Article

Language *and* Speech

Effects of Spectral Envelope and Fundamental Frequency Shifts on the Perception of Foreign-Accented Speech

Language *and* Speech
2022, Vol. 65(2) 418–443

DOI: 10.1177/00238309211029679
journals.sagepub.com/home/las
SAGE

Michelle R. Kapolowicz
University of California Irvine, USA

Daniel R. Guest
University of Minnesota, USA

Vahid Montazeri
University of Texas at Dallas, USA

Melissa M. Baese-Berk
University of Oregon, USA

Peter F. Assmann
University of Texas at Dallas, USA

Abstract

To investigate the role of spectral pattern information in the perception of foreign-accented speech, we measured the effects of spectral shifts on judgments of talker discrimination, perceived naturalness, and intelligibility when listening to Mandarin-accented English and native-accented English sentences. In separate conditions, the spectral envelope and fundamental frequency (F0) contours were shifted up or down in three steps using coordinated scale factors (multiples of 8% and 30%, respectively). Experiment 1 showed that listeners perceive spectrally shifted sentences as coming from a different talker for both native-accented and foreign-accented speech. Experiment 2 demonstrated that downward shifts applied to male talkers and the largest upward shifts applied to all talkers reduced the perceived naturalness, regardless of accent. Overall, listeners rated foreign-accented speech as sounding less natural even for unshifted speech. In Experiment 3, introducing spectral shifts further lowered the intelligibility of foreign-accented speech. When speech from the same foreign-accented talker was shifted to simulate five different talkers, increased exposure failed to produce an improvement in intelligibility scores, similar to

Corresponding author:
Michelle R. Kapolowicz, University of California Irvine, 114 Medical Sciences E, Irvine, CA 92697, USA.
Email: m.kapolowicz@uci.edu

the pattern observed when listeners actually heard five foreign-accented talkers. Intelligibility of spectrally shifted native-accented speech was near ceiling performance initially, and no further improvement or decrement was observed. These experiments suggest a mechanism that utilizes spectral envelope and F0 cues in a talker-dependent manner to support the perception of foreign-accented speech.

Keywords
Foreign-accented speech, fundamental frequency, spectral envelope, talker discrimination

Introduction

Listeners are sensitive to distinct properties of speech for perceiving phonetic and prosodic contrasts, such as the spectral envelope (related to vocal tract length and formant spacing) and the fundamental frequency (F0) (related to glottal pulse rate). In addition to assisting with the acoustic realization of different phonemes, spectral envelope/formant and F0 cues also aid listeners with accurate identification/familiarization of individual talkers (Kreiman & Sidtis, 2011). One functional benefit associated with talker familiarization is that it aids listeners with speech perception in both quiet (Nygaard & Pisoni, 1998) and noisy conditions (Johnsrude et al., 2013). Properties that allow listeners to recognize a voice as familiar may also help to resolve the problem of acoustic variability across talkers by means of a perceptual process known as talker normalization. Talker normalization is a mechanism that allows listeners to achieve perceptual constancy by attending to talker-specific properties such as the mean F0 and spectral envelope/formant pattern (Barreda, 2020). These properties provide a basis for interpreting acoustic variability that would otherwise result in ambiguity in the relationship between the acoustics and the perceptual realization of these sounds (Joos, 1948; Nearey, 1989; Wong & Diehl, 2003). In short, talker normalization helps listeners to recalibrate to each talker's speech patterns to assist with accurately interpreting subsequent utterances from that same talker (Nearey & Assmann, 2007), and this process seems to remain relatively stable unless a talker change is detected (Barreda, 2012; Magnuson & Nusbaum, 2007).

The relative ease of adapting to talker characteristics in native-accented speech is presumably facilitated by listeners' extensive familiarity with their native language; however, when perceiving foreign-accented speech, listeners often require longer exposure in order to adapt. Listeners perform better on intelligibility tasks when perceiving speech from a single foreign-accented talker versus several different foreign-accented talkers, at least in the early exposure period (Bent & Holt, 2013). In the multiple-talker scenario, where listeners are presented with speech from several foreign-accented talkers in succession, they need to adapt to each talker's characteristics within each utterance. Such adjustments are not necessary during the single-talker scenario, providing an explanation for the talker-dependent benefit. Specifically, the initial normalization process for a single talker would have applied to each successive utterance from that same talker, whereas when perceiving speech from different alternating talkers, the process of mapping the acoustic input onto stored representations becomes more difficult, since listeners must also adjust their perceptual maps for changes in vocal characteristics across talkers, as suggested by Bent and Holt (2013).

