MAY 01 2018

Musician effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents

Deniz Başkent; Christina D. Fuller; John J. Galvin, III; Like Schepel; Etienne Gaudrain; Rolien H. Free

J. Acoust. Soc. Am. 143, EL311–EL316 (2018)
https://doi.org/10.1121/1.5034489

Musician effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents

Deniz Başkent,[1,a),b)] Christina D. Fuller,[1,b)] John J. Galvin III,[2] Like Schepel,[1] Etienne Gaudrain,[1,c)] and Rolien H. Free[1,b)]

[1]*Department of Otorhinolaryngology/Head and Neck Surgery, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands*

[2]*House Ear Institute, Los Angeles, California 90057, USA*

d.baskent@umcg.nl, c.d.fuller@umcg.nl, jgalvin@hei.org, l.schepel@franciscus.nl, etienne.gaudrain@cnrs.fr, r.h.free@umcg.nl

Abstract: In adult normal-hearing musicians, perception of music, vocal emotion, and speech in noise has been previously shown to be better than non-musicians, sometimes even with spectro-temporally degraded stimuli. In this study, melodic contour identification, vocal emotion identification, and speech understanding in noise were measured in young adolescent normal-hearing musicians and non-musicians listening to unprocessed or degraded signals. Different from adults, there was no musician effect for vocal emotion identification or speech in noise. Melodic contour identification with degraded signals was significantly better in musicians, suggesting potential benefits from music training for young cochlear-implant users, who experience similar spectro-temporal signal degradations.

[DDO]

Date Received: December 22, 2017 **Date Accepted:** April 10, 2018

1. Introduction

Long-term music training has been shown to benefit music perception, pitch perception, and to some degree, speech perception.[1,2] A few studies have shown a musician advantage for speech understanding with competing speech, where voice pitch cues presumably play a role in segregating target and interfering speech.[3,4] Musicians appear to be better able to extract pitch cues under difficult listening conditions. For example, even when signals were spectro-temporally degraded, Fuller *et al.*[5] found a musician advantage for perception of melodic contours and vocal emotion. Such spectro-temporal degradation is experienced by cochlear implant (CI) users, who have difficulty with pitch-mediated listening tasks such as music perception, vocal emotion perception, voice gender perception, and speech prosody perception.[6–9]

Most previous studies examining the musician effect have been conducted with adults. It is unclear whether long-term music training would also provide an advantage for children. In this study, we explored musician effect on perception of speech, vocal emotion, and music with both unprocessed signals and degraded signals in adolescent normal-hearing listeners aged 11–14 yrs, similar to previous related studies with adults.[3,5] This age range might represent a "sweet spot" where music training may accelerate development of speech and music perception. For this age group, language development would not be entirely completed yet, but is in the later stages.[10] Under conditions of spectro-temporal degradation, perception of frequency sweeps (presumably related to the ability to use dynamic pitch cues in speech) have been shown to be poorer in children than in adults.[11] Adolescent musicians have had several years of musical training during this important developmental period, and thus might exhibit advantages for speech and music perception compared to similarly aged non-musicians.[12]

Testing with unprocessed and spectro-temporally degraded signals may differentiate musician effects in adolescents for relatively easy and difficult listening conditions, as well as investigate robustness of these effects.[5] Performance under conditions

[a)]Author to whom correspondence should be addressed.

[b)]Also at: Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, The Netherlands.

[c)]Also at: CNRS UMR 5292 Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics, Université Lyon, Lyon, France.

of spectro-temporal degradation is also relevant for pediatric CI users, as music training has been shown to improve pitch and speech prosody perception in this population.[13,14] Using a similar degradation will, hence, also provide insights into potential benefits for pediatric CI users.

