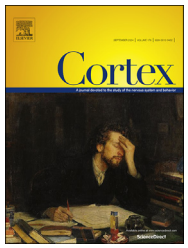
CORTEX 178 (2024) 213–222

Available online at www.sciencedirect.com

ScienceDirect

Journal homepage: www.elsevier.com/locate/cortex



Research Report

Voice actors show enhanced neural tracking of pitch, prosody perception, and music perception

Magdalena Kachlicka and Adam Tierney *

School of Psychological Sciences, Birkbeck, University of London, London, UK

ARTICLE INFO

Article history:
Received 6 March 2024
Reviewed 7 May 2024
Revised 28 May 2024
Accepted 26 June 2024
Action editor Sonja Kotz
Published online 10 July 2024

Keywords:
Auditory
Speech
Music
FFR
Actors

ABSTRACT

Experiences with sound that make strong demands on the precision of perception, such as musical training and experience speaking a tone language, can enhance auditory neural encoding. Are high demands on the precision of perception necessary for training to drive auditory neural plasticity? Voice actors are an ideal subject population for answering this question. Voice acting requires exaggerating prosodic cues to convey emotion, character, and linguistic structure, drawing upon attention to sound, memory for sound features, and accurate sound production, but not fine perceptual precision. Here we assessed neural encoding of pitch using the frequency-following response (FFR), as well as prosody, music, and sound perception, in voice actors and a matched group of non-actors. We find that the consistency of neural sound encoding, prosody perception, and musical phrase perception are all enhanced in voice actors, suggesting that a range of neural and behavioural auditory processing enhancements can result from training which lacks fine perceptual precision. However, fine discrimination was not enhanced in voice actors but was linked to degree of musical experience, suggesting that low-level auditory processing can only be enhanced by demanding perceptual training. These findings suggest that training which taxes attention, memory, and production but is not perceptually taxing may be a way to boost neural encoding of sound and auditory pattern detection in individuals with poor auditory skills.

1. Introduction

The auditory system is home to foundational human abilities, including speech and music. Nevertheless, there are large, reliable individual differences in the robustness of auditory neural encoding (Easwar, Scollie, Aiken, & Purcell, 2020; Hornickel, Knowles, & Kraus, 2012). This variability in auditory precision is a bottleneck for the development of skills crucial for communicating in everyday life, including pitch perception (Coffey, Colagrosso, Lehmann, Schonwiesner, & Zatorre, 2016; Krishnan, Bidelman, Smalt, Ananthakrishnan, & Gandour, 2012), temporal acuity (Purcell, John, Schneider, & Picton, 2004), second language speech perception (Kachlicka, Saito, & Tierney, 2019; Omote, Jasmin, & Tierney, 2017), and speech-in-noise perception (Anderson, Parbery-Clark, Yi, & Kraus, 2011; Ruggles, Bharadwaj, & Shinn-Cunningham, 2012). Understanding the factors which drive variability in auditory neural encoding, therefore, could lead

* *Corresponding author.*
E-mail address: a.tierney@bbk.ac.uk (A. Tierney).
https://doi.org/10.1016/j.cortex.2024.06.016

to targeted approaches to boost auditory system function in individuals who struggle to perceive sound.

Individual differences in auditory neural encoding reflect not only innate variability, but also the effects of experience with sound. For example, tone language speakers must track pitch on a finer scale, compared to non-tone-language speakers, in order to distinguish lexically-relevant tone contours. As a result, neural representation of the fundamental frequency (F0, an acoustic correlate of pitch) is enhanced in tone language speakers (Bidelman, Gandour, & Krishnan, 2010; Chandrasekaran, Krishnan, & Gandour, 2009; Krishnan, Xu, Gandour, & Cariani, 2005, 2009). This enhanced pitch representation could help explain the enhanced pitch processing skills demonstrated by tone language speakers. These include more precise pitch discrimination (Creel, Weng, Fu, Heyman, & Lee, 2018), enhanced melody memory (Liu, Hilton, Bergelson, & Mehr, 2023), more accurate singing (Pfordresher & Brown, 2009), and greater use of pitch as a cue to speech and music perception (Jasmin, Sun, & Tierney, 2021; Petrova, Jasmin, Saito, & Tierney, 2023).

Effects of experience on the robustness of auditory neural encoding are not limited to experiences begun in infancy and early childhood. For example, music perception requires tracking of pitch on a finer scale than does speech perception. Cross-culturally, music tends to feature discrete pitch variations which listeners map onto fixed musical intervals (Ozaki et al., 2024). Speech, on the other hand, features continuous changes in pitch, with intonational structure conveyed by changes in the direction of pitch contour (Zatorre & Baum, 2012). As a result, musicians must learn to very precisely perceive pitch: a difference of less than one semitone can separate a note which is in tune versus out of tune, while similarly sized pitch changes in non-tonal languages are inconsequential. The need to track such fine gradations in pitch may enhance auditory neural encoding. Indeed, musicians have more robust pitch representations, compared to non-musicians, as measured using the mismatch negativity (MMN; Fujioka, Trainor, Ross, Kakigi, & Pantec, 2004; Magne, Schön, & Besson, 2006; Chandrasekaran et al., 2009) and the frequency-following response (FFR; Musacchia, Sams, Skoe, & Kraus, 2007; Wong, Skoe, Russo, Dees, & Kraus, 2007; Skoe & Kraus, 2012). Although most prior research on musical experience and auditory neural encoding has been cross-sectional, longitudinal studies have demonstrated that music training can lead to enhancements of auditory function and sound perception (Nan et al., 2018; Tierney, Krizman, & Kraus, 2015).

What conditions must be met for an experience to enhance neural representation of sound and sound perception? Both musical training and tone language experience make strong demands on the precision of pitch perception. One possibility, therefore, is that high demands on the precision of perception drive auditory neural plasticity. An alternate possibility, however, is that auditory neural encoding can be enhanced by training that taxes sound perception without requiring high perceptual precision, instead requiring attention to sound, memory for sound features, and accurate sound production. This possibility would be consistent with Reverse Hierarchy Theory, which suggests that attention can enhance processing at low levels of perceptual systems (Ahissar & Hochstein, 2004).

