Evaluation of Two Algorithms for Detecting Human Frequency-Following Responses to Voice Pitch

Fuh-Cherng Jeng[1], Jiong Hu[1], Brenda Dickman[1], Ching-Yu Lin[2], Chia-Der Lin[2], Ching-Yuan Wang[2], Hsiung-Kwang Chung[2]

[1] School of Hearing, Speech and Language Sciences, Ohio University, USA

[2] Department of Otolaryngology-HNS, China Medical University Hospital, Taiwan

Keywords: frequency-following response, algorithm, voice pitch, human

Abbreviations: f0, fundamental frequency; FFR, frequency-following response

Corresponding Author:

Fuh-Cherng Jeng, MD, PhD

Grover Center W224

School of Hearing, Speech and Language Sciences

Ohio University

Athens, OH 45701

USA

Phone: (740) 593 4157

Fax: (740) 593 0287

E-mail: jeng@ohio.edu

ABSTRACT

Voice pitch carries important cues for speech perception in humans. Recent studies have shown the feasibility of recording the frequency-following response (FFR) to voice pitch in normal-hearing listeners. The presence of such a response, however, was dependent on subjective interpretation of experimenters. The purpose of this study was to develop and test an objective method, including a control-experimental protocol and response-threshold criteria suitable for detecting the presence of an FFR to voice pitch. Eleven normal-hearing adults were recruited. A set of four Mandarin tones (Tone 1 flat; Tone 2 rising; Tone 3 dipping; and Tone 4 falling) was prepared to reflect the four contrastive pitch contours. The stimulus tokens were presented monaurally at 55 dB nHL. Electrical activities of the brain were recorded using three surface electrodes placed on the mid forehead and mastoids. Two distinctive algorithms, short-term autocorrelation in the time domain and narrow-band spectrogram in the frequency domain, were used to estimate the Frequency Error, Slope Error, Tracking Accuracy and Pitch Strength of the recordings taken from individual listeners as well as the power, false-positive rate, probability of errors and efficiency of each algorithm. The results demonstrated that both algorithms were suitable for detecting the presence of an FFR to voice pitch.

INTRODUCTION

Developing neural indices of the brain's ability to process differences in the pitch of an incoming signal are important for measuring the experience-dependent brain plasticity and understanding the sensory-level processing of voice pitch at the brainstem level. Objective measures of a listener's responses to the changes in voice pitch, however, were not achieved until the past few years (Galbraith et al., 2004; Krishnan et al., 2004; Kraus & Nicol, 2005; Musacchia et al., 2007; Wong et al., 2007). These results supported the idea that when neurons in the human brainstem were passively activated by speech signals, the synchronized neural activities that reflected changes in voice pitch were preserved in the scalp-recorded frequency-following response (FFR) to voice pitch. Recent studies expanded the scope of the use of the FFR by showing the characteristics of such a response using speech (Aiken & Picton, 2006; Dajani et al., 2005; Krishnan et al., 2004) and non-speech (Krishnan et al., 2009a, 2009b; Swaminathan et al., 2008a, 2008b) stimuli in normal-hearing adults. Jeng and Schnabel (2009) also reported that the FFR recorded in young infants accurately reflected the pitch contours of acoustic stimuli. The presence of such a response, however, is dependent on subjective interpretation of the experimenters. If the FFR to voice pitch is meant to be an objective method to examine the pitch processing mechanisms in the human brainstem, development of objective methods and evaluation of automatic algorithms suitable for detecting the presence of such a response is needed. The primary aim of the present study was to develop an objective method using a control-experimental protocol and response-threshold criteria for objectively judging the existence of an FFR.

Pitch encoding in the auditory brainstem

Voice pitch is a psychological perception of the fundamental frequency (*f0*) of a speech signal and is determined by the vibration pattern of the human vocal folds. Although pitch perception ultimately resides in the neocortex, neurons in the brainstem play an important role in decoding features of the incoming signal, including the voice pitch. Recent advancements in electrophysiological techniques demonstrated that neurons in the human brainstem decode voice pitch with high temporal and spectral resolutions such that the *f0* and its harmonics of the incoming signal are preserved in the scalp-recorded FFRs (Galgraith et al., 2004; Krishnan et al., 2004; Kraus & Nicol, 2005).

Accurate encoding of voice pitch and its change over time is critical for listeners to perceive different lexical meanings and prosodic cues embedded in a speech signal. The human brainstem's ability to accurately follow the changes of voice pitch, as reflected by the scalp-recorded FFRs, has been reported in normal-hearing adults who spoke tonal (Krishnan et al., 2005, 2010; Swaminathan et al., 2008b) and non-tonal languages (Galbraith et al., 2004). The brainstem's ability to decode voice pitch accurately and deliver such information to the neocortex is also important for listeners to process and appreciate music. Recent studies (Musacchia et al., 2007; Wong et al., 2007) have shown that musical training enhanced the acuity of pitch tracking in the human brainstem, as reflected by the scalp-recorded FFRs to voice pitch in musicians versus non-musicians.

Development and evaluation of an objective method is particularly important when clinicians and researchers are trying to apply such a technique on populations who cannot provide reliable feedback such as infants, children, and difficult-to-test patients. For example, children with autism spectrum disorders showed deficient pitch-tracking accuracy compared to typically developed children (Russo et al., 2008). It is also reported that with short-term training on specific linguistic pitch contours, listeners not only were able to improve their behavioral response correctness, but also were able to express enhanced pitch-tracking accuracy reflected through scalp-recorded FFRs (Song et al., 2008). These findings support the notion that FFR to voice pitch can be a viable, objective, and non-invasive neurophysiological index of the brain's ability to process voice pitch. Most importantly, these findings also demonstrated potential clinical applications for diagnostic and remediation strategies for normal and pathological populations.

Pitch detecting algorithms

The secondary aim of the present study was to apply and evaluate the adequacy of this objective method by using two distinctive pitch detecting algorithms. One is the short-term autocorrelation algorithm that takes advantage of signal processing in the time domain; the other is the narrow-band spectrogram algorithm that analyzes the spectral components of the incoming signal.

Boersma (1993) adopted the short-term autocorrelation algorithm and described several techniques for improving the accuracy of pitch extraction. Briefly, this method employed an autocorrelation function on multiple time frames of a sampled signal. Fundamental frequency of each time frame was identified at the time shift that yielded the maximum value in the autocorrelation function. The f0 contour of the sampled signal was then constructed by concatenating the fundamental frequencies estimated from each of the time frames. Studies (Krishnan et al., 2004, 2005; Swaminathan et al., 2008a, 2008b; Wong et al., 2007) have shown the feasibility of using this short-term autocorrelation algorithm to extract the f0 contour of a response. Recordings taken from each individual, however, have less favorable signal-to-noise ratios than that of a grand-averaged recording from a group of participants. The outcome of the autocorrelation-based algorithm might consequently be compromised.

A second algorithm that can be used to detect the presence of an FFR is derived from a narrow-band spectrogram on the recordings. Briefly, this algorithm searches for the frequency that contains the largest spectral density in a pre-defined frequency range for each time frame in a recording. This technique is similar to the short-term autocorrelation algorithm, but the f0 of a sampled signal is determined by examining the distribution of the spectral energy of a recorded signal. When an FFR is present, spectral components of a recording that are in close proximity to the f0 contour of the stimulus would have relatively larger and distinguishable spectral energy than the frequency components that are further from the f0 contour of the stimulus. Thus, small spectral energies at the frequency range around the f0 contour of the stimulus token would indicate the presence of a response. Recent studies (Russo et al., 2008; Song et al., 2008) have shown the feasibility of using the narrow-band spectrogram algorithm to extract the f0 contours of the FFR to voice pitch.

