Differences in a Musician's Advantage for Speech-in-Speech Perception Based on Age and Task

Article *in* Journal of Speech Language and Hearing Research · February 2023

DOI: 10.1044/2022_JSLHR-22-00259

3 authors:

Michelle Cohn
University of California, Davis
40 PUBLICATIONS **288** CITATIONS
SEE PROFILE

Santiago Barreda
University of California, Davis
56 PUBLICATIONS **1,610** CITATIONS
SEE PROFILE

Georgia Zellou
University of California, Davis
81 PUBLICATIONS **634** CITATIONS
SEE PROFILE

Accepted to *Journal of Speech, Language, and Hearing Research.*
https://doi.org/10.1044/2022_jslhr-22-00259

Differences in a musician's advantage for speech-in-speech perception based on age and task

Michelle Cohn
mdcohn@ucdavis.edu
Phonetics Lab, Linguistics Department
University of California, Davis

Santiago Barreda
sbarreda@ucdavis.edu
Phonetics Lab, Linguistics Department
University of California, Davis

Georgia Zellou
gzellou@ucdavis.edu
Phonetics Lab, Linguistics Department
University of California, Davis

Abstract

Purpose: This study investigates the debate that musicians have an advantage in speech-in-noise perception from years of targeted auditory training. We also consider the effect of age on any such advantage, comparing musicians and non-musicians (age range: 18-66), all of whom had normal hearing. We manipulate the degree of fundamental frequency (f0) separation between the competing talkers, as well as use different tasks, to probe attentional differences that might shape a musician's advantage across ages.

Method: Participants (ranging in age from 18-66) included 29 musicians and 26 non-musicians. They completed two tasks varying in attentional demands: 1) a selective attention task where listeners identify the target sentence presented with a 1-talker interferer (Experiment 1), and 2) a divided attention task where listeners hear two vowels played simultaneously and identify both competing vowels (Experiment 2). In both paradigms, f0 separation was manipulated between the two voices (Δf0 = 0, 0.156, 0.306, 1, 2, 3 semitones (ST)).

Results: Results show that increasing differences in f0 separation lead to higher accuracy on both tasks. Additionally, we find evidence for a musician's advantage across the two studies. In the sentence identification task, younger adult musicians show higher accuracy overall, as well as a stronger reliance on f0 separation. Yet, this advantage declines with musicians' age. In the double vowel identification task, musicians of all ages show an across-the-board advantage in detecting two vowels — and use f0 separation more to aid in stream separation — but show no consistent difference in double vowel identification.

Conclusions: Overall, we find support for a hybrid *auditory encoding-attention account* of music-to-speech transfer: the musician's advantage includes f0, but the benefit also depends on the attentional demands in the task and listeners' age. Taken together, this study suggests a complex relationship between age, musical experience, and speech-in-speech paradigm on a musician's advantage.

1. Introduction

In everyday life, listeners often contend with sources of competing background noise to hear their interlocutor, known as speech-in-noise perception. A common, challenging listening scenario is trying to comprehend a talker when there are other overlapping speech signals, or speech-in-speech perception (e.g., listening to a friend in a crowded restaurant). Yet, for young adults with normal-hearing, the auditory system is surprisingly robust to environmental perturbations (Assmann & Summerfield, 2004). For example, listeners can use small differences in fundamental frequency (f0; Assmann & Summerfield, 1990; Bregman, 1990; Summers & Leek, 1998), onset timing (Lee & Humes, 2012), and vowel spectral peaks (Assmann & Summerfield, 1989) to separate competing speech signals.

A growing body of work has examined the extent to which musical training might drive changes in auditory perception (Bidelman & Yoo, 2020; Kraus & Chandrasekaran, 2010; Münte et al., 2002; Strait & Kraus, 2014). However, the search for a musician's advantage in speech-in-noise has produced decidedly mixed results in prior work (for a review, see Coffey et al., 2017). On the one hand, a cohort of studies have found a musicianship advantage for perceiving speech-in-speech, when a target talker's productions are obscured by one or more other talkers (Başkent & Gaudrain, 2016; Clayton et al., 2016; Kaplan et al., 2021; Morse-Fortier et al., 2017; Parbery-Clark, Skoe, Lam, et al., 2009; Parbery-Clark et al., 2011; Slater & Kraus, 2016; Zendel et al., 2015; Zendel & Alain, 2012). For example, musicians show higher accuracy in recognizing words embedded in 4-talker babble (e.g., Parbery-Clark, Skoe, Lam, et al., 2009). Yet, other studies, sometimes even using identical paradigms, have shown no difference between musicians and non-musicians (e.g., Anaya et al., 2016; Başkent et al., 2018; Boebinger et al., 2015; Couth et al., 2020; Deroche et al., 2017; Madsen et al., 2017; Mandikal Vasuki et al., 2016; Mussoi, 2021; Ruggles et al., 2014; Yeend et al., 2017). That we see differences across studies suggests that the musician's advantage may be relatively small and overwhelmed by between-listener and task-related variation (see Supplementary Data Table S1 for overview). Here, we will discuss possible sources of variation in any musician's advantage in the perception of speech-in-speech, both within- and between-listeners.

1.1. Age-related variation

Age-related changes are perhaps one of the largest contributors to between-speaker variation in speech-in-speech perception. Older adults with hearing loss face additional challenges in speech-in-speech perception (Arehart et al., 1997; Dubno et al., 1984; Helfer & Wilber, 1990; Lee & Humes, 2012; Lentz & Marsh, 2006; for a review, see Helfer et al., 2017). Even older adults with normal-hearing show increased difficulties perceiving a talker in the presence of a background talker, an effect attributed to age-related declines in centralized auditory processing, attention, and working memory (Heidari et al., 2020; Helfer & Freyman, 2014; for a review, see Akeroyd, 2008). For example, older adults (ages 67-81) show greater interference by linguistically meaningful maskers than younger adults (ages 17-19), suggesting that speech-in-speech difficulties might be attributed to possible declines in auditory inhibition (Tun et al., 2002). Age-related declines in speech perception also start to emerge in middle adulthood (Bergman et al., 1976; Helfer, 2015; Helfer & Jesse, 2021). For example, Başkent and colleagues (2014) found worse speech reception thresholds for adults ages 51-63 (all of whom had normal audiometric thresholds) than younger adults (ages 19-26) when listening to sentences with a competing talker. Accordingly, there is much interest in determining what types of experience might improve speech-in-speech perception across age, such as via musical training.

There have been some comparisons of musicians and non-musicians for specific age groups that suggest age-related factors in an advantage. For example, Başkent and colleagues (2018) tested adolescents (ages 11-14), and found no difference between the groups in perceiving sentences with a 1-talker interferer. Yet, using the identical paradigm, Başkent & Gaudrain (2016) found a musician's advantage in younger adults (ages 19-27), suggesting that the advantage might emerge with development.

Indeed, there is other support for a younger adult (YA) musician's advantage: Bidelman and Yoo (2020) found that YA musicians (ages 19-33) showed higher accuracy in recognizing a target sentence amidst an increasing number of competing talkers (or 'maskers'). Others have provided some evidence for a later-emerging advantage. Comparing adults across a wide age range (from ages 19-91), Zendel and Alain (2012) observed that non-musicians have a steeper decline in keyword perception in 4-talker babble with increasing age, relative to musicians. For younger adult listeners, on the other hand, musicians' and non-musicians' thresholds appear to be largely overlapping until after age 40 (see Zendel & Alain (2012) Figure 4, p. 415). Similarly, Tierney and colleagues (2020) found less of an age-related decline for musicians (ranging from ages 18-66) in perceiving a target sentence amidst a 1-talker interferer. Together, these findings suggest that a musician's advantage for speech-in-speech perception might not emerge until young adulthood or middle age, possibly due to cumulative years of musical experience.

1.2. The role of task on a musician's advantage

Most studies testing a musician's advantage for speech-in-speech perception examine a single type of task (e.g., sentence or words in multitalker babble), and the task most commonly tests selective attention, wherein listeners hone in on one target speaker while ignoring competing talker(s) (e.g., Boebinger et al., 2015; Parbery-Clark et al., 2011; Zendel & Alain, 2012). This requires that listeners are able to 1) separate the talkers, and 2) direct their attention to the target while inhibiting interfering speech. Indeed, a growing body of work has shown that musicians often show enhanced selective auditory attention (e.g., Medina & Barraza, 2019; Strait & Kraus, 2011; Zendel & Alain, 2014), which might underlie their improvements in speech-in-speech perception.

While less studied than selective attention, it might also be illuminating to compare musicians and non-musicians in tasks where attention is divided, such as when listeners are asked to recognize information from multiple speech streams simultaneously. For example, in double vowel paradigms, listeners hear two vowels simultaneously and are asked to identify both vowels they heard. Moreover, such a divided attention task might be particularly relevant for detecting a musician's advantage. Double vowel perception has been shown to be especially difficult for older adults (Vongpaisal & Pichora-Fuller, 2007), thought to be due to age-related difficulties in attending to multiple sources of incoming information at once. Meister et al. (2013) directly compared selective and divided attention by younger (ages 18-27) and middle-age/older (ages 58-79) listeners: in one task, participants were asked to repeat words from a target talker (selective), while in another task they were asked to repeat words from two talkers (divided). They found no difference by age in the selective attention task, but a sizable decrease for older listeners in the divided attention task. Therefore, we might be better able to detect differences in a musician's advantage in the current study in tasks that require divided attention.

1.3. Role of f0 difference between voices in a musician's advantage

In addition to age and task, the properties of the target and competing voice(s) themselves might play a role in a musician's advantage. As mentioned, listeners use f0 separation between voices to tease them apart, a critical first step in speech-in-speech perception (auditory stream separation, Bregman, 1990). F0 is related to the psychoacoustic perception of pitch: as f0 increases, listeners perceive an increase in pitch. Musicians, in particular, receive specific instruction, feedback, and training related to the accurate perception and discrimination of pitch (Schlaug, 2011), and musicians have higher perceptual acuity in perceiving small differences in f0 than non-musicians (Bianchi et al., 2016; Kishon-Rabin et al., 2001; Micheyl et al., 2006). While prior studies examining a musician's advantage compare talkers (e.g., a male target with a female masker in Boebinger et al., 2015), the majority do not control for differences in f0 between the voices. This might be one source of the mixed results observed. For example, previous

studies using large f0 differences (e.g., Δf0 = 0, 2, 4, 6, & 8 semitones (ST)[1] in Madsen et al., 2017) show no difference for musicians and non-musicians. Furthermore, natural fluctuations of f0 in speech intonation support stream segregation, making additional f0 separation unnecessary if f0 contours are sufficiently large (Darwin et al., 2003). When controlling for both f0 separation and fluctuation, Başkent & Gaudrain (2016) found evidence for a musician's advantage in younger listeners (ages 19-27) perceiving a target sentence with a 1-talker interferer. Similarly, Cohn (2018b) found that younger musicians (ages 18-40) showed an advantage in perceiving a sentence with a 1-talker interferer (the same talker) when controlling for f0 separation and fluctuation. An additional consideration, in the mixed results for the musician's advantage, is that listeners' ability to use f0 separation between voices changes by age. For example, Vongpaisal & Pichora-Fuller (2007) found that younger listeners (ages 21-34) could tease apart and identify double vowels at smaller f0 differences than older listeners (ages 65-83). Thus, f0 separation is a particularly relevant feature to examine when investigating changes in any potential musicianship advantage by age.