2. Methods

Perception of music (melodic contours), vocal emotion in spoken phrases, and speech in noise (words in noise, and sentences in competing speech) were measured with unprocessed signals or with an 8-channel sine-wave vocoder to (partially) simulate the spectro-temporal degradation experienced by CI users.[3,5]

2.1 Participants

Participants included 10 musicians (7 female; mean age at testing: 12.4 yrs, range 11–13 yrs) and 11 non-musicians (3 female; mean age at testing: 12.3 yrs, range 11–14 yrs), all recruited from local primary and middle schools, music schools, and orchestras. The general inclusion criteria for all participants were native Dutch language, normal hearing ($\leq$20 dB hearing level at audiometric test frequencies between 0.5 and 6 kHz), and no neurological, cognitive, or developmental disorders. Musician and non-musician criteria were similar to previous studies with adults,[2,3,5] except slightly modified to accommodate for the younger age. Musician inclusion criteria were (1) having begun musical training at or before age 7 yrs, (2) having more than 5 yrs of musical training, and (3) having musical training on a regular basis within the last 3 yrs. Non-musician inclusion criteria were (1) not meeting the musician criteria, and (2) having no musical training (i.e., no activities such as music or singing lessons, playing in an orchestra or band, or playing an instrument regularly) within the last 5 yrs. The study was approved by the Medical Ethical Committee of the University Medical Center Groningen. Parents and participants older than 12 yrs provided written informed consent, and all participants received a gift card for compensation according to departmental guidelines.

2.2 Stimuli and procedure for individual tests

Melodic contour identification. Melodic contour identification (MCI) was measured using a closed-set paradigm for one of the conditions in Fuller *et al.*:[5] organ target, presented simultaneously with a piano masker. Target stimuli consisted of nine 5-note melodic contours played by a MIDI organ sample (Roland Sound Canvas GS). The lowest note among the contours was A3 (220 Hz). The spacing between successive notes was 1, 2, or 3 semitones. The masker was a "flat" contour played by a MIDI piano sample; the same note (A3) was repeated five times. For both the target and masker contours, the duration of each note was 250 ms and the silent period between successive notes was 50 ms. MCI was measured in 2 repetitions of one test block consisting of 27 stimuli (2 blocks $\times$ 9 contours $\times$ 3 spacings = 54 stimuli in total). During testing, a target stimulus was randomly selected from the set (without replacement) and presented to the participant, who was asked to indicate the perceived melodic contour by pressing on one of the nine response boxes on a touchscreen monitor labeled according to the nine contours. In this and all others tests, participants heard each stimulus once only in each test block.

Vocal emotion identification. Vocal emotion identification was measured using a closed-set paradigm with a nonsense word ("nutohɔmsɛpikaŋ") stimulus, produced by two male and two female Dutch-speaking actors in four emotions: joy, anger, relief, and sadness.[15] The total duration and amplitude were normalized as in Fuller *et al.*[5] to encourage listeners to mainly rely on pitch contours for identification. Vocal emotion identification was measured in one test block consisting of 32 stimuli (4 emotions $\times$ 4 talkers $\times$ 2 utterances). During testing, a stimulus was randomly selected from the set and presented to the participant, who was asked to indicate the perceived vocal emotion by pressing on one of the four response boxes on a touchscreen monitor labeled according to the four target emotions.

Word identification in noise. Word identification was measured in steady, speech-shaped noise at 10, 5, and 0 dB signal-to-noise ratios (SNRs) using an open-set paradigm. Stimuli consisted of meaningful monosyllabic Dutch words in CVC format taken from the NVA corpus.[16] One randomly selected list of 12 words was tested for each SNR condition. During testing, a word was randomly selected from the list (without replacement) and participants were asked to verbally repeat the word they heard. No time limit was imposed; however, participants were asked to indicate if they could not produce a response. The experimenter entered the correctly identified words for each trial using the custom test software, and the software automatically calculated the overall percent correct.

