JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Who is singing? Voice recognition from spoken versus sung speech

Angela Cooper,[1] Matthew Eitel,[2] Natalie Fecher,[1] Elizabeth Johnson,[1] and Laura K. Cirelli[2,a)]
[1]Department of Psychology, University of Toronto Mississauga, Mississauga, Ontario, Canada
[2]Department of Psychology, University of Toronto Scarborough, Toronto, Ontario, Canada
a.kanita.cooper@gmail.com, matthew.eitel@utoronto.ca, fecherna@gmail.com, elizabeth.johnson@utoronto.ca, laura.cirelli@utoronto.ca

Abstract: Singing is socially important but constrains voice acoustics, potentially masking certain aspects of vocal identity. Little is known about how well listeners extract talker details from sung speech or identify talkers across the sung and spoken modalities. Here, listeners ($n = 149$) were trained to recognize sung or spoken voices and then tested on their identification of these voices in both modalities. Learning vocal identities was initially easier through speech than song. At test, cross-modality voice recognition was above chance, but weaker than within-modality recognition. We conclude that talker information is accessible in sung speech, despite acoustic constraints in song.

[Editor: Irina A. Shport] https://doi.org/10.1121/10.0026385

Received: 16 February 2024 **Accepted:** 3 June 2024 **Published Online:** 18 June 2024

1. Introduction

Human vocal behavior yields substantial acoustic-phonetic variability both between speakers and within a speaker (Lavan *et al.*, 2019; Mullennix *et al.*, 1995). Speakers dynamically alter their speech patterns depending on their audience (e.g., speaking to children vs elderly listeners, ManyBabies Consortium, 2020; Picheny *et al.*, 1986), environmental factors (e.g., in a noisy bar or a library, Ito *et al.*, 2005) and emotional contexts (e.g., laughing, shouting, Bachorowski *et al.*, 2001). Thus, voice recognition entails not only telling *apart* distinct voices and encoding voice characteristics in memory but also preserving perceptual constancy of the same voice across contexts ("telling people together," Lavan *et al.*, 2019). While listeners are adept at discriminating and recognizing different voices under optimal listening conditions (Johnson, 2005; Kreiman, 1997), considerably less is known about how listeners establish a speaker identity percept despite within-speaker variation. Perhaps one of the more challenging examples of within-speaker variability is introduced when singing, which not only creates acoustical variation (Sundberg, 1977) but also restricts and determines timing and pitch information.

Understanding how listeners can utilize and overcome within-speaker variability to form stable identity percepts is a challenge for which theoretical models for talker recognition must account (Lavan *et al.*, 2019). Across other kinds of within-speaker variation, voice identification is possible but difficult. For example, listeners can identify bilingual speakers across their two spoken languages better than chance, but this task is far more challenging across languages than within (Winters *et al.*, 2008). Other studies that explicitly introduce within-speaker variation (e.g., disguised vs undisguised speech, whispered or spontaneous laughter vs normal speech, negative vs neutral affect) consistently show a decrement in voice recognition performance with increased variability (Bartle and Dellwo, 2015; Lavan *et al.*, 2016; Lavan *et al.*, 2019).

Can listeners learn vocal identities despite the within-speaker variability generated across song and speech? Only one study, to our knowledge, provides some evidence to suggest that the singing and speaking voice retain identifying cues, but more questions remain (Peynircioğlu *et al.*, 2017). In an AX (same-different) voice discrimination task, Peynircioğlu *et al.* (2017) asked listeners to indicate whether a pair of vowels or phrases (both sung, both spoken, or cross-modal) were produced by the same voice or two different voices. Phrases were one of two well-known lines from famous shows ("doe, a deer, a female deer" from The Sound of Music, and "somewhere over the rainbow" from The Wizard of Oz). While performance was better for within-modality trials, it was still above chance when listeners were presented with one speech and one song token. This suggests that despite the within-speaker variability introduced through song, enough acoustical cues for voice identification are held constant for listeners to identify the same voice after a direct comparison. However, this AX task, which relies on short term memory and immediate comparisons, may overestimate listeners' abilities to recognize a speaker across modalities. More specifically, the acoustic cues that lead to successful performance on the AX task may not make it into stable long-term representations of vocal identity.

a)Author to whom correspondence should be addressed.