1.4. Theoretical accounts of a possible musician's advantage

There are varying accounts for possible mechanisms underlying the purported musician's advantage. *Shared auditory encoding* accounts (Bidelman et al., 2014; Shahin, 2011) propose that musical experience tunes how the brain perceives auditory features shared by both music and speech. For example, musicians show an improved frequency following response (FFR), indicating improved subcortical encoding of f0 for music and speech sounds (Bidelman & Krishnan, 2010; Wong et al., 2007). Others have shown that musicians show higher fidelity representations of harmonics in speech (Tierney et al., 2015) and duration of speech properties (e.g., voice onset time in Kühnis et al., 2013). Some *shared auditory encoding* accounts place restrictions on what features could transfer from music-to-speech. For example, Patel (2011, 2012, 2014) proposes that only features that have more fine-grained distinctions in music than in speech are possible candidates. Pitch is thought to be one such feature, as tonal distinctions are argued to be more fine-grained in music than in speech (Zatorre et al., 2002) (while the converse is argued for spectral distinctions). In the current study, a *shared auditory encoding* prediction is that musicians show better speech-in-speech perception on the basis of f0 differences across tasks.

Domain-general attention accounts (Besson et al., 2011; Strait & Kraus, 2011) propose that musical training strengthens general attentional mechanisms; that is, the benefits that come with musical training are not limited to the auditory domain. For example, Medina & Barraza (2019) found that musicians showed better performance than non-musicians, and that this was consistent with better executive attention in a vision task wherein they had to ignore an irrelevant stimulus (all younger adults; 17-33 years of age). Similarly, Tierney and colleagues (2020) found a relationship between an auditory attention task — attending to one stream of tones while inhibiting another — and improved speech perception in the presence of a 1-talker interferer. In the current study, a *domain-general attention* prediction is for an across-the-board musician's advantage in tasks that require greater attentional demands (e.g., divided attention).

A hybrid *auditory encoding - attention* account (Kraus & Nicol, 2014; Kraus & White-Schwoch, 2015) would predict an interaction between enhanced subcortical encoding of speech and top-down cognitive processes. For example, Kraus and Nicol (2014) conceive of auditory training, including musical experience, as an attentional "mixing board", increasing subcortical representation for certain types of inputs, while dampening others. In the current study, finding an advantage only for a cue (e.g., f0) in one type of attentional task, but not in the other would be in line with a hybrid *encoding-attention* account.

[1] A semitone is relative distance between two tones (in log-2 Hertz) in Western musical scales (e.g., C to a C#)

1.5. Current Study

The current study consists of two speech-in-speech experiments to test for a musician's advantage: 1) competing sentences, 2) competing vowels. In both, we use identical voices for both target(s) and masker and manipulate the degree of f0 separation for competing talkers, holding f0 fluctuation constant by monotonizing the stimuli. For the handful of previous studies that do control for f0, we see that sufficiently large f0 separation levels often show no musician's advantage (e.g., Δf0 = 0, 2, 4, 6, & 8 ST in Madsen et al., 2017; Δf0 = 0, 2, 8 ST in Deroche et al., 2017). Therefore, we selected smaller f0 separation (Δf0 = 0, 0.156, 0.306, 1, 2, 3 ST) based on the just-noticeable-difference (JND) in f0 in pure tones for musicians (Δf0 = 0.156 ST) and non-musicians (Δf0 = 0.306 ST) (Kishon-Rabin et al., 2001) (see Supplementary Material Table S2. for semitone calculations). While the majority of prior experiments examine one type of attentional demand and often one age group, the present study[2] tests how musicians and non-musicians (ranging in age from 18-66) use f0 differences across two speech perception tasks varying in attentional demands: 1) a selective attention task where listeners identify the target sentence presented with a 1-talker interferer (Experiment 1), and 2) a divided attention task where listeners hear two vowels played simultaneously and identify both competing vowels (Experiment 2). This cross-sectional approach can reveal if listeners, varying in musicianship and age, differ in their performance based on f0 separation of the voices across varying attentional demands.

2. General Methods

2.1. Participants (Experiments 1 and 2)

A total of n = 72 participants were recruited for the study, consisting of native English speakers in four groups based on their age and whether they received musical training or not. Based on related work showing an advantage emerging around age 40 (Zendel & Alain, 2012), we recruited younger adult (YA< 40 years) and middle-aged/older adult (OAs; ≥ 40 years) age groups. Musicians were recruited if they had at least 9 years of musical training and were practicing on a weekly basis at the time of the study. Non-musicians were recruited if they reported having minimal musical training (<1 year in duration that had occurred at least 7 years ago, following Parbery-Clark, Skoe, Lam, et al., 2009). While participants were recruited based on not having "hearing impairments or any auditory disorders", n = 2 participants (n = 1 OA musician, n = 1 OA non-musician) were excluded as they did not pass an in-lab pure tone hearing screening (described in Section 2.2. in more detail). Participants who did not complete both Experiment 1 and 2 (n = 15 participants[3]) were also excluded from analysis (described in more detail in Section 4.4.1.).

The retained participants consisted of n = 55 adults, ranging in age from 18-66 (median age = 40.0 years), who completed both experiments. Musician (n = 29) and non-musician (n = 26) groups did not differ in terms of age or years of education (shown in Table 1).

Table 1. Age and education of musician and non-musician groups.

	Musician group (n = 29)	Non-musician group (n = 26)	Group comparison t-test
Age	Mean = 39.7 years old (sd = 15.2)	Mean = 39.7 years old (sd = 14.8)	$t(60.82) = 0.24$, $p = 0.8$
Education	Mean = 16.6 years (sd = 2.6)	Mean = 16.4 years (sd = 2.2)	$t(56.21) = 0.39$, $p = 0.7$

[2] This project is an extension of material collected from a doctoral thesis (Cohn, 2018a) and an adaptation of a proceedings paper (Cohn, 2018b)

[3] n = 5 OA musicians, n = 2 OA non-musicians; n = 2 YA musicians, n = 6 YA non-musicians

Musicians had an average of 26.1 years of musical training (sd = 15.7, range = 9.5-63 years) and practiced on a weekly basis at the time of the study (mean = 10.7 hours/week, sd = 7.9). Musicians varied in the family of their primary instrument(s): n = 3 brass (e.g., trombone, french horn, trumpet), n = 8 keyboard (e.g., piano), n = 8 string (e.g., violin, guitar, cello, double bass), n = 10 woodwind (e.g., flute, clarinet, saxophone). Slightly more than half of musicians (58.6%) additionally had voice training (n = 17). None of the subjects reported prior experience with a tonal language (e.g., Mandarin Chinese, Thai, Punjabi, etc.). All participants completed informed consent in accordance with the UC Davis Institutional Review Board (IRB).

2.2. General Procedure

Participants came into the lab for an hour-long session in which they completed both Experiment 1 (Section 3) and Experiment 2 (Section 4) in a sound-attenuated booth wearing over-ear headphones (Sennheiser 280 PRO) (experiment order counterbalanced across subjects).

After completing the experiments, participants completed the hearing screening (adapted from Reilly et al., 2007). To pass the hearing screening, participants needed an average of (≤ 25 dB HL) at each of the frequencies tested (250-8000 Hz). Participants were compensated with a $15 giftcard for their time.

3. Experiment 1: Sentence perception with a competing sentence

In Experiment 1, listeners completed a sentence-in-speech task where they are instructed to attend to one signal and ignore the other. The task consisted of sentences from the Coordinate Response Measure (CRM) corpus (Bolia et al., 2000). CRM sentences all have the same form: “Ready <call sign> go to <color> <number> now.” Following Brungart (2001), target sentences used in the current study were cued by the call sign “baron”, and participants were asked to identify the color/number from that sentence (e.g., “Ready baron go to green three now.”), while ignoring a masking sentence that has a different call sign, color, and number (e.g., “Ready arrow go to red one now”). This paradigm has been widely used to assess speech-in-speech perception (Bidelman & Yoo, 2020; Carlile & Corkhill, 2015; Darwin et al., 2003; Johnsrude et al., 2013), including investigations of age and/or hearing loss (Gygi & Shafiro, 2014; Lee & Humes, 2012).

3.1. Stimuli

Stimuli consisted of sentences produced by a single male talker (Talker 1) from the CRM corpus (Bolia et al., 2000), monotonized at 100 Hz and amplitude normalized to 70 dB[4] in Praat (Boersma & Weenink, 2021). Target sentences (n = 16), indicated by the call sign “baron”, were monotonized at six f0 levels relative to 100 Hz (Δf0 = +0, 0.156, 0.306, 1, 2, 3 ST). Masker sentences, which contained 6 different call signs (“arrow”, “eagle”, “hopper”, “laker”, “ringo”, “tiger”) were monotonized at 100 Hz. We pseudo-randomly mixed the target sentences with the masker sentences spoken by the same talker, at a signal-to-noise (SNR) ratio[5] of 0 dB. Sentences were mixed with the constraint that the target and masker contained different call signs, colors, and numbers[6]. Each masker call sign, color, and number occurred an equal number of times. In total, 96 stimuli were generated (16 baron sentences * 6 f0 levels).

3.2. Procedure

Participants began with 12 pre-test trials, where they heard all possible “baron” target sentences in isolation (randomly presented) (see Figure 1.A). They were asked to click the color-number combination

[4] relative to 2e−05 Pascal, the “normative auditory threshold for a 1000-Hz sine wave” (Praat default).
[5] Also referred to as a target-to-masker ratio (TMR) for speech-in-speech.
[6] Note that as the sentences were naturally recorded and used different call signs, colors, and numbers, there are small differences in timing across the target/masker sentences.

from the target sentence. After subjects made a response, they were shown immediate feedback on their performance ("Correct" or "Incorrect") (inter-trial interval (ITI) = 1 s). Subjects' accuracy was calculated at the end of the pre-test block; in order to continue on to the experimental trials, subjects needed to correctly identify the target sentences at 90% accuracy or higher. If they did not reach the 90% requirement, they repeated the single sentence pre-test block again (up to 2 additional times).

Next, participants completed the experimental trials consisting of a target and masker sentence presented simultaneously (see Figure 1.B) (ITI = 1 s). Subjects began with a short practice block consisting of 4 stimuli randomly selected at each of the six f0 levels (total of 24 trials). No feedback on performance was provided. Next, they completed 192 experimental trials (16 sentences * 6 f0 levels *2 repetitions), presented across 8 blocks (24 trials each; order randomized) lasting roughly 20 minutes.

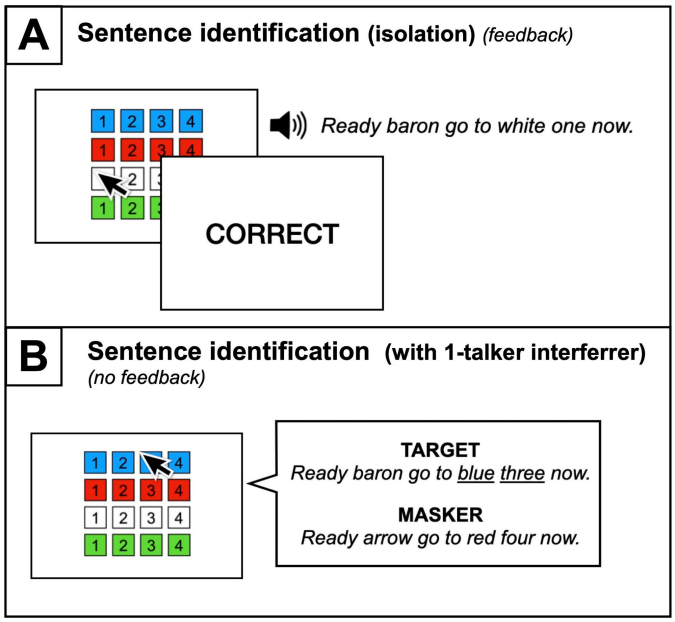

Figure 1. 1-sentence talker interference paradigm for Experiment 1, based on the Coordinate Response Measure (CRM) paradigm. **(A)** Participants started with sentence identification in isolation (i.e., without a masking sentence), where they clicked on the color/number from the target sentence. They heard 6 sentences and received immediate feedback on their accuracy after each trial. **(B)** Participants then completed sentence identification with a 1-talker masker (0 dB SNR). Their task was to click on the color/number box associated with the target (cued by the call sign "baron"). No feedback was given.

