2018

Effects of signal bandwidth and noise on individual speaker identification

Jeremy C. Schwartz
West Virginia University

Ashtyn T. Whyte
West Virginia University

Mohanad Al-Nuaimi
West Virginia University

Jeremy J. Donai
West Virginia University

Effects of signal bandwidth and noise on individual speaker identification

Jeremy C. Schwartz, Ashtyn T. Whyte, Mohanad Al-Nuaimi, and Jeremy J. Donai

Citation: The Journal of the Acoustical Society of America **144**, EL447 (2018); doi: 10.1121/1.5078770
View online: https://doi.org/10.1121/1.5078770
View Table of Contents: https://asa.scitation.org/toc/jas/144/5
Published by the Acoustical Society of America

JASA
THE JOURNAL OF THE
ACOUSTICAL SOCIETY OF AMERICA
SPECIAL ISSUE:
Noise-Induced Hearing Loss: Translating Risk
from Animal Models to Real-World Environments
READ NOW!

Effects of signal bandwidth and noise on individual speaker identification

Jeremy C. Schwartz,[1] **Ashtyn T. Whyte,**[1] **Mohanad Al-Nuaimi,**[2] **and Jeremy J. Donai**[1,a)]

[1] *Department of Communication Sciences and Disorders, West Virginia University, Morgantown, West Virginia 26506, USA*

[2] *Department of Mechanical and Aerospace Engineering, West Virginia University, Morgantown, West Virginia 26506, USA*

jeremy.donai@mail.wvu.edu

Abstract: Two experiments were conducted to evaluate the effects of increasing spectral bandwidth from 3 to 10 kHz on individual speaker recognition in noisy conditions (+5, 0, and −5 dB signal-to-noise ratio). Experiment 1 utilized h(Vowel)d (hVd) signals, while experiment 2 utilized sentences from the Rainbow Passage. Both experiments showed significant improvements in individual speaker identification in the 10 kHz bandwidth condition (6% for hVds; 10% for sentences). These results coincide with the extant machine recognition literature demonstrating significant amounts of individual speaker information present in the speech signal above approximately 3–4 kHz. Cues from the high-frequency region for speaker identity warrant further study.

[BHS]

Date Received: August 28, 2018 **Date Accepted:** October 28, 2018

1. Introduction

The utility of and cues present within the high-frequency region of speech have been studied over the previous decade (e.g., McClurg, 2018; Donai and Halbritter, 2017; Donai and Paschall, 2015; Monson *et al.*, 2014; Vitela *et al.*, 2015). The focus of these studies has primarily been consonant and vowel identification, gender identification, and mode of production (speaking vs singing) recognition. Together, these studies have documented the presence of perceptual information in the high-frequency portion of the speech signal (above approximately 4–5 kHz). Furthermore, Moore (2016) provided a review of studies highlighting the importance of speech energy above 3 kHz for sound-source localization and speech recognition in noise.

Additional studies have examined the use of high-frequency speech energy for automated speech and speaker recognition tasks. Donai *et al.* (2016) utilized a dataset of two male, two female, and two child speakers for classification. The authors extracted Mel-frequency cepstral coefficients (MFCCs) from vowel signals which were high-pass filtered at approximately 3 kHz and utilized them as inputs to a support vector machine classifier. Results showed that information above 3 kHz could be used to accurately classify vowel identity and talker type (male, female, or child). Earlier work by Hayakawa and Itakura (1994) showed continuous improvements in individual speaker identification as the bandwidth increased up to 16 kHz. In addition, Hayakawa and Itakura (1994) noted that for some speakers, the use of high-frequency energy improved speaker identification performance and that "a rich amount of speaker individual information is contained in the higher frequency band" (p. 140). Later, Hayakawa and Itakura (1995) studied the contributions of spectral regions to individual speaker identification by adding white noise to create a given signal-to-noise ratio (SNR) in the low- (0–4 kHz) and high-frequency (4–10 kHz) bands. Results showed the high-frequency band to be more resilient to the negative effects of competing noise, with fewer errors occurring in the 4–10 kHz band. Deshpande and Holambe (2011) investigated the potential benefits of additional high-frequency information for speaker identification in the presence of ecological car noise. Results suggested over 90% identification accuracy when using features from spectral energy between 4 and 8 kHz. Despite the aforementioned automated speaker recognition studies, little is known about the role and use of high-frequency speech energy in individual speaker identification among human listeners.

