All Faculty Scholarship   School of Health Sciences

6-19-2013

An automated procedure for detecting human frequency-following responses to voice pitch

Fuh Cherng Jeng
Ohio University

Jiong Hu
University of the Pacific, jhu@pacific.edu

An automated procedure for detecting human frequency-following responses to voice pitch

Fuh-Cherng Jeng, and Jiong Hu

Proceedings of Meetings on Acoustics

ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Psychological and Physiological Acoustics
Session 1pPPb: Psychoacoustics and Perception (Poster Session)

1pPPb14.   An automated procedure for detecting human frequency-following responses to voice pitch

Fuh-Cherng Jeng* and Jiong Hu

*Corresponding author's address: Communication Sciences and Disorders, Ohio University, Athens, OH 45701, jeng@ohio.edu

 The frequency-following response (FFR) to voice pitch has been widely examined in research laboratories and has demonstrated its potential to be transformed into a useful tool for patients with hearing, speech, and language disorders in the clinic. During the past decade, many aspects of the FFR have been reported. The presence of such a response, however, still relies on subjective interpretation of the observer. Aside from a recent study reporting two algorithms for detecting such a response, there has been limited number of studies reporting the development of an automated procedure for FFR. The purpose of this study is (1) to develop an automated procedure that utilizes the statistical properties of the temporal and spectral energy distributions in the recorded waveforms and (2) to examine the sensitivity and specificity of the automated procedure.

Published by the Acoustical Society of America through the American Institute of Physics

INTRODUCTION

Electrophysiological measurements have an important advantage in that they do not require awareness or active feedback from the participant, which make these measurements particularly suitable for populations that conventional behavioral tests could not be reliably tested on. Traditionally, these electrophysiological test procedures are complex in construction and technologically heavy, which consequently demands a fair amount of training to operate and to interpret the results. If some of the procedures of those electrophysiological measurements could be automated, it could dramatically promote the efficiency and portability of the tests for both research and clinical applications and most importantly, reduce the potential inconsistency in data interpretation due to subjective human judgments involved in the conventional data-interpretation process. Some electrophysiological measurements, such as the auditory brainstem response (ABR) and auditory steady-state response (ASSR), have fairly matured automated procedures readily available to researchers and clinicians. On the contrary, although the frequency-following response (FFR) to voice pitch has received a fair amount of attention in recent years, there has not been much of progress in the development of an automated procedure. Thus, the goal of this study was to develop an automated procedure for the advancement in the FFR research and its potential applications in clinical settings.

Voice pitch is an auditory perception that reflects the psychological representation of fundamental frequency (f0) of speech signal and carries important information for the perception of speech and music (Moore 2003). The accurate encoding of voice pitch and its temporal variation is essential and critical for listeners to perceive prosodic information and lexical cues in speech signals, as well as to appreciate melodies of music. The scalp-recorded FFR to voice pitch has been reported recently with results that correlate with behavioral tests of pitch perception (Akhoun et al., 2008; Skoe et al., 2010; Krishnan et al., 2012). Among the stimulus tokens that had been utilized in the FFR research, Mandarin Chinese syllables that contain different f0 contours and lexical meanings have been widely used (Krishnan et al., 2004, 2009; Xu et al., 2006; Wong et al., 2007; Song et al., 2008; Jeng et al., 2010, 2011a, 2011b, 2011c, 2011d; Li & Jeng, 2011). Several aspects of the voice pitch elicited FFR have been examined. For example, children with autism spectrum disorders (Russo et al., 2008) and reading and spelling disorders (Chandrasekaran et al., 2009) have shown to have degraded voice-pitch tracking accuracy and decreased FFR amplitudes. There is also an emerging interest in defining the developmental trajectory of the human scalp-recorded FFR to voice pitch in recent years. Studies have reported the characteristics of the FFR to voice pitch in newborn infants of different parental language backgrounds (Jeng et al., 2011d), in maturing infants (Jeng et al., 2010) and in normal-hearing children (Johnson et al., 2008). With the extensive reports of recent advancement in the field of the voice pitch elicited FFR, naturally the next progress we want to make is to explore and develop algorithms and procedures that would promote the efficiency and portability of the FFR technology.