Increased exposure to speech from a single native-accented talker over time leads to improved recognition accuracy (Magnuson & Nusbaum, 2007). Similarly, increased exposure to foreign-accented speech from a single talker leads to improved recognition accuracy (Kapolowicz et al., 2018). This suggests that talker normalization contributes to the initial process of adapting to foreign-accented speech, just as it does in native-accented speech. However, this process may be

different from the mechanism that drives adaptation at later time points. For example, listeners are also able to adapt to multiple foreign-accented talkers who share the same native language (Bradlow & Bent, 2008; Sidaras et al., 2009) and to multiple foreign-accented talkers who do not share the same native language in an accent-independent manner (Baese-Berk et al., 2013; Bent & Holt, 2013), but these two processes require additional time compared to adapting to a single foreign-accented talker (Xie et al., 2018). The additional exposure time needed for adaptation to alternating talkers with accents which are perceived as foreign may be explained by listeners attempting to adjust to stable vocal characteristics across talkers while simultaneously altering their expectations to account for differences in pronunciation that deviate from the expected target signal. This could impede the process of rapidly adapting to different foreign-accented talkers, such as over the course of a conversation.

Our earlier work sought to clarify the effects of talker variability by investigating the relative contribution of spectral cues and temporal envelope cues in this perceptual process. In that work, we utilized a tone vocoder that preserves the temporal envelope of speech, while limiting talker-specific spectral cues. By varying the number of channels available to each listener, we found that, compared to native-accented speech, intelligibility and accent detection of foreign-accented speech were more adversely affected by spectral reduction (Kapolowicz et al., 2016). These results reveal that listeners rely on spectral pattern information to a greater extent for foreign-accented speech than for native-accented speech. A follow-up study using a 9-channel vocoder (Kapolowicz et al., 2018) further revealed that intelligibility performance when perceiving spectrally reduced native-accented speech was near ceiling when listeners heard only one talker over time as well as when listeners heard multiple interleaved talkers. Intelligibility scores were much lower when perceiving spectrally reduced foreign-accented speech, and no rapid accent-dependent adaptation (i.e., no improvement with increased exposure) occurred in conditions where listeners were exposed to multiple foreign-accented talkers despite sharing the same native language. However, listeners were able to partially adapt when they were exposed to spectrally reduced speech from a single foreign-accented talker, though never to levels observed for unprocessed foreign-accented speech. In this case, since talker-specific spectral cues were limited by the vocoder, listeners had to largely depend on other cues, such as duration and amplitude, which are less able to facilitate this rapid adaptation process.

These previous studies suggest that: (a) listeners adapt to foreign-accented speech in a talker-dependent manner; and (b) these talker-dependent cues reside somewhere within the spectral domain. To further test the hypothesis that these talker-dependent cues reside somewhere within the spectral domain, the present work aims to investigate the role of the spectral envelope and F0 when adapting to foreign-accented speech in a talker-dependent manner. In the present work, we co-varied the spectral envelope and F0, two properties known to provide cues for talker identity (e.g., Kreiman & Sidtis, 2011), in ways that mirror the observed pattern of covariation found across talkers in natural speech. This technique allowed us to extend our previous work by testing listeners' combined reliance on spectral envelope and F0 cues when perceiving foreign-accented speech compared to native-accented speech. We tested for the presence of an interaction between spectral shifts and foreign accent with tokens from single talkers and measured whether listeners can rapidly adjust to these spectral changes. We implemented this method as a way to investigate whether adapting to several different foreign-accented talkers is a more challenging perceptual process because listeners need to cope with shifts that are perceived as coming from different talkers, in addition to adjusting to non-native patterns in speech production. Shifting the temporal trajectories of the formants and F0 up or down along the frequency scale allowed us to access the extent to which these talker-specific source and vocal-tract properties enable listeners to adapt to foreign-accented speech in a talker-dependent manner. Our main hypothesis was that perception of

spectrally shifted foreign-accented speech from a single talker in terms of sentence intelligibility scores would be lower compared to unshifted foreign-accented speech from that same talker. Our secondary hypothesis was that rapid adaptation, entailing an improvement in sentence intelligibility scores with increased exposure, would only be observed with perception of unshifted foreign-accented speech from a single talker. These findings would suggest that listeners implement a talker-dependent listening strategy that relies on spectral envelope and F0 cues when rapidly adapting to a talker who has a foreign accent.

Whereas the tone vocoder utilized in our previous studies lacks a voicing source and has a reduced number of spectral channels which smears the spectrum and, therefore, lowers intelligibility, the present study used a signal processing technique that does not reduce the spectral resolution of speech. Here, we utilized the STRAIGHT analysis–synthesis system (Kawahara et al., 1999) to manipulate F0 and the spectral envelope in sentences recorded from five different native-accented talkers and five different foreign-accented talkers. Previous studies using STRAIGHT to implement spectral shifts have shown that this method preserves the spectral pattern and (over the range explored here) largely preserves intelligibility (Kawahara et al., 2005).