3.3. Analysis

Trial responses were scored binomially as to whether participants correctly identified both the target color and number from the sentence (= 1), or not (= 0). We modeled accuracy with a Bayesian multilevel logistic regression model using the *brms* package (Bürkner, 2017) in R [version 4.0.5] (R Core Team, 2021) using the *bernoulli* family (8,340 iterations; warmup = 1000; thin = 3). Fixed effects included F0 Separation (centered), Age (centered), Group (musician, non-musician), and their interactions. We also included fixed effects of Block Number (centered) and Subject Single Sentence Accuracy (standardized) (model structure provided in Equation 1). Random effects included by-Sentence and by-Subject random intercepts, and by-Subject random slopes for F0 Separation and Block Number. Contrasts were sum coded.

$$\text{correct} \sim \text{F0} * \text{Age} * \text{Group} + \text{Block} + \text{SingleSentenceAcc} + (1|\text{Sentence}) + (\text{F0} + \text{Block}|\text{Subject}) \quad (1)$$

3.4. Results

All participants reached the requisite 90% accuracy for target word identification when listening to sentences in isolation. There was no difference in accuracy in single sentence identification across the musician (98.1%) and non-musician groups (96.9%) [X^2 (1, N = 55) = 0.51, p = 0.48]. Investigating group and age (< 40, 40+ years), based on the median age of our participants, we see highest average accuracy for younger non-musicians (100%), then older musicians (98.2%), followed by younger musicians (98.0%) and older non-musicians (94.4%).

Mean proportion of trials in which the target color/number were identified in the experimental trials is plotted in Figure 2.A. Figure 2.B. presents posterior means and credible intervals for all of the fixed effects in our model, and Table 2 presents the full model output. The model revealed an effect of F0 Separation, with higher accuracy with a larger f0 separation. The effect for Block Number indicated that participants improved over time. There was also an effect for Age, where participants' accuracy decreases with advanced age. An interaction between Age and Group revealed steeper age-related declines in accuracy for older musicians. Finally, there was a three-way interaction between Group, F0 Separation, and Age: older musicians show less of a benefit of f0 separation.

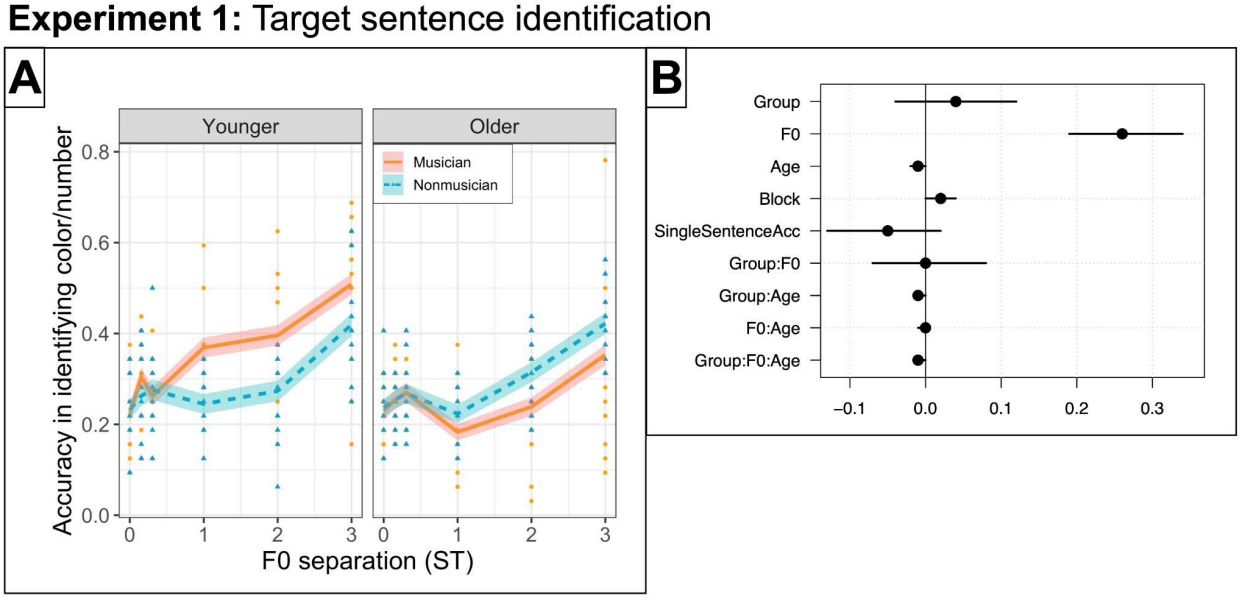

Figure 2. (A) Mean accuracy in correctly identifying the color and number from the target ("baron") sentence by Group (musicians = orange solid line, dots; non-musicians = blue dashed lines, triangle) at each F0 Separation level (in semitones, ST). Accuracy is faceted by Age Category (< 40 years, 40+ years based on median age of our sample). Shading indicates the standard error of the mean. **(B)** Posterior means and credible intervals for all of the fixed effects in the model.

Table 2. **Sentence identification** (Experiment 1): Posterior means (Estimate), standard deviation of the posterior (Error), 95% credible intervals (Q2.5, Q97.5), and percent of posterior distribution above or below zero, for fixed effects. Effects whose credible intervals do not include zero, or those with 95% of their distribution on one side of 0 are in bold.

					% Distribution	
	Estimate	Error	Q2.5	Q97.5	< 0	> 0
Intercept	**-0.79**	**0.26**	**-1.29**	**-0.29**	**100**	**0**
Group (Musician)	0.04	0.04	-0.04	0.12	16	84
F0	**0.26**	**0.04**	**0.19**	**0.34**	**0**	**100**
Age	**-0.01**	**0.00**	**-0.02**	**0.00**	**100**	**0**
Block	**0.02**	**0.01**	**0.00**	**0.04**	**1**	**99**
SingleSentenceAcc	-0.05	0.04	-0.13	0.02	92	8
Group(Musician):F0	0.00	0.04	-0.07	0.08	48	52
Group(Musician):Age	**-0.01**	**0.00**	**-0.01**	**0.00**	**98**	**2**
F0:Age	0.00	0.00	-0.01	0.00	88	12
Group(Musician):F0:Age	**-0.01**	**0.00**	**-0.01**	**0.00**	**98**	**2**

Num. observations = 10,560; *Num participants* = 55 ; *Num. sentences* = 16

3.5. Post hoc analyses & Results

3.5.1. Age Category

To ascertain whether the age-related effects for musicians in Experiment 1 reflect 1) a general decline with age, or 2) the presence of a YA musician's advantage, but one that is lost with age, we conducted a post hoc analysis examining effects across age categories. We modeled accuracy in the CRM task with a *brms* model (Bürkner, 2017) in R [version 4.0.5] (R Core Team, 2021) (bernoulli family; 8,340 iterations; warmup = 1000; thin = 3). The model structure was same as the main analysis, except for Age: here, we used an Age Category predictor (<40, 40+ years old, sum coded) based on the median age in our sample (median = 40.0 years; mean = 39.9 years)[7].

Model output is provided in Supplementary Data (Table S3). We see credible effects for F0 separation and Block, increasing accuracy with f0 separation and over the course of the experiment. Additionally there are several effects of Age Category: first, younger adults show higher accuracy overall in the task. An interaction between Age Category and Group showed that this boost was even higher for YA musicians. Furthermore, a 3-way interaction between Age Category, Group, and F0 showed that this YA musician's advantage increased with increasing f0 separation. No other effects or interactions were observed.

3.6. Interim discussion

[7] correct ~ F0 * Age Category * Group + Block + SingleSentenceAcc + (1|Sentence) + (F0 + Block|Subject) (1)

In Experiment 1, we find that f0 separation improves listeners' ability to identify a target sentence when presented alongside a 1-talker interferer, consistent with prior work showing intelligibility gains with increasing f0 separation for competing sentences (Lee & Humes, 2012) and vowels (Vongpaisal & Pichora-Fuller, 2007). In comparing performance by listener age, we see that younger adults overall show higher accuracy on the task than older adults, consistent with age-related declines in speech-in-speech perception (e.g., Heidari et al., 2020).

Additionally, we find some support for a musician's advantage in speech perception for perceiving a sentence with a competing talker. However, this advantage is modulated by age. Specifically, YA musicians perform the best, and part of their improvement is rooted in their ability to leverage the f0 separation between competing sentences. This finding differs from that of Zendel & Alain (2012), who observed an advantage that emerges after age 40. Our finding suggests that f0 separation might have played a role in younger musician's advantage observed in other studies that did not control for the voice characteristics of competing talkers (e.g., Morse-Fortier et al., 2017; Parbery-Clark et al., 2009). Taken together, we see a possible transfer for increased pitch sensitivity — from music to speech-in-speech perception — supporting *shared auditory encoding* accounts (Bidelman et al., 2014; Patel, 2014; Shahin, 2011).

Thus, while we do find evidence for a musician's advantage, there appear to be limitations to this benefit. For one, middle-aged/older adults in the current study do not exhibit a musician's advantage for sentence-in-sentence perception. While understudied, some work has shown a reduced ability for older adults to tease apart competing vowels at smaller f0 differences, compared to younger adults (Vongpaisal & Pichora-Fuller, 2007; though, musical background was not reported in that study). At the same time, older musicians (ages 65+) in other studies show improved frequency discrimination, compared to age-matched non-musicians (e.g., Grassi et al., 2017), suggesting that the type of task might shape whether an musician's advantage emerges for older adult listeners.

4. Experiment 2: Double vowel perception

While Experiment 1 investigated the ability of listeners to hone in on a target sentence, amidst an interfering sentence, Experiment 2 employs a double vowel paradigm (Assmann & Summerfield, 1990; Vongpaisal & Pichora-Fuller, 2007), where participants hear two synthetic vowels varying in degree of f0 separation. Given that the vowels are presented with very small f0 separation levels (the same as in Experiment 1), it is likely that the vowels could perceptually fuse into one 'auditory object'. Experiment 2 tests the extent to which musicians and non-musicians might leverage f0 differences to tease apart the competing vowels — and also if this varies by age. Additionally, we ask listeners to identify both vowels they heard, testing their divided attention (i.e., attending to both streams simultaneously).

Given prior work showing that musicians display enhanced subcortical representations of the spectrum in speech (e.g., /ba/ vs. /ga/ in Kraus et al., 2014; Parbery-Clark, Skoe, & Kraus, 2009), even with advanced listener age (Bidelman & Alain, 2015), we also take into account the spectral distance between the double vowels (F1-F2 Euclidean distance) (Bradlow et al., 1996). As listeners do not perceive isolated (naturally produced) vowels 100% correctly (e.g., Peterson & Barney, 1952), we began the study with a single vowel identification task. Furthermore, we account for each participant's accuracy in identifying each vowel in the full model to account for differences attributable to vowel identification in general.

4.1. Stimuli

Five steady-state vowels (260 ms; f0 = 100 Hz) were synthesized in R with the *phonTools* package (Barreda, 2015): /i, ɛ, æ, ɑ, u/ (formant frequency values are provided in Appendix A) based an acoustic analysis of California English vowels (Holland, 2014). We generated six versions of each vowel, varying in f0 separation levels from 100 Hz (Δf0 = +0, +0.156, +0.306, +1, +2, +3 ST), all at 60 dB. To create the

double vowels, all possible vowel combinations were combined (excluding combination with itself, e.g. no /u/ + /u/), with the vowel presentation levels matched (and double vowel stimuli amplitude normalized to 60 dB[8]). In each double vowel combination, one vowel had a higher f0 than the other, for a total of 120 stimuli (20 vowel pairs * 6 f0 levels).

4.2. Participants

The same participants from Experiment 1 completed Experiment 2 (see Section 2.1 for details).