The fact that listeners found cross-modality trials challenging in the previously-described AX task is perhaps unsurprising given how much the acoustic-phonetic characteristics of sung speech can differ from spoken speech, especially since many of the speech acoustics that are altered in sung speech are thought to be key to talker identification in classic models of speech processing (Mathias and von Kriegstein, 2014). For example, in singing, the larynx is often lowered, and the pharynx expanded relative to speaking, which has numerous acoustic consequences, including lowered formant frequencies and often the presence of an additional spectral peak between 2500 and 3000 Hz (Sundberg, 1977). Across languages and with both trained and untrained singers, vowels are often compressed in F1–F2 formant space for song compared to speech (Bloothooft and Plomp, 1984; Hansen *et al.*, 2020). Moreover, sung speech is required to align with a set musical melody and rhythm and is often accompanied by the presence of vocal vibrato.

On the other hand, singing is often part of listeners' daily social lives—across the lifespan, informal singing is part of family and community life. Singing communicates emotions, and singing together fosters group cohesion (Livingstone and Russo, 2018; Pearce *et al.*, 2015). Historically, across many cultures, orally-transmitted work songs helped laborers connect and coordinate—the Shan'ge "Mountain Songs" of China, for example, would often be sung across great geographic distances using call and response (Ōki and Santangelo, 2011). Given the social importance of song, it may be the case that listeners can flexibly rely on fewer acoustic features to identify a singer.

To summarize, the talker identification literature has traditionally minimized within-talker variation, resulting in an idealized test scenario that does not reflect typical real-world recognition contexts. Singing alters many properties of the human voice, leading to substantial variation between a talker's spoken and sung productions. To date, the only study examining listeners' ability to identify talkers across the sung and spoken modalities utilized an AX task, which involved making low-level, acoustic comparisons (Peynircioğlu *et al.*, 2017). The present study tests a larger sample size in a training-voice identification paradigm, which taps into long-term memory of speaker identity representations (as opposed to the short-term memory demands of a discrimination task, such as in Peynircioğlu *et al.*, 2017, and Levi, 2018). Additionally, to eliminate potential confounds with tune familiarity and to introduce melodic variation, we composed nine original melodies for use in this task. Listeners completed a three-phase paradigm, whereby they were (1) briefly exposed to four female speakers with either sung or spoken speech materials (pre-training) before (2) being trained to identify these speakers/singers (training), and (3) finally tested on their voice identification abilities across both modalities (test). Because singing constrains many of the acoustic dimensions thought to be important for talker recognition, we predicted (1) that it would be easier to train participants to recognize a voice via speech than song, and (2) that, after training within a single modality, cross-modal recognition during test would be possible but more challenging than within-modal recognition.

2. Method

2.1 Participants

In total, 160 Canadian participants took part in the experiment. Participants were randomly assigned to one of three training groups (Spoken; Sung; Mixed). Nine participants did not complete the experiment and two were excluded for not providing demographic information. This left 149 participants in the dataset [Spoken = 44; Sung = 52; Mixed = 53; 39 men, 108 women, 2 undisclosed; M age = 18.7 years, standard deviation (SD) age = 1.8]. All participants reported learning English prior to age 6-years and reported no speech and hearing deficits. Participants reported between 0 and 16 years of formal musical training (mean, M = 2.89 years, SD = 3.43 years) ("Have you had any formal music training, i.e., music classes, choir, etc.? If yes, what type of training—e.g., piano, voice, etc."). Participants were considered non-musicians if they reported fewer than 3 years of training ($n = 90$). Participants were recruited from our institute's undergraduate research participant pool and received course credit for their time.