A number of studies have shown that intelligibility can still be preserved (i.e., listeners can quickly adjust) when speech is shifted along the frequency scale (Sjerps et al., 2011; Smith et al., 2005). For instance, Smith et al. (2005) observed that intelligibility can remain high when the spectral envelope and F0 of speech are shifted even when these shifts extend beyond the typical human range. However, sufficiently large spectral shifts can reduce the intelligibility and naturalness of speech. Such effects can vary as a function of age and gender of the talker. For example, Assmann and Nearey (2008) reported that identification accuracy of vowels spoken by men declines with downward shifts of the spectral envelope and F0 compared to downward shifts in vowels spoken by women and children. Similarly, there is a greater decline in intelligibility with upward shifts in children's speech compared to adult speech (Assmann & Nearey, 2007 (decline observed for identification of words in connected speech); Fu & Shannon, 1999; Assmann & Nearey, 2008 (decline observed for identification of vowels). Additionally, when the spectral envelope and F0 are manipulated in ways that do not preserve their natural covariation in speech (i.e., shifting the spectral envelope and F0 in opposite directions), perceptual costs are higher than when they are coherently shifted (Assmann & Nearey, 2008). These results imply that declines in perceptual performance are related to the natural ranges of the formant frequencies (or other features of the spectral envelope) across age and gender classes, and that these constraints include the spectral envelope in combination with F0.

These previous studies were concerned with spectrally shifted native-accented speech. To our knowledge, no such investigations have been reported for spectrally shifted foreign-accented speech. For this reason, we performed two control experiments prior to our main experiment to ensure that spectral shifting using STRAIGHT would produce similar perceptual results for both accents in terms of talker variability and naturalness. Specifically, the first experiment was conducted to ensure that spectrally shifted sentences from a single talker would be perceived as different talkers for both accents. The second experiment was conducted to assess perception of naturalness for spectrally shifted sentences for both accents. Further description is provided below.

Our first control experiment tested the hypothesis that shifting the spectral envelope and F0 in ways that mirror the natural covariation of these properties across talkers will produce changes in perceived talker identity. When manipulations that shift the spectral envelope and F0 of speech from a single talker are implemented, these shifts may result in listeners perceiving the speech as coming from different talkers (Gaudrain et al., 2009; Kuwabara & Sagisaka, 1995;). Our motivation to ensure that listeners perceived different talkers stemmed from previous reports showing that talker normalization is disrupted when listeners perceive that they are hearing different talkers

(Barreda, 2012; Magnuson & Nusbaum, 2007). It was predicted that shifting the spectral envelope and F0 in the same direction (up or down) beyond a certain limit would induce the perception of a different talker, regardless of whether or not the talker was foreign-accented. To test this prediction, we adopted a scaling function (described below in Section 2) derived from acoustic measurements of natural speech from men, women and children (Assmann et al., 2008). The spectral envelope (formant frequencies) and F0 scale by different proportions across talkers. For example, F0 is nearly an octave higher in adult female voices compared to adult males, while formant frequencies increase by about 15% (Nearey, 1989). Additional studies have also reported the importance of these two acoustic parameters for talker discrimination, with an increased reliance on spectral envelope cues over F0 cues. For example, Baumann and Belin (2010) used multidimensional analysis to conclude that, for both male and female voices, these cues are sufficient to represent perceived talker similarity. Bachorowski and Owren (1999) also found evidence using statistical discriminant classifications of talker identity to support the role of spectral envelope and F0 cues, with spectral envelope cues being more reliable than F0 cues. Compatible with these findings, Kuwabara and Takagi (1991) and Gaudrain et al. (2009) demonstrated that vocal-tract length cues (related to the spectral envelope) weigh more heavily than source characteristics (related to F0) for perceived talker identity. It was, therefore, expected that covarying both of these parameters simultaneously would give listeners the impression that they were hearing different talkers.

Our second control experiment tested whether shifting the spectral envelope and F0 would affect the perceived "naturalness" of native-accented and foreign-accented speech. Naturalness judgments provide evidence that spectral shifts create novel voices that are representative of real voices. Assmann et al. (2006) showed that listeners judge spectrally shifted sentences as being more natural when formant frequencies and F0 were shifted in a fashion which matched their covariation in actual human voices. Since we utilized shift factors that have been shown to mirror the covariation found in the voices of men, women, and children (Assmann & Nearey, 2008), we predicted that naturalness might be reduced with shifts that produced formant frequencies and F0s outside of the typical ranges for human voices but would otherwise be preserved. However, both regional and foreign accents that differ from a native speaker's own dialect have been shown to affect listeners' ratings of naturalness (Mackey et al., 1997). For this reason, listeners were explicitly instructed to avoid judging naturalness based on the perceived accentedness or intelligibility of a talker; therefore, we did not expect to see an effect of the talker's accent (native or foreign) on perceived naturalness.

