Musician effect in cochlear implant simulated gender categorization

Fuller, Christina D.; Galvin, John J.; Free, Rolien H.; Başkent, Deniz

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/1.4865263

Publication date:
2014

Musician effect in cochlear implant simulated gender categorization

Christina D. Fuller,[a] John J. Galvin III,[b] Rolien H. Free,[c] and Deniz Başkent[c]

University of Groningen, University Medical Center Groningen, Department of Otorhinolaryngology/Head and Neck Surgery, Groningen, The Netherlands
c.d.fuller@umcg.nl, jgalvin@ucla.edu, r.h.free@umcg.nl, d.baskent@umcg.nl

Abstract: Musicians have been shown to better perceive pitch and timbre cues in speech and music, compared to non-musicians. It is unclear whether this "musician advantage" persists under conditions of spectro-temporal degradation, as experienced by cochlear-implant (CI) users. In this study, gender categorization was measured in normal-hearing musicians and non-musicians listening to acoustic CI simulations. Recordings of Dutch words were synthesized to systematically vary fundamental frequency, vocal-tract length, or both to create voices from the female source talker to a synthesized male talker. Results showed an overall musician effect, mainly due to musicians weighting fundamental frequency more than non-musicians in CI simulations.

PACS numbers: 43.71.Bp, 43.75.St, 43.64.Me [SGS]
Date Received: October 31, 2013 **Date Accepted:** January 19, 2014

1. Introduction

Identifying a talker's gender depends on two anatomically related vocal characteristics: (1) Fundamental frequency (*F*0), mainly related to vocal pitch, and (2) vocal-tract length (VTL), mainly related to the size of the speaker (Smith and Patterson, 2005). The ability to identify the voice of a talker is important to separate various talkers in a multi-talker environment and possibly improve speech intelligibility (Brungart, 2001). Recently Fuller *et al.* (2013) demonstrated that cochlear-implant (CI) users do not utilize both voice cues efficiently. Due to a diminished weighting of VTL cues and an over-reliance on *F*0 cues, CI users' gender categorization differs from that of normal hearing (NH) listeners, possibly leading to errors in categorization under certain conditions.

NH musicians have been shown to better understand speech in noise, better discriminate voices on the basis of timbre differences, and better perceive pitch in both speech and music, compared to non-musicians (e.g., Schon *et al.*, 2004; Chartrand and Belin, 2006; Parbery-Clark *et al.*, 2009). This "musician advantage" has been shown to enhance linguistic processing at brainstem, subcortical, and cortical levels, and is associated with better functional working memory and auditory attention (e.g., Besson *et al.*, 2011). Some of the musician advantage for speech-related tasks has been attributed to a better perception of acoustical cues, such as timbre or prosody (e.g., Deguchi *et al.*, 2012).

Based on these findings, musicians might be expected to better perceive both *F*0 and VTL cues compared to non-musicians. Fuller *et al.* (2013) showed CI users

[a]Author to whom correspondence should be addressed. Also at: Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Graduate School of Medical Sciences, Groningen, The Netherlands.
[b]Also at: Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Graduate School of Medical Sciences, Groningen, The Netherlands; Division of Communication and Auditory Neuroscience, House Research Institute, Los Angeles, CA 90057; Department of Head and Neck Surgery, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095-1624.
[c]Also at: Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Graduate School of Medical Sciences, Groningen, The Netherlands.

rely almost exclusively on *F*0 cues to categorize voice gender. NH non-musicians listening to acoustic CI simulations used both cues less efficiently, but relied more strongly on *F*0 cues in comparison to listening to normal acoustical stimuli. Under such conditions of spectro-temporal degradation, musicians may be better able to extract *F*0 and VTL information, compared to non-musicians. If so, past musical experience or active music training may benefit CI users' gender categorization that may help speech perception in noise. In this study, voice gender categorization was measured in NH musicians and non-musicians listening to unprocessed speech or to an acoustic CI simulation. Dutch words were synthesized to vary *F*0, VTL, or both, thereby systematically creating voices from the female source talker to a synthesized male talker. We hypothesized that musicians would be better able to utilize the *F*0 and VTL cues in a gender categorization task than non-musicians, especially with the CI simulation.