METHOD

Thirty native Mandarin Chinese speakers (13 males), averaged age of 25.4 years, were recruited. Four monosyllabic Mandarin Chinese syllables with different pitch contours (Tone 1 flat /yi1/ 'clothes', Tone 2 rising /yi2/ 'aunt', Tone 3 dipping /yi3/ 'chair', Tone 4 falling /yi4/ 'meaning') were prepared. These voice samples were sampled to a unified duration of 250 ms with a rising/falling time of 10 ms. Each stimulus token were presented at 70 dB SPL for blocks of 2000 accepted sweeps for each token. The inter-stimulus interval between adjacent stimulus tokens was set at 45 ms. To ensure that the stimulus artifact was appropriately eliminated from the recordings, in each testing session, a control condition was conducted at the end of each recording session to provide EEG wave forms that have no direct relationship to the stimulus presentation. Three gold-plated recording electrodes were applied to each participant on the forehead (Fpz, non-inverting), right mastoid (M2, inverting), and left mastoid (M1, ground). Continuous recordings were amplified using an Opti-Amp amplifier with a gain of 50k, online-filtered 100-3000 Hz 6 dB/octave and digitized at a rate of 20k samples per second. All data were digitally filtered using a brick-wall, linear-phase finite-impulse-response (FIR) band-pass filter (85-1500 Hz, 500th order). An artifact rejection limit of ±25 µV was used to ensure the quality of the recorded waveforms. For each condition, a total of 2000 accepted recording sweeps were collected.

Automated Response Detection Algorithms

Residue-Noise-Level based Temporal Algorithm

This study utilized two algorithms that focus on the temporal and spectral aspects: For the temporal properties, an algorithm that was based upon the residue noise level (RNL) index of the temporal property of the recorded waveform (Wong & Bickford., 1980). Briefly, this algorithm was implemented by first dividing all recorded sweeps into two separate banks - with all the odd-number sweeps stored in one bank and even-number sweeps in the other. Assuming each sweep contains a representation of the desired neural responses, the average of the two banks of sweeps would then represent the response waveform itself, while the difference would represent the noise.

The average of the two banks essentially is the same as the average of all sweeps collected, and thus can be expressed using Formula 3.

$$\overline{A_N(t)} \;=\; \frac{1}{N}[\Sigma_i\{EEG_i(t)\}]\,, \tag{1}$$

where A_N represents the averaged signal between the two banks, N is the number of sweeps, EEG indicates the recorded waveform for each sweep, and i denotes each individual sweep ranging from 1 to N.

Similarly, the difference between the two banks can be expressed using Formula 4.

$$\overline{D_N(t)} \;=\; \frac{1}{N}[\Sigma_i\{EEG_i(t)\cdot(-1)^i\}] \tag{2}$$

where D_N represents the difference between the two banks, N is the number of sweeps, EEG indicates the recorded waveform for each sweep, and i denotes each individual sweep ranging from 1 to N.

Alternatively, each recording can be constructed by two components:

$$\overline{A_N(t)} = s(t) + \overline{n(t)} \tag{3}$$

where s(t) represents the physiological response and n(t) represents noise.

With increasing number of collected sweeps, the averaged response $\overline{A_N(t)}$ would approach the real waveform of the elicited response s(t). On the other hand, the difference between the two banks (i.e., the noise $\overline{D_N(t)}$) would optimally approximate the true value of the noise n(t). Thus, if the recorded waveform were to contain an elicited neural response, which should then be temporal locked to the stimulus in each recorded sweep, the response component s(t) should not contribute in $\overline{D_N(t)}$, because it represents response and should have been cancelled out during the alternation of addition and subtraction. Thus, we would have two estimates that would approximate the elicited neural response $\overline{A_N(t)}$ and noise $\overline{D_N(t)}$, which is of the order of $1/\sqrt{N}$.

To determine the presence/absence of a response signal, computation of an appropriate signal-to-noise ratio (SNR) is almost inevitable (Fig. 11). Because variance is in essence a measure of the overall amplitude of a response signal, the ratio of the variance of the response signal (i.e., $\overline{A_N(t)}$) to the variance of the noise (i.e., DN(t)) was used to indicate the magnitude that the response signal stood out from the noise (i.e., SNR). This concept can be expressed using Formula 6.