Our third (main) experiment was designed to test the expectation that if listeners perceive spectrally shifted speech from a single talker as originating from different talkers (tested in Experiment 1), then their intelligibility scores should be lower in conditions where they heard spectrally shifted foreign-accented speech as opposed to unshifted speech from a single foreign-accented talker. This hypothesis was motivated by an explanation proposed by Bent and Holt (2013) describing how adapting to different foreign-accented talkers is a more challenging perceptual process than adapting to different native-accented talkers or even to a single foreign-accented talker: listeners need to reconcile with speech that they perceive as coming from different talkers in addition to adjusting to non-native patterns in speech production. Experiment 3 also investigated the ability of listeners to rapidly adapt to spectrally shifted foreign-accented speech with increased exposure. We predicted that intelligibility scores in the spectrally shifted condition would be similar to those in the actual multiple talker condition, namely, no rapid adaptation would be observed. This outcome would provide support for the idea that talker normalization is especially important in rapid adaptation to foreign-accented speech. Specifically, when listening to a single talker, listeners utilize talker-specific vocal-tract length and glottal pulse rate cues to aid with online processing of foreign-accented speech. When these cues vary

as they do across different talkers, rapid adaptation is impaired. Conversely, we did not expect spectral shifts to affect intelligibility of native-accented speech, since our previous work using the same paradigm found no difference in intelligibility scores across single-talker and multiple-talker conditions (Kapolowicz et al., 2018). The native-accented spectrally shifted condition was included for comparative purposes and to demonstrate that listeners have no trouble perceiving spectrally-shifted native-accented speech using the signal processing implemented in the present work, which resulted in clean speech signals, with minimal artifacts/distortions that could affect intelligibility (as tested in Experiment 2). Based on this reasoning, we predicted that intelligibility scores for spectrally shifted sentences would be similar to baseline intelligibility scores (reported below in Section 2) for unshifted sentences.

The motivation for the present work was to better understand the mechanism that enables normal-hearing listeners to adapt to foreign-accented speech, and to inform studies examining how this adaptation process may be impaired for cochlear implant (CI) users. Access to talker-specific voice cues such as the spectral envelope and F0 is limited in CIs (Fuller et al., 2014; Gaudrain & Başkent, 2018). Limited access to such cues hinders perception of native speech from different talkers for CI users (Chang & Fu, 2006; Kaiser et al., 2003). This detriment may be more pronounced when the talkers are different genders, as was demonstrated using cochlear-implant simulated hearing in normal-hearing listeners (Tamati et al., 2020). CI users also struggle with perception of foreign-accented speech compared to normal-hearing listeners (Ji et al., 2014; Tamati et al., 2021), and adaptation is impaired even when perceiving speech from only a single foreign-accented talker (Kapolowicz et al., 2020). The results from the present work, which attempts to emphasize the importance of talker-specific voice cues when adapting to foreign-accented speech, may also have direct implications to help explain the added difficulties experienced by CI users when they are perceiving foreign-accented speech.

2 Speech materials and signal processing

2.1 Talkers

Harvard sentences (IEEE, 1969) spoken by five native-accented and five foreign-accented American English talkers (three females and two males per accent condition) were utilized for the listening experiments. Sentences from one additional female native-accented talker were used for task familiarization in the intelligibility experiment. Native-accented talkers were monolingual speakers of American English who had only resided in Texas, ranging in age from 18 to 38 years (mean age: 23). Using a 9-point Likert scale, where 1 corresponded to having "no foreign accent," and 9 corresponded to "heavily foreign-accented," all native-accented talkers were given a rating of 1. Baseline intelligibility scores in quiet for native-accented talkers ranged from 95.8% to 96.9% (mean: 96%). The foreign-accented group consisted of Mandarin-accented talkers who had only resided in Taiwan and Texas, ranging from 1 month to 22 years in Texas. The age range for the foreign-accented talkers was from 18 to 47 years (mean age: 30.6). Using the same 9-point Likert scale as for the native-accented talkers, ratings for the foreign-accented talkers ranged from 7 to 8 (mean: 7.5). Baseline intelligibility scores in quiet for these talkers were as follows: females, 57.96%, 63.51%, and 73.77%; and males, 70.67%, and 73.87% (mean: 67.96%). Digital waveforms were stored at a sampling rate of 48 kHz and 16-bit resolution and root-mean-square-equalized across talkers and sentences. All talkers reported no hearing or speech impairments and passed a hearing screening at 20 dB hearing level (HL) at octave frequencies from 250 Hz to 8000 Hz in both ears. Talkers were students of the University of Texas at Dallas, and each talker was paid US$20 for participation. All recording procedures were reviewed and approved by the University of Texas at Dallas Institutional Review Board.