In a conference proceeding, Gallardo *et al.* (2013) reported results examining individual speaker identification through common communication interfaces varying in

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

bandwidth using eight male and eight female speakers. The authors reported that approximately half of the listeners participating in the study also served as speakers for the experiment and the remainder were colleagues working in the same office. As such, there was significant familiarity with the voices used in the study, even the listener's own voice. Results showed significant improvements in speaker identification when listeners were presented information through broadband devices. However, it is unknown how familiarity with the speakers used in the study influenced listener performance or interacted with signal bandwidth.

The purpose of the current experiments was to examine the effects of increasing signal bandwidth for individual speaker identification in competing noise conditions. Two bandwidth conditions (3 and 10 kHz) and three noise levels (+5, 0, and −5 SNR) were utilized. It was predicted that increasing signal bandwidth would improve speaker identification due to the presence of speaker identity cues above 3 kHz. The current project is one of the first to specifically address the topic from a behavioral perspective, but is theoretically founded in the machine recognition of speaker identity literature.

2. General methods

Two behavioral experiments were conducted to study the effects of signal bandwidth and noise on individual speaker identification. Experiment 1 utilized h/Vowel/d (hVd) signals produced by five males and five females between ages 18 and 35, whereas experiment 2 utilized three sentences from the Rainbow Passage produced by the same five males and five females. All signals were recorded at 44.1 kHz in a sound-treated audio suite using a Miktek C1 FET condenser microphone (Nashville, TN), with the talkers placed approximately 12 in. from the microphone. The speakers were instructed to maintain the mouth-to-microphone distance and speak using normal vocal effort throughout the recording session.

The signals were extracted and low-pass filtered using an Equiripple finite impulse response filter with a 500 Hz transition band, 120 dB attenuation, and 1 dB passband ripple. Filter cutoff frequencies were 3 kHz (narrowband condition) and 10 kHz (broadband condition). A pre-recorded 20-talker babble containing 10 male and 10 female talkers speaking in a general American dialect was utilized as the noise masker. Recordings were conducted in a sound-treated suite using a high-fidelity microphone (Shure SM81, Niles, IL) at a 44.1 kHz sampling frequency with 16-bit resolution. A noise masker was utilized to (1) assess the effects of various noise levels on speaker identification performance and (2) to reduce the likelihood of reaching a ceiling effect when presenting the signals in quiet. The use of 20-talker babble was intended to minimize the effects of informational masking due to listeners understanding segments of the masking noise. Prior to the experiment, the research staff listened to the masking noise and were unable to detect any portion of the linguistic message.

A graphical user interface (GUI) created in MATLAB and specifically designed for the task was utilized. Prior to the experiment, listeners were required to pass a hearing screening at 20 dB hearing level at octave frequencies from 250 to 8000 Hz in both ears. Listeners underwent training with the unfiltered signals in quiet until they could reach a minimum of 70% correct identification. During the training phase, listeners familiarized themselves with the unfiltered speech signals associated with the name on the GUI button selection by being able to click on the GUI and hear a random signal. The name associated with the signal turned red to indicate which talker produced the signal. Names for the male signals included: Mike, Chris, Matt, Josh, and Jake. Female names included: Jessica, Ashley, Emily, Sara, and Rachel. These names were selected because they were considered common names among the age group of listeners participating in the study. Prior to the experiment, study personnel and one additional listener from outside the study judged the voices of the male and female speakers from the experiment to be representative of male and female speaker categories.

During the experimental phase, listeners were presented the stimuli and were instructed to click on the name associated with the speaker. Upon making a selection, a randomly selected stimulus would play, and the process repeated until all signals were presented. The stimuli were presented diotically at approximately 65 dB sound pressure level via Sennheiser HD 380 Pro (Old Lyme, CT) circumaural headphones. Upon completion of the experiment, listeners were asked if they could identify any of the speakers from the listening session.

2.1 Experiment 1 (hVd)

Stimuli. The signals for the experiment were five hVds including: /æ/ as in had, /ɝ/ as in herd, /i/ as in heed, /ɔ/ as in hawd, and /u/ as in who'd produce by five male and five female speakers from a previously recorded corpus. Speakers were all previously judged to speak in a general American dialect. Prior to the experiment, the signals were checked for audible artifact or distortion by three study personnel, with none found. Mean fundamental frequency ($F0$) values extracted via PRAAT's "pitch tracking" feature were 113 Hz [standard deviation (SD) = 12.9] for the male signals and 212 Hz (SD = 17.6) for the female signals. These values were obtained by averaging $F0$ values from the steady-state portion of the hVd signal.