$$SNR = \frac{Var(\overline{A_N(t)})}{Var(\overline{D_N(t)})} \tag{4}$$

where SNR stands for signal-to-noise ratio, $Var(\overline{A_N(t)})$ represents the variance of the average of the two banks, $Var(\overline{D_N(t)})$ indicates the variance of the difference between the two banks. If we continue to break down the SNR calculation, we have

$$SNR = \frac{Var(\overline{A_N(t)})}{Var(\overline{D_N(t)})} = \frac{\text{mean power of s(t)}}{\text{mean power of n(t)}} + \frac{\text{mean power of n(t) in } \overline{A_N(t)}}{\text{mean power of n(t) in} \overline{D_N(t)}} \tag{5}$$

Denote the second term in Formula 7 by F and using $\sigma^2(n)/N$ as the denominator of the first term, we have

$$SNR \approx N\cdot\frac{\sigma^2(s)}{\sigma^2(n)} + F = Nf^2 + F \tag{6}$$

where f is the amplitude ratio of s(t) and n(t), i.e. the real SNR. Thus, when there is no response, where s(t)=0, and f=0, we have SNR=F, which is the F ratio of the noise. When there is a response, then $SNR \approx Nf^2 + 1$.

To implement the RNL algorithm on the automated response detection procedure for the FFR to voice pitch, the averaged recording $(\overline{A_N(t)})$ first went through a cross-correlation with the stimulus waveform to identify the time shift that produces the maximum cross-correlation value between the 3-10 ms response window. A 250 ms segment of data was then extracted from the averaged recording (starting from the maximum cross-correlation value) and was denoted as $\overline{A_N(t)}_c$. The same procedure was applied to the estimate of the noise and the result was denoted as $\overline{D_N(t)}_c$. Thus, for the realm of the FFR to voice pitch, the magnitude of the voice-pitch elicited response can be expressed in the form of SNR, which takes into account of the variance ratio between the estimated response and noise. In this study, the term pitch variance ratio (PVR) is introduced to indicate the SNR of each recording and will be used to determine the presence or absence of an FFR. In sum, the PVR can be expressed as Formula 10.

$$PVR = \frac{Var(\overline{A_N(t)})_c}{Var(\overline{D_N(t)})_c} \tag{7}$$

For each recording, such a calculation was performed and a PVR value was obtained. Because the process of calculating a PVR index is in essence an F-test of equal variance between the estimated signal and noise, a $PVR_{critical}$ value can be determined by the degrees of freedom of the two variances and the desired α level to set a criterion for the automated procedure. The underlying logic of the automated procedure was to use the $PVR_{critical}$ value as a response detection criterion. Any recording that has a PVR value greater than the $PVR_{critical}$ value will be considered to contain physiological energy that stands out from the background noise, thus will be considered to contain a voice pitch elicited FFR. Apparently the $PVR_{critical}$ value will vary depending on the choice of α level and the degrees of freedom of the estimated signal and noise. For our FFR recordings, the degree of freedom is a constant (i.e., 5000 – 1 = 4999) for both the estimated signal and noise. If α level is set at 0.05, an F-statistic table gives a $PVR_{critical}$ value of 1.05. On the other hand, if α level is set at 0.10, an F-statistic table gives a $PVR_{critical}$ value of 1.04. In this study, the more conservative $PVR_{critical}$ value of 1.05 was used as the detection criterion in the automated response detection procedure.

One-way repeated measure ANOVA was conducted between the experimental conditions and the control condition to test the null hypothesis that PVR values obtained from the experimental conditions are not significantly different than those obtained from the control condition.

To further explore the efficiency of the automated procedure, PVR values for all the experimental and control conditions were calculated as a function of number of sweeps. By gradually increasing the number of sweeps included in the averaged waveform (starting from the first recorded sweep), all recordings were reconstructed and re-calculated using the same data analysis procedures as mentioned above (including the averaging, filtering, segmentation, cross-correlation with the stimulus token, and the PVR computation).