4.3. Procedure

Participants first completed a vowel familiarization task, illustrated in Figure 3.A (Vongpaisal & Pichora-Fuller, 2007), with the labeled button box containing five example words for the vowels ("beat", "bet", "bat", "boot", "bought") (button-label correspondence was counterbalanced across participants). They heard each of the vowels (at each f0 level) presented individually in a total of 30 trials (5 vowels * 6 f0 levels; randomly presented). If, after 3 attempts, participants did not reach 90% accuracy in identifying the vowels, they did not participate in the experimental trials.

In the double vowel trials, listeners were told that they might hear 1 or 2 vowels (following Vongpaisal & Pichora-Fuller, 2007) and instructed to identify the vowel(s) they heard via two button presses (schematized in Figure 3.B). If they perceived two different vowels, they were instructed to identify each vowel in the pair. If they perceived just one vowel, they were instructed to press that vowel button twice. While there were always two vowels presented in the experimental trials, this allows us to test how the vowels might 'perceptually fuse' at small f0 separation levels.

The double vowel portion began with 24 practice trials (4 randomly selected vowel pairs from each of the 6 f0 levels); no feedback was given. Then, they saw the instructions repeated again before starting the experimental trials where they heard each of the 120 double vowel stimuli (20 vowel pairs * 6 f0 levels) twice, for a total of 240 trials presented across 8 blocks (30 trials per block). Assignment of stimuli to block was randomized. After each block, participants were shown their progress (e.g., "Block 1/8 Complete"). In total, the double vowel experiment took roughly 25 minutes to complete.

[8] Presentation level for stimuli was 70 dB SPL in Experiment 1; 60 dB SPL in Experiment 2. Both are within a reasonable range for comfortable listening in a sound-attenuating booth, wearing over-ear headphones. Additionally, as our research question aimed to test the impact of the relative difference in f0 between two sounds (identical in intensity), we would not expect these small differences in presentation level to shape the effects we observe.

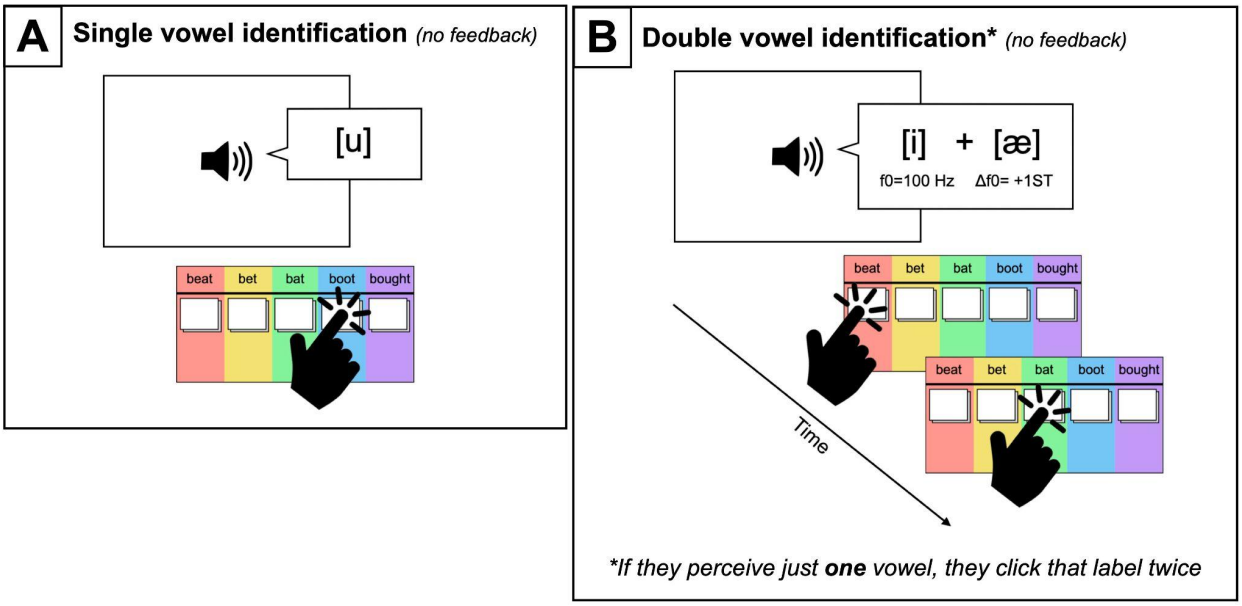

Figure 3. Vowel identification paradigms. **(A)** Participants begin with a single vowel identification block; they hear each synthesized vowel in isolation and select the representative word using a labeled button box. **(B)** Participants then complete a double vowel identification task. They hear a blend of two vowels varying in f0 separation. If they perceive 2 vowels, they identify each vowel with a button press (order of button presses does not matter). If they perceive just one vowel, they click that vowel button twice.

4.4. Single vowel identification: Analysis & Results

Of the total of n = 72 participants recruited, n = 14 participants (n = 4 OA musicians, n = 2 OA non-musicians; n = 2 YA musicians, n = 6 YA non-musicians) did not reach the required 90% accuracy in the single vowel identification after three blocks and therefore did not complete Experiment 2. A vowel confusion matrix (see Supplementary Data, Table S4) sheds some light on the source of this difficulty for these participants. They identify /ɑ/ as "bat" 66.7% of the time. Additionally, they show confusions about vowel height, identifying /æ/ as "bet" 22.7% of the time and /i/ as "bet" 15.9% of the time. Finally, they identify /u/ as "bought" 22.7% of the time (perhaps due to the "u" letter in the word).

All other participants passed the single vowel identification portion with an average accuracy of 90% or greater (in a single block, with three attempts). We did see differences in single vowel identification accuracy, which was higher for musicians (95.9%) than non-musicians (91.4%) [X^2 (1, N = 50) = 14.16, $p < 0.001$]. Investigating age groups (< 40, 40+ years), we see lower average accuracy for both younger non-musicians (90.5%) and older non-musicians (92.1%), than younger musicians (97.3%) and older musicians (94.3%). Vowel confusion matrices for each age/musician group (provided in Supplementary Data, Tables S5-S8) reveal sources for these differences, summarized in Table 3. For example, all groups show confusions in identifying /ɑ/ as "bat" (rather than "bought"). Vowel height confusions were also common, such as identifying /æ/ as "bet", indicating perception of a lowered vowel. YA non-musicians and OA musicians also mistook /u/ for "bet", attributing the fronted /u/ as a front vowel.

Table 3. Summary of single vowel confusions

Confusion	YA non-musician	YA musician	OA non-musician	OA musician
Selected “bat” for /ɑ/	23.1%	11.5%	10.7%	16.7%
Selected “bet” for /æ/	10.3%	2.5%	6.0%	2.4%
Selected “bet” for /u/	5.1%	0.0%	1.2%	4.8%

4.4. Double vowel identification: Analyses & Results

In addition to the participants who did not complete the experimental trials, data was excluded for n = 4 listeners who performed at floor in the double vowel identification task (mean accuracy < 5%) (resulting in removal of 2 middle-aged/older musicians and 2 middle-aged/older non-musicians). Data was also excluded due to a computer error for 1 participant (1 younger non-musician), where the single vowel portion crashed and they completed it more than 3 times. Accordingly, n = 50 participants[9] were included in the Experiment 2 analysis (summarized in Table 4).

Table 4. Participant breakdown

Originally recruited	n = 72	
Retained	n = 55	n = 2 Did not pass hearing screening n=14 Did not pass single vowel portion and did not complete Experiment 2 n = 1 Left study before the end of Experiment 2
Experiment 1	n = 55	
Experiment 2	n = 50	n = 4 Had double vowel accuracy at floor (< 5%) n = 1 Computer error

4.4.1. Stream separation

We coded stream separation binomially (identifying that two vowels were presented = 1, or not = 0) and modeled it with a Bayesian logistic regression using the *brms* R package (Bürkner, 2017) (8,340 iterations; warmup = 1000; thin = 3). The model included fixed effects of F0 Separation (centered), Age (centered), and Group (musicians, non-musicians), and all possible interactions. Additionally, we included a fixed effect of F1/F2 Vowel Euclidean Distance (log Hertz) to account for the degree of spectral difference between the vowels. The model also included interactions between F1/F2 Distance with Age and with Group. Furthermore, we included Block (centered) as a fixed effect to account for changes over time. Contrasts were sum coded. Random effects included random intercepts for Participants and Vowel Pair. We also included by-Participant random slopes for F0 Separation, F1/F2 Distance, and Block. The model syntax is shown in Equation 2.

F0*Age*Group + F1F2.Distance*Age*Group + Block + (F0 + F1F2.Distance + Block|Subject) + (1| VowelPair) (2)

Figure 4.A plots the proportion of trials in which two vowels were identified, the credible intervals are plotted in Figure 4.B., and the model output is provided in Table 4. The model revealed an

[9] All participants performed above 5% accuracy in Experiment 1.

effect of F0 Separation, where likelihood of perceiving two vowels increases with a larger f0 difference between the vowels. There was also an overall effect of Group, as seen in Figure 4.A, wherein musicians are more likely to detect two vowels (than just one vowel). Additionally, F1/F2 Distance was a predictor, such that a larger F1/F2 distance between the vowels was associated with a higher likelihood they hear two versus just one vowel. Over the course of the Block, participants also showed overall improvements in detecting two vowels. We also observe an interaction between Group and F0 Separation: musicians show stronger stream separation on the basis of increasing f0 separation. No other effects or interactions were observed.

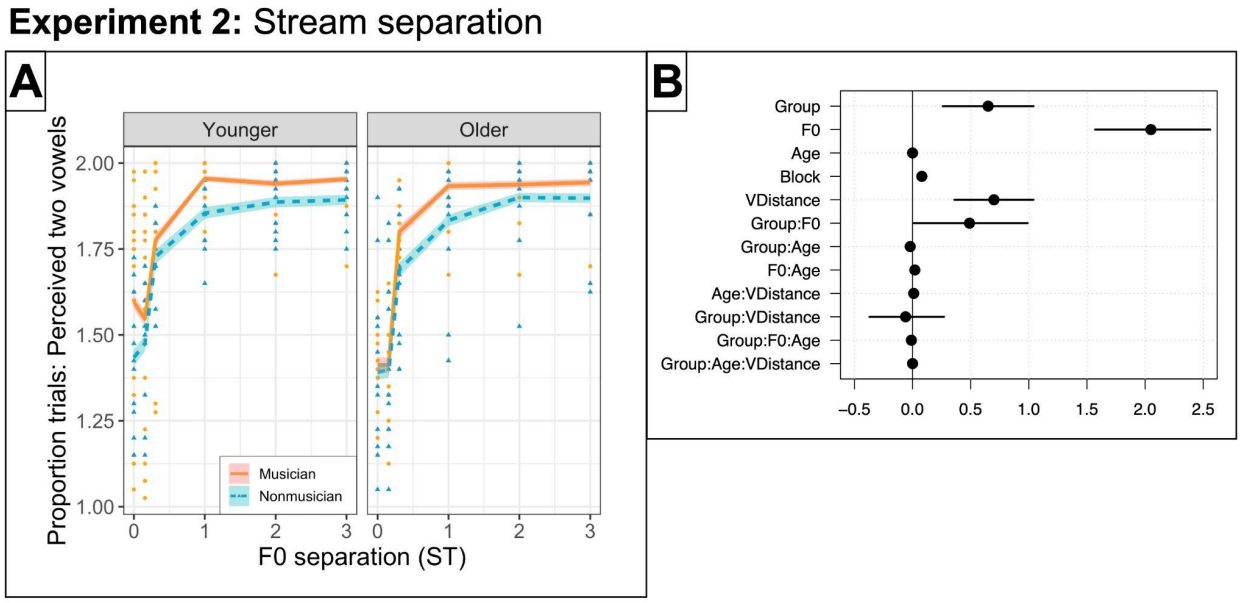

Figure 4. (A) Mean number of vowels perceived (1 or 2) by Group (musicians = orange solid line, dots; non-musicians = blue dashed lines, triangle) at each F0 Separation level (in semitones, ST). Accuracy is faceted by Age Category (< 40 years, 40+ years based on median age of our sample). Error ribbons show the standard error of the mean. **(B)** Posterior means and credible intervals for all of the fixed effects in the stream separation model.