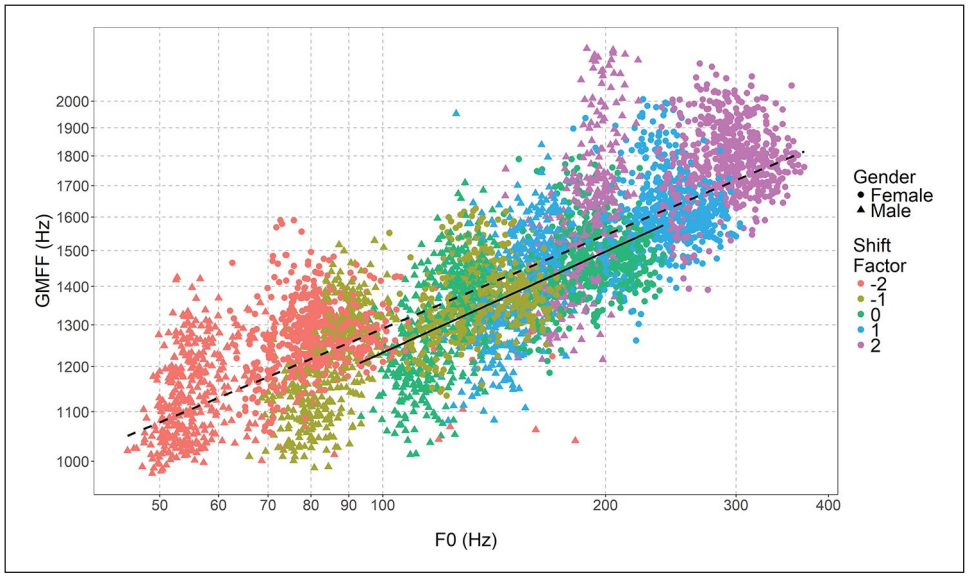

Figure 1. Geometric average of the lowest three formant frequencies (GMFF) plotted against fundamental frequency (F0) for shifted and unshifted stimuli. Decreasing the spectral envelope and F0 results in deeper-sounding voices, whereas increasing the spectral envelope and F0 results in voices sounding more female-like or child-like. The dotted line shows the regression fit for shifted stimuli, while the solid line displays the fit for unshifted stimuli only. Note that shifted stimuli largely match the natural covariance between GMFF and F0 found in unshifted stimuli.

2.2 Speech processing

Frequency scaled sentences were constructed by shifting the spectral envelope and F0 while preserving broadband slow modulation envelope cues using STRAIGHT (Kawahara et al., 1999). Seven versions of each sentence were synthesized, and in each version the spectral envelope and F0 were shifted in the same direction and were proportional in magnitude. The spectral envelope of each sentence was shifted up or down based on the following scale factor, denoted by sf_{ENV} as given by Equation (1):

$$sf_{ENV} = 1 + 0.08k, \tag{1}$$

where $k \in \{-3, -2, -1, 0, 1, 2, 3\}$ denotes the shift levels (8% step size). Similarly, F0 for each sentence was shifted up or down based on the following scale factor, denoted by sf_{F0} as given by Equation (2):

$$sf_{F0} = 1 + 0.296k, \tag{2}$$

where $k \in \{-3, -2, -1, 0, 1, 2, 3\}$ denotes the shift levels (approximately 30% step size). The scale factors were chosen to preserve the natural covariation between the spectral envelope and F0 observed in natural speech (Assmann & Nearey, 2008). The linear relationship between the geometric mean of the lowest three formant frequencies (GMFF) and F0 was maintained for shifted stimuli (data excludes extreme shifts of ± 3) as shown in Figure 1. The sample points show acoustic measurements of GMFF (reflecting the spectral envelope shift) and F0 for the shifted and

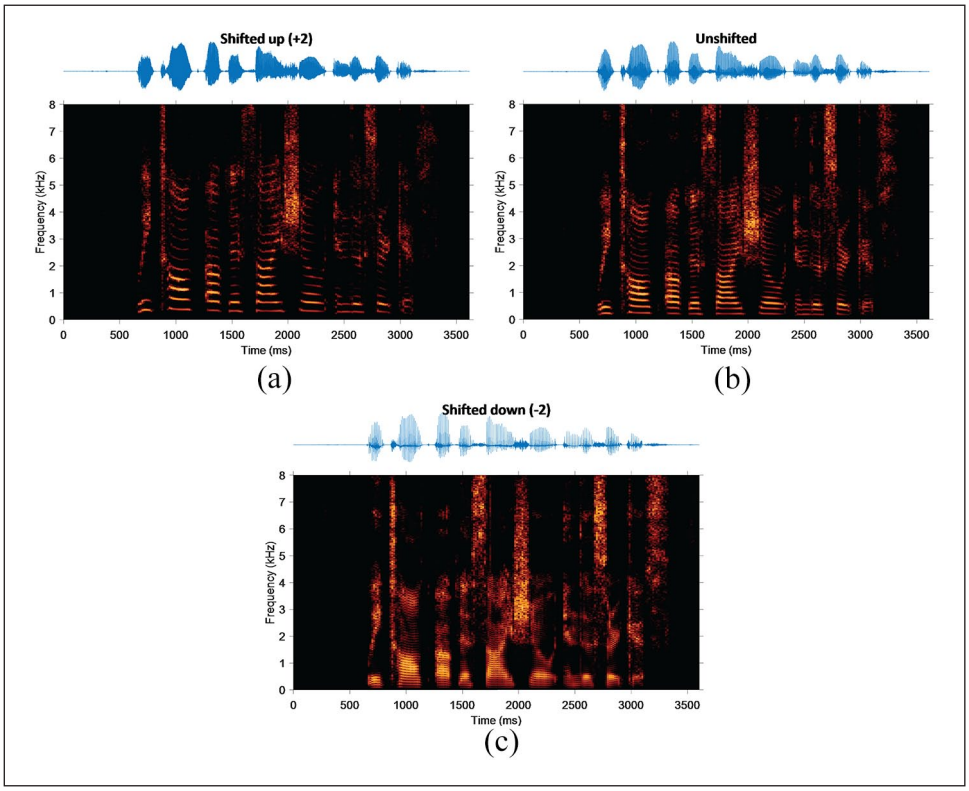

Figure 2. (a) Time-domain and spectrogram representations of the utterance "We talked of the side show in the circus" spoken by a female foreign-accented talker representing the signal when spectral envelope and F0 are shifted up by two increments (+2); (b) time-domain and spectrogram representations of the utterance "We talked of the side show in the circus" spoken by the same female foreign-accented talker representing the signal when spectral envelope and F0 are unshifted; and (c) time-domain and spectrogram representations of the utterance "We talked of the side show in the circus" spoken by the same female foreign-accented talker representing the signal when the spectral envelope and F0 are shifted down by two increments (-2).