2.2 Procedure

Testing took place in a quiet testing space in a psychology laboratory. Participants were first randomly assigned to either the spoken only, sung only, or mixed conditions, which determined the stimuli they received during pre-training and training phases. In a pre-training phase, listeners were exposed to 16 trials—each of the four voices speaking or singing four sentences while simultaneously viewing a cartoon face associated with each voice (see Fig. 1). Participants in the mixed training condition heard two spoken and two sung sentences from each voice during pre-training. Participants made no response during this phase.

Participants then underwent a training phase. For the mixed condition, each block contained an equal number of spoken and sung training trials, which were randomized within blocks. Each training block consisted of 48 randomized trials (four speakers each producing 12 sentences). In each trial, participants heard either a spoken or sung sentence along with two cartoon images. They were tasked to select which of the two images they thought were speaking/singing and received feedback on whether they were correct or incorrect in their response. If listeners achieved 70% accuracy by the end of a block, they moved on to the test phase. If they scored below 70%, they started a new training block (up to a maximum of six blocks).

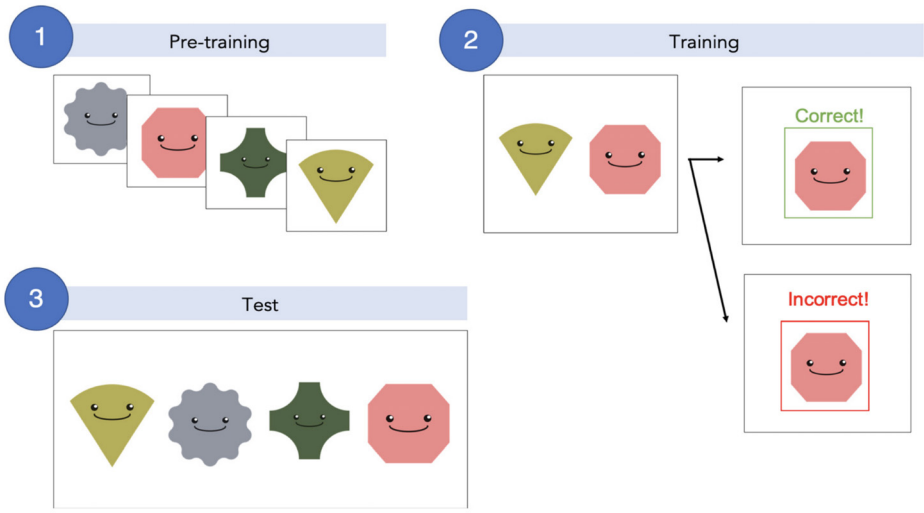

Fig. 1. Graphics used in the experimental paradigm for the three phases of the training-identification task: (1) pre-training, (2) training, and (3) test. Adapted from Cooper *et al.* (2020).

In the test phase, listeners heard a sentence while viewing all 4 cartoon faces, once again being tasked to select who they thought was producing the sentence. They received no accuracy feedback. The test phase consisted of 64 trials (48 novel sentences and 16 familiar sentences from the pre-training and training phases) comprised of sentences from all four speakers. Associations between speaker and cartoon, and which sentences were attributed to each voice, were counterbalanced across participants.

2.3 Stimuli

Sentence materials consisted of 64 Hearing-in-Noise Test (HINT) sentences (Soli and Wong, 2008), which are monoclausal declarative sentences containing basic-level English vocabulary (e.g., "The bananas are too ripe"). All materials were produced by eight female native Canadian English speakers with at least five years of musical/vocal training. These eight voices were divided into two stimulus sets with four voices each and counterbalanced across participants. A total of nine original melodies (in C major with a tempo of 100 BPM) were composed to fit the selected sentences in terms of syllable count and stress placement. To create the sung recordings, the speakers first listened to a piano rendition of each melody and then sang back the corresponding sentence without musical accompaniment. The mean duration for each spoken sentence was 1.91 s (SD = 0.20), which was significantly shorter than the sung sentences (M = 2.50 s, SD = 0.26) by about half a second, $p < 0.001$ (see supplementary material1 for more details about the length of stimuli per voice).