Determining whether high perceptual precision is a necessary ingredient of training for auditory neural enhancement would require testing a population who receives specialized sound training but without needing to perform precise sound discrimination. One such population is actors, who must exaggerate their use of prosodic cues (Berry & Brown, 2019; Jürgens, Grass, Drolet, & Fischer, 2015; Matharu, Berry, & Brown, 2022) to clearly communicate phrase structure and pragmatics, to produce speech with maximal emotional intensity and arousal (Whiting, Kotz, Gross, Giordano, & Belin, 2020), and to convey a character's personality. To learn which characteristics to exaggerate, actors must explicitly attend to the relevant acoustic characteristics, including pitch, duration, and amplitude (Breen Fedorenko, Wagner, & Gibson, 2010; Fear, Cutler, & Butterfield, 1995; Mattys, 2000). For example, judgments of a speaker's dominance versus trustworthiness are linked to distinctive pitch contours, with the former characterized by a rapid decrease in pitch, and the later by a moderate rise in pitch (Ponsot, Burred, Belin, & Aucouturier, 2018); exaggerating these features would clearly convey to a listener that a character is dominant or trustworthy, even in the absence of cues from facial expression, gesture, or gait. In order to learn to clearly convey dominance and trustworthiness, therefore, actors must selectively direct attention towards pitch and away from less relevant acoustic characteristics, detect the difference in pitch contour associated with the two characteristics, and exaggerate it.

Moreover, prosodic exaggeration requires explicit control over the production of cues to prosodic features. Prior neural evidence suggests that this enhanced control over the production of sound could also have consequences for auditory perception: professional actors showed greater premotor activation when listening to speech versus violin music, whereas the opposite was true for professional violinists (Dick, Lee, Nusbaum, & Price, 2011), suggesting the use of motor resources during sound perception. Acting training also taxes memory, as actors must memorize long passages of speech, including relevant prosodic cues to information structure and emotion; indeed, prior evidence suggests enhanced verbal memory in actors compared to non-actors (Groussard, Coppalle, Hinault, & Platel, 2020; Noice & Noice, 2009), although it remains unclear whether this advantage extends to memory for non-verbal sounds.

To test whether speech training which does not tax auditory precision can enhance auditory neural encoding and sound perception, we tested voice actors and non-actors matched for age and degree of musical training. (The need to exaggerate prosodic cues is especially strong for voice actors, who cannot use facial expression or gesture and so must rely on sound to communicate emotion, character, and phrase structure.) We assessed the consistency and amplitude of their neural encoding of sound using the FFR, as well as behaviourally assessing their acoustic discrimination, memory for sound patterns, and ability to categorize prosodic and musical patterns.

2. Methods

2.1. Participants

Participants were professional voice actors ($n = 21$, 13 female) or individuals without voice acting experience ($n = 21$, 7 female) living in London. All participants were native English speakers; 10 participants in the voice actor group and 6 in the non-actor group reported some degree of bilingual experience. The voice actors and non-actors did not significantly differ in gender ($\chi 2 = 3.44$, $p = .064$) or bilingualism ($\chi 2 = 1.62$, $p = .20$). The groups did not differ in age, with voice actors reporting a mean age of 36.3 years (sd = 9.2) while non-actors reported a mean age of 33.5 years (11.1; t (40) = .91, $p > .05$). The groups also did not differ in musical training, with voice actors reporting an average of 3.9 (8.9) years of musical training, while non-actors reported 4.0 (7.7) years of musical training (t (40) = −.06, $p > .05$). Voice actors reported 9.1 (6.8) years of acting experience (range = 3 to 30) and had obtained some formal vocal training (e.g., acting performance, accent imitation). Voice actors reported that they began their training, on average, in early adulthood (mean = 18.6 years, SD = 7.5). The control group reported no voice acting training or experience.

We report below how we determined our sample size, all data exclusions (if any), all inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. Sample size was determined as the largest number of participants we could recruit, given time/resource limitations, with sample size primarily limited by the feasibility of recruiting voice actors. No data were excluded. Inclusion/exclusion criteria were established prior to data analysis.

Detailed information on acting training programmes was available from only a subset of participants. Within this subset, there was a substantial degree of variability in the type of training they obtained. Most of them completed programs in arts, drama, theatre, or stage performance. Along with modules focused on voice control, they also completed additional specialized voice acting courses, seminars, and professional development practices (e.g., voice and accent coaching, pronunciation).

2.2. Behavioural measures

2.2.1. Test of musical aptitude

The Musical Ear Test (Wallentin, Hojlund Nielsen, Friis-Olivarius, Vuust, & Vuust, 2010) measures musical abilities across two fundamental aspects of music – melody (including pitch and contour) and rhythm (Krumhansl, 2000). On each trial participants decided whether two short melodic or rhythmic phrases they heard were identical by clicking an appropriate button on the screen. The test consisted of 104 trials, 52 in each subtest, half of which were "same" trials and half of which were "different" trials. Each subtest started with two examples with feedback and there was no feedback thereafter. The melodic phrases contained 3–8 tones played with sampled piano sounds. The "different" trials had one pitch violation, half of which (13 trials) involved a contour violation while the remainder did not. 25 trials contained non-diatonic tones, while 20 were in a major and 7 in a minor key. The rhythmic phrases consisted of 4–11 wood block beats. The "different" trials had one rhythmic change. Rhythmic complexity differed across trials – there were triplets in 21 trials and the remaining 31 trials only contained even beat subdivisions. 37 trials began on the downbeat whereas the rest began later. All melodic and rhythm sequences had a duration of one measure and were played at 100 beats per minute. The order of trials within each subtest was randomized. Performance was assessed as percent correct and averaged across the melody and rhythm sub-tests to form an overall measure of musical aptitude.

2.2.2. Musical phrase perception test

A musical phrase perception test (Jasmin, Dick, Holt, & Tierney, 2020) was used to assess participants' use of pitch and temporal cues in perception of musical phrase boundaries. The stimuli were 150 musical phrases taken from a corpus of folk songs (Schaffrath & Park, 1995) synthesised as sequences of six harmonic complex tones with 10 ms on/off cosine ramps. To include trials with a variety of difficulty levels there were three conditions that manipulated the available acoustic information – a Pitch Only condition, a Duration Only condition, and a Both Cues condition (50 trials in each condition). Musical phrases in the Both Cues condition contained both pitch and duration information. Phrases in the Pitch Only condition had natural pitch variations, but the durations were set to be isochronous and equal to the mean duration of the notes in the original melody. Conversely, the Duration Only condition retained the original note durations while setting the pitch to a monotone at 220 Hz. Additionally, half of the stimuli in each condition formed a complete musical phrase with the notes in an unmodified sequential order—these could be perceived as a complete musical phrase. The remaining stimuli were made to sound incomplete by combining the second half of one musical phase and the first half of the next musical phrase in the song. The order of the notes within the two halves was preserved, resulting in stimuli with a phrase boundary in the middle of the sequence, rather than at the end.

Phrases were presented in the same condition order across participants (Both Cues, then Duration Only, and then Pitch Only condition) with short breaks between the conditions. Within each condition, half of the trials were complete, and half incomplete. On each trial participants heard a single phrase once, and then indicated how complete the phrase sounded by clicking on a red bar on the screen. The red bar represented the degree of perceived phrase completeness on a continuum from incomplete (left side of the bar) to complete (right side of the bar). Participants could click anywhere on the bar to indicate their responses. We opted for a continuous measure of completeness to avoid potential ceiling/floor effects attributable to listeners' bias to hear phrases as complete or incomplete. The main outcome measure was the raw rating difference between complete and incomplete trials for each condition. This outcome measure was averaged across the three conditions to form a composite measure of musical phrase perception.