Development of an objective method

Development of an objective method for detecting the presence of an FFR to voice pitch requires the use of a control-experimental protocol. That is, one can assure the presence of such a response in the experimental condition and no response in the control condition. For scalp-recorded FFRs from human listeners, it cannot be assumed that all participants will produce measureable responses with the same characteristics. Thus, pre-defining a set of recordings with ‘known’ responses becomes a challenge. In ideal situations, one could simulate the existence of an FFR by using a set of known mathematical equations to ‘generate’ a known response and inject the known responses in a controlled condition to create a set of recordings with known responses. This approach, however, is not applicable until mechanisms of pitch encoding in the human brainstem is thoroughly understood and the FFR to voice pitch can be readily derived using a set of mathematical equations. Thus, an alternative solution that has been commonly adopted in clinical settings is to recruit experienced human observers to determine the existence of a response and use that set of recordings as the ‘gold’ standard. This alternative approach provides a realistic solution for the development of an objective for detecting the existence of an FFR. Moreover, such an alternative approach has been used in clinical applications such as the automated ALGO algorithm (Natus Medical Inc., San Carlos, California USA) that has been commonly used for newborn hearing screenings to detect the presence of an auditory brainstem response (ABR) to clicks stimuli. During the development of objective method for detecting the presence of an ABR, a set of ABR waveforms was judged by experienced human observers and rated with known responses. These ABR waveforms were then used as a ‘gold standard’ template in the experimental protocol. Recordings obtained from individual neonates were then compared to the ABR template in order to determine the existence of a response.

To accommodate the requirements for developing an objective method for detecting the presence of an FFR, a control-experimental protocol was adopted. Specifically, the experimental condition was conducted with acoustic stimuli delivered monaurally to each listener's ear; whereas the control condition was performed when the sound tube was occluded and moved away from the participant's ear at the end of each testing session. All recordings were judged by three experienced observers to subjectively determine the presence of a response. All recordings were also analyzed using two objective pitch-detecting algorithms. Results obtained using the objective algorithms were then compared with subjective human judgments to derive the power, false-positive rate and probability of errors of this objective method.

MATERIALS AND METHODS

Experimental protocols and procedures used in this study were approved by the China Medical University Hospital (Taichung, Taiwan) Institutional Review Board. All recordings were obtained in an acoustically-treated chamber in the Auditory Electrophysiology Laboratory at China Medical University Hospital.

Participants

Eleven adult participants (5 males; mean ± S.D. = 31.4 ± 4.7 years) with hearing sensitivity $\leq$ 25 dB HL at octave frequencies from 125 to 8000 Hz were recruited. All participants were native speakers of Mandarin Chinese.

Preparation of acoustic tokens

A set of four monosyllabic Mandarin Chinese syllables was recorded by an adult Chinese male to create four contrastive pitch contours (Tone 1 flat /yi^1/ *clothes*, Tone 2 rising /yi^2/ *aunt*, Tone 3 dipping /yi^3/ *chair*, Tone 4 falling /yi^4/ *meaning*). These stimulus tokens were recorded in a sound-treated booth with an Audio-technica AT825 field recording microphone, connected through a preamplifier and an analog-to-digital converter (USBPre microphone interface) to an IBM-compatible computer. The recording of the stimulus tokens was digitally sampled using the Brown Lab Interactive Speech System v7 (BLISS, Providence, RI) at 40 kHz with 14-bit quantization. Each stimulus token was normalized to a duration of 250 ms with a rising/falling time of 10 ms using Praat v5.1 (Boersma and Weenink, 2009). Frequency ranges of the f0 contours for the four stimulus tokens (Tones 1, 2, 3 and 4) were 163-180, 116-157, 98-125 and 105-156 Hz, respectively.

Stimulus presentation

Presentation of the stimulus tokens and trigger synchronization were controlled by custom-made software written in LabView 8.0 (National Instruments, Austin, TA). All stimulus tokens were presented through a 12-bit digital-to-analog converter (National Instruments, DAQ 6062E) and a GSI 61 audiometer. All stimulus tokens were presented monaurally to the right ear through an electromagnetically-shielded insert earphone (Bio-logic Systems Corp., Mundelein, Illinois USA) at a stimulus level of 55 dB nHL. Two trials of 1200 sweeps were recorded using each stimulus token. The inter-stimulus interval was set at 50 ms. The four Mandarin tones were presented in a random order across participants. A control condition (sound tube occluded and moved away from the participant’s ear) was conducted at the end of each testing session to provide waveforms with no physiological responses to the stimuli. The control condition was needed not only to establish the experimental-control protocol, but also to ensure that stimulus artifact was appropriately eliminated from recordings.

Recording parameters

Three gold-plated recording electrodes were applied to all participants at the midline of the forehead at the hairline (non-inverting), right mastoid (inverting), and left mastoid (ground). All electrode impedances were under 3 kOhm at 10 Hz. Recordings were amplified (Neuroscan SynAmps^2, 24-bit resolution, least significant bit: 0.15 nV), bandpass filtered (0.05–3500 Hz, 6 dB/octave), and digitized at a rate of 20000 samples/s. Continuous data were recorded using Neuroscan Acquire 4.4 software (Compumedics, Charlotte, NC) and stored on a computer for offline analysis.

Data analysis

All data were analyzed using MatLab 2008a (MathWorks, Natick, MA) and EEGLab 6.01b (Swartz Center for Computational Neuroscience, San Diego, CA). To better isolate spectral energies at the f0 contours, continuous recordings were digitally bandpass filtered using a brick-wall, linear-phase finite-impulse-response (FIR) filter (85-1500 Hz, 3000 dB/octave). Filtered recordings were segmented into sweeps of 300 ms in length. A total of 1200 sweeps were collected for each condition. An individual sweep was rejected if it contained voltages greater than ±25 μV. During each recording condition, the typical rejection rate was less than 100 sweeps per trial. The remaining sweeps were averaged. To identify the onset of the response, the stimulus tokens were down-sampled to 2000 samples/s so that all stimuli and recordings had the same sampling rate. The down-sampled stimulus tokens were used throughout data analyses. Cross-correlation of the stimulus and recorded waveforms was performed to identify the time shift that produced the maximum cross-correlation value between the 3-10 ms response window. A 250 ms segment of the recorded waveform was extracted from the originally recorded waveform starting from the maximum cross-correlation value. The same analytical procedures were applied to all recordings obtained in the experimental and control conditions. Data obtained from each trial were analyzed separately. Test-retest reliability was determined from results of the two trials obtained from each participant.

Extraction of fundamental frequency (f0) contours

Two distinctive methods were used to estimate the *f0* contours of the stimulus tokens and recordings. All the stimulus tokens and recordings were first segmented using a 50-ms Hanning window with a step size of 1 ms. This resulted in a total of 201 windowed segments to be analyzed. In the spectrograms and pitch-tracking plots, the time shown on the abscissa indicates the midpoint of each 50-ms time window.

The first method used the short-term autocorrelation algorithm (Krishnan et al., 2004, 2005; Swaminathan et al., 2008a, 2008b; Wong et al., 2007) which extracted the pitch based on a time-domain analysis of the signal. Specifically, autocorrelation was conducted on each of the windowed segments. The time shift (τ_{max}) which yielded the maximum autocorrelation value between 5 and 13 ms was identified. This time range corresponded to 75–200 Hz that covered the frequency range of the *f0 contours*. The fundamental frequency of each windowed segment was calculated as $f0 = 1 / \tau_{max}$.