Equalization. Following digital filtering, the speech and noise signals were root-mean-square (RMS) equalized separately by filter cutoff. To create the various noise conditions, the level of the noise was adjusted to create the appropriate SNR ratio (+5, 0, and −5 dB) and added to the speech signals in MATLAB. The combined signals were then visually inspected and verified for sound quality by a listening check completed by the research staff.

Presentation. The experimental signals were blocked by talker gender and counterbalanced by filter cutoff and SNR. Within each presentation, the signals were randomized by the MATLAB GUI. Male and female stimulus sets were tested separately. This was done due to concern that combining the male and female talkers (10 voices instead of 5) would substantially increase the difficulty associated with learning the voice paired with the name on the GUI selection screen without adding value to the experiment. Twenty listeners (15 females) with normal hearing (mean age = 24.9 yrs, SD = 3.57) participated in the experiment.

2.2 Experiment 2 (Sentences)

Stimuli. Signals for the experiment were three sentences produced by the speakers described above from the Rainbow Passage, a commonly used speech production screening measure. The sentences were: "When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow," "The rainbow is a division of white light into many beautiful colors," "Throughout the centuries people have explained the rainbow in various ways." These sentences were chosen because they were all of similar duration and between 3 and 5 s in length between the talkers. Mean fundamental frequency ($F0$) values extracted via PRAAT's "pitch tracking" feature were 136 Hz (SD = 18.5) for the male signals and 229 Hz (SD = 18.4) for the female signals. These values represent an average $F0$ across the duration of each sentence.

Equalization. Following digital filtering, the speech and noise signals were RMS equalized separately by filter cutoff. To create the various noise conditions, the level of the noise was adjusted to create the appropriate SNR ratio (+5, 0, and −5 dB) and then added to the speech signals. The combined signals were then visually inspected and verified for sound quality by a listening check completed by three study personnel.

Presentation. The experimental signals were blocked by talker gender and counterbalanced by filter cutoff and SNR. Within each presentation, the signals were randomized by the MATLAB GUI. Male and female stimulus sets were tested separately. Twenty listeners (17 females) with normal hearing (mean age = 23.3 yrs, SD = 2.90) participated in the experiment.

3. Results

3.1 Experiment 1 (hVds)

Mean speaker identification performance for the female hVds was 47% [standard error (SE) = 1.77] and 54% (SE = 1.81) for the male hVds. For filter cutoff, mean speaker identification performance was 48% (SE = 1.87) in the 3 kHz cutoff and 54% (SE = 1.48) in the 10 kHz cutoff. Mean speaker identification performance was 60% (SE = 1.60) in +5 dB SNR condition, 53% (SE = 1.83) in the 0 dB SNR condition, and 40% (SE = 1.70) in the −5 dB SNR condition.

A mixed-model analysis of variance (ANOVA) was conducted. Mauchly's test of sphericity was not significant ($p > 0.05$) for noise condition, suggesting the assumption of sphericity was met. Results are shown in Fig. 1. A significant main effect was found for filter cutoff, $F(1, 19) = 16.63$, $p < 0.01$, $\eta_p^2 = 0.47$, gender, $F(1, 19) = 15.31$, $p < 0.01$, $\eta_p^2 = 0.45$, and for noise condition, $F(2, 38) = 112.27$, $p < 0.001$, $\eta_p^2 = 0.86$. Speaker identification was significantly better in the 10 kHz filter cutoff compared to the 3 kHz filtered condition. Additionally, speaker identification of the male talkers was significantly higher than the female talkers. *Post hoc* testing using Bonferroni