Sentence identification with competing speech. Sentence identification was measured in the presence of a competing talker, using an open-set paradigm and meaningful Dutch sentences with rich semantic context taken from Plomp and Mimpen.[17] The target sentence was always produced by a female talker. Similar to Başkent and Gaudrain,[3] the masker was a concatenation of partial sentences, produced by the same female talker or a different, male talker, and always preceded the target. Speech reception thresholds (SRTs) were measured using an adaptive 1-up/1-down procedure.[5,17] One list of 13 sentences was randomly selected (without replacement) to test each masker condition. During testing, a sentence was randomly selected from within the list (without replacement) and presented at the initial target SNR (typically 0 dB for unprocessed speech, 15 dB for degraded speech). The participants verbally repeated the sentence as accurately as possible, and the experimenter marked the response within the custom test software. The first test sentence was presented repeatedly until the participant repeated all words correctly, with the SNR increased for each wrong answer. For the remaining sentences, if the entire sentence was repeated correctly, the SNR was reduced by 2 dB, and if the entire sentence was not repeated correctly (i.e., even one incorrect word), the SNR was increased by 2 dB. The SRT was automatically calculated by the software as the average SNR across the reversals during the final nine sentences.

2.3 Vocoder degradation

For the vocoder, the input frequency range (0.2–7.9 kHz) was divided into 8 channels according to Greenwood's formula[18] (fourth order Butterworth bandpass filters). The temporal envelope was extracted from each analysis band (fourth order Butterworth lowpass filter, cutoff frequency = 160 Hz) and used to modulate sine waves whose frequency corresponded to the center frequency of the analysis bands. The modulated sine waves were then summed and the level of the stimuli was adjusted to have the same long-term root-mean-square amplitude.

2.4 General procedures

All stimuli were presented at 65 dBA in sound field in an anechoic chamber via single loudspeaker (Tannoy Precision 8D; Tannoy Ltd., North Lanarkshire, United Kingdom) connected to a computer soundcard (Asus Virtuoso; ASUSTeK Computer Inc., Fremont, CA). Participants were seated directly facing the loudspeaker, positioned 1 m away. All stimuli were processed and presented via custom software (iStar: www.tigerspeech.com/istar;[21] AngelSound: angelsound.tigerspeech.com[22]). To familiarize participants with the stimuli and procedures for each listening task, participants were given a brief training session with auditory and visual feedback using stimuli different from those used for testing; during testing, no feedback was provided.

3. Results

Figure 1 shows boxplots of musician and non-musician scores for individual tests. For both participant groups, mean performance was much poorer with degraded signals

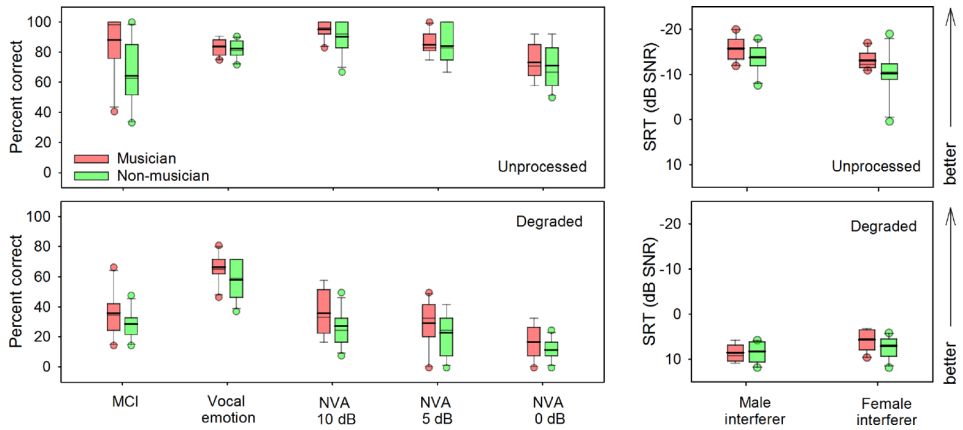

Fig. 1. (Color online) Boxplots for MCI, vocal emotion identification, NVA word identification in noise (left panels), and SRTs for sentence identification with competing speech (right panels) with unprocessed (top row) or degraded signals (bottom row), for adolescent musicians and non-musicians. The boxes show the 25th and 75th percentiles, the error bars show the 5th and 95th percentiles, the circles show the data points outside of this interval, the solid line shows the median, and the dashed line shows the mean. Note that the SRT panels are scaled such that better performance is at top.