The second method used a "narrow-band spectrogram" algorithm (Krishnan et al., 2005; Krishnan et al., 2009b; Russo et al., 2008; Song et al., 2008; Wong et al., 2007) which extracted the pitch based on a spectral-domain analysis. The spectrogram was calculated with a frequency resolution of 4.88 Hz. For each windowed segment, this algorithm searched for the frequency corresponding to the maximal peak of the spectral density within a pre-defined frequency range. To ensure proper inclusion of f0 estimates, an additional measure, Pitch-Noise Ratio, was conducted. Pitch-Noise Ratio provides an estimate of the response amplitude to voice pitch relative to the amplitude of the ongoing neural responses that are not synchronized to the stimulus. Pitch was calculated by finding the spectral amplitude corresponding to the f0 of each time bin and averaging the spectral amplitudes across the 201 time bins. To obtain an estimate of the background physiological noise, waveforms were extracted from the 45-ms prestimulus interval to determine the amount of brain activities not synchronized to the stimuli. Spectral amplitudes corresponding to the response f0 frequencies were extracted from the prestimulus waveforms for each stimulus presentation. Pitch-Noise Ratio was then calculated as the dB ratio of the Pitch spectral amplitude relative to that of the Noise. Spectral peaks that contained Pitch-Noise Ratio$\leq$ 0 dB were excluded from possible candidates of *f0* estimates. The frequency that corresponded to the narrow-band spectrogram was determined as the f0 estimate for that windowed segment. This procedure was repeated for all windowed segments. All f0 estimates were concatenated to constitute the f0 contour of a recording. Different pre-defined frequency ranges (120-200 Hz for Tone 1, 90-180 Hz for Tone 2, 75-170 Hz for Tone 3, and 100-165 Hz for Tone 4) were used for each stimulus token and its associated recordings. The same techniques were applied to the stimulus waveforms.

Objective measures

f0 contours were extracted from each recording and were analyzed with respect to *f0* contours of the stimulus waveforms. Pitch-tracking accuracy and phase-locking magnitude were described by measures of Frequency Error, Slope Error, Tracking Accuracy, and Pitch Strength. Frequency Error represented the accuracy of pitch-encoding over the duration of stimulus presentation. Slope Error indicated the degree to which the shapes of the pitch contours were preserved in the human brainstem. Tracking Accuracy (i.e., the regression r values) denoted the overall faithfulness of pitch tracking between the stimulus and response f0 contours. Pitch Strength measured the magnitude of neural phase-locking to the *f0* contours of the stimulus waveforms.

To obtain estimates of the four objective measures, Frequency Error was computed by finding the absolute Euclidian distance between the *f0* contours of the stimuli and recordings and averaging the errors across the 201 windowed segments. Slope Error was derived by subtracting the slopes of the regression lines of the stimulus *f0* contours from the regression slopes of the recording *f0* contours. Slope estimates of the four stimulus tokens Tones 1-4 were 61, 268, 99, -265 Hz/s using the short-term autocorrelation algorithm and were 62, 272, 114, -372 Hz/s when using the narrow-band spectrogram algorithm. To obtain an estimate of Tracking Accuracy, linear regression was first conducted on a recording-versus-stimulus *f0* contours plot. Regression r value was then denoted as the Tracking Accuracy between the stimulus and recording *f0* contours.

The fourth measure, Pitch Strength, was derived from an autocorrelation function. Autocorrelation allowed the measurement of overall periodicity of a sampled signal. Specifically, each recording was multiplied by a copy of itself with increasing time shifts. For each time shift, an autocorrelation value was calculated and expressed between -1 and 1. Fundamental frequency was calculated from the output of the autocorrelation function by finding the time shift that yielded the maximum autocorrelation value and took the inverse of the time shift (i.e., frequency = 1/periodicity; e.g., 200 Hz = 1/5 ms). Pitch Strength was calculated from the autocorrelation function by finding the peak-to-trough amplitude starting from the maximum positive peak (within the 5-13 ms time shifts) to the following negative trough in the normalized autocorrelation output. Because the f0 contours of the four stimulus tokens used in this study fell within the frequency range of 75-200 Hz, the time shifts were limited to 5-13 ms when searching for the location of the maximum peak in the autocorrelation output.

Subjective judgment

To evaluate the power and false-positive rate of each algorithm, the presence of an FFR to voice pitch was also determined by visual inspection of experienced human observers. Three human observers who had at least 2 years experience in electrophysiological recordings and their morphologies were recruited to determine whether a recording contained an FFR to voice pitch. These observers were familiar with spectrograms of speech tokens, but were blind to the experimental setup and data collection protocol. All recordings obtained in the 8 experimental conditions (4 pitch contours + 4 control conditions) were used. These recordings were counterbalanced across stimulus tokens and were presented in a random order to each observer. The spectrogram of each recording was presented alongside in pairs with the spectrogram of the stimulus token. Each observer was instructed to determine the presence of an FFR by using the following criteria:

(1) f0 contour of the recording was clearly identifiable within the pre-defined frequency range;

(2) f0 contour energy of the recording exceeded the energy of the background noise located at other frequencies;

(3) f0 contour of the recording followed the general trend (i.e., flat, rising, dipping, or falling) of the f0 contour of the stimulus token; and

(4) f0 contour of the recording had no more than 2 disruptions along the f0 contour.

If all the criteria were met, the human observer would press on a “Yes” button on the computer monitor to indicate the presence of a response. If any criterion was not met, the human observer would press on a “No” button to indicate the absence of such a response. In this study, we used a 2-alternative-forced-choice paradigm. The presence of an FFR to voice pitch was determined if a recording received a “Yes” from two observers or more. The absence of an FFR was determined if a recording received a “No” from two observers or more.

The FFR elicited by voice pitch was visualized by plotting the distribution of spectral energies of the recordings as a function of time. Figure 1 shows a typical example of the spectrograms of the stimuli, responses and controls for the four different pitch contours used in this study. Spectrograms of the stimulus (top row) showed clear spectral energy located at the fundamental frequency and its harmonics. Spectrograms of typical recordings (middle row) taken from an adult participant showed FFRs that followed the f0 contours of the stimuli. For recordings taken from individual participants, disruptions of the FFR along the f0 contour were often observed. Spectrograms of the recordings obtained in the control condition (bottom row) showed energy randomly distributed within the pre-defined frequency range and no FFR was observed for any of the all four tones used in this study. Due to the relatively strict criteria used in human judgment, recordings with weak FFRs would likely be determined as “no response” waveforms. During the process of developing an objective method, this imperfection was inevitable until the existence of a weak response could be objectively identified precisely without involvement of human judgments.

Development of an objective method

By comparing the distribution curves for the experimental and control conditions, a response threshold value was calculated for each of the four Tones and the four objective measures. Due to the distinctive nature of the four objective measures, each measure was applied differently than the others. For Frequency Error, waveforms with Frequency Error values at or below the response thresholds were classified as responses. The threshold corresponded to the intersection between the experimental and control curves, the point at which the experimental condition had a higher probability of occurrence than the control conditions, was used to determine the presence of an FFR. Due to the nature of the bell-shaped distribution of Slope Error (i.e., Slope Error can be positive or negative numbers), two response thresholds were used to determine the presence of a response. Waveforms with Slope Error values within the two response thresholds were classified as responses. Similarly, waveforms with Slope Error values that fell outside of the two response thresholds were considered as no responses. For Tracking Accuracy, waveforms with Tracking Accuracy values at or above response thresholds were classified as responses. Pitch Strength used the same logistics as that used in Tracking Accuracy.