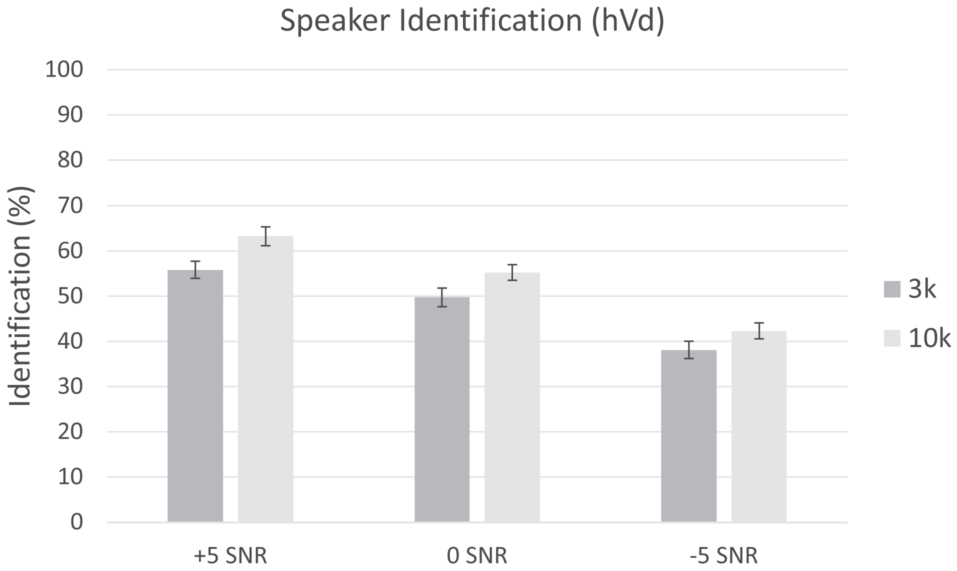

Fig. 1. Results from experiment 1 by noise condition and filter cutoff. Mean identification scores with standard error bars.

corrections showed significantly different ($p < 0.05$) speaker identification between all conditions (+5, 0, and -5 dB SNR), with performance decreasing with increasing noise level. The only significant interaction occurred between gender and noise condition, ($F(2, 38) = 6.44$, $p < 0.01$, $\eta_p{}^2 = 0.25$), with better performance on the male talkers in more favorable conditions (+5 and 0 dB SNR). All other interactions, including the three-way interaction, were non-significant ($p > 0.05$).

3.2 Experiment 2 (sentences)

Mean speaker identification performance for the female sentences was 65% (SE = 3.05) and 77% (SE = 2.44) for the male sentences. For filter cutoff, mean speaker identification performance was 66% (SE = 2.76) in the 3 kHz cutoff and 76% (SE = 2.30) in the 10 kHz cutoff. Mean speaker identification performance was 72% (SE = 2.32) in the +5 dB SNR condition, 73% (SE = 2.70) in the 0 dB SNR condition, and 69% (SE = 2.39) in the -5 dB SNR condition.

A mixed-model ANOVA was conducted. Mauchly's test of sphericity was not significant ($p > 0.05$) for noise condition, suggesting the assumption of sphericity was met. Results are shown in Fig. 2. A significant main effect was found for filter cutoff, $F(1, 19) = 15.92$, $p < 0.01$, $\eta_p{}^2 = 0.46$, gender, $F(1, 19) = 14.22$, $p < 0.01$, $\eta_p{}^2 = 0.43$, but not for noise condition, $F(2, 38) = 2.50$, $p = 0.095$, $\eta_p{}^2 = 0.12$. Results showed significantly better speaker identification in the 10 kHz filter cutoff condition compared to the 3 kHz filtered condition. Additionally, speaker identification for the male talkers was significantly higher than the female talkers. Significant interactions occurred between gender and noise condition [$F(2, 38) = 3.44$, $p < 0.05$, $\eta_p{}^2 = 0.15$] and cutoff and noise condition [$F(2, 38) = 8.26$, $p < 0.01$, $\eta_p{}^2 = 0.30$]. In the former, identification performance decreased for the female talkers in the noisier conditions (0 and -5 dB SNR). In the latter, better identification performance was observed for the 10 kHz condition for the noisier conditions (0 and -5 dB SNR). All other interactions, including the three-way interaction, were non-significant ($p > 0.05$).

4. Discussion

Overall, increasing the signal bandwidth significantly improved individual speaker identification for both hVds and sentences. The following is a summary of important findings from the study:

(1) Increasing the signal bandwidth from 3 to 10 kHz resulted in a 6% improvement in speaker identification for the hVd signals and 10% improvement for sentences.
(2) Speaker identification among the male signals was significantly higher than the female signals for both hVds and sentences.
(3) In experiment 2, the addition of high-frequency energy resulted in no decrement in speaker identification performance when the SNR decreased, whereas in the 3 kHz condition performance decreased with decreasing SNR.