2. Materials and methods

2.1 Participants

Twenty-five NH musicians and 25 NH non-musicians were recruited for this study. "Musician" inclusion criteria were defined as: (1) Having begun musical training before or at the age of 7 yrs; (2) having had musical training for 10 yrs or more; and (3) having received some musical training within the last 3 yrs (Micheyl *et al.*, 2006; Parbery-Clark *et al.*, 2009). In addition to not meeting the musician criteria, non-musicians were defined as not having received musical training within the 7 yrs before the study. All participants had pure tone thresholds better than 20 dB HL at audiometric test frequencies between 250 and 4000 Hz, and all were native Dutch speakers with no neurological disorders.

The study was approved by the Medical Ethical Committee of the University Medical Center Groningen. Participants were given detailed information about the study and written informed consent was obtained. A financial reimbursement was provided.

2.2 Stimuli

Four meaningful Dutch words in consonant-vowel-consonant format ("bus," "vaak," "leeg," and "pen," meaning "bus," "often," "empty," and "pencil," respectively) were used as sources for subsequent speech synthesis. The source speech tokens were taken from the NVA corpus (Bosman and Smoorenburg, 1995) and produced by a single, female Dutch talker. The naturally spoken tokens were systematically manipulated to produce voices that ranged from the female to a male talker, using the STRAIGHT software (v40.006b), implemented in MATLAB and developed by Kawahara *et al.* (1999). The *F*0 was decreased by an octave in five steps, 0, 3, 6, 9, or 12 semitones (st), and the VTL was increased by 23% (resulting in a downward spectral shift of 3.6 st) in six steps, 0.0, 0.7, 1.6, 2.4, 3.0, or 3.6 st, relative to the female voice. All combinations were generated, resulting in 30 synthesized "voices" and a total of 120 stimuli (4 words $\times$ 5 *F*0 values $\times$ 6 VTL values); note that all 120 stimuli were synthesized. Mm. 1 contains the word "bus" with four example voice manipulations.

Mm. 1. The word "bus" for (1) 0 semitone change in *F*0 and a 0 semitone change in VTL (female voice); (2) 12 semitone change in *F*0 and 0 semitone change in VTL; (3) 0 semitone change in *F*0 and 3.6 semitone change in VTL; and (4) 12 semitone change in *F*0 and 3.6 semitone change in VTL (male voice).

2.3 CI simulations

Eight-channel, sine-wave vocoded acoustic CI simulations were generated using Angelsound™ software (Emily Shannon Fu Foundation, http://www.angelsound.tigerspeech.com/). The acoustical input was first band limited to a frequency range of 200 to 7000 Hz, and then bandpass-filtered into 8 frequency analysis bands [fourth order Butterworth filters with band cutoff frequencies according to Greenwood (1990)

frequency-place formula]. For each channel, the temporal envelope was extracted using half-wave rectification and low-pass filtering (fourth order Butterworth filter with cutoff frequency = 160 Hz). These envelopes modulated a sinusoidal carrier that was equal to the center frequency of the analysis filter. The modulated carriers were summed to produce the final stimulus and the overall level was adjusted to be the same level as the original signal. Figure 1 shows the spectrums for the word "bus". The middle row shows the original stimulus resynthesized in STRAIGHT, with the original parameters of the recorded female voice. In the top row, only the *F*0 was changed, by an octave down. In the bottom row, only the VTL was changed to be made 23% longer, which results in shifting all the formants down by 3.6 st. The left panels show the non-simulated stimulus and the right panel shows the CI-simulated stimulus.

2.4 Procedure

The stimuli were presented using AngelSound™ software (Emily Shannon Fu Foundation, http://www.angelsound.tigerspeech.com/) and were played from a PC with an Asus Virtuoso Audio Device soundcard (ASUSTeK Computer Inc., Fremont, CA). Participants were seated in an anechoic chamber facing the speaker (Tannoy Precision 8D; Tannoy Ltd., North Lanarkshire, United Kingdom) at 1 m distance. After conversion to an analog signal via a DA10 digital-to-analog converter of Lavry Engineering