Table 1. Results of split-plot RM ANOVAs performed for each test in Fig. 1, with signal processing as the within-subject factor and group as the between subject factor. The asterisks indicate significant effects.

	Signal processing		Group		Signal processing × group	
Test	$F(1,19)$	p	$F(1,19)$	p	$F(1,19)$	p
MCI	84.8	<0.001*	7.8	0.012*	3.0	0.099
Emotion	81.1	<0.001*	2.7	0.117	2.3	0.146
NVA 10 dB	335.5	<0.001*	0.3	0.584	0.3	0.584
NVA 5 dB	227.2	<0.001*	3.1	0.090	0.5	0.486
NVA 0 dB	352.2	<0.001*	1.0	0.340	0.3	0.614
SRT—Male	947.7	<0.001*	1.4	0.256	2.7	0.115
SRT—Female	498.7	<0.001*	3.6	0.074	0.8	0.396

than with unprocessed signals. In general, mean performance was better for musicians than non-musicians, although differences were sometimes small. A split-plot Repeated Measures Analysis of Variance (RM ANOVA) was performed for each of the tests shown in Fig. 1, with signal processing (unprocessed, degraded) as the within-subject factor and group (musician, non-musician) as the between-subject factor; results are shown in Table 1. For all tests, performance was significantly better with unprocessed signals. There was a significant effect of participant group only for MCI.

The present adolescent data were compared to adult data from Fuller *et al.*[5] Figure 2 shows boxplots of MCI and vocal emotion identification scores for adolescent and adult musicians and non-musicians. Mean performance was generally better for adults than adolescents both with unprocessed or degraded signals. Because the distribution of data was sometimes not normal, non-parametric Mann-Whitney tests were used to compare age effects in musicians and non-musicians; the results are shown in Table 2. Significant age effects were observed for MCI performance in musicians listening to unprocessed or degraded signals; there were no significant age effects for MCI performance for non-musicians. Significant age effects were observed for vocal emotion identification in musicians listening to unprocessed signals, and in non-musicians listening to unprocessed or degraded signals.

4. Discussion

Similar to the adult data from Fuller *et al.*,[5] a significant musician effect was observed for adolescents for MCI (within-domain effect). In both studies, the musician effect for MCI was robust, and persisted even when the signal was spectro-temporally degraded. There were no other significant musician effects for adolescents for vocal emotion, word identification in noise, or sentence identification in competing speech (no cross-domain effects). Although the musician effect was not significant for vocal emotion identification when the signal was degraded, mean performance was 8.3 points better for musicians than non-musicians, slightly larger than the mean musician advantage for MCI (7.3 points). Though not significant, similar musician advantages in mean performance were observed for SRTs with unprocessed (male interferer; −2.0 dB) and degraded signals (female interferer; −1.5 dB). This is somewhat in agreement with the musician effect observed by Başkent and Gaudrain[3] for adults listening to unprocessed speech, suggesting that musicians may be better able to take advantage of voice pitch and spectral envelope cues to segregate competing speech.

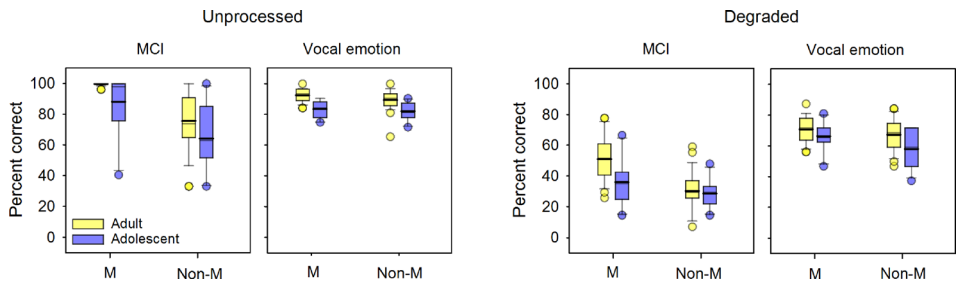

Fig. 2. (Color online) Boxplots for MCI and vocal emotion identification with unprocessed (left two panels) or degraded signals (right two panels), for adolescent and adult musicians (M) and non-musicians (Non-M); the adult data from Fuller *et al.* (Ref. 5).