The pre-training phase included four HINT sentences produced by each of the four speakers (the same four sentences across participants). For sung pre-training, these sentences were produced with one melody. The training phase consisted of 12 different HINT sentences (the same 12 sentences across participants). For sung training, these sentences were divided between six melodies (two different sentences sung with each melody) for sung training. For the mixed training, participants heard six spoken sentences and six sung sentences (divided across three unique melodies).

The test phase contained the 16 familiar sentences described previously, half of which were sung and half spoken. Additionally, 48 novel HINT sentences were also included, half of which were sung and half spoken. The 24 novel sung sentences were divided between three melodies not heard during familiarization/training (eight sentences per melody).

3. Results

3.1 Analyses

The data were analyzed in R (version 3.6.3, R Core Team, 2023). Linear mixed-effects models (glmmTMB package, Brooks *et al.*, 2017) were used to examine training accuracy and test performance, using Poisson distributions for count data, beta distributions for proportion scores, and binomial distributions for yes/no accuracy.

3.2 Training

To test our first prediction (that learning a voice through speech will be easier than learning a voice through song), we examined how many training blocks participants required to reach the 70% accuracy end-block criterion. See Fig. 2. A linear mixed-effects model with fixed effects of Training Group (spoken, sung, mixed) and Musicianship (musician, non-musician)

and their interactions were included. We also included a random intercept for stimulus set (voice set 1, voice set 2). There were no effects or interactions with Musicianship (p-values > 0.31). Confirming our hypothesis, participants learned voice identity through speech more easily (after M = 2.02 blocks, SD = 1.64) than through song (M = 3.37 blocks, SD = 1.85), B = 0.61, standard error (SE) = 0.16, $z = 3.70$, $p < 0.001$, or mixed training (M = 3.42 blocks, SD = 1.84), B = 0.61, SE = 0.16, $z = 3.88$, $p < 0.001$. This pattern of results (faster learning in the speech than song or mixed conditions) is the same when we explore proportion correct scores at the end of block 1 using a similar model with a beta family distribution for proportion scores (p-values < 0.001). *Post hoc* t-tests on both dependent measures (block 1 proportion correct and number of training blocks required) confirm that performance from participants in the sung training condition matches those in the mixed training condition (p-values > 0.730). These patterns of results are also consistent when participants who did not reach the criterion (70% accuracy by the end of the 6th block, $n = 26$) are removed from the analyses.

3.3 Test

To avoid overestimating performance if listeners learned sentence-specific idiosyncrasies, our pre-planned analyses on performance at test focused only on the 48 novel sentences (the 16 familiar sentences were excluded). All statistical decisions are unchanged if these familiar sentences are retroactively included. We then tabulated the proportion of correct voice identification responses, broken down by training group (mixed, spoken, sung) and trial type (spoken, sung; see Fig. 3). Listeners' proportion of correct voice identification was above chance (25%) for each trial type and in each training group (all p-values ≤ 0.002).

Our second hypothesis was that within-modality voice recognition would be easier than cross-modality voice recognition. To address this question, we used linear mixed-effects models with voice identification accuracy as the dependent variable, focusing only on the participants in the sung and spoken training conditions (the participants in the mixed training condition were, by definition, trained on both modalities so have no scores for cross-modal performance). Contrast-coded fixed effects of Training Group (sung, spoken), Test Trial type (sung, spoken), and Musicianship (musician, non-musician), and their interactions were included in the model. We also included end block accuracy scores as a covariate. Random intercepts for participant and item nested within voice and stimulus set were also included, as well as a by-participant random slope for Trial type and a by-item random slope for the Training Group.