There were four possible outcomes based on the human judgments and test results. Two of them were correct interpretations and two were incorrect (Figure 2). To estimate the power and false-positive rate of the objective method for each algorithm used in this study, procedures published by Altman and Bland (1994) were followed. Specifically, the presence of an FFR was first determined by experienced human observers and treated as the “gold” standard; whereas the presence of an FFR determined by each algorithm was treated as the “test” result. Procedures used in this study were consistent with those employed in automatic ABR detection algorithms used for newborn hearing screenings. For example, the ALGO algorithm used a set of recordings where the presence of an ABR waveform was pre-determined by experienced human observers. These waveforms were then treated as the “gold” standard in deriving power and false-positive rate, for the use of developing an automatic ABR detecting algorithm

The bottom portion of Figure 2 shows the calculation of the commonly used operating characteristics of a test, including power, false-positive rate, efficiency and likelihood of types I and II errors. Power (i.e., sensitivity, true-positive rate) was calculated as the proportion percentage of actual positives (i.e., presence of an FFR determined by human observers) which were correctly identified as having a response. False-positive rate (i.e., 1-specificity, alpha error) was classified as the proportion percentage of true-negatives that were incorrectly identified as having a response.

RESULTS

Extraction of pitch contours

Short-term autocorrelation algorithm

Figure 3 shows a typical example of the f0 contours of the stimuli and responses recorded from a participant (Subject_010). The f0 contours, extracted using the short-term autocorrelation algorithm, are shown in Figure 3**a**. Tone 1 had a relatively flat pitch contour at around 174 Hz. Tone 2 had a rising pitch starting from 116 Hz to 158 Hz. Tone 3 had a dipping pitch contour with the highest frequency at 125 Hz and the lowest frequency at 98 Hz. Tone 4 had a falling voice pitch ranging from 185 Hz to 133 Hz. Using the short-term autocorrelation algorithm, the f0 contour estimates of responses generally followed the f0 contours of the stimuli. In some instances, the f0 contour estimates of the responses deviated from the f0 contours of the stimuli (e.g., Tone 3 around 140-160 ms and Tone 4 around 110-160 ms after stimulus onset). This was likely due to less favorable signal-to-noise ratios for recordings taken from individual listeners.

Narrow-band spectrogram algorithm

The f0 contours of the stimuli and responses extracted using the narrow-band spectrogram algorithm (Figure 3**b**) showed comparable results. The f0 contour estimates of the responses generally followed the f0 contours of the stimuli. In some instances, the f0 contour estimates of the responses showed deviations from the f0 contours of the stimuli (e.g., Tone 1 around 20-50 ms and Tone 2 around 0-5 ms after stimulus onset). This was likely due to either disruptions of the FFR along the f0 contour or background noise in the pre-defined frequency range which exceeded the response energy.

Objective measures of pitch-encoding

Figure 4 presents the distribution (i.e., histograms) of Frequency Error, Slope Error, Tracking Accuracy and Pitch Strength that were derived by using the short-term autocorrelation (left four columns) and narrow-band spectrogram (right four columns) algorithms.

Frequency Error

Distributions of the Frequency Error (Figure 4**a**) obtained in the experimental condition were skewed to the left, whereas the distributions of the Frequency Error obtained in the control condition were skewed to the right. Most importantly, distributions of the Frequency Error obtained in the experimental condition were distinctive from those obtained in the control condition. The differences between the peak locations of the Frequency Error distributions between the experimental (i.e., solid lines) and control (i.e., shaded areas) conditions were 21, 18, 18 and 15 Hz for Tones 1-4, respectively. When the short-term autocorrelation algorithm was used, distributions of the Frequency Error obtained in the experimental condition showed a maximum occurrence at 0, 0, 3 and 6 Hz for Tones 1-4, respectively. Distributions of the Frequency Error obtained in the control condition showed maximum occurrences at 21, 18, 21 and 21 Hz for Tones 1-4, respectively.

The narrow-band spectrogram algorithm showed comparable results. Distribution histograms of the Frequency Error obtained in the experimental condition showed a peak separation of 12, 12, 12 and 15 Hz for Tones 1-4, respectively, when compared with those obtained in the control condition. Tone 4 showed a bimodal distribution and, therefore, decreased the power (or sensitivity) of this algorithm.

Slope Error

Despite the use of four contrastive pitch contours with different slopes, distributions of the Slope Error (Figure 4**b**) showed prominent peaks for recordings in response to the four stimulus tokens. When the short-term autocorrelation algorithm was used, Slope Errors obtained in the experimental condition were clustered at around 0 Hz for the four pitch contours. In contrast, Slope Errors obtained in the control condition were scattered. Due to the bell-shaped distribution of Slope Error, two threshold lines were used for each tone. Waveforms with Slope Errors that fell outside of the two threshold lines were classified as having no response, whereas waveforms with Slope Errors that fell within the two threshold lines were considered responses.

When the narrow-band spectrogram algorithm was used, Slope Errors obtained in the experimental condition showed concentrated peaks at around -60 Hz for all pitch contours, whereas the Slope Errors obtained in the control condition were relatively scattered with rounded peaks at 0, -360, -180 and 360 Hz for Tones 1-4, respectively. This was likely due to the randomness of f0 estimates, and thus flatness of the slope estimates, derived from waveforms recorded in the control condition.

Tracking Accuracy

Figure 4**c** shows the distributions of the Correlation Coefficient between the stimulus and response f0 contours. When the short-term autocorrelation algorithm was used, distributions of the Tracking Accuracy in response to the four tones were skewed to the right, whereas the distributions of the Tracking Accuracy obtained in the control condition were relatively dispersed and skewed to the left. Distributions of the Tracking Accuracy demonstrated a maximum occurrence at 0.8, 0.9, 0.3 and 0.9 for Tones 1-4, respectively. The narrow-band spectrogram algorithm demonstrated comparable results. Distributions of the Tracking Accuracy revealed a maximum occurrence at 0.7, 0.9, 0.6 and 0.9 for Tones 1-4, respectively. Tone 4 showed a bimodal distribution and therefore decreased the power (or sensitivity) of this algorithm.

Pitch Strength

Distributions of the Pitch Strength (Figure 4**d**) obtained in the control condition showed separable trends than those obtained in the control condition. Although the Pitch Strength obtained in the experimental condition ranged from 0.1-0.9, the Pitch Strength obtained in the control condition were all smaller than 0.4. This finding indicated the potential usefulness of Pitch Strength in developing an objective method to evaluate the brain’s ability to follow the changes in pitch over time for individual listeners.

Integration of objective measures and subjective judgment

To better illustrate the process of developing an objective method for detecting the presence of an FFR, data from Figure 4 were replotted in Figure 5 as plots according to the results of human judgment. To estimate the power and false-positive rate of each algorithm, waveforms recorded in the experimental and control conditions were pooled together and plotted according to human judgment results. Subjective response thresholds are plotted as horizontal dotted lines in each panel. Due to the bell-shaped distribution of Slope Error, two threshold lines were used for each tone. Consistent with the procedures used in Figure 4, waveforms with Slope Errors that fell outside the two threshold lines were classified as having absent response, whereas waveforms with Slope Errors within the two threshold lines were considered responses. Percentage proportions of each of the integration results (i.e., objective measures versus subjective judgment) are denoted numerically across the horizontal dotted lines in each panel.

