Deniz Başkent and Etienne Gaudrain: JASA Express Letters [http://dx.doi.org/10.1121/1.4942628] Published Online 2 March 2016

Musician advantage for speech-on-speech perception

Deniz Başkent[a] **and Etienne Gaudrain**[b]
Department of Otorhinolaryngology/Head and Neck Surgery, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
d.baskent@umcg.nl, etienne.gaudrain@cnrs.fr

Abstract: Evidence for transfer of musical training to better perception of speech in noise has been mixed. Unlike speech-in-noise, speech-on-speech perception utilizes many of the skills that musical training improves, such as better pitch perception and stream segregation, as well as use of higher-level auditory cognitive functions, such as attention. Indeed, despite the few non-musicians who performed as well as musicians, on a group level, there was a strong musician benefit for speech perception in a speech masker. This benefit does not seem to result from better voice processing and could instead be related to better stream segregation or enhanced cognitive functions.

[DOS]
Date Received: July 3, 2015 **Date Accepted:** January 20, 2016

1. Introduction

Musicians perform better than non-musicians in a wide range of auditory perceptual tasks. This musician advantage may be due to better processing of acoustic features, such as pitch or fundamental frequency (F0), which is a primary dimension in music (Micheyl *et al.,* 2006) or better stream segregation (Zendel and Alain, 2013). Alternatively, or perhaps in addition, the advantage could also be due to enhanced auditory cognitive abilities, such as better attention or extended working memory capacity, at least in the auditory modality (Strait *et al.,* 2010; Carey *et al.,* 2015).

Recently, it was proposed that the enhanced auditory skills developed through musical training could transfer to better perception of speech in noise, as the neural structures involved in music and speech processing seem to partially overlap (Besson *et al.,* 2011; Miendlarzewska and Trost, 2014). However, evidence for such a transfer has been mixed, and when observed, the effect has been minimal: studies measuring speech perception in steady or multi-talker babble noise showed either no, or only small, musician advantage (Parbery-Clark *et al.,* 2009; Fuller *et al.,* 2014b; Ruggles *et al.,* 2014). Perception of speech in such background noises is not a situation recognized as relying on fine F0 processing. Moreover, it is generally considered primarily driven by energetic masking, the release from which is thought to be less dependent on cognitive abilities than informational masking. Perception of speech masked by background speech, namely, speech-on-speech perception, on the other hand, has been shown to directly depend on F0 differences between the two competing voices (Darwin *et al.,* 2003) and involves informational rather than energetic masking (Gaudrain and Carlyon, 2013), which mobilizes more cognitive resources (Zekveld *et al.,* 2013). Speech-on-speech could therefore be a more suitable test condition than speech-in-noise to investigate the potential musician advantage for speech perception. However, research on this has also produced mixed results: while one study showed stronger musician advantage for conditions with more informational masking (Swaminathan *et al.,* 2014), another showed no such advantage for speech on speech masking (Boebinger *et al.,* 2015).

Here we have aimed to provide a more definitive answer to whether speech perception in background speech masker would show a strong musician advantage. Better perception of pitch, stream segregation, and increased attention to subtle acoustic cues should all play a prominent role in separating and perceiving a target speech from a single interfering talker, more so than speech in steady noise or multi-talker babble. Further, to investigate what underlying factors may contribute to such

[a]Author to whom correspondence should be addressed. Also at: Graduate School of Medical Sciences, Research School of Behavioral and Cognitive Neurosciences, University of Groningen, Groningen, The Netherlands.
[b]Also at: Lyon Neuroscience Research Center, Auditory Cognition and Psychoacoustics, CNRS UMR 5292, Inserm U1028, Université Lyon 1, Lyon, France.

advantage, if it exists, we manipulated the competing voices of the target and masker speech such that they differed in their vocal characteristics. F0, but also vocal tract length (VTL), were systematically varied to differentiate the voice of the target sentence from that of the masker sentence. This manipulation was done based on a previous observation: unlike F0, VTL has been found to be less utilized by musicians than by non-musicians for sine-wave vocoded voice gender categorization (Fuller *et al.*, 2014a). Hence if the musician advantage is primarily based on better perception of F0 cues, then musicians should show more improvement in intelligibility in conditions where the voices differed in their F0s.