2.2.3. Auditory processing battery

Pitch and duration discrimination thresholds were estimated using the adaptive maximum likelihood procedure (Green, 1993) as implemented in the MLP Matlab toolbox (Grassi & Soranzo, 2009). The tests use an adaptive three-alternative forced-choice design, with the difficulty decreasing after every incorrect response and increasing after every correct response. The difference between the tones was adaptively reduced until the smallest discriminable difference between the frequencies and duration was established.

The stimuli for the pitch discrimination test were 250-ms complex tones with four harmonics (gated on and off with 10-ms raised cosine ramps). The standard tone was 330 Hz and the range of comparison tones extended to 390.1 Hz. The stimuli for the duration discrimination test were complex tones with four harmonics and F0 set at 330 Hz (ramped with 10-ms cosine onset and offset gates). The durations ranged from 250 ms (standard tone) to 450.1 ms. In the pitch discrimination task, participants heard three tones in each trial and were asked to identify the tone which had the highest pitch by pressing the appropriate number on the keyboard ('1', '2', or '3'). In the duration discrimination task, participants heard two tones and had to indicate which tone was longer by pressing '1' or '2' on the keyboard.

In each condition, participants performed three blocks of 30 trials. In addition, 20% of the trials across the three blocks were catch trials, in which there was no change in the duration or frequency of the tones. Stimulus level thresholds corresponded to the midpoint of the psychometric curve defined as a probability of correct responses across trials (66% for pitch discrimination and 63.1 % for duration discrimination; Grassi & Soranzo, 2009).

2.2.4. Prosody perception test

A prosody perception task was used to assess participants' ability to distinguish between phrases varying in placement of linguistic focus (first *vs* second word emphasized) and phrase boundaries (early *vs* late closure). Stimuli were taken from the Multidimensional Battery of Prosody Perception (MBOPP; Jasmin, Dick, & Tierney, 2020). All samples were recorded by a native British English-speaking male. The stimuli were pairs of recorded phrases which were identical lexically but differed in prosody. For phrase boundary stimuli silent pauses after phrase breaks were removed prior to further stimulus manipulation. These pairs of phrases were morphed onto one another with the speech morphing software STRAIGHT (Kawahara & Irino, 2005), by varying the degree to which pitch and duration were informative of linguistic focus or phrase boundary placement. We synthesized pairs of stimuli in three conditions in which different types of acoustic cues to prosodic categorization were present, to ensure a wide range of difficulty across trials. In the Pitch Only condition, the stimulus pair had the same duration as the average durations of the two original recordings but the pitch information was set at the morphing level of 80% versus 20%. (A morphing level of 100% would indicate that the pitch information was taken entirely from the late closure or second word emphasized recordings, while a level of 0% would indicate that the pitch contour exactly matched the early closure or first word emphasized recordings.) (2) In the Duration Only condition, pitch was set to the mean of the two original versions but the duration was set at the morphing levels of 80% versus 20%. (3) in the Both Cues condition, both pitch and time cues were available simultaneously (both set at 80% *vs* 20% morphing levels). The morphed samples varied only in duration and pitch, while other amplitude and spectral characteristics other than pitch were kept constant at 50% between the two original recordings and therefore remained uninformative.

There were 42 trials in the phrase boundary test and 47 trials in the linguistic focus test. Each test included three conditions: Pitch Only, Duration Only, and Combined Cues. During each trial a single sentence appeared on the screen for participants to read. They were asked to imagine how the sentence might sound if they were to read them aloud. After 5 s, participants heard a man speaking the first part of the sentence in two different ways. The task was to choose which recording best matched the sentence they just read by pressing either '1' or '2' on the keyboard for the first or second recording respectively. Prior to the main task, participants completed a short practice run to familiarize themselves with the task procedure. Performance was scored as portion correct, averaged across the Both Cues, Pitch Only, and Duration Only conditions, and averaged across the focus and phrase perception sub-tests to form an overall measure of prosody perception.

2.3. Electrophysiology

2.3.1. EEG data acquisition and pre-processing

The stimulus was the consonant-vowel syllable /da/ synthesised with a Klatt-based synthesiser. The syllable was 190 ms in duration with an F0 contour that was flat at 100 Hz for the initial 110 ms of the sound and then rose to 143 Hz over the course of the subsequent 80 ms. The stimulus began with a 5 ms onset burst. Between 5 and 50 ms F1 rose from 400 to 720 Hz, F2 fell from 1700 to 1240 Hz, and F3 fell from 2580 to 2500 Hz. Between 50 and 190 ms F1, F2, and F3 were stable at 720 Hz, 1240 Hz, and 2500 Hz, respectively. F4, F5, and F6 were constant between 5 and 190 ms at 3300 Hz, 3750 Hz, and 4900 Hz, respectively.

The stimulus was presented 6000 times over the course of 25 min at alternating polarities and at a rate of 3.8 Hz. The sound was delivered diotically through E-A-RTONE 3A insert earphones at 80 dB SPL. During the recording, participants were allowed to read a magazine or a book of their choice. They were asked to relax and stay still during the sound presentation. The data were recorded using a BioSemi ActiveTwo EEG system at a 16,384 Hz sample rate and with open filters in ActiView (BioSemi) acquisition software. A montage of five electrodes with a sintered Ag–AgCl pallet was used. One active electrode was placed at Cz, reference electrodes on both earlobes and two ground electrodes above the left and right eyebrows. Electrode contact impedance was kept below 20 kΩ throughout the session.

The mean of the reference channels was subtracted from the signal at Cz during offline data processing. Data were then bandpass-filtered using a first-order Butterworth filter with a low-pass of 70 Hz and a high-pass of 2000 Hz. Next, neural responses were divided into epochs beginning 10 ms before the onset of the sound and ending 240 ms after. Trials with an

amplitude of greater than 35 μV were rejected as artifacts, and 5000 total artifact-free sweeps were selected, 2500 from each polarity.

A windowed FFT analysis was then used to extract the phase consistency of the response over time. 40-ms segments of each epoch were extracted, starting at a center time point of 10 ms after the onset of the sound and ending 220 ms after sound onset, with a 1-ms step size between segments. For each segment, a Hanning windowed fast Fourier transform was used to calculate the time frequency spectrum. This procedure generates, for each trial, an amplitude value and a phase value. The resulting vectors were then transformed into unit vectors, which discards the amplitude but retains the phase value, and averaged. The length of the resulting vector was calculated as a measure of inter-trial phase locking, which varies from zero (no consistency) to one (perfect consistency). The result is a measure of the phase consistency of the response across trials for each time-frequency point.