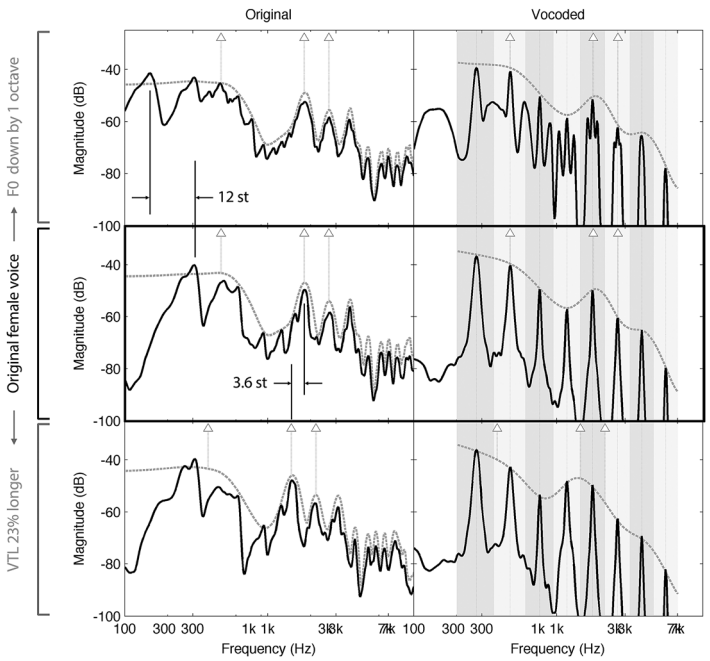

Fig. 1. Power spectrum and waveform of the vowel /u/ in "bus". Each row represents a different voice. The *middle row* shows the stimulus resynthesized, in STRAIGHT, with the original parameters of the recorded female voice. In the *top row*, only the *F*0 was changed, by an octave down. In the *bottom row*, only the VTL was changed to be made 23% longer, which results in shifting all the formants down by 3.6 st. The *left panel* shows the spectra over the duration of the vowel, for the vocoded (*right column*) and non-vocoded (*left column*, noted "Original") versions of the stimulus. The black solid line represents the spectrum itself, making the harmonics and/or the sinusoidal carriers (and sidebands) of the vocoder visible. The dashed gray line represents the spectral envelope, as extracted by STRAIGHT on the left, and interpolating between the carriers for the vocoded sounds on the right. The triangles and stems point to the location of the first three formants, as defined by visual inspection of the STRAIGHT envelope, both for the left and right columns. In the right column, the vocoder analysis filter bands are shown with grayed areas. The frequency of the sine-wave carrier is marked with a dotted line.

Inc. (Washington, USA), the stimuli were played at 65 dB sound pressure level in the free field. All stimuli were randomly selected from the stimulus set (without replacement) and played to the subject once. The subject indicated whether the talker was a man or woman by selecting one of two response buttons shown on an A1 AOD 1908 touch screen (GPEG International, Woolwich, United Kingdom) and labeled "man" or "vrouw" (i.e., "man" and "woman"). Subject responses were recorded by the testing software. No feedback was provided. All of the NH listeners were familiar with CI simulations as they had participated in similar experiments before, but otherwise no specific training for the gender recognition task was provided. The gender categorization task lasted for 10 min, resulting in a total testing time of approximately 20 min for all participants.

2.5 Cue weighting

To quantify how efficiently musicians and non-musicians used voice cues, perceptual weighting of *F*0 and VTL was calculated using a generalized linear mixed model based on a binomial distribution (logit link function). *F*0 and VTL were fixed factors and subject was the random intercept. The model was applied to normalized dimensions defined as $F0 = -\Delta F0/12$ and $\mathrm{VTL} = \Delta\mathrm{VTL}/3.6$, where $\Delta F0$ and $\Delta\mathrm{VTL}$ represent the *F*0 or VTL difference in st relative to the source talker. With these normalized dimensions, the point (0,0) represents the synthesized female talker and the point (1,1) represents the synthesized male talker. The cue weights were then expressed as a and b in the equation $\mathrm{logit}(\mathrm{score}) = a\ F0 + b\ \mathrm{VTL} + \varepsilon$, where ε is the random intercept that is subject-dependent.

3. Results

Figure 2 shows the mean voice gender categorization with unprocessed (but synthesized) stimuli for non-musicians (left panels) and musicians (right panels), as a function of *F*0 difference (top plots) or VTL difference (bottom plots) relative to the female