2. Methods

2.1 Participants

A total of 38 participants, comprising 18 musicians and 20 non-musicians, participated in the study. Details of education level, sex and age for each group are given in Table 1. Selection criteria for musician group were identical to Fuller *et al.* (2014b): (1) having had 10 or more years of musical training, (2) having begun musical training before or at the age of 7, and (3) having received musical training within the last 3 yr prior to the study. Non-musician criteria were not meeting the musician criteria as well as not having received musical training within the past 7 yr prior to the study. The non-musicians were further selected to roughly match the age, sex and education level of the musician group. All participants, musician or non-musician, were native speakers of Dutch with no known neurological disorders and had audiometric thresholds $\leq$20 dB hearing level (HL) at audiometric test frequencies between 500 Hz and 6 kHz.

The groups did not differ significantly in age [Welch t-test: $t(35.6) = 1.21$, $p = 0.24$] or sex (Fisher's exact test: $p = 0.50$). However, they did differ in education level (Fisher's exact test: $p = 0.027$). This difference seems mostly due to the fact that there are four Master of Science (MSc) students amongst the musicians, but none amongst the non-musicians; and there are four vocational school ("Middelbaar beroepsonderwijs"—MBO) students amongst the non-musicians, but none amongst the musicians. When considering only the higher level vocational school ("Hoger beroepsonderwijs"—HBO) and Bachelor of Science (BSc) students, there is no significant difference between the two groups anymore (Fisher's exact test: $p = 0.47$).

The study was approved by the Medical Ethical Committee of the University Medical Center Groningen. Before starting the experiment written consent was obtained from the participants. All participants received financial compensation for their participation.

2.2 Stimuli

The stimuli consisted of target sentences presented concurrently with masker sentence sequences at the target to masker ratio (TMR) of −6 dB with the mixture fixed at a presentation level of 65 dB sound pressure level (SPL). All sentences were taken from the lists of digital recordings of grammatically simple, meaningful sentences (13

Table 1. Contingency table for education level and sex, and descriptive statistics for age, shown for each group. MBO, vocational school ("Middelbaar beroepsonderwijs"); HBO, higher level vocational school ("Hoger beroepsonderwijs"); BSc, Bachelor of Science; MSc, Master of Science.

Education level (n)	Non-musician	Musician
MBO	4	0
HBO	6	8
BSc	9	6
MSc	0	4
Other	1	0
Sex (n)	**Non-musician**	**Musician**
Female	14	10
Male	6	8
Age (years)	**Non-musician**	**Musician**
Min	19	19
Max	27	25
Mean	22.75	21.89
Standard deviation	2.43	1.97
Median	23	21.5

sentence per list), spoken by a male speaker (see Versfeld *et al.,* 2000, for details). The sentences contained 4–9 words each with an average of 6.1 words. The target sentences used for training were taken from lists 1–2 and for data collection from lists 3–11 (for the nine test conditions), and the masker sentences were taken from lists 27–31.

The masker sentence sequences were created by randomly selecting a sentence from masker sentences and extracting excerpts of random durations (a minimum of 1 s duration, starting from the end of the sentence). No attention was paid to retaining first words or special sections of the masker sentences. This was done to somewhat reduce the meaning of the masker sentences, as the main emphasis of the present study was on the voice manipulations, but also to ensure that the masker sequence was never the same from one target stimulus to the other. The masker always started 500 ms before the target sentence and was ramped for the entire duration of 500 ms. If the duration of masker sentence sequence was shorter than the 500 ms + target sentence duration, a second (or third, if necessary) masker sentence was added to the masker sentence sequence following the same procedure. The masker sentence sequence was such that it ended either as the same time as the target sentence, or up to a maximum of 500 ms after the target sentence offset.

The difference in the voices of the target and masker sentences was created by manipulating the voice characteristics F0 and VTL of the masker sentence using the STRAIGHT software implemented in MATLAB (developed by Kawahara *et al.,* 1999) in a manner similar to Fuller *et al.* (2014a). F0 was shifted up by 0, 4, and 8 semitones, and VTL was shifted down by 0, 0.75, and 1.5 semitones, resulting in nine different voice conditions. The test order of the conditions was random; however, the same list of target sentences was used for the same condition from one participant to the other.