To measure the amplitude spectrum, an average waveform was first calculated by taking the mean across all 5000 artifact-free sweeps. A windowed FFT analysis was then used to extract the amplitude of the response at each frequency over time. FFT settings were identical to those used for the phase consistency analysis (see last paragraph for details).

2.4. *General procedure*

Data were collected at the Department of Psychological Sciences at Birkbeck, University of London. All procedures were approved by the departmental Research Ethics Committee. Each testing session lasted approximately 3 h (with breaks between the tasks). All participants were reimbursed £25 for their time. All participants provided written consent prior to the testing session. Data and scripts are provided at https://osf.io/jevbm/. Study procedures and analyses were not preregistered. Research materials are provided at https://osf.io/eaqbj/.

3. Results

To test for a relationship between voice acting experience and the robustness of auditory processing, a series of linear regressions were run using the lm function in R (R Core Team, 2023). There were seven dependent measures, including prosody perception, musical phrase perception, musical aptitude, duration discrimination threshold, pitch discrimination threshold, FFR F0 consistency, and FFR F0 amplitude. Predictors included group (voice actor *vs* non-actor), as well as age and years of musical training. Fig. 1 displays individual data points for each dependent measure across voice actors and non-actors. Full regression tables are supplied in the Supplementary Information at https://osf.io/jevbm/, with significant predictors summarized below.

Voice actors performed better than non-actors at classifying patterns in both speech prosody and music, even after accounting for age and years of musical training. For prosody perception, the overall model predicted 37.4% of the variance (F (3,38) = 7.57, $p < .001$). Voice actors (M = 83.6% correct, SD = 9.9%) performed significantly better than non-actors (M = 74.3%, SD = 13.9%; $\beta = .110$, $p < .001$). The only additional significant predictor was age, indicating that older participants performed worse on the task ($\beta = -.006$, $p = .001$). For musical phrase perception, the overall model predicted

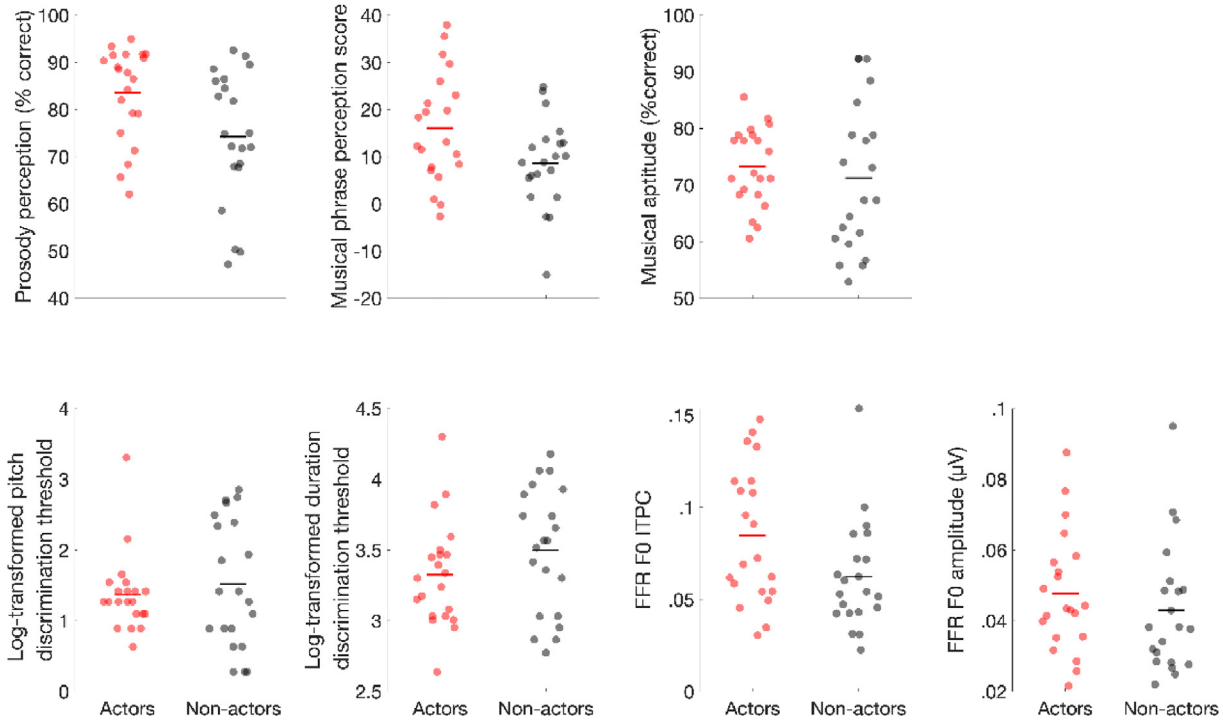


Fig. 1 – Performance across all seven main outcome variables in voice actors (left, red) and non-actors (right, black). The horizontal line indicates the mean.

16.7% of the variance (F (3,38) = 2.54, p = .07). Voice actors (M = 16.1, SD = 11.7) performed significantly better than non-actors (M = 8.7, SD = 9.3; β = 8.12, p = .018).

A few participants performed worse than chance on the prosody perception (n = 2) and musical phrase perception (n = 5) tasks, suggesting that they may have either struggled severely with these tasks or did not understand the instructions. To ensure that these few participants with very poor performance were not driving our results, we re-ran the prosody perception and musical phrase perception regressions after removing the participants who performed below chance on each task (performance less than .5 and less than 0, respectively). For prosody perception, voice actors once again performed significantly better than non-actors (β = .087, p = .007). Similarly, for musical phrase perception, voice actors again performed significantly better than non-actors (β = 7.36, p = .020).

Voice actors did not out-perform non-actors on more basic auditory processing tasks. For musical aptitude, the overall model predicted 10.5% of the variance (F (3,38) = 1.48, p = .24). No predictors were significant, although there was a trend for more years of musical training to predict better performance (β = .376, p = .068). For duration discrimination, the overall model predicted 9.3% of the variance (F (3,38) = 1.31, p = .29). No predictors were significant. For pitch discrimination, the overall model predicted 12.3% of the variance (F (3,38) = 1.87, p = .15). The only significant predictor was years of musical training (β = −.031, p = .029), indicating that a greater degree of musical experience was linked to more precise pitch perception.

Voice actors showed more consistent early auditory neural encoding of sound. To measure the consistency of neural encoding of the F0, the inter-trial phase coherence (ITPC) was averaged across a 10-Hz window around the stimulus F0 at each time point (after accounting for a 10-ms stimulus-response lag) and averaged across time points between 21 and 190 ms. The result was an overall measure of the consistency of encoding of the F0 across the entire response (See Fig. 2 for a depiction of ITPC across time and frequency points for voice actors and non-actors, as well as a spectrogram of the stimulus). The overall model predicted 12.3% of the variance in F0 encoding consistency (F (3,38) = 1.78, p = .17). Voice actors had more consistent neural encoding of the fundamental frequency compared to non-actors (β = .023, p = .034).