Residue-Noise-Level based Temporal Algorithm

Extended from the "narrow-band spectrogram" algorithm (Jeng et al., 2011c), a distinctive statistics-amended response detection algorithm was used on the averaged EEG waveform and its corresponding spectral density distribution representation. With the stimulus (one of the four Mandarin syllables) information, the algorithm first read the pre-stored f0 contour array of the stimulus. It extends its searching range based upon the frequency value of the stimulus' f0 contour. Within in each time segment, the algorithm has a range, ±5 Hz in this study, centered at the frequency value of the stimulus' f0, within which it calculates the mean spectral energy value and it represents the energy of the signal. Outside the signal range, the algorithm also have two ranges that are used to calculate the mean spectral energy for the background noise, on above the upper frequency limit of signal range, 20 Hz in this study, and one below the lower frequency limit, 10 Hz in this study. With a spectral resolution of 1 Hz, the pooled noise part leaves the algorithm a total number of 30 frequency points for the estimation of the noise. A one-sample Student's t-test is then conducted upon the mean spectral energy values of the signal and the noise, i.e. the t-test is carried out between the signal part and the combined noise part in every windowed segments. Thus, for each t-test, the algorithm tests the null hypothesis that the averaged energy in noise part is similar to the signal energy, with a confidence level of 95% and degree of freedom at 29. If a significant difference is detected by the t-test, that windowed segment at the moment will be marked as a "response" segment; otherwise, it will be marked as a "non-response" segment. The algorithm then continues along the temporal axis and repeats the above procedures until the end of the recorded waveform. As there are a total of 201 windowed segments, same t-tests are carried out in all of these segments. The results of these 201 t-tests form a vector of response marks: some would have a mark of "response" and others would be "non-response". A response detection index, Relative Significance Level (RSL), was calculated at this point to be further used in the response detection procedure. For the windowed segments that were marked as a "response" segment, they were given a value of 1, whereas those marked "non-response" were given a value of 0. The RSL value was calculated as

$$RSL = \sum_{i=1}^{201} Segment_i \tag{8}$$

where the value of Segment is either 1 or 0, depending on the result of the t-test in that windowed segment. Thus, a strong voice pitch elicited FFR recording would have a perfect RSL score of 201, while the RSL value obtained from a control condition, i.e. no response, would have a score that was close to 0.

The response detection criterion that was used in the automated procedure was determined by selecting a desired RSL score ($RSL_{critical}$), i.e. the percentage of "significance" segments in the 201 total segments. The automated response detection procedure uses the criterion as a decision maker: if a recording provides a RSL value that is greater than $RSL_{critical}$, it would be considered to contain a voice pitch elicited FFR, otherwise it would be considered as a recording that contains no physiological response.

RESULTS AND DISCUSSION

Pitch-Variance-Ratio based Temporal Algorithm

The PVR values obtained in both the experimental and control conditions are shown in Figure 1. A group comparison (Figure 1A) between the experimental and control conditions revealed that the PVR values obtained from the experimental conditions were significantly larger than those obtained in the control condition ($p < 0.001$). The PVR values obtained from the four experimental conditions had a mean of 1.738 (SD = 1.651), whereas the PVR values obtained from the control condition had a mean of 0.261 (SD = 0.042). Note the mean PVR value was approximately 6.7 times larger than that of the control condition. For the control condition, its PVR values were clustered within a much smaller rage than that of the experimental conditions ($p < 0.001$).
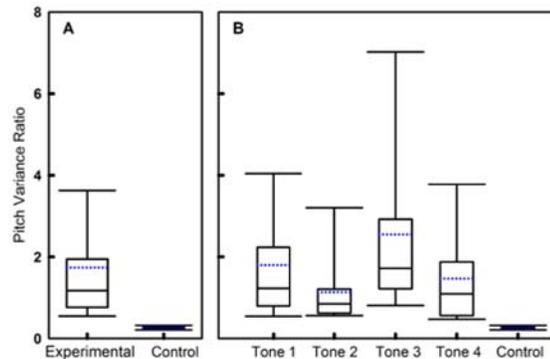
FIGURE 1. Pitch Variance Ratio (PVR) values obtained in the experimental and condition conditions. Group comparison (panel A) shows that the PVR values obtained in the four experimental conditions were significantly larger than those obtained in the control condition. Comparison for each individual tone (panel B) shows the PVR values obtained from each individual tone was also significantly larger than that of the control condition. Blue dotted horizontal lines indicate the mean PVR value for each condition; black solid lines indicate the medians. The upper and lower boundaries of the box indicate the 25th and 75th percentiles, respectively. Whisker bars indicate the 10th and 90th percentile.