Response thresholds for subjective measures

When the short-term autocorrelation algorithm was used, response thresholds for Frequency Error were 9.0, 7.2, 9.8 and 12.0 Hz for Tones 1-4, respectively. The two response thresholds for Slope Error were -145/50, -75/60, -180/90 and -60/100 Hz/s for Tones 1-4, respectively. Tracking Accuracy response thresholds for the four tones were 0.68, 0.74, 0.65 and 0.66, respectively. Pitch Strength response thresholds for the four tones were 0.36, 0.28, 0.27 and 0.29, respectively. When the narrow-band spectrogram algorithm was used, response thresholds for Frequency Error were 8.1, 8.0, 10.8 and 9.0 Hz for Tones 1-4, respectively. The two response thresholds for Slope Error were -110/35, -150/50, -150/0 and -180/190 Hz/s for Tones 1-4, respectively. Tracking Accuracy response thresholds for the four tones were 0.54, 0.70, 0.50 and 0.74, respectively.

Power and false-positive rate

The presence of an FFR as determined by the subjective interpretation of human observers was compared to the outcomes of the objective algorithms in order to construct the power, false-positive rate and receiver-operating-characteristics curves in Figure 6.

For Frequency Error (Figure 6**a**), when the short-term autocorrelation algorithm was used, the power values (i.e., sensitivity of this algorithm in agreement with visual inspection of experienced human observers) for the four different pitch contours were 68%, 62%, 70% and 74%, respectively; while the false-positive rates (i.e., 100-specificity) were 0%, 0%, 4% and 5%, respectively. Tone 4 had the best sensitivity, but its false-alarm rate was also the largest. When the narrow-band spectrogram algorithm was used, the power values were 91%, 76%, 65% and 44% for the four different pitch contours, respectively; while the false-positive rates (i.e., 100-specificity) were 0%, 4%, 4% and 5%, respectively. Tone 1 had the best sensitivity and the smallest false-positive rate. It was noted that the narrow-band spectrogram algorithm improved the power for detecting the presence of a response (i.e., agreement with visual subjective judgments) for Tones 1 and 2 by 23% and 14%, respectively. Such improvement, however, was not observed for Tones 3 and 4.

For Slope Error (Figure 6**b**), when the short-term autocorrelation algorithm was used, the power values for the four different pitch contours were 73%, 62%, 85% and 43%, respectively; while the false-positive rates were 27%, 17%, 46% and 29%, respectively. When the narrow-band spectrogram algorithm was used, the power values were 91%, 86%, 70% and 87% for the Tones 1-4, respectively; while the false-positive rates were 50%, 17%, 58% and 14%, respectively. In both algorithms, Slope Error had larger false-positive rates than those for Frequency Error, Tracking Accuracy and Pitch Strength. This was likely due to the effect of occasional disruptions of the slope estimates along the response f0 contours. It was also noted that the narrow-band spectrogram algorithm improved the power for detecting the presence of a response for Tones 1, 2 and 4 by 18%, 24% and 44%, respectively. Such improvement, however, was not observed for Tone 3.

For Tracking Accuracy (Figure 6**c**), when the short-term autocorrelation algorithm was used, the power values for the four different pitch contours were 50%, 67%, 35% and 74%, respectively; while the false-positive rates were 5%, 9%, 4% and 19%, respectively. Tone 4 had the best sensitivity, but its false-alarm rate was also the largest. When the narrow-band spectrogram algorithm was used, the power values were 64%, 76%, 60% and 48% for the four different pitch contours, respectively; while the false-positive rates were 14%, 9%, 8% and 14%, respectively. Tone 2 had the best sensitivity and the smallest false-positive rate. It was noted that the narrow-band spectrogram algorithm improved the power for detecting the presence of an FFR for Tones 1, 2 and 3 by 14%, 9%, and 25%, respectively. Such improvement, however, was not observed for Tone 4.

For Pitch Strength (Figure 6**d**), when the short-term autocorrelation algorithm was used, the power values for the four different pitch contours were 100%, 91%, 85% and 74%, respectively; while the false-positive rates were 5%, 13%, 13% and 10%, respectively. Tone 1 had the best sensitivity and the smallest false-alarm rate.

Predictive values and efficiency

To better illustrate the performance of each objective index, operating characteristics of the four objective indices were tabulated for short-term autocorrelation algorithm (Table 1) and narrow-band spectrogram algorithm (Table 2). The use of the narrow-band spectrogram algorithm improves the predictive values and efficiency for Frequency Error, Slope Error and Tracking Accuracy for the four tones used in this study.

Test-retest reliability

As one ultimate goal of developing objective methods is to apply measurements in difficult-to-test populations, it is of interest to examine the stability of the test results across the two trials of recordings. Paired *t*-test revealed no significant changes in the measurements of Frequency Error (t = -0.28, p = 0.61), Slope Error (t = 1.05, p = 0.29), Tracking Accuracy (t = 0.12, p = 0.45) and Pitch Strength (t = -0.35, p = 0.64), across the four pitch contours and the two subjective algorithms used in the present study. That is, for a given participant, both trials had similar pitch-tracking accuracy values and were rated very similarly by human observers, evidenced by the high power and low false-positive rate of the two algorithms. This finding, in addition to the test-retest reliability, confirms the clinical applicability of the FFR to human voice pitch and supports the use of the response detection algorithms described in the present study.

DISCUSSION

This study evaluates an objective method for detecting the presence of an FFR to voice pitch. Previously, this judgment was made subjectively by an experienced human observer, which limited the clinical utility of the FFR in the assessment of pitch processing. This objective method is applied to two pitch-extracting algorithms (short-term autocorrelation and narrow-band spectrogram) to determine which method as well as which stimulus pitch contour produces greater sensitivity and fewer false-positives. It is observed that the narrow-band spectrogram algorithm provides greater sensitivity and efficiency, thereby promoting its clinical use.

Pitch-detecting algorithms

The overall periodicity (i.e., fundamental frequency) of a sampled signal can be extracted either in the time or frequency domain. The short-term autocorrelation algorithm estimates the overall periodicity by correlating a sampled signal to itself with shifted data points in the time domain. This approach is sound when the signal contains a robust and continuous pitch contour over time, such as a stimulus token produced by a human subject or a grand-averaged waveform recorded from a group of participants. Recordings taken from individual listeners tend to have a less favorable signal-to-noise ratio which greatly reduces the accuracy of utilizing the short-term autocorrelation algorithm in estimating the f0 contour of a recording taken from an individual listener. One possible solution to minimize the effect of background noise on the extraction of the FFR f0 contour is to examine the spectral energy of a response in a pre-defined frequency range that contains most of the FFR f0 contour and excludes the frequency components of the background noise. The fact that the narrow-band spectrogram algorithm improved the power (i.e., sensitivity) of detecting the presence of an FFR to Tones 1, 2, and 3 proved the potential of utilizing this algorithm to detect the presence of an FFR in individual listeners.

Although both algorithms produce useful results, there are advantages and disadvantages of each algorithm. The short-term autocorrelation algorithm estimates f0 candidates in the time domain and it takes the advantage of measuring the overall periodicity of a sampled signal, including the energies located at the f0 and its harmonics. Thus, this algorithm is likely to provide robust f0 estimates when the response energies at the harmonics are substantial. In addition, pitch is calculated as 1 / τ_{max} in this algorithm; therefore, it gives better f0 estimates at lower frequencies than higher frequencies. In contrast, the narrow-band spectrogram algorithm derives f0 estimates in the spectral domain and its frequency resolution is stable across the spectrum. Another advantage of the narrow-band spectrogram algorithm is that this algorithm is less compromised by the poor signal-to-noise ratios that are commonly observed in waveforms recorded from individual listeners. Other advantages and disadvantages of using one algorithm over the other include things such as octave jumps in short-term autocorrelations and incorrect f0 estimates in narrow-band spectrograms. This weakness can be improved by using octave jumps in the autocorrelation algorithm as well as finding the spectral peaks in the spectrogram that are closest to the expected stimulus frequency. These advantages and refinements produced improvements in detecting the presence of an FFR in this study, as evidenced by the increased sensitivity and decreased false-positive rates for some of the pitch contours.