Table 5. **Stream Separation** (Experiment 2). Posterior means (Estimate), standard deviation of the posterior (Error), 95% credible intervals (Q2.5, Q97.5), and percent of posterior distribution above or below zero, for fixed effects. Effects whose credible intervals do not include zero, or those with 95% of their distribution on one side of 0 are in bold.

					% Distribution	
	Estimate	Error	Q2.5	Q97.5	< 0	> 0
Intercept	1.88	0.24	1.41	2.37	0	100
Group(Musician)	**0.65**	**0.20**	**0.26**	**1.04**	**0**	**100**
F0	**2.05**	**0.25**	**1.57**	**2.56**	**0**	**100**
Age	0.00	0.01	-0.03	0.03	45	55
Block	**0.08**	**0.02**	**0.05**	**0.12**	**0**	**100**
VowelDistance	**0.70**	**0.17**	**0.36**	**1.04**	**0**	**100**
Group(Musician):F0	**0.49**	**0.25**	**0.01**	**0.99**	**2**	**98**
Group(Musician):Age	-0.02	0.01	-0.04	0.01	87	13
F0:Age	0.02	0.02	-0.01	0.06	9	91
Age:VowelDistance	0.01	0.01	-0.02	0.03	24	76
Group(Musician):VowelDistance	-0.06	0.16	-0.37	0.27	64	36
Group(Musician):F0:Age	-0.01	0.02	-0.04	0.03	70	30
Group(Musician):Age:VowelDistance	0.00	0.01	-0.02	0.02	56	44

Num. observations = 12,000; *Num. participants* = 50; *Num. vowel pairs* = 20

4.4.3. Double vowel identification

Participants' identifications of the two vowels in the experimental trials was binomially coded (1 = both vowels correctly identified, 0 = not) and modeled with a Bayesian multilevel logistic regression model using the *brms* R package (Bürkner, 2017) (8,340 iterations; warmup = 1000; thin = 3). Fixed effects included F0 Separation (centered), Age (centered), Group (musician, non-musician), and F1/F2 Vowel Euclidean Distance (log Hertz). The 3-way interaction between F0 Separation, Age, and Group was included, as well as one between F1/F2 Vowel Distance, Age, and Group. We also included a predictor of Joint Single Vowel Accuracy for each vowel in the pair based on pre-experiment single vowel identification accuracy as a measure for how well they perceived each of the synthetic vowels (logit of product of the Vowel-1 and Vowel-2 probabilities). Finally, we included the predictor of Block Number (centered). Contrasts were sum coded. Random effects included by-Subject random intercepts and by-Subject random slopes for F0 Separation, F1/F2 Distance, and Block Number, and random intercepts for Vowel Combination. The model syntax is provided in Equation 3.

correct ~ F0*Age*Group + F1F2.Distance*Age*Group + JointSingleVowelAcc + Block + (F0 + F1F2.Distance + Block|Subject) + (1|VowelPair) (3)

Mean accuracy for the task is plotted in Figure 5.A. Table 5 and Figure 5.B present posterior means and credible intervals (between the 2.5 and 97.5th percentiles) for all of the fixed effects in our model. Effects whose credible intervals do not include zero or have 95% of their distribution on one side of 0 are bolded. Results indicate a credible effect for F0 Separation, where identification accuracy increases as a function of f0 separation between voices. Similarly, F1-F2 Vowel Distance reliably affected accuracy, wherein listeners are more accurate in identifying both vowels the more spectrally distinct they were. There was also an effect for Block Number resulting in an increase in accuracy over time for all groups. Furthermore, listeners' performance identifying each of the vowels presented in isolation (prior to the double vowel experimental trials) is positively related to their ability to perceive those same vowels in the double-vowel stimuli (Joint Single Vowel Accuracy). We also see an interaction between Group and Age, where musicians show lower accuracy with increasing age; as Group is sum coded, the converse is also true: non-musicians show higher accuracy with increasing age. Finally, Age and F0 Separation interacted, such that the effect of f0 separation is weaker for younger adults. No other effects or interactions whose 95% credible intervals did not overlap with zero were observed.

While Figure 5.A appears to show a musician's advantage for YAs, when we account for listeners' accuracy in correctly identifying each of the synthetic vowels in isolation, the model confirms that this is not a reliable musician group-level difference.

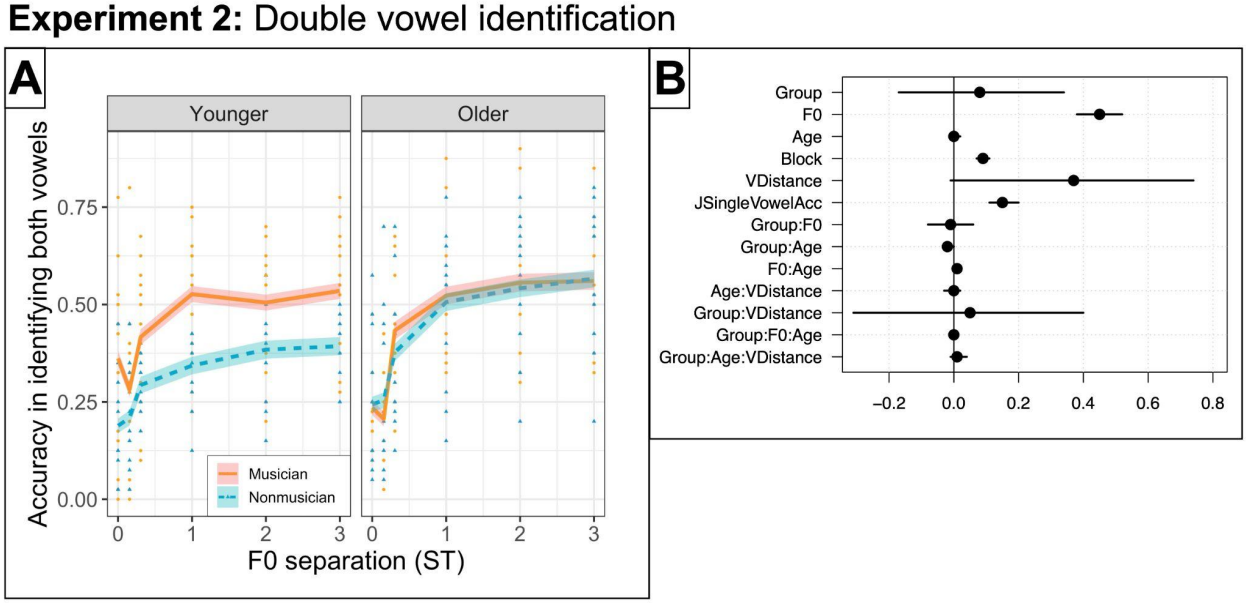

Figure 5. (A) Mean accuracy for correctly identifying both vowels by Group (musicians = orange solid line, dots; non-musicians = blue dashed lines, triangle) at each F0 Separation level (in semitones, ST). Accuracy is faceted by Age Category (< 40 years, 40+ years based on median age of our sample). Error ribbons show the standard error of the mean. **(B)** Credible intervals for double vowel identification.

Table 6. **Double vowel identification** (Experiment 2): Posterior means (Estimate), standard deviation of the posterior (Error), 95% credible intervals (Q2.5, Q97.5), and percent of posterior distribution above or below zero, for fixed effects. Effects whose credible intervals do not include zero, or those with 95% of their distribution on one side of 0 are in bold.

					% Distribution	
	Estimate	Error	Q2.5	Q97.5	< 0	> 0
Intercept	**-1.17**	**0.21**	**-1.59**	**-0.75**	**100**	**0**
Group (Musician)	0.08	0.13	-0.17	0.34	26	74
F0	**0.45**	**0.04**	**0.38**	**0.52**	**0**	**100**
Age	0.00	0.01	-0.01	0.02	30	70
Block	**0.09**	**0.01**	**0.07**	**0.11**	**0**	**100**
VowelDistance	**0.37**	**0.19**	**-0.01**	**0.74**	**3**	**97**
JointVowelAcc	**0.15**	**0.02**	**0.11**	**0.20**	**0**	**100**
Group(Musician):F0	-0.01	0.04	-0.08	0.06	59	41
Group(Musician):Age	**-0.02**	**0.01**	**-0.03**	**0.00**	**96**	**4**
F0:Age	**0.01**	**0.00**	**0.00**	**0.01**	**0**	**100**
Age:VowelDistance	0.00	0.01	-0.03	0.02	53	47
Group(Musician):VowelDistance	0.05	0.18	-0.31	0.4	40	60
Group(Musician):F0:Age	0.00	0.00	-0.01	0.00	63	37
Group(Musician):Age:VowelDistance	0.01	0.01	-0.01	0.04	11	89

Num. observations = 12,000; *Num. participants* = 50; *Num. vowel pairs* = 20

4.4.4. Double vowel identification: Post hoc analysis & Results

To confirm there is no YA musician's advantage in double vowel identification, we conducted a post hoc analysis, fitting accuracy with a *brms* model (bernoulli family; 8,340 iterations; warmup = 1000; thin = 3). We used the identical model structure as in the main analysis, but with Age Category (<40, 40+ years), in lieu of continuous Age[10].

The model (provided in Supplementary Material, Table S9) revealed largely parallel results as in the main model, including less of a reliance of f0 separation for YAs than OAs. Yet, we did not see an effect of Group or its interaction with Age Category. Indeed, as previously mentioned (Section 4.4.), YA non-musicians in the study — while ultimately able to pass the 90% accuracy for single vowel identification — still have lower single vowel accuracy than the other groups. Accounting for this lower

[10] correct ~ F0*Age*Group + F1F2.Distance*Age*Group + JointSingleVowelAcc + Block + (F0 + F1F2.Distance + Block|Subject) + (1|VowelPair)

accuracy allowed us to correctly attribute their difficulty perceiving the vowels in general, rather than a musician's advantage per se.

4.7. Interim discussion

In Experiment 2, we see that increasing f0 separation between the voices improves listeners' ability to tease apart and identify two competing vowels, in line with prior work (de Cheveigné et al., 1997; Vongpaisal & Pichora-Fuller, 2007). Furthermore, degree of F1-F2 Euclidean distance between the vowels also supports vowel separation and identification: vowels that had larger vowel-space differences (e.g., /i/ + /ɑ/) are better recognized than vowels that are closer together (e.g., /i/ + /ɛ/). This finding aligns with reduced intelligibility for speech produced with a reduced vowel space observed in other studies (Bradlow et al., 1996).

We see some support for a musician's advantage, but critically only for stream separation. That is, musicians (all ages) are more likely to perceive two separate vowels, relative to non-musicians. This finding aligns with work showing musicians' enhanced ability to tease apart two complex (non-speech) harmonic sounds (Zendel & Alain, 2009). Furthermore, we see that musicians are better at separating the two vowels with increasing f0. Yet, for both separation and identification, we saw no difference in how musicians use spectral differences (here, degree of vowel space expansion). Together, these findings support *shared auditory* accounts that propose more constrained transfer from music-to-speech: for pitch, but not for features that music places less 'precision' on, such as the spectrum (Patel, 2012; Zatorre et al., 2002).

Age category also played a role in the double vowel experiment independently of musicianship, specifically in how listeners are able to identify the competing vowels. In particular, younger adults leverage f0 separation less than middle-aged/older adults. That is, with increasing age, adults show a steeper increase in accuracy as f0 separation increases. Why is this the case? We might predict the opposite based on the prior literature: older adults (ages 65-83) show less of a benefit of f0 separation (Vongpaisal & Pichora-Fuller, 2007). Here, it is important to note that our older adults ranged from ages 40-66, largely occupying middle-adulthood, which could mean that declines in f0 encoding are not as prevalent as those for elderly listeners. Indeed, some work has shown similar or decreased speech-in-speech perception for middle aged adults (ages 49-59) compared to college-age adults (ages 19-24), but better than older adults (ages 60-83) (Helfer & Freyman, 2009). While speculative, our OAs (ages 40-66) might be using f0 as a compensatory strategy to offset the start of age-related difficulties perceiving speech-in-speech. Taken together, our pattern of results suggests that both age and musical training shape the way listeners both separate and identify competing vowel sounds in independent, but nuanced, ways.