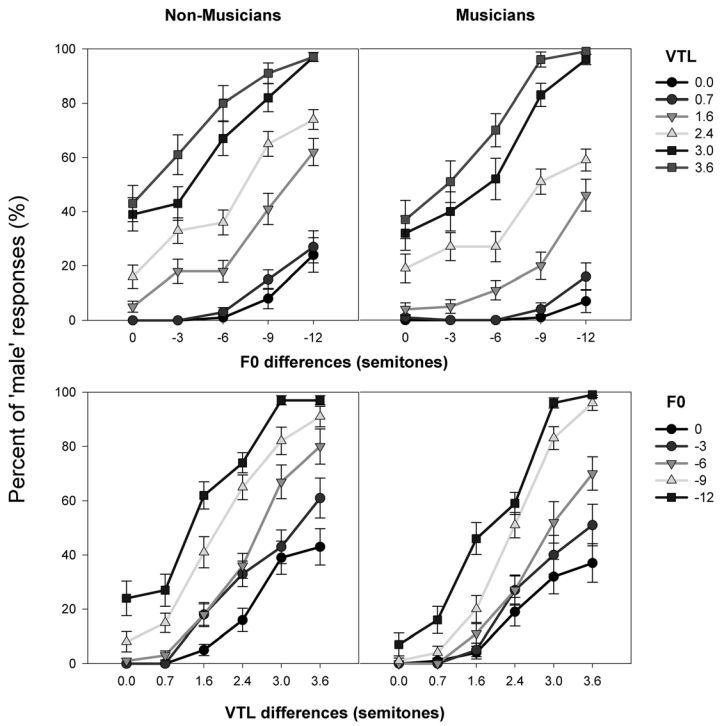

Fig. 2. Mean gender categorization (across subjects and test words) shown for non-musicians (left panels) and musicians (right panels) tested with unprocessed (but synthesized) stimuli, as a function of the difference in *F*0 (top plots) or VTL (bottom plots) relative to the female source talker. Error bars denote one standard error of the mean.

source talker. The percentage of male responses was averaged across the four words and across subjects. Both musicians and non-musicians seemed to utilize both *F*0 and VTL cues similarly, as confirmed by the cue weighting analysis (musicians: $F0 = 3.77$, $VTL = 6.99$; non-musicians: $F0 = 3.72$, $VTL = 5.56$).

Figure 3 shows similar data for musicians and non-musicians as in Fig. 2, but with the acoustic CI simulation. In general, both subject groups seemed to use both *F*0 and VTL less efficiently compared to the unprocessed condition (Fig. 2). This is shown by the flatter performance lines and by the lower cue weighting (musicians: $F0 = 2.03$, $VTL = 0.31$; non-musicians: $F0 = 1.76$, $VTL = 0.99$). The pattern of results with the *F*0 cue was more diffuse for non-musicians than for musicians, who scored similarly regardless of the VTL cue. The cue weighting analysis suggests that musicians utilized *F*0 cues more and VTL cues less when compared to non-musicians.

A three-way repeated measures, split-plot analysis of variance (ANOVA) was performed on all data using a Greenhouser-Geisser correction to ensure sphericity assumption (Table 1). The within-subject factors were *F*0 (five levels), VTL (six levels), and listening condition (two levels: Unprocessed, CI simulated); the between-subject factor was musical experience (two levels: Musician, non-musician). Results confirm a significant overall musician effect on gender categorization. There were significant interactions between *F*0 and VTL, the listening condition and VTL, and the listening condition, *F*0 and VTL. However, there were no significant interactions between musical experience and any other factors.

4. Discussion

Based on previous studies in which a positive musician effect in NH listeners had been observed in speech and music-related tasks (Kraus and Chandrasekaran, 2010; Besson *et al.*, 2011; Patel, 2014), we hypothesized that musicians would utilize the voice cues for gender categorization more effectively than non-musicians, especially in