However, voice actors did not have larger frequency-following responses. To measure the amplitude of the frequency-following response, spectral amplitudes were averaged across a 10-Hz window around the stimulus F0 at each time point (after accounting for a 10-ms stimulus-response lag) and averaged across time points between 21 and 190 ms. The result was an overall measure of the amplitude of the neural encoding of F0 across the entire response. The overall model predicted only 3.6% of the variance in F0 encoding (F (3,38) = .47, p = .70). Voice actors and non-actors had frequency following responses with equivalent amplitude (β = .005, p = .342).

Finally, to determine the extent to which non-verbal auditory processing skill helped predict prosody perception ability, we ran two regressions with prosody perception scores from the linguistic focus and phrase boundary sub-tests as the dependent measures and duration discrimination, pitch discrimination, musical aptitude, the consistency of neural encoding of F0, and musical phrase perception as predictors. For phrase perception, the overall model predicted 31.5% of the variance (F (5,36) = 3.31, p = .015). Participants who were more successful at phrase perception had lower duration discrimination thresholds (β = −.160, p = .004). No other predictors were significant. For linguistic focus perception, the overall model predicted 38.9% of the variance (F (5,36) = 4.59, p = .002). Participants who were more successful at focus perception had more consistent neural encoding of F0 (β = 1.427, p = .016) and performed better at musical phrase perception (β = .005, p = .008).

4. Discussion

To clearly communicate emotion, character, and phrase structure, voice actors must attend to and explicitly exaggerate acoustic cues to prosodic features, a process which does not require precise sound perception. Nevertheless, we find that voice actors benefit from more consistent neural encoding of the fundamental frequency of speech, relative to matched non-actors. That the consistency of neural encoding of pitch is enhanced in actors suggests that precise demands on perception are not a necessary pre-condition for auditory neural enhancement, which can instead be driven by training attention to sound, memory for sound features, and accurate

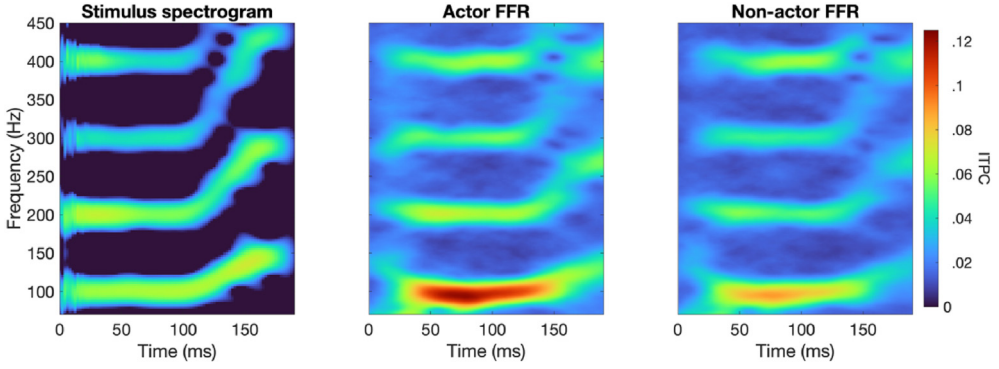


Fig. 2 – Stimulus spectrogram (left) and neural encoding of speech in voice actors (middle) and non-actors (right). The color at each time/frequency point indicates the ITPC at that frequency in a 40-ms window centered on that time point.

sound production. Our finding of frequency-following-response enhancement in voice actors, therefore, is consistent with Reverse Hierarchy Theory, which suggests that attention can have effects which extend to lower levels of perceptual systems (Ahissar & Hochstein, 2004). Our findings also suggest that developing training which is not perceptually demanding but taxes listeners' attention, memory, and production may be a way of boosting auditory processing in listeners with poor auditory skills who would struggle to complete more demanding auditory tasks. For example, individuals could be presented with highly exaggerated prosodic features such as greatly lengthened durations before phrase boundaries or large pitch excursions during emphasized words and explicitly trained to reproduce these patterns.

We find that voice acting training is linked to behavioural enhancements as well, with voice actors performing better on tests of both prosody and musical phrase perception. Prior research on the effects of acting experience has focused on changes in social and cognitive skills directly relevant to acting, including episodic memory recall (Banducci et al., 2017), verbal memory (Groussard et al., 2020; Noice & Noice, 2009), social relationships (Joronen, Konu, Rankin, & Astedt-Kurki, 2011), and theory of mind (Goldstein, Wu, & Winner, 2010). One prior study reported that drama training can lead to enhanced emotional prosody perception in 6-year-old children (Thompson, Schellenberg, & Husain, 2004), suggesting that there are perceptual as well as cognitive effects of theatre experience. Here, we find that the voice actor advantage is not limited to speech perception and neural encoding of speech but extends to music perception as well, suggesting that acting can lead to far transfer of learning to other domains (Barnett & Ceci, 2002). It remains to be seen whether the effects of voice acting training on the neural encoding of pitch extend to musical as well as speech stimuli.

Why would voice acting training have consequences for music perception, a task which is not directly relevant to voice acting? One possible explanation is that there is overlap between the acoustic cues used to convey certain prosodic features in language and in music. For example, in both domains notes/syllables feature increased duration and pitch movements just before a phrase boundary (De Pijper & Sanderman, 1994; Palmer & Krumhansl, 1987; Tierney, Russo, & Patel, 2011). By learning to explicitly attend to the acoustic cues conveying the presence of phrase boundaries in order to clearly produce and exaggerate them, voice actors may become primed to detect them even in another domain. A similar explanation could underlie the finding from longitudinal intervention studies that musical training can lead to gains in identification of emotional prosody (Mualem & Lavidor, 2015; Thompson et al., 2004), given overlap in the cues to emotion in language and music (Ilie & Thompson, 2006; Cuotinho & Dibben, 2013). If acoustic overlap between speech and music drives learning transfer from voice acting to music, these effects may extend beyond musical phrase perception to other musical features. For example, voice actors may be better able to perceive emotion in music, compared to non-actors. The acoustic cues to the presence of musical beats and stressed syllables are also strikingly similar, including changes in pitch, amplitude, and duration (Chrabaszcz, Winn, Lin, & Idsardi, 2014; Ellis & Jones, 2009; Fear et al., 1995; Hannon, Snyder, Eerola, & Krumhansl, 2004), and so voice actors may demonstrate advantages in perception and neural encoding of musical beats as well.