Frequency Error is a measure of the acuity of pitch encoding in the brainstem and it demonstrates the fewest false-positive rates than the other three objective measures. When the narrow-band spectrogram algorithm was used, Frequency Error showed improved high hit rates of 90% for Tone 1 and 76% for Tone 2. These high hit rates, accompanied with the fewest false-positives, indicate the potential usefulness of Frequency Error (together with the narrow-band spectrogram) in detecting the presence of FFR from individual listeners. This improvement is also indicative of the advantages and refinements used in the spectrogram algorithm, likely due to the fact that the narrow-band spectrogram algorithm is stable across the spectrum and less compromised by the less favorable signal-to-noise ratios that are commonly observed in waveforms recorded from individual listeners. These high sensitivities and low false-positive rates support the idea that the Frequency Error can be used as a viable index to detect the existence of an FFR.

Slope Error indicates the extent to which the shape of a pitch contour is preserved in the human brainstem. Although Slope Error shows concentrated peaks at around 0 Hz/s, its distribution overlaps with wide-spread distribution of the Slope Errors obtained from the control condition. Substantial overlaps between the distributions of the experimental and control conditions compromise the power and false-positive rate of this measurement and its potential usefulness in detecting the presence of an FFR. For Slope Error, the use of the spectrogram algorithm does not appear to increase the power or decrease the false-positive rate of this measurement. This is likely due to the fact that incorrect f0 estimates, even just a few data points, embedded in individual recordings could adversely affect the Slope Error estimate much more than the other objective measures.

Tracking Accuracy, an index of the overall faithfulness of how the response f0 contours follow the stimulus contours, showed moderate hit rates and relatively low false-positive rates. It is worth noting that the use of the narrow-band spectrogram algorithm increases the power of Tracking Accuracy to 76% for Tone 2. In addition to a false-positive rate of 9%, Tracking Accuracy (with the spectrogram algorithm) may be a useful index to include when developing a set of objective measures for detecting the existence of an FFR. This finding, however, does not conclude that the human brainstem is better at tracking dramatic pitch changes than flat tones. On the other hand, it is possible that the human brainstem's ability to follow extreme changes in voice pitch (or unnaturally exaggerated changes in voice pitch) would decline when the slope of the pitch change approaches the human brainstem's limits. This, however, would require a separate study to purposely manipulate the slope of pitch changes and to examine the brainstem's limitations in following them.

The human brain is better at tracking the pitch contours that are specific to the listener's native language. Krishnan et al. (2009a) recently reported that the human brainstem is more sensitive only to a naturally rising pitch contour that is specific to the listener's native language, and is less sensitive to unnatural pitch contours that were manipulated and deviated from the natural pitch contours. The human brainstem also has a differential preference to specific pitch contours over others. Krishnan et al. (2009b) recorded FFRs in response to Tone 2 with a naturally rising, linear-ramping and inverted-curvilinear pitch contours and found that the Tracking Accuracy in a Chinese group of participants was larger than the English group in response to the naturally rising pitch contour only, but not in the linear-ramping and inverted-curvilinear pitch contours. Data obtained in the present study showed highest Tracking Accuracy for Tone 2 with a naturally rising pitch (Figure 3). This finding is consistent with the data reported in the above-mentioned literature.

Pitch-tracking accuracy (i.e., Tracking Accuracy used in the present study) has been used as a viable index to represent the degree to which a response followed the pitch contour of the stimulus token.

Krishnan et al. (2005) examined pitch-tracking accuracy by measuring the ranked cross-correlation coefficients between the recorded waveforms and stimulus signals; they found that the f0 contours of the FFRs recorded in native Chinese speakers were less variable and followed the f0 contours of the stimuli with greater precision than those recorded in native English speakers. Dajani et al. (2005) recorded human evoked potentials to voice pitch of a natural vowel /a/ and reported a great accuracy of the response pitch contours relative to the pitch contour of the stimulus token. Jeng and Schnabel (2009) measured the FFR from infants and reported similar pitch-tracking accuracy to that in adults. All of these results indicated the adequacy of using pitch-tracking accuracy to indicate the presence of such a response. Results obtained from the current study took advantage of this phenomenon and reported the power and false-positive rates of the short-term autocorrelation and narrow-band spectrogram algorithms. It is important to point out that the use of the narrow-band spectrogram algorithm improves the power for Tones 1, 2 and 3.

Pitch Strength is a measure of the response periodicity and it demonstrates the largest hit rate (100% for Tone 1 and 91% for Tone 2) and relatively low false-positive rates (5% for Tone 1 and 13 % for Tone 2). This finding, in our view, is a reflection of the clear separation between the distributions of Pitch Strength that are derived from waveforms recorded in experimental and control conditions. This finding illustrates the potential usefulness of Pitch Strength in developing objective methods to detect the existence of an FFR to voice pitch. Krishnan et al. (2004; 2005) reported that Tone 2 with a rising pitch contour was able to elicit a response with the largest pitch strength than the other three pitch contours (Tone 1, Tone 3 and Tone 4). Swaminathan et al. (2008b; 2009) divided the FFR recordings into six non-overlapping sections and found that the pitch strength of the response was highly correlated with the slope (i.e., acceleration or deceleration) of the pitch changes within each section.

Results obtained from this study indicated that Frequency Error is the most sensitive index over the others in detecting the existence of an FFR. It is possible that a combination of two or more indices may produce better sensitivity and specificity and, therefore, improve the accuracy of the proposed algorithms. In addition, one would use one index (e.g., Pitch Strength ≥ 0.3) as an initial criterion, followed by another index (e.g., Frequency Error $\leq$ 10 Hz) as decision criterion, or a combination of several indices as a time or in sequence.

Statistical considerations and disease epidemiology

From the viewpoint of statistics, it is readily apparent that there is a tradeoff between the power and false-positive rate of a test. A sensitive test takes preference whenever the probability of false-positive rates is high or whenever it is needed to reduce the likelihood of errors. That is, sensitivity is primarily used to "rule-out" the existence of a disease. On the other hand, specificity is often used to confirm an existing diagnostic impression; highly specific tests indicate low false-positive rates. For any specific objective method, we need to weigh and balance between the advantages and disadvantages of each method and choose one that best suits clinical practice. For example, when screening for diseases that carry serious consequences if misdiagnosed or not treated early such as tumors and severe-profound hearing loss at birth, it would be wise to use stringent criteria and follow-up with patients on a regular basis. Similarly, when treating diseases that do not carry serious consequences, it would be appropriate to use a test that yields the largest power and an acceptable false-positive rate.

