these two experiments after task familiarization. In tone identification, subjects matched the auditory stimulus with visually presented arrows depicting the pitch trajectory. In tone discrimination, subjects made a same-different judgment on the pitch patterns of stimulus pairs.
REFERENCES