Our finding of enhanced consistency of neural responses to speech and musical phrase perception in voice actors suggests that precise perceptual demands are not a necessary precondition for certain types of auditory learning, given that actors produce exaggerated cues to prosody (Berry & Brown, 2019; Jürgens et al., 2015; Matharu et al., 2022). The effectiveness of voice actor experience in driving auditory learning may be due to the roles that attention, emotion, and repetition play in acting (Patel, 2011, 2014): emotion is commonly evoked during acting practice and performance, attention must be directed to acoustic cues in order to exaggerate them, and lines are repeated over and over again as actors learn a scene. However, the lack of precision required by voice acting training may limit the extent of the perceptual advantages which result. Here, we find that the voice actor advantage was limited to musical phrase perception, with the actor and non-actor groups performing similarly on tests of pitch and duration discrimination, as well as music aptitude. Actors' lack of enhanced perceptual precision distinguishes them from tone language speakers and musicians, both of whom demonstrate enhanced pitch discrimination (Creel et al., 2018; Micheyl, Delhommeau, Perrot, & Oxenham, 2006; Pfordresher & Brown, 2009); indeed, in our study we found that years of musical training were linked to pitch discrimination. Precise perceptual demands, therefore, may not be necessary for enhancement of auditory neural encoding or transfer of learning across domains, but may be necessary for boosting auditory discrimination.

Another possible explanation for the actor advantage for music perception, prosody perception, and consistency of neural sound encoding is that actors need to exercise explicit control over their production of prosodic features to exaggerate them. This specialized speech production experience could have consequences for sound perception, as acting training has been shown to modulate premotor activation during perception, even in the absence of overt movement (Dick et al., 2011). However, there is some inconsistency in the literature as to the extent to which pitch perception and production are related (Hutchins & Moreno, 2013); for example, pitch discrimination training has been reported to have no effect on the accuracy of melody production (Zarate, Delhommeau, Wood, & Zatorre, 2010), and difficulties with pitch perception can sometimes occur alongside preserved pitch production (Loui, Guenther, Mathys, & Schlaug, 2008). The Linked Dual Representation model (Hutchins & Moreno, 2013) accounts for these findings by proposing that there is a direct link between low-level pitch perception and vocal-motor production, a second indirect link with production mediated by categorical perception, and a "feedback" connection from production to low-level pitch perception. Our finding of more consistent frequency-following responses in voice actors is consistent with this model, as this enhancement could reflect an influence of feedback from pitch production to low-level encoding of pitch.

Although we find that voice actors show more consistent neural encoding of the F0 of speech, we find no difference between voice actors and non-actors in the amplitude of the F0 response. This contrasts somewhat with previous research

comparing musicians and non-musicians; musicians have been reported to show not only enhanced consistency of the FFR (Parbery-Clark, Anderson, Hittner, & Kraus, 2012; Skoe & Kraus, 2013; Tierney et al., 2015) but also greater F0 amplitude (Mussachia et al., 2007; Parbery-Clark, Strait, & Kraus, 2011). If F0 amplitude is enhanced in musicians but not voice actors, this could suggest that musical training enhances auditory neural encoding to a greater extent than voice acting training. However, this finding of greater F0 amplitude in musicians has not always been replicated, with other studies reporting F0 amplitude to be similar between musicians and non-musicians (Parbery-Clark, Skoe, & Kraus, 2009), or even enhanced in non-musicians relative to musicians (Parbery-Clark et al., 2012). To our knowledge, only two prior studies have directly compared neural encoding of sound in voice actors versus musicians. Dick et al. (2011) found voice actor training was linked to greater activation in the right planum temporale for dramatic speech relative to violin music, while the opposite was true for violin training, suggesting possible domain-specific effects of training type on auditory neural encoding of speech versus music. Rosslau et al. (2016) found greater pitch detection accuracy for singers than for actors, as well as prolonged late activity in right temporal and left parietal areas for singers. Based on the current literature, therefore, it remains unclear whether musical training has a greater effect on early auditory neural encoding, compared to voice acting training, a question which could be addressed by future research.

All our participants were native English speakers. However, we found large individual differences across participants in the ability to perceive prosodic features in English, with performance ranging from 45 to 95%. This test featured sound cues that were smaller than those present in naturalistic speech, to limit ceiling effects. Nevertheless, this wide variability suggests that not all individuals are equally able to perceive prosody in their native language. Moreover, we find that variability in prosody perception is linked to the robustness of encoding of the acoustic cue most relevant to a given feature. For example, perception of phrase boundaries, for which duration is the primary cue (Jasmin et al., 2021), was linked to the precision of duration perception, while perception of linguistic focus, for which pitch is the primary cue (Symons & Tierney, 2023), was linked to the consistency of neural encoding of F0. Prior research has indicated a link between auditory processing and individual differences in speech perception in childhood (Cumming, Wilson, & Goswami, 2015) as well as in adults learning a second language (Kachlicka et al., 2019). Our results here suggest that auditory processing continues to drive speech perception in one's native language throughout the lifespan.

An alternate explanation for our results is that individuals with pre-existing consistent neural encoding of sound, excellent prosody perception, and better-than-usual musical phrase perception are more likely to choose to engage in voice acting as a profession. Indeed, prior research has shown that perceptual and cognitive factors can predict whether individuals will engage in music training (Corrigall, Schellenberg, & Misura, 2013; Corrigall & Schellenberg, 2015), a finding which could extend to acting training as well. We would argue that this interpretation is not very parsimonious, as it seems unlikely that F0 consistency and prosody and musical phrase perception could innately co-vary in the absence of variation in more basic acoustic discrimination and musical aptitude skills. It seems particularly unlikely that musical phrase perception skill would be a major factor determining whether an individual pursues acting experience; this cross-domain relationship between voice acting and music perception is better explained as resulting from a transfer of learning from speech to music. Nevertheless, only future research using random assignment of interventions to participants and a longitudinal design could definitively establish whether voice acting can cause auditory neural enhancements.

Here we treat voice acting training as a homogenous category; this is in line with standard practice in research on the cognitive and perceptual profiles of musicians and non-musicians, which generally characterizes any individual with at least six years of training as a musician (Zhang, Susino, McPherson, & Schubert, 2020). In practice, however, the experience of being trained in music can differ widely depending on instrument (voice, wind, percussion, stringed) or genre (classical, jazz, rock, electronic). Similarly, voice acting training is heterogenous, with training varying depending on genre, including commercials, animation, and video games. Future work should compare and contrast different types of acting training and experience to determine the factors driving improvement in auditory neural encoding.

Overall, we find that although voice acting training does not require precise auditory perception, it is linked to enhancements in neural encoding of sound, as well as detection of musical and prosodic patterns. As a result, training which requires attention to, memory for, and production of exaggerated sound features may provide a method of boosting auditory neural function in individuals with poor auditory perception who might struggle to complete more demanding auditory training programs requiring fine precision.