From the practical point of disease epidemiology, it should be clear that detecting the existence of a disease is an imperfect process that results in the estimates of likelihood of errors (i.e., type I or II errors) rather than absolute certainty. In ideal situations, we like to have a perfect power with zero likelihood of errors. However, there is always a likelihood of error. Another important factor in evaluating the power and false-positive rate of a test is the predictive values, i.e., the probability of a disease (e.g., inability to process voice pitch) being present (positive predictive value) versus not being present (negative predictive value). Predictive values are directly associated with prevalence of the disease and decrease systemically with increasing prevalence (Schwartz, 1987). It is important to note that the power, false-positive rate and likelihood of errors represent mathematical properties of a test that clinicians must account for when making a decision of whether to administer a test or not. This consideration is particularly relevant to the discussion of developing a test for identifying the existence of an FFR – indicating the listener's ability to process changes in voice pitch. Although the prevalence of pitch-processing disorders has not been reported, it is likely to climb when considering patients with central auditory processing disorders, autism spectrum disorders, sensorineural hearing loss as well as and patients fit with hearing aids or cochlear implants.

Ultimately, we would need to make recordings with 'known' responses and use them as the gold standard. However, the achievement of a 'known' response in the case of FFR is not as straightforward as surgical confirmation of the existence of a tumor on the eighth nerve. To date, the existence of an FFR still relies on the subjective interpretation of human judgment, as well as it is for determining the presence of an ABR wave V. It is possible that in the near future, when characteristics of FFR are fully understood and the techniques of recording FFR have been improved that the existence of an FFR can be modeled through its 'known' characteristics. Rejection versus acceptance of the results from objective methods can then be considered based on that 'known' model and characteristics of an FFR. Until then, the best we can do is to associate the results of objective measures and presence of an FFR determined by experienced human observers. It is anticipated that the objective method proposed in the present study could be implemented in the clinic if a 'normative' data set of FFR waveforms were obtained the normal-hearing population and recognized by a group of experienced human observers.

Clinical implications

This study evaluated two algorithms suitable for detecting the presence of an FFR and compared the power and false-positive rates of each algorithm. While previous work had described the use of the short-term autocorrelation (Krishnan et al., 2004, 2005) and the narrow-band spectrogram (Russo et al., 2008; Song et al., 2008; Wong et al., 2007) algorithms to measure the pitch-tracking accuracy of a response, none of them compared the results of the two objective algorithms used in this study. The objective method used in the current study, including the use of the control-experimental protocol and response thresholds used for each of the four objective measures, can be used for difficult-to-test patients and may prove to be useful as an assessment and diagnostic method in both clinical and basic research efforts. Specifically, these techniques open a door to help assess the pitch processing mechanisms at the brainstem level and diagnose abnormal signal processing of voice pitch in patients who cannot provide reliable behavioral feedback. Data reported in the present study provide results about the FFR to voice pitch in normal-hearing adults. These results can serve as the basic knowledge to help patients with communication disorders, such as patients with autism spectrum disorders (Russo et al., 2008), central auditory processing disorders and hearing loss. It is also important that, although Mandarin tones are used to elicit FFRs in this study, the objective method could realistically be applied to any complex sound with a variable pitch contour; thus it has utility beyond a Mandarin speaking population and can be useful to any clinician interested in obtaining an objective method of pitch processing. It is worthwhile noting that different populations (e.g., Mandarin versus non-Mandarin speaking populations, adults versus infants) may have differential response characteristics on each of the four indices and, therefore, may require different response-threshold criteria to be used. It is anticipated that this improvement can be made by including specific populations (e.g., non-Mandarin speakers or infants) in future studies.

Additionally, because the current study investigates a response that is evoked by the pitch contour of a speech token, results may advance the knowledge base regarding pitch encoding mechanisms in normal and pathological populations. For example, further knowledge gained from this and similar studies could help us better understand how cochlear implant designs can be changed to help those who communicate using tonal and non-tonal languages. Cochlear implants have provided benefits to more than 188,000 people worldwide (National Institutes of Health, 2009) who are deaf or hard-of-hearing. One of the major challenges in the most recent cochlear implant research is to find a way to improve pitch-contour perception and music appreciation. Both the perception of voice pitch and the appreciation of music require the human brainstem’s ability to process the changes in pitch of a speech signal or a musical melody (Deutsch et al., 2004; Gandour et al., 1998; Lee & Lee, 2010; Wong, et al., 2007). A technique that allows an objective measurement of the brainstem's response to changes in pitch may provide a greater understanding of the neural underpinnings of pitch perception which may ultimately enhance communication and music appreciation for thousands of cochlear implant users who speak tonal and non-tonal languages around the world.

ACKNOWLEDGMENTS

This study was supported in part by (1) Advancing Academic-Research Career (AARC) Award, American Speech-Language-Hearing Association and (2) Scholarly Activity Award, Ohio University. Preliminary results of this study were presented at the American Auditory Society Annual Meeting held at the Chaparral Suites Resort, Arizona, March 5-7, 2009. The authors thank Dr. Ananthanarayan Krishnan for his insightful comments on the preliminary results of this work as well as Dr. Monita Chatterjee and three anonymous reviewers for their invaluable feedback that substantially enhanced the quality of this manuscript.

REFERENCES

Aiken, S.J. & Picton, T.W. 2006. Envelope following responses to natural vowels. *Audiol Neurootol*, 11, 213-32.

Altman, D.G. & Bland, J.M. 1994. Diagnostic tests. 1: Sensitivity and specificity. British Medical Journal, 308(6943), 1552.

Boersma, P. 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proc Inst of Phon Sci*, 17, 97-110.

Boersma, P. & Weenink, D. 2009. PRAAT: doing phonetics by computer.

Dajani, H.R., Purcell, D., Wong, W., Kunov, H. & Picton, T.W. 2005. Recording human evoked potentials that follow the pitch contour of a natural vowel. *IEEE Trans Biomed Eng*, 52, 1614-8.

Deutsch, D., Henthorn, T. & Dolson, M. 2004. Absolute pitch, speech, and tone language: Some experiments and a proposed framework. *Music Percept*, 21, 339-56.

Galbraith, G.C., Amaya, E.M., de Rivera, J.M., Donan, N.M., Duong, M.T., et al. 2004. Brain stem evoked response to forward and reversed speech in humans. *Neuroreport*, 15(13), 2057-60.

Gandour, J., Wong, D. & Hutchins, G. 1998. Pitch processing in the human brain is influenced by language experience. *NeuroReport*, 9, 2115-9.

Jeng, F.C., & Schnabel, E.A. 2009. Frequency-Following Responses to Voice Pitch in Infants. In: Abstracts of American Auditory Society Annual Meeting, page 21 (#62), Scottsdale, Arizona.

Kraus, N. & Nicol, T. 2005. Brainstem origins for cortical 'what' and 'where' pathways in the auditory system. *Trends Neurosci*, 28(4), 176-81.

Krishnan, A., Gandour, J.T., Bidelman, G.M. & Swaminathan, J. 2009a. Experience-dependent neural representation of dynamic pitch in the brainstem. *Neuroreport*, 20(4), 408-13.

Krishnan, A., Gandour, J.T. & Bidelman, G.M. 2010. The effects of tone language experience on pitch processing in the brainstem. *J Neurolinguist*, 23, 81-95.

Krishnan, A., Swaminathan, J. & Gandour, J.T. 2009b. Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *J Cognitive Neurosci*, 21, 1092-105.

Krishnan, A., Xu, Y., Gandour, J.T. & Cariani, P. 2004. Human frequency-following response: Representation of pitch contours in Chinese tones. *Hear Res*, 189, 1-12.

Krishnan, A., Xu, Y., Gandour, J.T. & Cariani, P. 2005. Encoding of pitch in the human brainstem is sensitive to language experience. *Cogn Brain Res*, 25, 161-8.