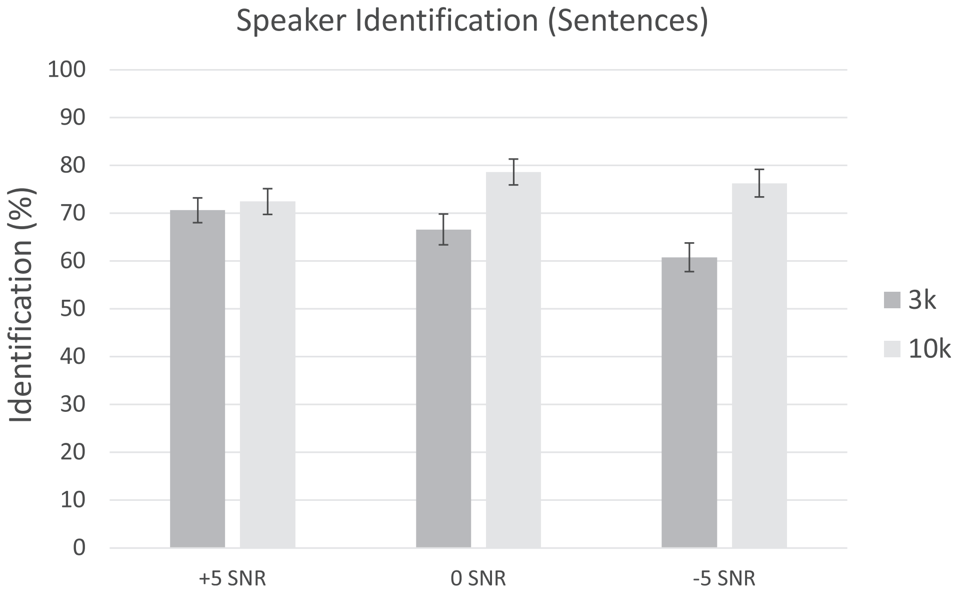

Fig. 2. Results from experiment 2 by noise condition and filter cutoff. Mean identification scores with standard error bars.

(4) The significant interaction between SNR and filter cutoff for sentences suggested that as noise level increased, increasing signal bandwidth significantly improved speaker identification accuracy. In other words, high-frequency energy became more beneficial for speaker identification at lower SNRs, which has been observed in the machine recognition literature. Deshpande and Holambe (2011) developed an algorithm that highly weighted high-frequency features (Teager Energy Operator based Cepstral Coefficients) and compared the results to the commonly used MFCCs, which have better resolution in the low-frequency region. Results showed that at low SNRs (i.e., 0 dB SNR), the proposed system outperformed MFCCs by nearly 20%, but at higher SNRs (i.e., +20–30 dB SNR) little difference in speaker identification was observed.

(5) Results of the current experiments represent the first (to our knowledge) highly controlled behavioral study to investigate the effects of adding (or conversely limiting) high-frequency energy on speaker identification abilities.

Upon completion of each experiment, the listeners were asked if they recognized and could name any of the speakers from the study. None of the 20 listeners from each experiment could do so. This ensured that there was limited *a priori* familiarity with the individuals serving as speakers in the study. The authors considered this to be an important methodological consideration given that the only other comparable study (Gallardo *et al.*, 2013) did not control for this, and it is therefore unknown how familiarity with the speakers (including their own voice) influenced the findings in the previous study.

The current studies have potential implications for the process of attending to a speaker of interest in degraded acoustic environments. It is possible that high-frequency energy provides redundant cues when following a speaker of interest in environments degraded by low-frequency noise. This is plausible due to the fact that high-frequency signals are highly directional (e.g., Carlile and Schonstein, 2006; Kocon and Monson, 2018; Monson *et al.*, 2012) and have been shown to provide important cues for localization (e.g., Best *et al.*, 2005; Carlile and Schonstein, 2006).

As previously described, it is clear from the literature that high-frequency energy plays an important role in the automated recognition of speaker identity, particularly in noisy environments. Additional research by Hu and Wang (2004) utilized amplitude modulation rate information from the high-frequency region to aid in extracting speech from a speech-in-noise mixture. The authors devised a system that utilized spectral energy from the resolved, low-frequency region, and amplitude modulations from the unresolved, high-frequency region. An approximate 5 dB improvement in SNR was found when utilizing amplitude modulation information from the high-frequency region compared to low-frequency information alone.