Open practices

The study in this article has earned Open Data, and Open Materials for transparent practices. The data, and materials are available at: https://osf.io/jevbm/.

CRediT authorship contribution statement

Magdalena Kachlicka: Writing – review & editing, Writing – original draft, Visualization, Investigation, Formal analysis. **Adam Tierney:** Writing – review & editing, Writing – original draft, Visualization, Supervision, Project administration, Methodology, Formal analysis, Conceptualization.

REFERENCES

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*, 457–464.

Anderson, S., Parbery-Clark, A., Yi, H., & Kraus, N. (2011). A neural basis of speech-in-noise perception in older adults. *Ear and Hearing, 32*, 750–757.

Banducci, S., Daugherty, A., Biggan, J., Cooke, G., Voss, M., Noice, T., et al. (2017). Active experiencing training improves episodic memory recall in older adults. *Frontiers in Aging Neuroscience, 9*, 133.
Barnett, S., & Ceci, S. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612–637.
Berry, M., & Brown, S. (2019). Acting in action: Prosodic analysis of character portrayal during acting. *JEP:G, 148*, 1407–1425.
Bidelman, G., Gandour, J., & Krishnan, A. (2010). Cross-domain effects of music and language experience on the representation of pitch in the auditory brainstem. *Journal of Cognitive Neuroscience, 23*, 425–434.
Breen, M., Fedorenko, E., Wagner, M., & Gibson, E. (2010). Acoustic correlates of information structure. *Language and Cognitive Processes, 25*, 1044–1098.
Chandrasekaran, B., Krishnan, A., & Gandour, J. (2009). Relative influence of musical and linguistic experience on early cortical processing of pitch contours. *Brain and Language, 108*, 1–9.
Chrabaszcz, A., Winn, M., Lin, C., & Idsardi, W. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research, 57*, 1468–1479.
Coffey, E., Colagrosso, E., Lehmann, A., Schonwiesner, M., & Zatorre, R. (2016). Individual differences in the frequency-following response: Relation to pitch perception. *Plos One, 11*, Article e1052374.
Corrigall, K., & Schellenberg, E. (2015). Predicting who takes music lessons: Parent and child characteristics. *Frontiers in Psychology, 6*, 282.
Corrigall, K., Schellenberg, E., & Misura, N. (2013). Music training, cognition, and personality. *Frontiers in Psychology, 4*, 222.
Creel, S., Weng, M., Fu, G., Heyman, G., & Lee, K. (2018). Speaking a tone language enhances musical pitch perception in 3-5-year-olds. *Developmental Science, 21*, Article e12503.
Cumming, R., Wilson, A., & Goswami, U. (2015). Basic auditory processing and sensitivity to prosodic structure in children with specific language impairments: A new look at a perceptual hypothesis. *Frontiers in Psychology, 6*, 972.
Cuotinho, E., & Dibben, N. (2013). Psychoacoustic cues to emotion in speech prosody and music. *Cognition & Emotion, 27*, 658–684.
De Pijper, J., & Sanderman, A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *JASA, 96*, 2037–2047.
Dick, F., Lee, H., Nusbaum, H., & Price, C. (2011). Auditory-motor expertise alters "speech selectivity" in professional musicians and actors. *Cerebral Cortex, 21*, 938–948.
Easwar, V., Scollie, S., Aiken, S., & Purcell, D. (2020). Test-retest variability in the characteristics of envelope following responses evoked by speech stimuli. *Ear and Hearing, 41*, 150–164.
Ellis, R., & Jones, M. (2009). The role of accent salience and joint accent structure in meter perception. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 264–280.
Fear, B., Cutler, A., & Butterfield, S. (1995). The strong/weak syllable distinction in English. *JASA, 97*, 1893–1904.
Fujioka, T., Trainor, L., Ross, B., Kakigi, R., & Pantec, C. (2004). Musical training enhances automatic encoding of melodic contour and interval structure. *Journal of Cognitive Neuroscience, 16*, 1010–1021.
Goldstein, T., Wu, K., & Winner, E. (2010). Actors are skilled in theory of mind but not empathy. *Imagination, Cognition and Personality, 29*, 115–133.
Grassi, M., & Soranzo, A. (2009). Mlp: A MATLAB toolbox for rapid and reliable threshold estimation. *Behavior Research Methods, 41*, 20–28.
Green, D. (1993). A maximum-likelihood method for estimating thresholds in a yes-no task. *Journal of the Acoustical Society of America, 93*, 2096–2105.
Groussard, M., Coppalle, R., Hinault, T., & Platel, H. (2020). Do musicians have better mnemonic and executive performance than actors? Influence of regular musical or theatre practice in adults and in the elderly. *Frontiers in Human Neuroscience, 14*, Article 557642.
Hannon, E., Snyder, J., Eerola, T., & Krumhansl, C. (2004). The role of melodic and temporal cues in perceiving musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 956–974.
Hornickel, J., Knowles, E., & Kraus, N. (2012). Test-retest consistency of speech-evoked auditory brainstem responses in typically-developing children. *Hearing Research, 284*, 52–58.
Hutchins, S., & Moreno, S. (2013). The Linked Dual Representation model of vocal perception and production. *Frontiers in Psychology, 4*, 825.
Ilie, G., & Thompson, W. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception, 23*, 319–329.
Jürgens, R., Grass, A., Drolet, M., & Fischer, J. (2015). Effect of acting experience on emotion expression and recognition in voice: Non-actors provide better stimuli than expected. *J Nonverbal Behav, 39*, 195–214.
Jasmin, K., Dick, F., Holt, L., & Tierney, A. (2020). Tailored perception: Individuals' speech and music perception strategies fit their perceptual abilities. *Journal of Experimental Psychology: General, 149*, 914–934.
Jasmin, K., Dick, F., & Tierney, A. (2020). The multidimensional Battery of prosody perception (MBOPP). *Wellcome Open Research, 5*, 4.
Jasmin, K., Sun, H., & Tierney, A. (2021). Effects of language experience on domain-general perceptual strategies. *Cognition, 206*, Article 104481.
Joronen, K., Konu, A., Rankin, H., & Astedt-Kurki, P. (2011). An evaluation of a drama program to enhance social relationships and anti-bullying at elementary school: A controlled study. *Health Promotion International, 27*, 5–14.
Kachlicka, M., Saito, K., & Tierney, A. (2019). Successful second language learning is tied to robust domain-general auditory processing and stable neural representation of sound. *Brain and Language, 192*, 15–24.
Kawahara, H., & Irino, T. (2005). Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation. In P. Divenyi (Ed.), *Speech separation by humans and machines* (pp. 167–180). Boston, MA: Kluwer Academic Publishers. https://doi.org/10.1007/0-387-22794-6_11.
Krishnan, A., Bidelman, G., Smalt, C., Ananthakrishnan, S., & Gandour, J. (2012). Relationship between brainstem, cortical and behavioral measures relevant to pitch salience in humans. *Neuropsychologia, 50*, 2849–2859.
Krishnan, A., Swaminathan, J., & Gandour, J. (2009). Experience-dependent enhancement of linguistic pitch representation in the brainstem is not specific to a speech context. *Journal of Cognitive Neuroscience, 21*, 1092–1105.
Krishnan, A., Xu, Y., Gandour, J., & Cariani, P. (2005). Encoding of pitch in the human brainstem is sensitive to language experience. *Cognitive Brain Research, 25*, 161–168.
Krumhansl, C. L. (2000). Rhythm and pitch in music cognition. *Psychological Bulleting, 126*(1), 159–179.
Liu, J., Hilton, C., Bergelson, E., & Mehr, S. (2023). Language experience predicts music processing in a half-million speakers of fifty-four languages. *Current Biology, 33*, 1916–1925.
Loui, P., Guenther, F., Mathys, C., & Schlaug, G. (2008). Action-perception mismatch in tone-deafness. *Current Biology, 18*, R331–R332.
Magne, C., Schön, D., & Besson, M. (2006). Musician children detect pitch violations in both music and language better than