5. Discussion

This study investigated listeners' speech-in-speech perception across two tasks: perception of a sentence with a 1-talker interferer (selective attention) and perceiving two competing vowels (divided attention). We compared musicians and non-musicians varying in age to test the purported musician's advantage, investigating their reliance on f0 separation to perform both tasks.

Across both experiments, one of the strongest predictors of speech-in-speech perception is f0 separation. All listeners — musicians and non-musicians alike — use f0 differences to separate competing voices, and identify the target stream(s), consistent with prior work (Lee & Humes, 2012; Summers & Leek, 1998; Vongpaisal & Pichora-Fuller, 2007). Furthermore, when separating and identifying two vowels, all listeners use their spectral distance (in F1/F2 vowel space).

Additionally, in both studies, we see support for a musician's advantage for speech-in-speech perception (e.g., Parbery-Clark et al., 2011; Parbery-Clark, Skoe, Lam et al., 2009; Zendel & Alain, 2012). In the case of sentence perception (Experiment 1), YA musicians consistently show higher

accuracy. In stream separation of vowels (Experiment 2), we see a consistent musician's effect across ages. However, we do not see *identical* musician's advantages across the experiments. Both f0 separation, age, task shape the way an advantage emerges, highlighting potential sources for the mixed results observed in the literature for speech-in-speech perception (cf. Coffey et al., 2017).

Why do we see a musicians' advantage for speech-in-speech in the current study, while other studies report no difference between musicians and non-musicians? Our results suggest that one factor is f0. F0 appears to be a particularly useful cue for musicians, likely due to their enhanced perception of pitch (e.g., Kishon-Rabin et al., 2001). Supporting our predictions, we see that our musicians' advantage is supported by their ability to leverage f0 separation between the voices: to identify the target sentence (Experiment 1) or separate the competing streams (Experiment 2). These findings point to the importance of controlling f0 characteristics of competing voices when assessing speech-in-speech (cf. Başkent et al., 2018; Başkent & Gaudrain, 2016; Deroche et al., 2017; Madsen et al., 2017) and suggest that an advantage might only emerge when two voices have very similar f0 values. In studies that examine voices that vary widely in f0 separation and other speaker indexical characteristics (e.g., Boebinger et al., 2015; Ruggles et al., 2014), it is possible that both musicians and non-musicians alike are able to use these cues equivalently given sufficiently large differences.

Another possible source of the mixed findings in the speech-in-speech literature is due to listener age. In the current study, we see a younger musician's advantage (ages 18-39) in Experiment 1 (perceiving a sentence with a 1-talker interferer), and part of their advantage is in using f0 separation between the target and masker sentences. This finding is consistent with related work with younger adults (ages 19-27) that also controlled for f0 separation and fluctuation and found an advantage (Başkent & Gaudrain, 2016). In a related study of slightly younger participants (ages 11-14) who began training early (around age 7) and had more than 5 years of musical training, no advantage was detected (Başkent et al., 2018). One possibility is that a musician's advantage *emerges* with age, based on cumulative years of musical training. For example, Zendel & Alain (2012) find an advantage emerging around age 40, while others show it occurring earlier. It might be fruitful to think of an advantage as emerging around young adulthood (from around 18 through around 40), but with the caveat that there is no precise "age" at which an advantage occurs for any one person. This idea of a gradual — and highly individualistic — emergence of a musician's advantage can also help explain the mixed results in the literature: studies examining college-age cohorts (ages 18-22) might be sampling too young of participants to catch it.

On the flip side, our results suggest an advantage might also *wane* with increased age, particularly for selective attention tasks. For example, the ability to inhibit a competing talker decreases with age (Tun et al., 2002). At the same time, other work has shown that musician's advantages persist for older adults. For example, Parbery-Clark and colleagues (2011) found that musicians (ages 45-65) showed consistent advantages in perceiving both words and sentences in 4-talker babble, relative to their non-musician (age-matched) counterparts. In our study, we do see a consistent musician's advantage for f0 separation for the competing vowels. Therefore, it might be more appropriate to think of multiple types of musician advantages — ones based on acoustic properties and attentional demands — that emerge and wane with age. Indeed, as nearly all studies are cross-sectional, the role of individual differences among participants cannot be understated. However, that is not to say that beneficial effects of musical training are limited to a younger age range. For example, Zendel and colleagues (2019) found that non-musician older adults (ages 55-75) who received 6 months of musical training show greater improvement in speech-in-speech (relative to control groups), suggesting that with new training, neuroplastic changes are possible across a wide range of ages.

That we see a nuanced musician's advantage — one that is shaped by f0, participant age, as well as task — sheds light on theories of music-to-speech transfer. In both experiments, musicians closely attend to f0 to improve performance. That the advantages we see are tied to f0 (but less so for other acoustic properties) supports accounts that propose a music-to-speech transfer for distinctions that have

greater ‘precision’ in music (i.e., pitch) than speech (Patel, 2011, 2012, 2014), but not the other way around (e.g., speech has more spectral distinctions than music; Zatorre et al., 2002). Additionally, the type of task, in providing varying attentional demands (e.g., selective attention), shapes the way listeners leverage acoustic cues (here, f0) (Kraus & Nicol, 2014; Kraus & White-Schwoch, 2015). Taken together, these findings support a hybrid *shared auditory encoding* and *domain-general attentional* account of music-to-speech transfer.

5.1. Limitations and Future Directions

There are several limitations of the present study that can serve as avenues for future research. First, we see that participants show difficulty in perceiving synthetic vowels in isolation. While we modeled the vowels based on the California Vowel System (CVS) (Eckert, 2008; Holland, 2014; Podesva, 2011; Villarreal, 2018), the English variety our participants were most familiar with, we see systematic confusions. One feature of the CVS is the backing of the /æ/ vowel in “bat” (known as TRAP-backing). Indeed, we see confusions of /æ/ for /ɑ/ for participants who completed the task (23.1% YA non-musicians, 11.5% YA musicians, 10.7% OA non-musicians, 16.7% OA musicians) and by far the most common confusion for the excluded individuals who did not pass with 90% accuracy (66.7%). Another feature of the CVS is front lax vowel lowering (e.g., for the vowels /ɪ/ and /ɛ/ in “bit” and “bet”, respectively). In the present study, non-musicians confused /æ/ as /ɛ/ (10.7%) at a much higher rate than YA musicians (2.5%), OA non-musicians (6%), and OA musicians (2.4%), suggesting that they are hearing a further CVS-shifted vowel. Indeed, perceiving a CVS-lowered vowel was the source of many ‘errors’ for the participants who did not pass with 90% accuracy: 22.7% confused /æ/ as “bet”. A third feature of CVS is back vowel fronting (e.g., the vowel /u/ in “boot” is fronted). While all age/musician groups display above 93% accuracy in identifying the intended /u/ vowel as “boot”, the most common confusion was with “bet”, indicating that they did not always perceive it as a back vowel (5.1% YA non-musicians, 0% YA musicians, 1.2% OA non-musicians, 4.8% OA musicians). Therefore, we see that perception of CVS features from the synthesized vowels is not consistent across the vowel space, and might also vary by both age and musical background. Future work can test whether a musician’s advantage might be present for more peripheral vowels, as is common in singing (e.g., for female speakers/singers in Marczyk et al., 2022). Furthermore, future work providing more phonetic context (e.g., playing longer samples of the talker) can better signal the speaker as belonging to a particular language/dialect variety than isolated words. Finally, our findings suggest that strict ‘cut-off’ points (e.g., 90% single vowel identification accuracy) and lack of feedback can result in a large number of participants who are excluded from the task.

Furthermore, the difficulty YA non-musicians in particular faced with the synthetic vowels underscores the importance of accounting for accuracy in perceiving vowels in isolation in models of speech-in-noise perception. To our knowledge, studies generally do not include a baseline accuracy measure directly in the models (e.g., Madsen et al., 2017; Parbery-Clark, Skoe, & Kraus, 2009). Here, without accounting for YA non-musicians’ difficulty, we could have incorrectly attributed their lower accuracy to a YA musicians’ advantage. We additionally control for single sentence accuracy in Experiment 1; indeed, we see that YA non-musicians perceive naturally recorded (but flattened f0) sentences equally well as their musician counterparts, suggesting that the unnatural stimuli made the double vowel identification task even more challenging. Future work using naturally recorded vowels can further probe whether a musicians’ advantage in stream separation might extend to identification.

Another limitation was our age range. Our division of the YA and OA age groups was centered around age 40, consistent with age differences found in related work (Zendel & Alain, 2012), giving us more of a middle-aged ‘OA’ group (Alain et al., 2001). While understudied, the interaction between age, musicianship, and task appears to be complex. In addition, it appears to be large enough to have a meaningful effect on observed outcomes and influences efforts to replicate findings across age groups and

experimental tasks. Most studies examine college-age students, and those that examine older cohorts vary in their age ranges. The current study suggests that age-related differences might not always go in the expected direction — and more cross-age category research is needed to better elucidate these differences.

Furthermore, we used small f0 separation levels based on the just-noticeable-differences (JND) for pure tones for musicians and non-musicians (Δf0 = 0.156 and 0.306 ST, respectively) (Kishon-Rabin et al., 2001). We see that increasing f0 separation beyond these levels — up to 3 ST — confers a benefit in sentence-in-speech perception (Experiment 1). As in other work (e.g., Assmann & Summerfield, 1990; Vongpaisal & Pichora-Fuller, 2007), increasing f0 separation beyond 1 ST for double vowels is less advantageous (Experiment 2). Yet, we see that at musician's JND for tones (Δf0 = 0.156 ST), musicians appear to show a small dip in both stream separation and double vowel identification, potentially indicating that the unexpected interval led to interference. Indeed, one musician participant reported that the "dissonance" between talkers was distracting at times, suggesting that the musician's advantage might be constrained to musical intervals they have trained on (e.g., a half step (1 ST), but not +0.156 ST). Still, another possibility is that these f0 separation levels were too small for linguistic stimuli. Related work has shown JNDs for participants (musical training not reported) that are larger for syllables, ranging from Δf0 = 1.23 STs for a child voice and Δf0 = 2.68 ST for a male voice (Gaudrain & Başkent, 2015). Taken together, these findings suggest that the JNDs vary for pure tones and across different levels of linguistic content (vowel, syllable, word, etc.). Future work directly comparing f0 separation, with and without the presence of meaningful stimuli (e.g., tones vs. vowels) and at an individual's JND and speech reception threshold (SRT), can start to tease this apart.

In addition, we see that while younger adults show an advantage for the 1-talker interferer (Experiment 1), older adult musicians appear to show a disadvantage. While the majority of studies examining speech-in-speech report a musician's advantage (e.g., Başkent & Gaudrain, 2016; Parbery-Clark, Skoe, & Kraus, 2009), or equal performance across musician and non-musician groups (e.g., Başkent et al., 2018, 2018; Boebinger et al., 2015), there is some work suggesting possible sources for a disadvantage. For example, Tufts and Skoe (2018) found that college-age musicians have greater noise exposure than non-musicians, particularly due to their experience in orchestras and bands. While all participants in the current study passed a pure tone hearing screening (Reilly et al., 2007), some work has shown that older adults and musicians have increased noise exposure and subclinical hearing loss, resulting in lower performance in speech-in-noise tasks (Drennan, 2022; Skoe et al., 2019). In the current study, some participants mentioned that they perform in front of loud instruments (e.g., brass, woodwind) in an orchestra or marching band setting, which is consistent with this interpretation. Future work both measuring subclinical hearing loss (e.g., using auditory brainstem responses; Skoe & Tufts, 2018), as well as directly asking participants about their noise exposure (e.g., Guest et al., 2018), can shed light on possible sources of an age-related disadvantage for musicians.