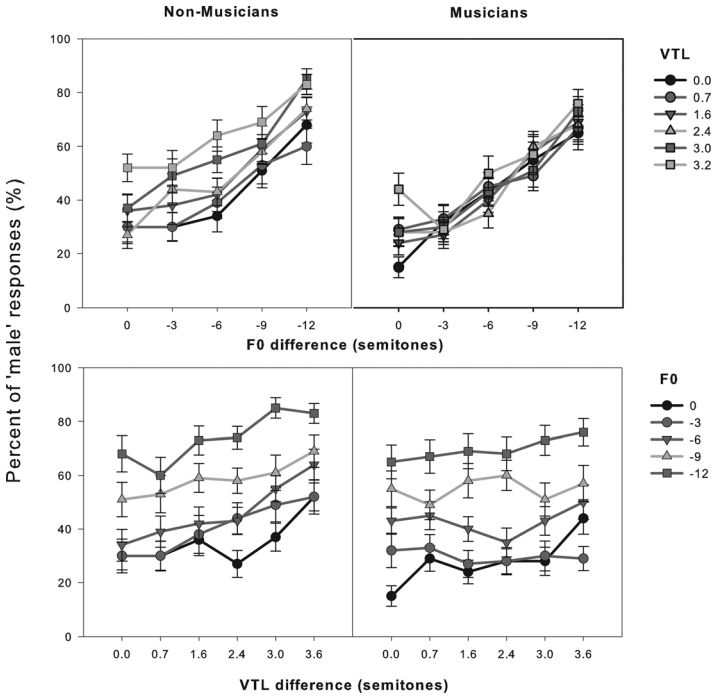

Fig. 3. Mean gender categorization (across subjects and test words) shown for non-musicians (left panels) and musicians (right panels) tested with the CI simulation, as a function of the difference in *F*0 (top plots) or VTL (bottom plots) relative to the female source talker. Error bars denote one error of the mean.

Table 1. Results of split-plot, three-way repeated measures ANOVA. The shading indicates significant effects.

Factors	*F*-ratio, *p*-value
Musical experience (musician, non-musician)	$F(1,1) = 4.47$, $p = 0.040$*
*F*0	$F(2.23,107.17) = 148.05$, $p < 0.001$ **
VTL	$F(2.68,128.72) = 197.73$, $p < 0.001$ **
Listening condition (unprocessed, CI simulation)	$F(1,1) = 41.09$, $p < 0.001$ **
*F*0 × Musical experience	$F(2.23,107.17) = 0.240$, $p = 0.811$
VTL × Musical experience	$F(2.68,128.72) = 1.81$. $p = 0.154$
Listening condition × Musical experience	$F(1,1) = 0.02$, $p = 0.889$
*F*0 × VTL	$F(13.11,629.47) = 6.70$, $p < 0.001$**
*F*0 × VTL × Musical experience	$F(13.11,629.47) = 1.67$, $p = 0.062$
Listening condition × *F*0	$F(1.82,87.53) = 0.58$. $p = 0.549$
Listening condition × VTL	$F(2.52,120.92) = 108.12$, $p < 0.001$**
Listening condition × VTL × Musical experience	$F(1.82,120.92) = 2.36$, $p = 0.086$
Listening condition × *F*0 × Musical experience	$F(1.82,87.53) = 1.23$, $p = 0.296$
Listening condition × *F*0 × VTL	$F(13.46,645.93) = 7.30$, $p < 0.001$**
Listening condition × *F*0 × VTL × Musical experience	$F(13.46,645.93) = 1.26$, $p = 0.23$

spectrally degraded conditions like the CI simulation. This study showed an overall musician effect, mainly in the CI simulation, and that the perceptual weighting of the two voice cues differed between musicians and non-musicians. Musicians perceptually weighted *F*0 more, but VTL less, than non-musicians in the CI simulation. It is possible that the CI simulation delivered *F*0 cues more reliably than VTL cues, and musicians made better use of the more reliable cue. Alternatively, musicians may have been more sensitive to *F*0 cues, and therefore relied on *F*0 cues more strongly than on VTL cues. If this is the case, musicians would appear to perform similarly to CI users, who have been shown to rely almost exclusively on *F*0 cues for gender categorization (Fu *et al.*, 2005). This may be a coincidence, as CI users generally do not have extensive musical experience due to hearing impairment (Fuller *et al.*, 2012). On the other hand, in CI simulations, as well as in actual CIs, VTL cues are likely less reliable. This is perhaps the reason for the overall low weighting of VTL in CI simulations, by musicians and non-musicians, as well as CI users. Hence, it may be more advantageous to rely on the more robust cue of *F*0.

While an overall musician effect was observed, note that the perceptual weighting of *F*0 and VTL cues was similar for musicians and non-musicians with unprocessed speech. Indeed, performance differences were quite small between subject groups with unprocessed speech (Fig. 2). Previous studies have shown better voice timbre recognition and pitch perception in both speech and music by musicians listening to unprocessed acoustic stimuli (Chartrand and Belin, 2006; Parbery-Clark *et al.*, 2009). The present gender categorization task may have been too easy with unprocessed stimuli, compared to a voice discrimination task (Chartrand and Belin, 2006). As such, gender categorization with unprocessed speech may not have been as sensitive to musical experience. Furthermore, the *F*0 steps used to synthesize the present "talkers" may have been too large to elicit differences in performance between musicians and non-musicians observed in previous studies (Chartrand and Belin, 2006; Micheyl *et al.*, 2006; Parbery-Clark *et al.*, 2009). Nevertheless, the musician effect on VTL perception or on combined VTL and *F*0 perception has not been studied previously. The present data do not show a difference in these between musicians and non-musicians under normal listening situations.