unshifted versions, sampled at 2 milliseconds intervals from the voiced portions of the sentences used in the experiment. Figure 2 provides time-domain and spectrogram representations of unprocessed and spectrally shifted stimuli. The largest shifts of +3 and -3 were not included in the talker discrimination experiment or the intelligibility experiment. These extreme shifts were presented in the naturalness ratings experiment to provide listeners with examples of shifts which were outside the range that is typical for human speech sounds.

3 Experiment 1: Talker discrimination

3.1 Listeners

One hundred and sixty monolingual, native speakers of English (age range: 18–52 years, mean: 21.5 years) were recruited for participation ($n = 80$ per accent condition, $n = 32$ per shift factor condition; conditions are described below). Participants had only ever resided in Texas and

reported variable exposure to foreign-accented speech. All participants reported normal hearing and passed a hearing screening at 20 dB HL at octave frequencies from 250 Hz to 8000 Hz in both ears. Participants were students at the University of Texas at Dallas and were compensated with course credit. The experiment was conducted in a sound-attenuating booth. Stimuli were presented at a comfortable level (around 65 dB sound pressure level (SPL)) through Sennheiser HD-598 headphones using Tucker-Davis System 3 and RP2.1 hardware. Stimuli and conditions were randomized and presented using custom scripts created in MATLAB (The MathWorks, Inc.). All experimental procedures were reviewed and approved by the University of Texas at Dallas Institutional Review Board.

3.2 Design and procedure

A two-alternative forced choice procedure was employed wherein listeners heard two sentences in sequence (referred to as the baseline sentence and the test sentence, respectively) in each trial and were asked to judge whether the sentences were spoken by the same or different talkers. The talker discrimination experiment was a $2 \times 2 \times 5 \times 5$ mixed design: talker accentedness (native-accented or foreign-accented) by same/different talker (same or different talker across the two sentences) by baseline shift factor (-2, -1, 0, 1, or 2) by test shift factor (-2, -1, 0, 1, or 2). Accentedness and baseline shifts were treated as between-subjects factors while the remaining two factors were considered as within-subjects factors. This design allowed for every listener to provide responses for all levels of same/different talker and test shift factors. The dependent variable measured was listeners' judgments of the talkers as being either the same or different; their choice was dichotomous. Our rationale for implementing a mixed design in this experiment was due to a limitation of having only 100 sentences in our talker database. By utilizing a mixed design approach here, we ensured that listeners never heard the same sentence more than once, as sentence familiarity could confound talker identity (i.e., being familiar with a sentence might be perceived as also being familiar with the talker). The choice of stimuli was a random effect, as the sentences were selected randomly, therefore, it was not necessary to pair the stimuli with the identity factor levels. The above design resulted in a total of 10 conditions per session. Listeners heard a total of 100 sentences (10 sentences per condition) for this experiment throughout a single session. Sentences were selected randomly without replacement, and the experiment was completed in one session that lasted approximately 30 minutes.

3.3 Results

A mixed effects logistic regression model was used to fit the listeners' binary responses (whether the stimuli in each trial were spoken by the same or different talkers). Model fitting was carried out in R version 3.6.1 using the lme4 package (Bates et al., 2015). All levels of the shift difference, talker difference, accentedness, as well as their interactions were included as fixed effects with participants included as a random effect (i.e., random intercept) in the regression model. The baseline and test factors were transformed into the absolute value of the difference between the two. For example, the baseline and test shifts of 1 and 3 were transformed into a shift difference of ($|3-1| = 2$). Likewise, the baseline and test shifts of 3 and 1 were transformed into a shift difference of ($|1-3| = 2$). Such transformations lead to five levels of shift difference: 0, 1, 2, 3, and 4.

The logistic regression model is summarized in Table 1. The sign and value of the estimated coefficients in Table 1 show the association of each factor level with the log odds ratio of listeners judging the talkers in the baseline and test stimuli as being the *same*. The summary model shows

Table 1. Logistic regression model summary.