Lee, C.-Y. & Lee, Y.-F. 2010. Perception of musical pitch and lexical tones by Mandarin-speaking musicians. *J Acoust Soc Am*, 127(1), 481-90.

Musacchia, G., Sams, M., Skoe, E. & Kraus, N. 2007. Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proc natl Acad Sci*, 2007, 104(40), 15894-8.

National Institutes of Health, 2009. http://www.nidcd.nih.gov/health/hearing/coch.asp

Russo, N.M., Skoe, E., Trommer, B., Nicol, T., Zecker, S., et al. 2008. Deficient brainstem encoding of pitch in children with autism spectrum disorders. *Clin Neurophysiol*, 119(8), 1720-1731.

Schwartz, D.M. 1987. Neruodiagnostic audiology: contemporary perspectives. *Ear Hearing*, 8(4), Suppl, 43-48.

Song, J.H., Skoe, E., Wong, P.C.M. & Kraus, N. 2008. Plasticity in the adult human auditory brainstem following short-term linguistic training. *J Cognitive Neurosci*, 20(10), 1892-1902.

Swaminathan, J., Krishnan, A. Gandour, J.T. & Xu, Y. 2008a. Applications of static and dynamic iterated rippled noise to evaluate pitch encoding in the human auditory brainstem. *IEEE Trans Biomed Eng*, 55(1), 281-7.

Swaminathan, J., Krishnan, A. & Gandour, J.T. 2008b. Pitch encoding in speech and nonspeech contexts in the human auditory brainstem. *Neuroreport*, 19, 1163-7.

Wong, P.C.M., Skoe, E., Russo, N.M., Dees, T. & Kraus, N. 2007. Musical experience shapes human brainstem encoding of linguistic pitch patterns. Nat Neurosci, 420-422.

FIGURE LEGENDS

Figure 1 A typical example of spectrograms of the stimuli and recordings in response to a set of four monosyllabic Mandarin tokens that reflect the four different contours of voice pitch (Tone 1 flat, Tone 2 rising, Tone 3 dipping, and Tone 4 falling). The control condition was conducted when the sound tube was occluded and removed from the listener's ear.

Figure 2 A typical example of fundamental frequency contours of the stimuli and recordings extracted using the short-term autocorrelation algorithm (**a**) and narrow-band spectrogram algorithm (**b**).

Figure 3 Distribution histograms of Frequency Error (**a**), Slope Error (**b**), Tracking Accuracy (**c**), and Pitch Strength (**d**) extracted using the short-term autocorrelation (left four columns) and narrow-band spectrogram (right four columns) algorithms. Histograms of the recordings taken during the experimental conditions (solid lines) and control conditions (shaded areas) for each of the four tones are plotted in the same panel for comparison. Response thresholds for determining the presence of a frequency-following response are plotted as vertical dotted lines in each panel. Numeric symbols on the two sides of the vertical dotted lines indicate the percentages of the waveforms occurring to the left and right of the response thresholds. Percentage numbers of the waveforms obtained from the control condition are noted in parentheses. Due to the bell-shaped distribution of Slope Error, two threshold lines were used for each tone. Pitch Strength is available only in the short-term autocorrelation algorithm.

Figure 4 Distributions of Frequency Error (**a**), Slope Error (**b**), Tracking Accuracy (**c**), and Pitch Strength (**d**) are plotted according to the results of human judgment. Response thresholds for determining the power and false-positive rates of each algorithm are plotted as horizontal dotted lines in each panel. Numeric symbols above and below the horizontal dotted lines indicate the percentages of waveforms occurring in the upper and lower portions of the response thresholds, respectively. Due to the bell-shaped distribution of Slope Error, two threshold lines were used for each tone. Pitch Strength is available only in the short-term autocorrelation algorithm.

Figure 5 Power and false-positive rates of the two algorithms used to determine the presence of a frequency-following response to each of the four different pitch contours (Tone 1 flat, Tone 2 rising, Tone 3 dipping and Tone 4 falling). Numbers inside the panel indicate the four pitch contours used in this study. Data obtained using short-term autocorrelation (plain numbers) and narrow-band spectrogram (underlined numbers) algorithms are plotted in the same panel for comparison. The oblique dotted line indicates the equal power and false-positive-rate boundary.

Figure 6 Averaged pitch-tracking accuracies derived from two sequential recording trials is shown for each of the four different pitch contours. The upper panel displays data obtained using the short-term autocorrelation algorithm; the lower panel displays data obtained using the narrow-band spectrogram algorithm. In each panel, the group mean is shown along with one standard error above the mean value. Note the mean difference in two pitch-tracking accuracies between the two sequential trials is less than 0.12 at the four different pitch contours.

Figure 2 A schematic 2x2 contingency table describing the calculation of the operating characteristics of a test.

		Human Judgment	
		Present	**Absent**
Test Result	**Positive**	True Positive (***a***)	False Positive (***b***)
	Negative	False Negative (***c***)	True Negative (***d***)

Operating Characteristics:

Sensitivity (Power) = $\frac{a}{a+c}$ %

Specificity = $\frac{d}{b+d}$ %

False-Negative Rate = $\frac{c}{a+c}$ %

False-Positive Rate = $\frac{b}{b+d}$ %

Positive-Predictive Value = $\frac{a}{a+b}$ %

Negative-Predictive Value = $\frac{d}{c+d}$ %

Efficiency = $\frac{a+d}{a+b+c+d}$ %

Type I (alpha) error = False-Positive Rate = (100 – specificity) %

Type II (beta) error = False-Negative Rate = (100 – sensitivity) %

Table 1 Operating characteristics of four objective measures for detecting the existence of an FFR to voice pitch using the short-term autocorrelation algorithm

									Predictive Value											
	Power				False-Positive Rate				Positive result				Negative result				Efficiency			
Objective Measure	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
Frequency Error	**68**	**68**	**74**	**63**	**0**	**0**	**4**	**5**	**100**	**100**	**93**	**93**	**76**	**81**	**83**	**72**	**84**	**86**	**86**	**80**
Slope Error	**73**	**68**	**89**	**68**	**27**	**16**	**44**	**23**	**73**	**76**	**61**	**75**	**73**	**78**	**88**	**71**	**73**	**77**	**70**	**73**
Tracking Accuracy	**50**	**74**	**37**	**59**	**5**	**8**	**4**	**5**	**92**	**88**	**88**	**93**	**66**	**82**	**67**	**70**	**73**	**84**	**70**	**77**
Pitch Strength	**100**	**95**	**89**	**82**	**5**	**16**	**12**	**9**	**96**	**82**	**85**	**90**	**100**	**95**	**92**	**83**	**98**	**89**	**89**	**86**

Table 2 Operating characteristics of three objective measures for detecting the existence of an FFR to voice pitch using the narrow-band spectrogram algorithm

									Predictive Value											
	Power				False-Positive Rate				Positive result				Negative result				Efficiency			
Objective Measure	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4	T1	T2	T3	T4
Frequency Error	**91**	**84**	**68**	**86**	**0**	**4**	**4**	**0**	**100**	**94**	**93**	**100**	**92**	**89**	**80**	**88**	**95**	**91**	**84**	**93**
Slope Error	**91**	**95**	**74**	**68**	**50**	**16**	**56**	**5**	**65**	**82**	**50**	**94**	**85**	**95**	**69**	**75**	**70**	**89**	**57**	**82**
Tracking Accuracy	**64**	**84**	**63**	**82**	**14**	**8**	**8**	**5**	**82**	**89**	**86**	**95**	**70**	**88**	**77**	**84**	**75**	**89**	**80**	**89**