As expected, participants who had high accuracy at the end of their last training block had high accuracy during test, B = 4.23, $SE = 0.49$, $z = 8.71$, $p < 0.001$. No main effects or interactions with Musicianship reached significance (p-values > 0.521). The main effects of Training Group ($z = -0.19$, $p = 0.851$) and Trial type ($z = 0.48$, $p = 0.630$) did not reach significance. Testing our main prediction, a significant interaction between Training Group and Trial type was found (B = 0.40, SE = 0.05, z = 8.07, $p < 0.001$). Simple effects were analyzed by constructing similar models for each Training Group. After accounting for training block accuracy (p-values < 0.001), a significant effect of Trial type was found for both the spoken training group (B = 0.43, SE = 0.08, $z = 5.32$, $p < 0.001$) and the sung training group (B = −0.37, SE = 0.06,

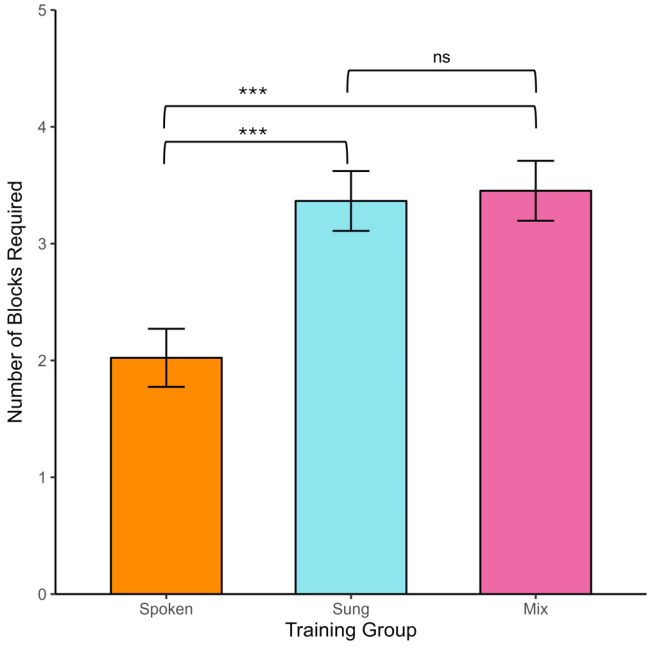

Fig. 2. Number of training blocks required to achieve 70% end-block accuracy before moving on to test trials, across the three training conditions (spoken, sung, mixed). Error bars represent the SE of the mean. *** = $p < .001$.

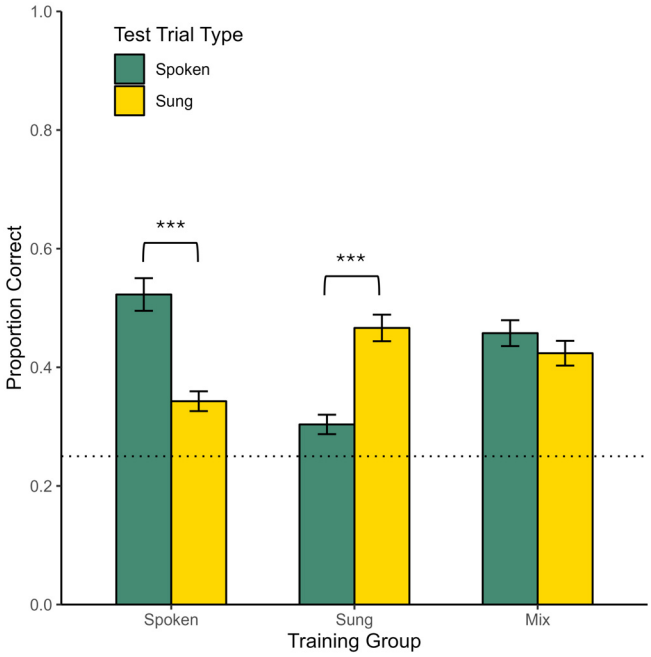

Fig. 3. Proportion correct on voice identification during test across trial type (spoken and sung) for participants across the three training conditions (spoken, sung, mixed). Error bars represent the SE of the mean. All bars are significantly above chance level (25%), p-values < 0.002. Cross-modal performance is compared to within-modal performance for unimodally trained participants. $^{***} = p < 0.001$.