Relatedly, while there is increased interest in effects of specific types of musical training (e.g., pianists and violinists in Carey et al., 2015), musician participants in the current study were heterogeneous in terms of the instruments they play, how they play (alone vs. in group ensembles), as well as their background in singing (i.e., voice training). Future work examining a specific type of experience (e.g., singing only, string instruments only, etc.) might better translate to f0 separation in speech perception, particularly for vowels (which are the focus of singing as the most sonorant elements).

While this study examined the role of musical experience by non-tonal language speakers, other work has shown the impact of linguistic experience on f0 perception (Bidelman et al., 2013). For example, a recent paper provided support for a Cantonese advantage for prosodic prominence perception in English (Choi, 2022), particularly for falling pitch. The extent to which we observe similar language-based advantages for stream separation (e.g., on the basis of f0 separation) are avenues for future study.

Finally, while we can think of an advantage as 'emerging' around young adulthood (sometime between ages 18-39), we know that children show tremendous plasticity in learning language and music (e.g., Chobert et al., 2012; Ilari et al., 2016). Increasingly, there are parallels drawn between 'sensitive periods' for language and music acquisition, as well as cross-domain transfer (music-to-speech and vice versa; Chen et al., 2022; White et al., 2013). While speculative, these possibilities are ripe for future research and can shed light on the nature — and interplay — of differing types of complex auditory experience.

6. Conclusion

A musician's advantage for speech-in-speech perception depends on many factors, including the f0 properties of the competing voices, as well as the age of listeners and type of task. This work broadens our understanding of the impact — as well as limitations — of nonlinguistic experience on speech perception, and contributes to our understanding of cross-domain plasticity (e.g., music-to-speech).

Acknowledgments.

Thanks to Steven Ilogan for help with data collection and to Antoine Shahin for feedback on earlier stages of this project. Thank you also to Melissa Baese-Berk, Deniz Başkent, and an anonymous reviewer for their feedback on the paper. This work is supported by a research award to MC from the UC Davis Interdisciplinary Graduate and Professional Symposium.

Data Availability Statement

All data generated or analyzed during this study are included as Supplemental Material in this published article.

References

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(sup2), S53–S71.https://doi.org/10.1016/j.neurobiolaging.2019.05.015

Alain, C., McDonald, K. L., Ostroff, J. M., & Schneider, B. (2001). Age-related changes in detecting a mistuned harmonic. *The Journal of the Acoustical Society of America*, *109*(5), 2211–2216.

Anaya, E. M., Pisoni, D. B., & Kronenberger, W. G. (2016). Long-term musical experience and auditory and visual perceptual abilities under adverse conditions. *The Journal of the Acoustical Society of America*, *140*(3), 2074–2081.

Arehart, K. H., King, C. A., & McLean-Mudgett, K. S. (1997). Role of fundamental frequency differences in the perceptual separation of competing vowel sounds by listeners with normal hearing and listeners with hearing loss. *Journal of Speech, Language, and Hearing Research*, *40*(6), 1434–1444.

Assmann, P. F., & Summerfield, Q. (1989). Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *The Journal of the Acoustical Society of America*, *85*(1), 327–338.

Assmann, P., & Summerfield, Q. (1990). Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *The Journal of the Acoustical Society of America*, *88*(2), 680–697. https://doi.org/10.1121/1.399772

Assmann, P., & Summerfield, Q. (2004). The perception of speech under adverse conditions. In *Speech processing in the auditory system* (pp. 231–308). Springer.

Barreda, S. (2015). *phonTools: Functions for phonetics in R.* (0.2-2.1).

Başkent, D., Fuller, C. D., Galvin III, J. J., Schepel, L., Gaudrain, E., & Free, R. H. (2018). Musician

effect on perception of spectro-temporally degraded speech, vocal emotion, and music in young adolescents. *The Journal of the Acoustical Society of America*, *143*(5), EL311–EL316.

Başkent, D., & Gaudrain, E. (2016). Musician advantage for speech-on-speech perception. *The Journal of the Acoustical Society of America*, *139*(3), EL51–EL56. https://doi.org/10.1121/1.4942628

Başkent, D., van Engelshoven, S., & Galvin III, J. J. (2014). Susceptibility to interference by music and speech maskers in middle-aged adults. *The Journal of the Acoustical Society of America*, *135*(3), EL147–EL153.

Bergman, M., Blumenfeld, V. G., Cascardo, D., Dash, B., Levitt, H., & Margulies, M. K. (1976). Age-related decrement in hearing for speech: Sampling and longitudinal studies. *Journal of Gerontology*, *31*(5), 533–538.

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of Training between Music and Speech: Common Processing, Attention, and Memory. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00094

Bianchi, F., Santurette, S., Wendt, D., & Dau, T. (2016). Pitch Discrimination in Musicians and Non-Musicians: Effects of Harmonic Resolvability and Processing Effort. *Journal of the Association for Research in Otolaryngology*, *17*(1), 69–79. https://doi.org/10.1007/s10162-015-0548-2

Bidelman, G. M., & Alain, C. (2015). Musical Training Orchestrates Coordinated Neuroplasticity in Auditory Brainstem and Cortex to Counteract Age-Related Declines in Categorical Vowel Perception. *Journal of Neuroscience*, *35*(3), 1240–1249. https://doi.org/10.1523/JNEUROSCI.3292-14.2015

Bidelman, G. M., Hutka, S., & Moreno, S. (2013). Tone language speakers and musicians share enhanced perceptual and cognitive abilities for musical pitch: Evidence for bidirectionality between the domains of language and music. *PloS One*, *8*(4), e60676.

Bidelman, G. M., & Krishnan, A. (2010). Effects of reverberation on brainstem representation of speech in musicians and non-musicians. *Brain Research*, *1355*, 112–125.

Bidelman, G. M., Weiss, M. W., Moreno, S., & Alain, C. (2014). Coordinated plasticity in brainstem and auditory cortex contributes to enhanced categorical speech perception in musicians. *European Journal of Neuroscience*, *40*(4), 2662–2673.

Bidelman, G. M., & Yoo, J. (2020). Musicians Show Improved Speech Segregation in Competitive, Multi-Talker Cocktail Party Scenarios. *Frontiers in Psychology*, *0*. https://doi.org/10.3389/fpsyg.2020.01927

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *The Journal of the Acoustical Society of America*, *137*(1), 378–387. https://doi.org/10.1121/1.4904537

Boersma, P., & Weenink, D. (2021). *Praat: Doing phonetics by computer* (6.1.40). http://www.praat.org/

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066. https://doi.org/10.1121/1.428288

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, *20*(3–4), 255–272.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound* (Vol. 159). MIT press Cambridge, MA.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

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

Carlile, S., & Corkhill, C. (2015). Selective spatial attention modulates bottom-up informational masking of speech. *Scientific Reports*, *5*, 8662.

Chen, J., Scheller, M., Wu, C., Hu, B., Peng, R., Liu, C., Liu, S., Zhu, L., & Chen, J. (2022). The relationship between early musical training and executive functions: Validation of effects of the sensitive period. *Psychology of Music*, *50*(1), 86–99. https://doi.org/10.1177/0305735620978690

Chobert, J., François, C., Velay, J.-L., & Besson, M. (2012). Twelve months of active musical training in 8-to 10-year-old children enhances the preattentive processing of syllabic duration and voice onset time. *Cerebral Cortex*, *24*(4), 956–967.

Choi, W. (2022). Theorizing positive transfer in cross-linguistic speech perception: The Acoustic-Attentional-Contextual hypothesis. *Journal of Phonetics*, *91*, 101135.

Clayton, K. K., Swaminathan, J., Yazdanbakhsh, A., Zuk, J., Patel, A. D., & Kidd Jr, G. (2016). Executive function, visual attention and the cocktail party problem in musicians and non-musicians. *PloS One*, *11*(7), e0157638.

Coffey, E. B. J., Mogilever, N. B., & Zatorre, R. J. (2017). Speech-in-noise perception in musicians: A review. *Hearing Research*, *352*, 49–69. https://doi.org/10.1016/j.heares.2017.02.006

Cohn, M. D. (2018a). *Investigating the Effect of Musical Training on Speech-in-Speech Perception: The Role of f0, Timing, and Spectral Cues* [Ph.D., University of California, Davis]. https://search.proquest.com/docview/2132930892/abstract/798232B0A5874FB1PQ/1

Cohn, M. D. (2018b). Investigating a possible “musician advantage” for speech-in-speech perception: The role of f0 separation. *Proceedings of the Linguistic Society of America*, *3*(1), 24-1–9. https://doi.org/10.3765/plsa.v3i1.4311

Couth, S., Prendergast, G., Guest, H., Munro, K. J., Moore, D. R., Plack, C. J., Ginsborg, J., & Dawes, P. (2020). Investigating the effects of noise exposure on self-report, behavioral and electrophysiological indices of hearing damage in musicians with normal audiometric thresholds. *Hearing Research*, *395*, 108021.

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913–2922.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., & Aikawa, K. (1997). Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. *The Journal of the Acoustical Society of America*, *101*(5), 2839–2847. https://doi.org/10.1121/1.418517

Deroche, M. L., Limb, C. J., Chatterjee, M., & Gracco, V. L. (2017). Similar abilities of musicians and non-musicians to segregate voices by fundamental frequency. *The Journal of the Acoustical Society of America*, *142*(4), 1739–1755.

Drennan, W. R. (2022). Identifying Subclinical Hearing Loss: Extended Audiometry and Word Recognition in Noise. *Audiology and Neurotology*, *27*(3), 217–226.

Dubno, J. R., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, *76*(1), 87–96. https://doi.org/10.1121/1.391011

Eckert, P. (2008). Where do ethnolects stop? *International Journal of Bilingualism*, *12*(1–2), 25–42.

Gaudrain, E., & Başkent, D. (2015). Factors limiting vocal-tract length discrimination in cochlear implant simulations. *The Journal of the Acoustical Society of America*, *137*(3), 1298–1308.

Grassi, M., Meneghetti, C., Toffalini, E., & Borella, E. (2017). Auditory and cognitive performance in elderly musicians and nonmusicians. *PLOS ONE*, *12*(11), e0187881. https://doi.org/10.1371/journal.pone.0187881

Guest, H., Dewey, R. S., Plack, C. J., Couth, S., Prendergast, G., Bakay, W., & Hall, D. A. (2018). The Noise Exposure Structured Interview (NESI): An instrument for the comprehensive estimation of lifetime noise exposure. *Trends in Hearing*, *22*, 2331216518803213.

Gygi, B., & Shafiro, V. (2014). Spatial and temporal modifications of multitalker speech can improve speech perception in older adults. *Hearing Research*, *310*, 76–86.

Heidari, A., Moossavi, A., Yadegari, F., Bakhshi, E., & Ahadi, M. (2020). Effect of vowel auditory training on the speech-in-noise perception among older adults with normal hearing. *Iranian Journal of Otorhinolaryngology*, *32*(111), 229.

Helfer, K. S. (2015). Competing speech perception in middle age. *American Journal of Audiology*, *24*(2), 80–83.

Helfer, K. S., & Freyman, R. L. (2009). Lexical and indexical cues in masking by competing speech. *The Journal of the Acoustical Society of America*, *125*(1), 447–456. https://doi.org/10.1121/1.3035837

Helfer, K. S., & Freyman, R. L. (2014). Stimulus and listener factors affecting age-related changes in competing speech perception. *The Journal of the Acoustical Society of America*, *136*(2), 748–759. https://doi.org/10.1121/1.4887463

Helfer, K. S., & Jesse, A. (2021). Hearing and speech processing in midlife. *Hearing Research*, *402*, 108097.

Helfer, K. S., Merchant, G. R., & Wasiuk, P. A. (2017). Age-related changes in objective and subjective speech perception in complex listening environments. *Journal of Speech, Language, and Hearing Research*, *60*(10), 3009–3018.