Table 2. Results of Mann-Whitney tests comparing age effects for the data shown in Fig. 2; adult data are from Fuller *et al.* (Ref. 5). The asterisks indicate significant effects.

	Unprocessed				Degraded			
	MCI		Emotion		MCI		Emotion	
	U	p	U	p	U	p	U	p
Musician	75.0	0.006*	23.0	<0.001*	53.5	0.009*	87.0	0.168
Non-musician	88.0	0.091	41.5	<0.001*	122.5	0.615	79.5	0.047*

Significant age effects were observed between the present adolescents and the adults from Fuller *et al.*[5] (Fig. 2 and Table 2). This suggests that the ability to use pitch and/or spectral envelope cues for challenging tasks continues to mature beyond adolescence. Alternatively, the listening tasks may have been more cognitively demanding for adolescents than for adults. Complex tasks such as melodic pitch perception, vocal emotion identification, and segregation of competing talkers may fully develop later in life.[19] Interestingly, MCI performance with unprocessed or degraded signals was similar between adolescent musicians and adult non-musicians, suggesting that music training may partly compensate for developmental differences between adults and adolescents. Among musicians, mean MCI performance with unprocessed signals was better and variance was reduced for adults compared to adolescents. With unprocessed signals, vocal emotion identification was generally better for adult non-musicians than for adolescent musicians, suggesting that music training alone may not offset developmental differences for some listening tasks. For the remaining tasks and conditions, mean performance was generally better for adults, but the variability in performance was comparable between adult and adolescent listeners.

Although the data with adolescents showed a significant musician effect only for music perception, there were some instances musician performance appeared to be generally better than non-musician performance (e.g., vocal emotion and word identification with degraded stimuli, SRTs with unprocessed speech). Note that there were a smaller number of participants compared to similar previous studies with adults,[3,5] due to the limited availability of the local adolescent population (which could not be tested at the school sites, different from young children). A greater number of participants would increase confidence in the present pattern of results, and might even demonstrate further musician effect in adolescents. Extended audiometric threshold measures out to 16 kHz might also be necessary to identify participants with potential hidden hearing loss, which may affect speech understanding in noise. Still, the present results suggest within-domain and potential cross-domain effects for music training in adolescents that appear to persist even when signals are spectro-temporally degraded.

Note that in this study, normal-hearing listeners were exposed to spectro-temporal degradation for a relatively short time during testing. In contrast, CI users have a much longer experience with spectro-temporally degraded signals, and perhaps a greater motivation to make use of the available cues to understand speech and music. Previous studies have shown significant benefits for music training in pediatric CI users.[13,14,20] Along with the present data, this suggests that music training may be a valuable addition to habilitation of pediatric CI users.

Acknowledgments

We thank Dr. Qian-jie Fu and the Emily Shannon Fu Foundation for providing the experimental software. We also thank three anonymous reviewers for helpful comments. This study was supported by VIDI Grant No. 016.096.397 and VICI Grant No. 918.17.603 from the Netherlands Organization for Scientific Research (NWO) and the Netherlands Organization for Health Research and Development (ZonMw), funds from Heinsius Houbolt Foundation, and the Rosalind Franklin Fellowship from University Medical Center Groningen. E.G.'s contribution was conducted in the framework of the LabEx CeLyA ("Centre Lyonnais d'Acoustique," ANR-10-LABX-0060/ANR-11-IDEX-0007) operated by the French National Research Agency. This research is part of the Healthy Aging and Communication research program of the Otorhinolaryngology Department of University Medical Center Groningen.

References and links

[1]C. Micheyl, K. Delhommeau, X. Perrot, and A. J. Oxenham, "Influence of musical and psychoacoustical training on pitch discrimination," Hear. Res. **219**, 36–47 (2006).