For each individual tone (Figure 1B), the mean of its corresponding PVR values (Tones 1, 2, 3 and 4 mean ± SD = 1.799 ± 1.597, 1.134 ± 0.858, 2.550 ± 2.221 and 1.468 ± 1.351, respectively) was also significantly larger than that of the control condition ($p < 0.001$ for Tones 1, 3 and 4, $p = 0.031$ for Tone 2). A post hoc Tukey-Cremer analysis showed no statistical difference among the four tones with the exceptions that Tones 3 versus Tone 2 ($p < 0.001$) and Tones 3 versus Tone 4 are significantly different from each other ($p = 0.004$).
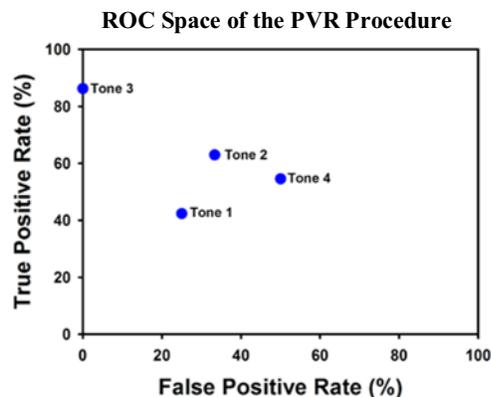
Figure 2. Receiver Operator Characteristics (ROC) space, with marked points obtained from the outputs from PVR based procedure compared with human judgments. The abscissa represents the False Positive Rate, i.e. 1-specificity, whereas the ordinate represents the True Positive Rate, i.e., sensitivity.

Residue-Noise-Level based Spectral Algorithm

The RSL values obtained in both the experimental and control conditions are shown in Figure 3. A group comparison (Fig. 3A) between the experimental (Tones 1, 2, 3, and 4 combined) and control conditions revealed that the RSL values obtained from the experimental conditions were significantly larger than those obtained in the control condition ($p < 0.001$). The RSL values obtained from the four experimental conditions had a mean of 0.91 (SD = 0.11), whereas the RSL values obtained from the control condition had a mean of 0.24 (SD = 0.25). Note the mean RSL value was approximately 4 times larger than that of the control condition. For the control condition, its RSL values were clustered within a much smaller rage than that of the experimental conditions ($p < 0.001$).
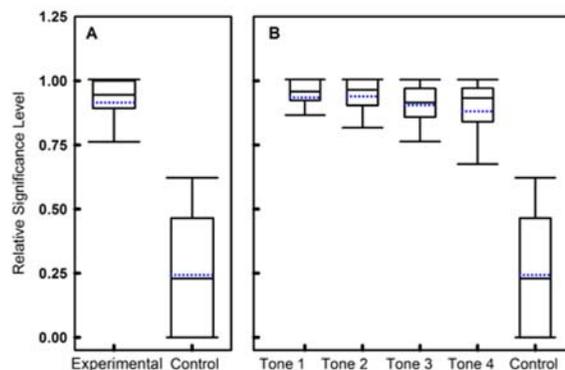
FIGURE 3. Relative Significance Level (RSL) values obtained in the experimental and condition conditions. Group comparison (panel A) shows that the RSL values obtained in the four experimental conditions were significantly larger than those obtained in the control condition. Comparison for each individual tone (panel B) shows the RSL values obtained from each individual tone was also significantly larger than that of the control condition. Blue dotted horizontal lines indicate the mean RSL value for each condition, black solid lines indicate the medians. The upper and lower boundaries of the box indicate the 25th and 75th percentiles, respectively. Whisker bars indicate the 10th and 90th percentile.

For each individual tone (Fig. 3B), the mean of its corresponding RSL values (Tones 1, 2, 3 and 4 mean ± SD = 0.93 ± 0.11, 0.93 ± 0.09, 0.90 ± 0.09 and 0.88 ± 0.13, respectively) was also significantly larger than that of the control condition ($p < 0.001$ for all four tones). A post hoc Tukey-Cremer analysis showed no statistical difference among the four tones.
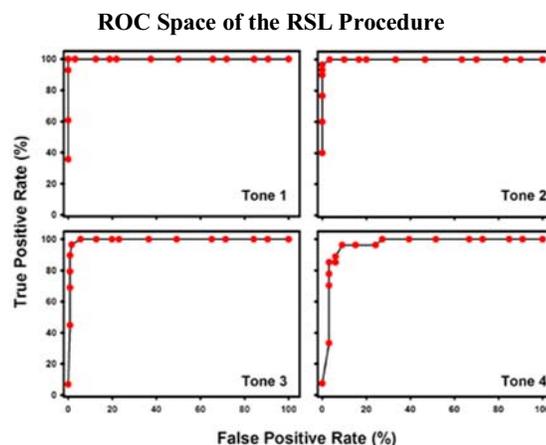
Figure 4. Receiver Operator Characteristics (ROC) space, with marked points obtained from the outputs from RSL based procedure compared with human judgments. The abscissa represents the False Positive Rate, i.e. 1-specificity, whereas the ordinate represents the True Positive Rate, i.e., sensitivity.