One interesting finding from the study was the fact that in both experiments listeners were more accurate in determining the identity of male speakers. The reason for this is unknown, but perhaps the male stimuli contained greater intergroup voice quality differences that assisted listeners in making perceptual judgements of individual speaker identity. The precise cues from both the low- and high-frequency portion of the spectrum utilized by listeners in this study warrants further analysis.

Taken together, the results of the current studies in conjunction with the previously reported studies described here suggest that useful information regarding the identity of a speaker is contained above 3–4 kHz. As such, future research should examine how hearing aids and other telecommunication devices preserve and transmit *speaker identity information* within this region. Additionally, because the highest filter cutoff utilized in the current experiments was 10 kHz, future studies should include a higher filter cutoff condition of approximately 14–16 kHz to determine if the presence of additional high-frequency energy positively influences performance. While the current study utilized headphones for stimuli presentation, future research should consider utilizing sound-field transducers to create a more ecological signal presentation.

5. Conclusions

This is the first (to our knowledge) highly controlled study to investigate the effects of signal bandwidth on identification of individual speakers among listeners with normal hearing. The current results coincide with the extant machine recognition literature in that the presence and use of high-frequency information from the speech signal improves the recognition of individual speakers, particularly in noisy environments. These findings have implications for hearing aids and other telecommunication devices, as preserving the cues within this frequency region appear to be of importance. The precise cues and information contained within the high-frequency portion of the spectrum warrants further study.

Acknowledgment

Portions of this project were funded by a Faculty Senate Research Grant provided by West Virginia University to J.J.D.

References and links

Best, V., Carlile, S., Jin, C., and van Schaik, A. (**2005**). “The role of high frequencies in speech localization,” J. Acoust. Soc. Am. **118**, 353–363.

Carlile, S., and Schonstein, D. (**2006**). “Frequency bandwidth and multi-talker environments,” in *Proceedings of the 120th Audio Engineering Society Convention*, Paris, France, pp. 1–8.

Despande, M. S., and Holambe, R. S. (**2011**). “Speaker identification using admissible wavelet packet based decomposition,” Int. J. Signal Process. **6**(1), 20–23.

Donai, J. J., and Halbritter, R. (**2017**). “Gender identification using high-frequency speech energy: Effects of increasing the low-frequency limit,” Ear Hear. **38**(1), 65–73.

Donai, J. J., Motiian, S., and Doretto, G. (**2016**). “Automated classification of vowel category and speaker type in the high-frequency spectrum,” Aud. Res. **6**(137), 1–5.

Donai, J. J., and Paschall, D. D. (**2015**). “Identification of high-pass filtered male, female, and child vowels: The use of high-frequency cues,” J. Acoust. Soc. Am. **137**(4), 1971–1982.

Gallardo, L. F., Möller, S., and Wagner, M. (**2013**). “Human speaker identification of known voices transmitted through different user interfaces and transmission channels,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 7775–7779.

Hayakawa, S., and Itakura, F. (**1994**). “Text dependent speaker recognition using the information in the higher frequency band,” *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 137–141.

Hayakawa, S., and Itakura, F. (**1995**). “The influence of noise on the speaker recognition performance using the higher frequency band,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 321–324.

Hu, G., and Wang, D. (**2004**). “Monaural speech segregation based on pitch tracking and amplitude modulation,” IEEE Trans. Neur. Net. **15**(5), 1135–1150.

Kocon, P., and Monson, B. B. (**2018**). “Horizontal directivity patterns differ between vowels extracted from running speech,” J. Acoust. Soc. Am. **144**(1), EL7–EL13.

McClurg, M. (**2018**). “Effect of high-pass filtering on perception of dialect and talker sex,” unpublished thesis, Ohio State University.

Monson, B. B., Lotto, A. J., and Story, B. H. (**2012**). “Directivity of low- and high-frequency energy in speech and singing,” J. Acoust. Soc. Am. **132**(1), 433–441.

Monson, B. B., Lotto, A. J., and Story, B. H. (**2014**). “Gender and vocal production mode discrimination using the high frequencies for speech and singing,” Front. Psych. **135**(1), 400–406.

Moore, B. C. J. (**2016**). “A review of the perceptual effects of hearing loss for frequencies above 3 kHz,” Int. J. Aud. **55**(12), 707–714.

Vitela, A. D., Monson, B. B., and Lotto, A. J. (**2015**). “Phoneme categorization and relying solely on high-frequency energy,” J. Acoust. Soc. Am. **137**(1), EL65–EL70.