2.3 Procedure

Stimuli were presented via a MATLAB GUI. Participants listened to stimuli diotically over headphones in a single-wall sound-treated booth and repeated the words they identified from the target sentence. The participants set the pace of testing by pressing the space bar for the next sentence presentation. An experimenter scored the correctly identified words during testing, and a later offline check was made again using digital recordings of participants' verbal responses. In this scoring, all words were counted with equal importance and without regard to the order.

The entire procedure consisted of a short training where participants passively listened to eight target sentences in quiet to become familiar with the target voice and another training where participants practiced the nine test conditions. During the training phase, visual feedback was provided after each sample by presenting the target sentence text on their screen. Data collection was similar to the second training except no feedback was provided. Data collection lasted around 20 min, and the entire procedure was completed in a single session.

3. Results

Figure 1 shows the intelligibility performance, calculated in percentage of correctly identified number of words to the total number of words presented in all the target sentences, for each condition and each group. In each and all conditions, the average intelligibility score of the musicians was greater than that of the non-musicians. On average, the musicians had intelligibility scores 11 percentage points higher than the non-musicians.

To analyze the intelligibility scores, a generalized linear mixed model (gLMM) was fitted to the binary (correct/incorrect) scores for individual words, following a model selection procedure described by Jaeger (2008). The models were implemented in R using the LME4 package (Bates *et al.,* 2013), using a *logit* link function. The initial step was the full model with ΔF0 (noted f0), ΔVTL (vtl), group (grp), and all possible interactions as fixed factors, and allowing a random intercept per subject, i.e., in lme4 syntax: `score ~ f0 * vtl * grp+(1|subject)`. From this model, the factor with the least significant Wald statistic was removed, and the new model was compared to the initial one using a chi-square test based on the log-likelihood difference. If the new model was found not to be significantly different from the initial one, it was kept. The procedure was repeated, finally resulting in the following model, where all interactions were removed: `score ~ f0+vtl+grp+(1|subject)`.

The fitted model corresponds to the following equation:

$$\mathrm{logit}(score) = 1.17\frac{f_0}{8} + 0.51\frac{vtl}{1.5} + 0.63\,grp, \tag{1}$$

where *grp* is 0 for non-musicians and 1 for musicians and f_0 and *vtl* are expressed in semitones.