$z = -5.65$, $p < 0.001$). Listeners in the spoken training group had more accurate voice identification on spoken trials relative to sung trials, and those in the sung training group were more accurate on sung trials relative to spoken trials (see Fig. 3). This trial type by Training Group interaction and the results of the *post hoc* analyses were similar if participants who did not reach criterion (70% accuracy by the end of the 6th block, $n = 26$) were included.

3.4 Acoustic analyses

A unique property of sung speech, as discussed in the Introduction, is that pitch (and more specifically *F*0) is constrained by melody. Melodies that force a speaker to shift further from their natural speaking pitch may make it more challenging for listeners to identify them. Here, speakers' average spoken *F*0 was 212 Hz (range = 177 to 269 Hz) and the sung renditions of these phrases had an average *F*0 of 286 Hz (range = 211–360 Hz). In exploratory analyses, we investigated whether voice identification accuracy across modality (those trained on speech identifying singers and those trained on song identifying speakers) would be worse when test trials contained mean pitch levels that were farther from those encountered during training. For each of the eight voices, mean spoken *F*0 and sung *F*0 were calculated using MIRToolbox (version 1.8) (Lartillot and Toivianen, 2007) running on MATLAB R2023b. A difference score for each cross-modal test trial (sung test trials for spoken trained participants, spoken test trials for sung trained participants) was then computed by subtracting this speaker-specific average from the mean pitch of the test trial. A linear mixed-effects model using a binomial distribution was used to explore voice identification accuracy on cross-modal test trials. Fixed effects and the interaction between contrast-coded Training Group (sung, spoken) and the linear effect of absolute pitch difference from mean pitch per speaker during training were included in the model, along with a random intercept for a participant. No interactions ($p = 0.20$) or main effects ($p = 0.07$) of the Training Group were significant, but voice identification accuracy was lower on cross-modal test trials when the absolute difference in pitch from that speaker's mean pitch during training was greater, $B = -0.01$, $SE = 0.003$, $z = -2.53$, $p = 0.011$.

4. Discussion

Singing involves the modification of many vocal features that are thought to be used in talker recognition, such as F0, F0 range, segmental properties, and intonation patterns (Marchenko, 2020; Sundberg, 1977). Can the human perceptual system recognize singing voices despite these modifications, and use stable long-term representations to identify a voice across modality (singing and speaking)? The present study demonstrates that the answer to both questions is yes. Two additional takeaways are offered by this study. First, based on performance during training, learning to identify singing voices is

more challenging than learning to identify speaking voices. Second, based on performance during test, cross-modal speaker recognition is more challenging than within-modality recognition. Next, we elaborate on each of these points.

Listeners trained on song performed as well on speech recognition at test as listeners trained on speech did for song recognition at test. However, listeners found it more challenging to learn to recognize a singer's identity than a speaker's identity during training, requiring more training blocks to reach the end-block 70% accuracy threshold. This speech advantage may be driven by cumulative input and experience with identifying speakers in the real world, reflecting that speech (and not song) is adults' dominant source of vocal input. Bottom-up effects may also explain this speech advantage, given that idiosyncratic features in speech may become constrained through song when temporal and pitch features are prescribed.

Across speech and song, the average *F*0 of talkers at training seemed particularly important for creating a representation of that talker when performing on cross-modal test trials. That is, cross-modal voice identification was hardest when the mean pitch of a test trial was farther from the mean pitch of that voice during training. This was the case regardless of the Training Group, suggesting that pitch is a cue that is relied on when identifying speakers across vocalization styles. This is intuitive if one hypothesizes that, during training, listeners formulate a representation of mean pitch and pitch range for each talker and use this association to make decisions about talker identity during test—a large deviation at test from the mean pitch of a particular voice could be interpreted as evidence for a different speaker.