The overall findings add to the previous reports of cross-domain effect of musical experience to speech-related tasks, such as voice timbre recognition and voice discrimination (Chartrand and Belin, 2006). In general, musical experience has been shown to enhance performance in a number of listening tasks. Music training has also been shown to improve CI users' music perception (e.g., Galvin *et al.*, 2007). Based on

these observations, music training may benefit CI users' speech perception, especially when pitch cues are important, for example, for separating foreground speech from masking speech and better understanding speech in multi-talker environments (Brungart, 2001).

Acknowledgments

We would like to thank the participants of this study. Furthermore, we would like to thank the research assistants for their help with collecting the data and Etienne Gaudrain for the help with the figures. We thank Qian-jie Fu and the Emily Fu Foundation for software support. J.J.G. is supported by Grant No. NIH R01-DC004792. R.H.F. is supported by an otological/neurotological stipendium from the Heinsius-Houbolt Foundation. D.B. is supported by a Rosalind Franklin Fellowship from the University Medical Center Groningen, University of Groningen, and the VIDI Grant No. 016.096.397 from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw). The study is part of the research program of our department: Healthy Aging and Communication. There is no conflict of interest regarding this manuscript.

References and links

Besson, M., Chobert, J., and Marie, C. (**2011**). "Transfer of training between music and speech: Common processing, attention, and memory," Front. Psychol. **2**, 94.

Bosman, A. J., and Smoorenburg, G. F. (**1995**). "Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment," Audiol. **5**, 260–284.

Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**, 1101–1109.

Chartrand, J. P., and Belin, P. (**2006**). "Superior voice timbre processing in musicians," Neurosci. Lett. **3**, 164–167.

Deguchi, C., Boureux, M., Sarlo, M., Besson, M., Grassi, M., Schon, D., and Colombo, L. (**2012**). "Sentence pitch change detection in the native and unfamiliar language in musicians and non-musicians: behavioral, electrophysiological and psychoacoustic study," Brain Res. **1455**, 75–89.

Fu, Q. J., Chinchilla, S., Nogaki, G., and Galvin, J. J., III (**2005**). "Voice gender identification by cochlear implant users: The role of spectral and temporal resolution," J. Acoust. Soc. Am. **118**(3), 1711–1718.

Fuller, C., Gaudrain, E., Clarke, J., Galvin, J. J., Fu, Q. J., Free, R., and Başkent, D. (**2013**). "Little Red Riding Hood was a Cochlear Implant User!!," Poster session presented at the *16th Conference on Implantable Auditory Prostheses*, July 14–19 (2013), Lake Tahoe, California. Accessible via: http://dx.doi.org/10.6084/m9.figshare.871408.

Fuller, C., Maat, A., Free, R., and Başkent, D. (**2012**). "Musical background not associated with self-perceived hearing performance or speech perception in postlingual cochlear-implant users," J. Acoust. Soc. Am. **132**, 1009–1016.

Galvin, J. J., III, Fu, Q. J., and Nogaki, G. (**2007**). "Melodic contour identification by cochlear implant listeners," Ear Hear. **3**, 302–319.

Greenwood, D. D. (**1990**). "A cochlear frequency position function for several species—29 years later," J. Acoust. Soc. Am. **87**(6), 2592–2605.

Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (**1999**). "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Commun. **3**, 187–207.

Kraus, N., and Chandrasekaran, B. (**2010**). "Music training for the development of auditory skills," Nat. Rev. Neurosci. **8**, 599–605.

Micheyl, C., Delhommeau, K., Perrot, X., and Oxenham, A. J. (**2006**). "Influence of musical and psychoacoustical training on pitch discrimination," Hear. Res. **1–2**, 36–47.

Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009**). "Musician enhancement for speech-in-noise," Ear Hear. **6**, 653–661.

Patel, A. D. (**2014**). "Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis," Hear. Res. **308**(2), 98–108.

Schon, D., Magne, C., and Besson, M. (**2004**). "The music of speech: music training facilitates pitch processing in both music and language," Psychophysiology **3**, 341–349.

Smith, D. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgments of speaker size, sex, and age," J. Acoust. Soc. Am. **5**, 3177–3186.