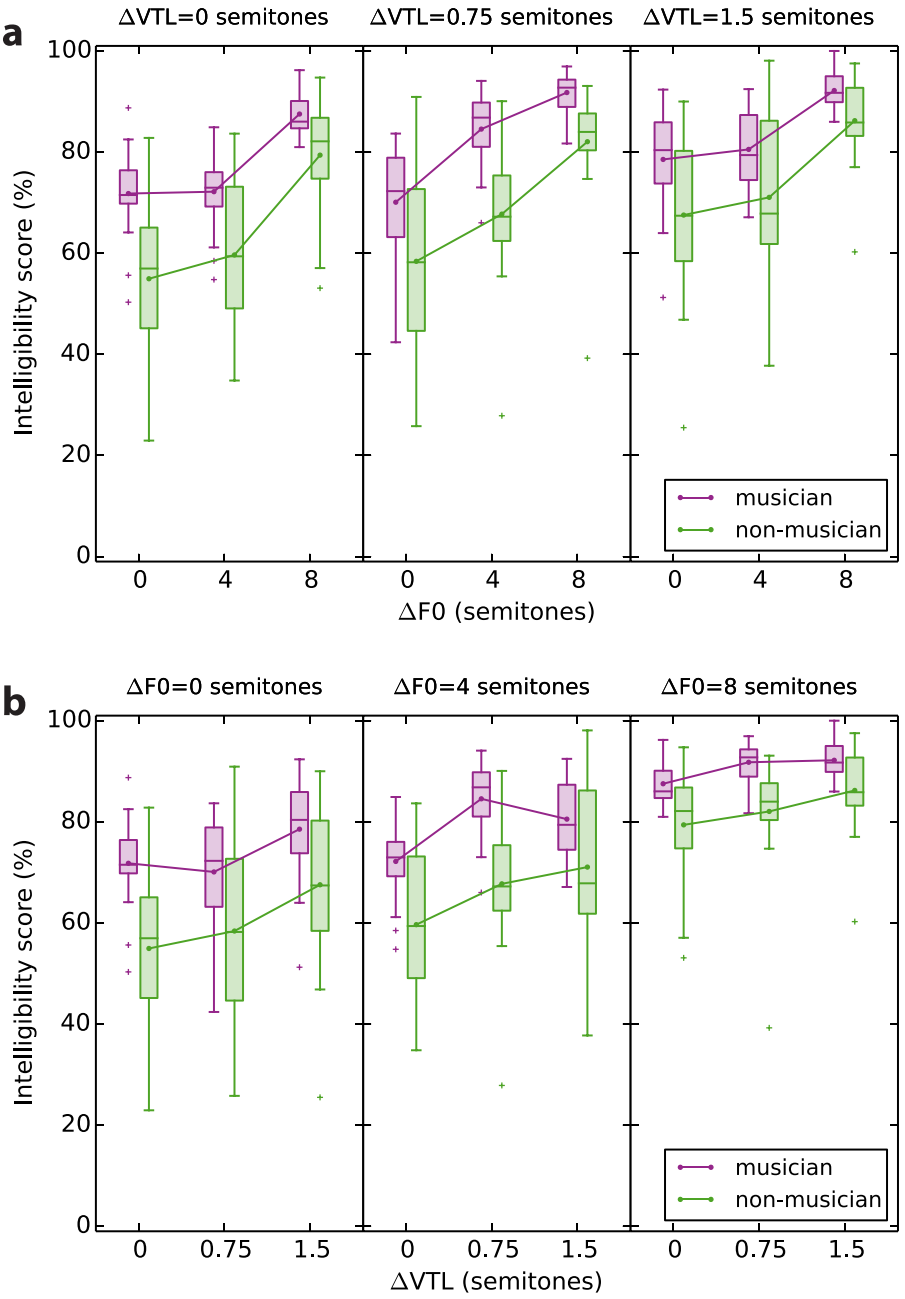

Fig. 1. (Color online) Top row: Each panel shows the intelligibility performance as a function of the average F0 difference between the target and masker voices and for musicians and non-musicians. The three panels each correspond to a given VTL difference between target and masker. The connected line shows the mean for each condition and group. The boxes extend from the lower to the upper quartile, and the middle line shows the median. The whiskers show the range of the data within 1.5 times the inner quartile. The + symbols show the individual data outside of 1.5 times the inner quartile range. Bottom row: The same data is displayed as a function of VTL difference while the panels show results for various F0 differences.

Based on the Wald statistics of this final model, both increasing ΔF0 ($z = 44.15$, $p < 0.001$) and ΔVTL ($z = 19.82$, $p < 0.001$) significantly improved intelligibility. Most importantly, musicians obtained significantly higher scores than non-musicians in all conditions ($z = 4.07$, $p < 0.001$). The gLMM fitting indicates that on average the odds ratio was nearly twice as large for musicians than for non-musicians.

To rule out that the difference in education level between the two groups could be responsible for the musician advantage, the same model was fitted to a subset including only participants with HBO or BSc education level. The fit obtained with this education-matched subset is very similar to the one obtained with the full data: ΔF0 ($z = 38.44$, $p < 0.001$) and ΔVTL ($z = 15.08$, $p < 0.001$) both significantly improved performance, and musicians performed better than non-musicians ($z = 3.21$, $p = 0.001$).

4. Discussion

As was observed before (Darwin *et al.,* 2003), F0 and VTL differences between the competing sentences both contributed to improved intelligibility. The musicians showed overall better intelligibility than non-musicians, confirming a musician advantage for speech-on-speech perception. In contrast to many of the previous studies (Parbery-Clark *et al.,* 2009; Fuller *et al.,* 2014b; Ruggles *et al.,* 2014; Boebinger *et al.,* 2015), but in line with one study on informational masking (Swaminathan *et al.,* 2015), this effect was strong and robust across all voice conditions. Hence the results confirmed that indeed the musician advantage for better speech perception may strongly depend on a specific task where the task relies on skills that are improved by musical training (also in line with Fuller *et al.,* 2014b), such as better pitch perception or stream segregation.