Formula: response ~ talker_difference + shift_difference + foreign_accent + shift_difference * talker_difference + shift_difference * foreign_accent + talker_difference * foreign_accent + shift_difference * talker_difference * foreign_accent + (1 | listener)

Random effects

Groups	Variance	Standard error
Listener (intercept)	1.680	1.296

Fixed effects

Factor Level	Estimated coefficient	Standard error	*Z* value	Pr(>|z|)	†
(Intercept)	−1.805	0.207	−8.738	< 2e–16	***
Talker difference (same)	5.526	0.318	17.359	< 2e–16	***
Shift difference (1)	−0.927	0.205	−4.515	6.34e–06	***
Shift difference (2)	−1.128	0.232	−4.867	1.13e–06	***
Shift difference (3)	−1.724	0.295	−5.837	5.31e–09	***
Shift difference (4)	−1.746	0.386	−4.524	6.06e–06	***
Foreign accent (native)	0.535	0.284	1.881	0.06	x
Talker difference (same): shift difference (1)	−4.346	0.364	−11.944	< 2e–16	***
Talker difference(same): shift difference (2)	−5.775	0.416	−13.875	< 2e–16	***
Talker difference (same): shift difference (3)	−5.180	0.466	−11.106	< 2e–16	***
Talker difference (same): shift difference (4)	−5.409	0.579	−9.336	< 2e–16	***
Shift difference (1): foreign accent (native)	−0.619	0.288	−2.146	0.032	*
Shift difference (2): foreign accent (native)	−0.339	0.315	−1.076	0.282	
Shift difference (3): foreign accent (native)	−0.554	0.428	−1.294	0.196	
Shift difference (4): foreign accent (native)	−0.993	0.595	−1.669	0.095	x
Talker difference (same): foreign accent (native)	0.0919	0.505	0.182	0.856	
Talker difference (same): shift difference (1): foreign accent (native)	−0.003	0.566	−0.006	0.996	
Talker difference (same): shift difference (2): foreign accent (native)	0.391	0.615	0.635	0.526	
Talker difference (same): shift difference (3): foreign accent (native)	−0.437	0.725	−0.603	0.546	
Talker difference (same): shift difference (4): foreign accent (native)	0.110	0.901	0.122	0.903	

Note: † significance codes: '***', $p < 0.001$; '**', $p < 0.01$; '*', $p < 0.05$; 'x', $p < 0.1$.

that the listeners more often judged the talkers as being the same when the talkers in the baseline and test stimuli were, in fact, the same. More importantly, listeners judged the talkers as being the same less often when there was a shift difference between the two stimuli. This likelihood further decreases with increasing the shift difference level.

To compare the above regression model against a null model (a model that only includes the intercept), a Chi-squared (χ^2) test of goodness of fit was applied. To do this, we performed an analysis of deviance on the above model. Results showed that adding the shift difference and talker difference factors to the model significantly improved the model fit compared to the null model, $\chi^2(4) = 583.576$, $p < 0.001$ and $\chi^2(1) = 208.597$, $p < 0.001$, respectively. In addition, the interaction between the shift difference and talker difference as well as the interaction between the shift difference and talkers' foreign-accentedness resulted in a significant improvement in the model fit, $\chi^2(4) = 379.287$, $p < 0.001$ and $\chi^2(4) = 13.396$, $p < 0.05$, respectively. Noticeably, foreign-accentedness, alone, did not have a significant increase in the model fit, $\chi^2(1) = 0.6607$, $p = 0.416$. Nor did the interaction between the talker difference and talker accentedness, $\chi^2(1) = 0.6699$, $p = 0.413$.

For the native and foreign-accented conditions, we calculated the d-prime values as the difference between the Z-transforms of the hit and false alarm rates ($d' = \mathrm{Z}(\mathrm{Hit}) - \mathrm{Z}(\mathrm{False\,alarm})$). A larger d-prime indicates more reliable sensitivity to talker differences while the d-prime (d′) values close to zero indicate near chance performance in the talker discrimination task. For the conditions in which the baseline and test shift factors were *different*, the d′ values were 0.275 and 0.312 for the foreign-accented and native talkers, respectively. For the conditions in which the baseline and test shift factors were the *same*, the d′ values were 2.66 and 2.70. These results are indicative that, when the shift factors were the *same* across baseline and test sentences, the spectral shifting implemented was effective at allowing listeners to reliably detect whether each talker was the same or different for each comparison. However, when the shift factors were *different*, the d′ values dropped, indicating that the listeners were no longer able to tell whether or not the talkers were the same (specifically, shifted speech from the same talker was perceived as being from a different talker).

To compare different levels of the variables than those presented in the fitted model, post-hoc tests based on the above logistic regression model were performed using the glht function in the "multcomp" package in R (Hothorn et al., 2008, p. 12). Results are reported where *M* corresponds to the mean difference, *S* corresponds to the standard error, and *Z* corresponds to the Z score. Results revealed a significant effect of shift difference on listeners' judgments of talker discrimination for trials wherein the two talkers were the same, $M = 9.428$, $S = 0.470$, $z = 20.043$, $p < 0.001$. Also, using the same analysis, foreign accent did not reveal a significant effect, $M = 0.535$, $S = 0.284$, $z = 1.881$, $p = 0.116$. Figure 3(a) displays listeners' judgments for the conditions in which the talkers in each trial were different. Figure 3(b) shows listeners' judgments for the conditions in which the talkers in each trial were the same. The relevant findings from the present study indicate that shifting the spectral envelope and F0 information of speech stimuli in 8% and 30% increments, respectively, results in listeners' judgments of the same talkers as being convincingly *different* talkers regardless of talker accent.