Helfer, K. S., & Wilber, L. A. (1990). Hearing Loss, Aging, and Speech Perception in Reverberation and Noise. *Journal of Speech, Language, and Hearing Research*, *33*(1), 149–155. https://doi.org/10.1044/jshr.3301.149

Holland, C. L. (2014). *Shifting or Shifted? The state of California vowels* [Ph.D., University of California, Davis]. https://search.proquest.com/docview/1665571797/abstract/690B582BA9FF4E1CPQ/1

Ilari, B. S., Keller, P., Damasio, H., & Habibi, A. (2016). The Development of Musical Skills of Underprivileged Children Over the Course of 1 Year: A Study in the Context of an El Sistema-Inspired Program. *Frontiers in Psychology*, *7*. https://www.frontiersin.org/articles/10.3389/fpsyg.2016.00062

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004.

Kaplan, E. C., Wagner, A. E., Toffanin, P., & Başkent, D. (2021). Do Musicians and Non-musicians Differ in Speech-on-Speech Processing? *Frontiers in Psychology*, *0*. https://doi.org/10.3389/fpsyg.2021.623787

Kishon-Rabin, L., Amir, O., Vexler, Y., & Zaltz, Y. (2001). Pitch discrimination: Are professional musicians better than non-musicians? *Journal of Basic and Clinical Physiology and Pharmacology*, *12*(2), 125–144.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews Neuroscience*, *11*(8), 599.

Kraus, N., & Nicol, T. (2014). The cognitive auditory system: The role of learning in shaping the biology of the auditory system. In *Perspectives on auditory research* (pp. 299–319). Springer.

Kraus, N., Slater, J., Thompson, E. C., Hornickel, J., Strait, D. L., Nicol, T., & White-Schwoch, T. (2014). Music enrichment programs improve the neural encoding of speech in at-risk children. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *34*(36), 11913–11918. https://doi.org/10.1523/JNEUROSCI.1881-14.2014

Kraus, N., & White-Schwoch, T. (2015). Unraveling the biology of auditory learning: A cognitive-sensorimotor-reward framework. *Trends in Cognitive Sciences*, *19*(11), 642–654.

https://doi.org/10.1016/j.tics.2015.08.017
Kühnis, J., Elmer, S., Meyer, M., & Jäncke, L. (2013). The encoding of vowels and temporal speech cues in the auditory cortex of professional musicians: An EEG study. *Neuropsychologia*, *51*(8), 1608–1618. https://doi.org/10.1016/j.neuropsychologia.2013.04.007
Lee, J. H., & Humes, L. E. (2012). Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background. *The Journal of the Acoustical Society of America*, *132*(3), 1700–1717. https://doi.org/10.1121/1.4740482
Lentz, J. J., & Marsh, S. L. (2006). The effect of hearing loss on identification of asynchronous double vowels. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1354–1367.
Madsen, S. M. K., Whiteford, K. L., & Oxenham, A. J. (2017). Musicians do not benefit from differences in fundamental frequency when listening to speech in competing speech backgrounds. *Scientific Reports*, *7*(1), 12624. https://doi.org/10.1038/s41598-017-12937-9
Mandikal Vasuki, P. R. M., Mridula Sharma, Demuth, K., & Arciuli, J. (2016). Musicians' edge: A comparison of auditory processing, cognitive abilities and statistical learning. *Hearing Research*, *342*, 112–123. https://doi.org/10.1016/j.heares.2016.10.008
Marczyk, A., Belley, É., Savard, C., Roy, J.-P., Vaillancourt, J., & Tremblay, P. (2022). Learning transfer from singing to speech: Insights from vowel analyses in aging amateur singers and non-singers. *Speech Communication*, *141*, 28–39.
Medina, D., & Barraza, P. (2019). Efficiency of attentional networks in musicians and non-musicians. *Heliyon*, *5*(3), e01315.
Meister, H., Schreitmüller, S., Grugel, L., Ortmann, M., Beutner, D., Walger, M., & Meister, I. G. (2013). Cognitive resources related to speech recognition with a competing talker in young and older listeners. *Neuroscience*, *232*, 74–82. https://doi.org/10.1016/j.neuroscience.2012.12.006
Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*(1–2), 36–47.
Morse-Fortier, C., Parrish, M. M., Baran, J. A., & Freyman, R. L. (2017). The Effects of Musical Training on Speech Detection in the Presence of Informational and Energetic Masking. *Trends in Hearing*, *21*, 2331216517739427. https://doi.org/10.1177/2331216517739427
Münte, T. F., Altenmüller, E., & Jäncke, L. (2002). The musician's brain as a model of neuroplasticity. *Nature Reviews Neuroscience*, *3*(6), 473.
Mussoi, B. S. (2021). The Impact of Music Training and Working Memory on Speech Recognition in Older Age. *Journal of Speech, Language, and Hearing Research*, *64*(11), 4524–4534. https://doi.org/10.1044/2021_JSLHR-20-00426
Parbery-Clark, A., Skoe, E., & Kraus, N. (2009). Musical Experience Limits the Degradative Effects of Background Noise on the Neural Processing of Sound. *Journal of Neuroscience*, *29*(45), 14100–14107. https://doi.org/10.1523/JNEUROSCI.3256-09.2009
Parbery-Clark, A., Skoe, E., Lam, C., & Kraus, N. (2009). Musician Enhancement for Speech-In-Noise. *Ear and Hearing*, *30*(6), 653. https://doi.org/10.1097/AUD.0b013e3181b412e9
Parbery-Clark, A., Strait, D. L., Anderson, S., Hittner, E., & Kraus, N. (2011). Musical Experience and the Aging Auditory System: Implications for Cognitive Abilities and Hearing Speech in Noise. *PLOS ONE*, *6*(5), e18082. https://doi.org/10.1371/journal.pone.0018082
Patel, A. D. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology*, *2*, 142.
Patel, A. D. (2012). The OPERA hypothesis: Assumptions and clarifications. *Annals of the New York Academy of Sciences*, *1252*(1), 124–128.
Patel, A. D. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research*, *308*, 98–108.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184.
Podesva, R. J. (2011). The California vowel shift and gay identity. *American Speech*, *86*(1), 32–51.
R Core Team. (2021). *R: A language and environment for statistical computing (Version 4.0. 5)[Programming language]*.
Reilly, J., Troiani, V., Grossman, M., & Wingfield, R. (2007). An introduction to hearing loss and screening procedures for behavioral research. *Behavior Research Methods*, *39*(3), 667–672. https://doi.org/10.3758/BF03193038
Ruggles, D. R., Freyman, R. L., & Oxenham, A. J. (2014). Influence of Musical Training on Understanding Voiced and Whispered Speech in Noise. *PLOS ONE*, *9*(1), e86980. https://doi.org/10.1371/journal.pone.0086980
Schlaug, G. (2011). Music, musicians, and brain plasticity. *The Oxford Handbook of Music Psychology*, 197–208.
Shahin, A. J. (2011). Neurophysiological Influence of Musical Training on Speech Perception. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00126
Skoe, E., Camera, S., & Tufts, J. (2019). Noise Exposure May Diminish the Musician Advantage for Perceiving Speech in Noise. *Ear and Hearing*, *40*(4), 782–793. https://doi.org/10.1097/AUD.0000000000000665
Skoe, E., & Tufts, J. (2018). Evidence of noise-induced subclinical hearing loss using auditory brainstem responses and objective measures of noise exposure in humans. *Hearing Research*, *361*, 80–91. https://doi.org/10.1016/j.heares.2018.01.005
Slater, J., & Kraus, N. (2016). The role of rhythm in perceiving speech in noise: A comparison of percussionists, vocalists and non-musicians. *Cognitive Processing*, *17*(1), 79–87.
Strait, D. L., & Kraus, N. (2011). Can you hear me now? Musical training shapes functional brain networks for selective auditory attention and hearing speech in noise. *Frontiers in Psychology*, *2*, 113.
Strait, D. L., & Kraus, N. (2014). Biological impact of auditory expertise across the life span: Musicians as a model of auditory learning. *Hearing Research*, *308*, 109–121. https://doi.org/10.1016/j.heares.2013.08.004
Summers, V., & Leek, M. R. (1998). F0 Processing and the Seperation of Competing Speech Signals by Listeners With Normal Hearing and With Hearing Loss. *Journal of Speech, Language, and Hearing Research*, *41*(6), 1294–1306. https://doi.org/10.1044/jslhr.4106.1294
Tierney, A., Rosen, S., & Dick, F. (2020). Speech-in-speech perception, nonverbal selective attention, and musical training. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(5), 968.
Tierney, A. T., Krizman, J., & Kraus, N. (2015). Music training alters the course of adolescent auditory development. *Proceedings of the National Academy of Sciences*, *112*(32), 10062–10067. https://doi.org/10.1073/pnas.1505114112
Tufts, J. B., & Skoe, E. (2018). Examining the noisy life of the college musician: Weeklong noise dosimetry of music and non-music activities. *International Journal of Audiology*, *57*(sup1), S20–S27.
Tun, P. A., O'kane, G., & Wingfield, A. (2002). Distraction by Competing Speech in Young and Older Adult Listeners. *Psychology and Aging*, *17*(3), 453–467. https://insights.ovid.com/psychology-aging/psyag/2002/09/000/distraction-competing-speech-young-older-adult/9/00002004
Villarreal, D. (2018). The construction of social meaning: A matched-guise investigation of the California Vowel Shift. *Journal of English Linguistics*, *46*(1), 52–78.
Vongpaisal, T., & Pichora-Fuller, M. K. (2007). Effect of Age on F0 Difference Limen and Concurrent

Vowel Identification. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1139–1156. https://doi.org/10.1044/1092-4388(2007/079)

White, E. J., Hutka, S. A., Williams, L. J., & Moreno, S. (2013). Learning, neural plasticity and sensitive periods: Implications for language acquisition, music training and transfer across the lifespan. *Frontiers in Systems Neuroscience*, *7*, 90.

Wong, P. C., Skoe, E., Russo, N. M., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience*, *10*(4), 420.

Yeend, I., Beach, E. F., Sharma, M., & Dillon, H. (2017). The effects of noise exposure and musical training on suprathreshold auditory processing and speech perception in noise. *Hearing Research*, *353*, 224–236. https://doi.org/10.1016/j.heares.2017.07.006

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, *6*(1), 37–46.

Zendel, B. R., & Alain, C. (2009). Concurrent sound segregation is enhanced in musicians. *Journal of Cognitive Neuroscience*, *21*(8), 1488–1498.

Zendel, B. R., & Alain, C. (2012). Musicians experience less age-related decline in central auditory processing. *Psychology and Aging*, *27*(2), 410–417. https://doi.org/10.1037/a0024816

Zendel, B. R., & Alain, C. (2014). Enhanced attention-dependent activity in the auditory cortex of older musicians. *Neurobiology of Aging*, *35*(1), 55–63. https://doi.org/10.1016/j.neurobiolaging.2013.06.022

Zendel, B. R., Tremblay, C.-D., Belleville, S., & Peretz, I. (2015). The impact of musicianship on the cortical mechanisms related to separating speech from background noise. *Journal of Cognitive Neuroscience*, *27*(5), 1044–1059.

Zendel, B. R., West, G. L., Belleville, S., & Peretz, I. (2019). Musical training improves the ability to understand speech-in-noise in older adults. *Neurobiology of Aging*, *81*, 102–115. https://doi.org/10.1016/j.neurobiolaging.2019.05.015

Appendix A. Formant frequency values used to synthesize vowel tokens.

Vowel	F1	F2	F3	F4	F5	Reference Word
/i/	350	2400	2500	3500	4500	beat
/ɛ/	550	1850	2500	3500	4500	bet
/æ/	800	1780	2500	3500	4500	bat
/ɑ/	850	1380	2500	3500	4500	bought
/u/	400	1600	2250	3500	4500	boot