However, the musician advantage did not seem to change as a function of the voice difference. The two groups drew equivalent benefit from F0 differences between concurrent voices [when fitting Eq. (1) individually for each group, the coefficients found for F0 were 1.18 and 1.17 for musicians and non-musicians, respectively]. Musicians seemed to benefit slightly less from VTL differences (VTL coefficient: 0.46) than non-musicians (VTL coefficient: 0.55), but this difference was not significant. This observation is also in line with Fuller *et al.* (2014a), who observed less reliance on VTL voice cues by musicians than non-musicians for speaker gender categorization when the stimuli were sine-wave vocoded but not when they were intact.

What was most surprising in these results was that most of the musician advantage was shown in a condition where there was no difference in average vocal characteristics of the target and the masker. This observation hints that the musician advantage may not be related to the processing of these vocal characteristics *per se*. On the other hand, even when the average F0s were identical, because the prosodic F0 contours of the concurrent voices differed, there would still be a difference in instantaneous F0. The musician advantage could thus also be derived from an enhanced ability to process and disentangle fast changing F0 differences.

Overall, perhaps the strong speech-on-speech perception advantage observed with musicians is not a direct result of better pitch perception, but instead more associated with other factors related to auditory perception, such as better stream segregation, better rhythm perception, or even better auditory cognitive abilities (Zendel and Alain, 2013; Miendlarzewska and Trost, 2014; Carey *et al.,* 2015).

Acknowledgments

This study was conducted as part of Bachelor's and Master's theses by Nikki Tahapary and Michael Chesnaye, respectively, and was supported by a VIDI Grant No. 016.096.397 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. The study is part of the Healthy Aging and Communication research program.

References and links

Bates, D., Maechler, M., Bolker, B., and Walker, S. (**2013**). "lme4: Linear mixed-effects models using Eigen and S4," R package version 1.

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

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., and Scott, S. K. (**2015**). "Musicians and non-musicians are equally adept at perceiving masked speech," J. Acoust. Soc. Am. **137**, 378–387.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., and Dick, F. (**2015**). "Generality and specificity in the effects of musical expertise on perception and cognition," Cognition **137**, 81–105.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Fuller, C. D., Galvin, J. J., III, Free, R. H., and Başkent, D. (**2014a**). "Musician effect in cochlear implant simulated gender categorization," J. Acoust. Soc. Am. **135**, EL159–EL165.

Fuller, C. D., Galvin, J. J., III, Maat, B., Free, R. H., and Başkent, D. (**2014b**). "The musician effect: Does it persist under degraded pitch conditions of cochlear implant simulations?," Front. Neurosci. **8**, 179.

Gaudrain, E., and Carlyon, R. P. (**2013**). "Using Zebra-speech to study sequential and simultaneous speech segregation in a cochlear-implant simulation," J. Acoust. Soc. Am. **133**, 502–518.

Jaeger, T. F. (**2008**). "Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models," J. Mem. Lang. **59**, 434–446.

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. **27**, 187–207.

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

Miendlarzewska, E. A., and Trost, W. J. (**2014**). "How musical training affects cognitive development: Rhythm, reward and other modulating variables," Front. Neurosci. **7**, 279.

Parbery-Clark, A., Skoe, E., Lam, C., and Kraus, N. (**2009**). "Musician enhancement for speech-in-noise," Ear Hear. **30**, 653–661.

Ruggles, D. R., Freyman, R. L., and Oxenham, A. J. (**2014**). "Influence of musical training on understanding voiced and whispered speech in noise," PLoS One **9**, e86980.

Strait, D. L., Kraus, N., Parbery-Clark, A., and Ashley, R. (**2010**). "Musical experience shapes top-down auditory mechanisms: Evidence from masking and auditory attention performance," Hear. Res. **261**, 22–29.

Swaminathan, J., Mason, C., Streeter, T., Best, V., Kidd, Jr., G., and Patel, A. (**2015**). "Musical training, individual differences and the cocktail party problem," Sci. Rep. **5**, 11628–11628.

Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671–1684.

Zekveld, A. A., Rudner, M., Johnsrude, I. S., and Rönnberg, J. (**2013**). "The effects of working memory capacity and semantic cues on the intelligibility of speech in noise," J. Acoust. Soc. Am. **134**, 2225–2234.

Zendel, B. R., and Alain, C. (**2013**). "The influence of lifelong musicianship on neurophysiological measures of concurrent sound segregation," J. Cognit. Neurosci. **25**, 503–516.