nonmusician children: Behavioral and electrophysiological approaches. *Journal of Cognitive Neuroscience, 18*, 199–211.
Matharu, K., Berry, M., & Brown, S. (2022). Storytelling as a fundamental form of acting. *Psychology of Aesthetics, Creativity, and the Arts, 16*, 272–289.
Mattys, S. (2000). The perception of primary and secondary stress in English. *Perception & Psychophysics, 62*, 253–265.
Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research, 219*, 36–47.
Mualem, O., & Lavidor, M. (2015). Music education intervention improves vocal emotion recognition. *International Journal of Music Education, 33*, 413–425.
Musacchia, G., Sams, M., Skoe, E., & Kraus, N. (2007). Musicians have enhanced subcortical auditory and audiovisual processing of speech and music. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 15894–15898.
Nan, Y., Liu, L., Geiser, E., Shu, H., Gong, C., Dong, Q., et al. (2018). Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children. *Proceedings of the National Academy of Sciences of the United States of America, 115*, E6630–E6639.
Noice, H., & Noice, T. (2009). An arts intervention for older adults living in subsidized retirement homes. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 16*, 56–79.
Omote, A., Jasmin, K., & Tierney, A. (2017). Successful non-native speech perception is linked to frequency following response phase consistency. *Cortex; a Journal Devoted To the Study of the Nervous System and Behavior, 93*, 146–154.
Ozaki, Y., et al. (2024). Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech: A registered report. *Science Advances, 10*, Article adm9797.
Palmer, C., & Krumhansl, C. (1987). Independent temporal and pitch structures in determination of musical phrases. *JEP:HPP, 13*, 116–126.
Parbery-Clark, A., Anderson, S., Hittner, E., & Kraus, N. (2012). Musical experience strengthens the neural representation of sounds important for communication in middle-aged adults. *Frontiers in Aging Neuroscience, 4*, 30.
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*, 14100–14107.
Parbery-Clark, A., Strait, D., & Kraus, N. (2011). Context-dependent encoding in the auditory brainstem subserves enhanced speech-in-noise perception in musicians. *Neuropsychologia, 49*, 3338–3345.
Patel, A. (2011). Why would musical training benefit the neural encoding of speech? The OPERA hypothesis. *Frontiers in Psychology, 2*, 142.
Patel, A. (2014). Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis. *Hearing Research, 308*, 98–108.
Petrova, K., Jasmin, K., Saito, K., & Tierney, A. (2023). Extensive residence in a second language environment modifies perceptual strategies for suprasegmental categorization. *JEP:LMC, 49*, 1943–1955.
Pfordresher, P., & Brown, S. (2009). Enhanced production and perception of musical pitch in tone language speakers. *Attention, Perception, & Psychophysics, 71*, 1385–1398.
Ponsot, E., Burred, J., Belin, P., & Aucouturier, J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences of the United States of America, 115*, 3972–3977.
Purcell, D., John, S., Schneider, B., & Picton, T. (2004). Human temporal auditory acuity as assessed by envelope following responses. *JASA, 116*, 3581–3593.
R Core Team. (2023). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. URL https://www.R-project.org/.
Rosslau, K., Herholz, S., Knief, A., Ortmann, M., Deuster, D., Schmidt, C., et al. (2016). Song perception by professional singers and actors: An MEG study. *Plos One, 11*, Article e0147986.
Ruggles, D., Bharadwaj, H., & Shinn-Cunningham, B. (2012). Why middle-aged listeners have trouble hearing in everyday settings. *Current Biology, 22*, 1417–1422.
Schaffrath, H., & Park, D. H. M. (1995). The Essen folksong collection in kern format [Computer database]. Retrieved from http://essen.themefinder.org/.
Skoe, E., & Kraus, N. (2012). A little goes a long way: How the adult brain is shaped by musical training in childhood. *Journal of Neuroscience, 32*, 11507–11510.
Skoe, E., & Kraus, N. (2013). Musical training heightens auditory brainstem function during sensitive periods in development. *Frontiers in Psychology, 4*, 622.
Symons, A., & Tierney, A. (2023). Informational masking influences segmental and suprasegmental speech categorization. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-023-02364-5
Thompson, W., Schellenberg, E., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion, 4*, 46–64.
Tierney, A., Krizman, J., & Kraus, N. (2015). Music training alters the course of adolescent neural development. *Proceedings of the National Academy of Sciences of the United States of America, 112*, 10062–10067.
Tierney, A., Russo, F., & Patel, A. (2011). The motor origins of human and avian song structure. *Proceedings of the National Academy of Sciences of the United States of America, 108*, 15510–15515.
Wallentin, M., Hojlund Nielsen, A., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The Musical Ear Test, a new reliable test for measuring musical competence. *Learning and Individual Differences, 20*, 188–196.
Whiting, C., Kotz, S., Gross, J., Giordano, B., & Belin, P. (2020). The perception of caricatured emotion in voice. *Cognition, 200*, Article 104249.
Wong, P., Skoe, E., Russo, N., Dees, T., & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10*, 420–422.
Zarate, J., Delhommeau, K., Wood, S., & Zatorre, R. (2010). Vocal accuracy and neural plasticity following micromelody-discrimination training. *Plos One, 5*, Article e11181.
Zatorre, R., & Baum, S. (2012). Musical melody and speech intonation: Singing a different tune? *Plos Biology, 7*, Article e1001372.
Zhang, J. D., Susino, M., McPherson, G. E., & Schubert, E. (2020). The definition of a musician in music psychology: A literature review and the six-year rule. *Psychology of Music, 48*(3), 389–409.