Aligning with research highlighting the social relevance of song, listeners were able to recognize voices despite the challenging levels of within-speaker variability introduced when voices shifted from speech to song. Cross-modal perception was more successful for some voices than others (see supplementary material1), highlighting that some vocalists may retain more identifying acoustic cues when shifting from speech to song than others. Singing style, training background, and pitch production ability may all factor into cross-modal recognizability.

When cross-modal identification is successful, what acoustic properties might listeners be relying on? The timbral qualities restricted by the physical dimensions of the vocal tract may be especially important for identifying a voice across speech and song. Formant frequencies, which depend on vocal tract dimensions, contribute to the timbral quality of one's singing voice (Cleveland, 1977). Harmonic, vibrato, and timbral information from song allows for reasonably high levels of singer identification accuracy by computer learning models (Khine *et al.*, 2008). When considering higher-level properties, a growing body of research highlights how talker recognition is harder when listening to accented speech (Stevenage *et al.*, 2012; Yu *et al.*, 2021), but that accented speakers sound less accented when singing compared to speaking (Chan *et al*, 2023; Mageau *et al.*, 2019). Future directions combining these areas of research with our results could explore whether talker recognition for accented speakers may be easier through song than speech.

Supplementary Material

See the supplementary material for additional stimulus details and voice by voice analyses.

Acknowledgments

This work was supported by SSHRC Insight Grants to L.K.C. and E.J. Thanks to Lisa Hotson, Tamim Fattah, and other members of the Child Language and Speech Studies lab directed by E.J. for assistance with data collection. Thanks to the voice actors who helped us create our stimuli. Thanks to Yingjia Wan for insightful discussions about Shan'ge Mountain Songs.

Author Declarations

Conflict of Interest

The authors have no conflicts to disclose.

Ethics Approval

The research was approved by the University of Toronto Research Ethics Board and was conducted in compliance with recognized standards for experimentation with human subjects. Informed consent was obtained from all participants.

Data Availability

Following blinded review, the data will be made public on OSF and a DOI will be created.

References

Bachorowski, J. A., Smoski, M. J., and Owren, M. J. (**2001**). "The acoustic features of human laughter," J. Acoust. Soc. Am **110**(3), 1581–1597.

Bartle, A., and Dellwo, V. (**2015**). "Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech," Int. J. Speech Lang. **22**, 229–248.

Bloothooft, G., and Plomp, R. (**1984**). "Spectral analysis of sung vowels. I. Variation due to differences between vowels, singers, and modes of singing," J. Acoust. Soc. Am. **75**(4), 1259–1264.