In summary, this study developed and evaluated an automated response detection procedure that was based on the statistical properties of the recorded data itself. The results were compatible with those obtained from a traditional method of human interpretation. The results also showed that the automated procedure could be used to

replace the work that traditionally was completed by a human observer. By eliminating the observer's input in the decision making process, this automated procedure is independent from the observer's subjective interpretation and thus it would require less amount of training for an end user to record and determine the presence or absence of an FFR to voice pitch. This advantage promotes the efficiency and portability of the proposed technology and thus reduces the cost in its research and clinical applications.

REFERENCES

Akhoun, I., Galle´go S., Moulin, A., Me´nard, M., Veuillet, E., et al. **(2008)**. The temporal relationship between speech auditory brainstem responses and the acoustic pattern of the phoneme /ba/ in normal-hearing adults. Clin Neurophysiol, **119**, 922–933.

Chandrasekaran, B., Hornickel, J., Skoe, E., Nicol, Trent. and Kraus, N. **(2009)**. Context dependent encoding in the human auditory brainstem relates to hearing speech in noise: Implications for developmental dyslexia. Neuron, **64**, 311–319.

Jeng, F.-C., Schnabel, E. A., Dickman, B. M., Hu, J., Li, X., et al. **(2010)**. Early maturation of frequency-following responses to voice pitch in infants with normal hearing. Percept Mot Skills, **111(3)**, 765-784.

Jeng, F.-C., Chung, H.-K., Lin, C.-D., Dickman, B. M. and Hu, J. **(2011a)**. Exponential modeling of human frequency-following responses to voice pitch. Int J Audiol, **50**, 582-593.

Jeng, F.-C., Costilow, C. E., Stangherlin, D. P. and Lin, C.-D. **(2011b)**. Relative power of harmonics in human frequency-following responses associated with voice pitch in American and Chinese adults. Percept Mot Skills, **113(1)**, 67-86.

Jeng, F.-C., Hu, J., Dickman, B. M., Lin, C.-Y., Lin, C.-D., et al. **(2011c)**. Evaluation of two algorithms for detecting human frequency-following responses to voice pitch. Int J Audiol, **50(1)**, 14-26.

Jeng, F.-C., Hu, J., Dickman, M. B., Montgomery-Reagan, K., Tong., et al. **(2011d)**. Cross-linguistic comparison of frequency-following responses to voice pitch in American and Chinese neonates and adults. Ear Hear, **32(6)**, 699-707.

Johnson, K. L., Nicol, T., Zecker, S. G., Bradlow, A. R., Skoe, E., et al. **(2008)**. Developmental plasticity in the human auditory brainstem. J Neurosci, **28**, 4000-4007.

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

Krishnan, A., Gandour, J. T., Bidelman, G. M. and Swaminathan, J. **(2009)**. Experience-dependent neural representation of dynamic pitch in the brainstem. Neuroreport, **20(4)**, 408-413.

Krishnan, A., Bidelman, G. M., Smalt, C. J., Ananthakrishnan, S. and Gandour, J. T. **(2012)**. Relationship between brainstem, cortical and behavioral measures relevant to pitch salience in humans. Neuropsychologia, **50(12)**, 2849-2859.

Li, X. and Jeng, F.-C. **(2011)**. Noise tolerance in human frequency-following responses to voice pitch. J Acoust Soc Am, **129(1)**, EL21-26.

Moore, B. C. J. **(2003)**. Pitch perception. In B. C. J. Moore (ed.) An Introduction to the Psychology of Hearing, Fifth Edition. London: Academic Press, pp. 158-193.

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

Skoe, E. and Kraus, N. **(2010)**. Auditory brain stem response to complex sounds: A tutorial. Ear Hear, **31**, 302-324.

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

Xu, Y., Krishnan, A. and Gandour, J. T. **(2006)**. Specificity of experience dependent pitch representation in the brainstem. Neuroreport, **17**, 1601-1605.

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T. and Kraus, N. **(2007)**. Musical experience shapes human brainstem encoding of linguistic pitch patterns. Nat Neurosci, **10(4)**, 420-422.