[2]A. Parbery-Clark, E. Skoe, C. Lam, and N. Kraus, "Musician enhancement for speech-in-noise," Ear Hear. **30**, 653–661 (2009).
[3]D. Başkent and E. Gaudrain, "Musician advantage for speech-on-speech perception," J. Acoust. Soc. Am. **139**, EL51–EL56 (2016).
[4]J. Swaminathan, C. Mason, T. Streeter, V. Best, G. Kidd, Jr., and A. Patel, "Musical training, individual differences and the cocktail party problem," Sci. Rep. **5**, 11628 (2014).
[5]C. D. Fuller, J. J. Galvin, B. Maat, R. H. Free, and D. Başkent, "The musician effect: Does it persist under degraded pitch conditions of cochlear implant simulations?," Front. Neurosci. **8**, 179 (2014).
[6]M. Chatterjee and S. Peng, "Processing F0 with cochlear implants: Modulation frequency discrimination and speech intonation recognition," Hear. Res. **235**, 143–156 (2008).
[7]C. D. Fuller, E. Gaudrain, J. N. Clarke, J. J. Galvin, Q.-J. Fu, R. Free, and D. Başkent, "Gender categorization is abnormal in cochlear implant users," J. Assoc. Res. Otolaryngol. **15**, 1037–1048 (2014).
[8]J. J. Galvin III, Q. J. Fu, and G. Nogaki, "Melodic contour identification by cochlear implant listeners," Ear Hear. **28**, 302–319 (2007).
[9]X. Luo, Q. J. Fu, and J. J. Galvin III, "Vocal emotion recognition by normal-hearing listeners and cochlear implant users," Trends Amplif. **11**, 301–315 (2007).
[10]S. Goorhuis-Brouwer and A. Schaerlaekens, *Handbook of Language Development, Language Pathology, and Language Therapy in Dutch-Speaking Children*, 2nd ed. (De Tijdstroom, Utrecht, The Netherlands, 2000).
[11]M. L. D. Deroche, A. M. Kulkarni, J. A. Christensen, C. J. Limb, and M. Chatterjee, "Deficits in the sensitivity to pitch sweeps by school-aged children wearing cochlear implants," Front. Neurosci. **10**, 73 (2016).
[12]C. Magne, D. Schön, and M. Besson, "Musician children detect pitch violations in both music and language better than nonmusician children: Behavioral and electrophysiological approaches," J. Cognitive Neurosci. **18**, 199–211 (2006).
[13]J. Chen, A. Y. C. Chuang, C. McMahon, J.-C. Hsieh, T.-H. Tung, and L. Li, "Music training improves pitch perception in prelingually deafened children with cochlear implants," Pediatrics **125**, e793–e800 (2010).
[14]A. Good, K. A. Gordon, B. C. Papsin, G. Nespoli, T. Hopyan, I. Peretz, and F. A. Russo, "Benefits of music training for perception of emotional speech prosody in deaf children with cochlear implants," Ear Hear. **38**, 455–464 (2017).
[15]M. Goudbeek and M. Broersma, "Language specific effects of emotion on phoneme duration," in *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Makuhari, Japan (2010), pp. 2026–2029.
[16]A. Bosman and G. Smoorenburg, "Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment," Int. J. Audiol. **34**, 260–284 (1995).
[17]R. Plomp and A. M. Mimpen, "Speech-reception threshold for sentences as a function of age and noise level," J. Acoust. Soc. Am. **66**, 1333–1342 (1979).
[18]D. D. Greenwood, "A cochlear frequency-position function for several species—29 years later," J. Acoust. Soc. Am. **87**, 2592–2605 (1990).
[19]L. J. Leibold, "Speech perception in complex acoustic environments: Developmental effects," J. Speech Lang. Hear. Res. **60**, 3001–3008 (2017).
[20]K. Gfeller, "Music-based training for pediatric CI recipients: A systematic analysis of published studies," Eur. Annals Otorhinol. Head Neck Dis. **133**, S50–S56 (2016).
[21]iStar: www.tigerspeech.com/istar (Last viewed April 23, 2018).
[22]AngelSound: angelsound.tigerspeech.com (Last viewed April 23, 2018).