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Maechler, M., and Bolker, B. M. (**2017**). "glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling," R J. **9**(2), 378–400, available at https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Chan, K. Y., Hall, M. D., and Herr, R. (**2023**). "What accounts for foreign accent reduction in singing?," Aud. Percept. Cognit. **6**(3–4), 233–249.
Cleveland, T. F. (**1977**). "Acoustic properties of voice timbre types and their influence on voice classification," J. Acoust. Soc. Am. **61**(6), 1622–1629.
Cooper, A., Fecher, N., and Johnson, E. K. (**2020**). "Identifying children's voices," J. Acoust. Soc. Am. **148**(1), 324–333.
Hansen, J. H., Bokshi, M., and Khorram, S. (**2020**). "Speech variability: A cross-language study on acoustic variations of speaking versus untrained singing," J. Acoust. Soc. Am **148**(2), 829–844.
Ito, T., Takeda, K., and Itakura, F. (**2005**). "Analysis and recognition of whispered speech," Speech Commun. **45**, 139–152.
Johnson, K. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. Remez (Blackwell Publishing Ltd, Oxford, UK), pp. 363–389.
Khine, S. Z. K., Nwe, T. L., and Li, H. (**2008**). "Exploring perceptual based timbre feature for singer identification," in *Proceedings of Computer Music Modeling and Retrieval (CMMR 2007)*, Copenhagen, Denmark (May 19–23), pp. 159–171.
Kreiman, J. (**1997**). "Listening to voices: Theory and practice in voice perception research," in *Talker Variability in Speech Research*, edited by K. Johnson and J. Mullennix (Academic Press, New York), pp. 85–108.
Lartillot, O., and Toivianen, P. (**2007**). "A Matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, France, edited by S. Marchand, pp. 237–244, available at http://cms2.unige.ch/fapse/neuroemo/pdf/ArticleLartillot2007 Bordeaux.pdf.
Lavan, N., Burton, A., Scott, S., and McGettigan, C. (**2019**). "Flexible voices: Identity perception from variable vocal signals," Psychon. Bull. Rev. **26**, 90–102.
Lavan, N., Scott, S. K., and McGettigan, C. (**2016**). "Impaired generalization of speaker identity in the perception of unfamiliar and familiar voices," J. Exp. Psychol. Gen. **145**, 1604–1614.
Levi, S. (**2018**). "Methodological considerations for interpreting the Language Familiarity Effect in talker processing," WIREs Cognit. Sci. **10**(2), e1483.
Livingstone, S. R., and Russo, F. A. (**2018**). "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS One **13**(5), e0196391.
Mageau, M., Mekik, C., Sokalski, A., and Toivonen, I. (**2019**). "Detecting foreign accents in song," Phonetica **76**(6), 429–447.
ManyBabies Consortium (**2020**). "Quantifying sources of variability in infancy research using the infant-directed-speech preference," Adv. Methods Pract. Psychol. Sci. **3**(1), 24–52.
Marchenko, V. (**2020**). "Speech intonation and music: A look at their dynamics within the song format," J. Lang. Linguist. Stud. **16**(2), 822–834.
Mathias, S. R., and von Kriegstein, K. (**2014**). "How do we recognise who is speaking?," Front. Biosci. **S6**(1), 92–109.
Mullennix, J., Johnson, K., Topcu-Durgun, M., and Famsworth, L. (**1995**). "The perceptual representation of voice gender," J. Acoust. Soc. Am. **98**, 3080–3095.
Pearce, E., Launay, J., and Dunbar, R. I. (**2015**). "The ice-breaker effect: Singing mediates fast social bonding," R Soc. Open Sci. **2**(10), 150221.
Picheny, M. A., Durlach, N. I., and Braida, L. D. (**1986**). "Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech," J. Speech Lang. Hear. R **29**(4), 434–446.
Peynircioğlu, Z., Rabinovitz, B., and Repice, J. (**2017**). "Matching speaking to singing voices and the influence of content," J. Voice **31**, 256.e13–256.e17.
R Core Team (**2023**). *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), available at https://www.R-project.org/.
Ōki, Y., and Santangelo, P. (**2011**). "Shan'ge, the 'mountain songs': Love songs in Ming China," in *Emotions and States of Mind in East Asia* (Brill Academic, Leiden, the Netherlands).
Soli, S., and Wong, L. (**2008**). "Assessment of speech intelligibility in noise with the hearing in noise test," Intl. J. Audiol. **47**, 356–361.
Stevenage, S. V., Clarke, G., and McNeill, A. (**2012**). "The 'other-accent' effect in voice recognition," J. Cognit. Psychol. **24**(6), 647–653.
Sundberg, J. (**1977**). "The acoustics of the singing voice," Sci. Am. **236**, 82–116.
Winters, S. J., Levi, S. V., and Pisoni, D. B. (**2008**). "Identification and discrimination of bilingual talkers across languages," J. Acoust. Soc. Am. **123**(6), 4524–4538.
Yu, M., Schertz, J., and Johnson, E. K. (**2021**). "The other accent effect in talker recognition: Now you see it, now you don't," Cognit. Sci. **45**(6), e12986.