4 Experiment 2: Naturalness ratings

4.1 Listeners

One hundred and sixty monolingual, native speakers of English (age range: 18–52 years, mean: 21.5 years) participated in the naturalness ratings. These same participants completed Experiment 2 immediately after Experiment 1. The experiment was conducted in a sound-attenuating booth, and stimuli were presented at a comfortable level (around 65 dB SPL) through Sennheiser HD-598 headphones using Tucker-Davis System 3 and RP2.1 hardware. Stimuli and conditions were randomized and presented using custom scripts created in MATLAB (The MathWorks, Inc.). The experimental procedures were reviewed and approved by the University of Texas at Dallas Institutional Review Board.

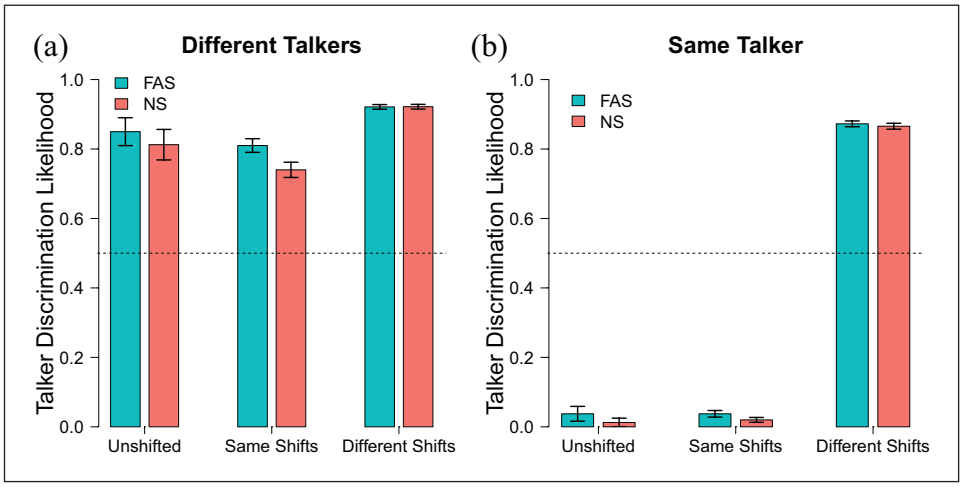

Figure 3. (a) Listeners' likelihood of judging talkers as being different when listeners compared two sentences from two different talkers in each trial. Values of 0 and 1 correspond to "same" and "different," respectively (FAS = foreign-accented speech, NS = native speech). Judgments were binary (same or different talkers). The Unshifted condition represents stimuli for which the talkers were different in a given trial and their speech was unshifted. The Same Shifts condition comprises trials for which the stimuli were processed with the same shift factor and the talkers for each stimulus were different. The Different Shifts condition represents when the first and second stimuli in each trial were processed with different shift factors and the talkers for each stimulus were different. Chance performance is shown by the horizontal dotted line. Error bars show the standard error of the mean (*SEM*) across listeners; and (b) listeners' likelihood of judging talkers as being different when listeners compared two sentences from the same talker in each trial. Values of 0 and 1 correspond to "same" and "different," respectively. Judgments were binary (same or different talkers). The Unshifted condition represents stimuli for which the talker was the same in a given trial and the talker's speech was unshifted. The Same Shifts condition comprises trials for which the stimuli were processed with the same shift factor and the talker for each stimulus was the same. The Different Shifts condition represents when the first and second stimuli in each trial were processed with different shift factors and the talker for each stimulus was the same. Chance performance is shown by the horizontal dotted line. Error bars show the *SEM* across listeners.

4.2 Design and procedure

In this experiment, listeners were asked to judge the naturalness of talkers using a 6-point Likert scale: extremely natural; natural; somewhat natural; somewhat unnatural; unnatural; and extremely unnatural. The naturalness experiment was a $2 \times 5 \times 7$ nested repeated measures design (Abdi et al., 2009, p. 373): accent by talker by shift factor ($\pm 3, \pm 2, \pm 1, 0$), where each variable had the same levels as in Experiment 1.[1] Listeners rated the naturalness of 70 sentences total throughout a single session. Listeners were informed that they would hear computer-processed speech that could come from males and females of any age range. They were also informed that the speech could range from native-accented speech to heavily foreign-accented speech. Listeners were instructed *not* to rate the level of naturalness for each sentence based on how intelligible the words were, nor on how heavily a foreign accent was perceived, but rather to focus on distinguishing between voices which sound like they could come from an actual human being (which should be rated as more natural) and voices that sound more fictitious, such as a cartoon character or a monster (which should be rated as less natural). The duration to complete the experiment was approximately 30 minutes, and the experiment was completed in a single session.

4.3 Results

A nested factor repeated measures analysis of variance (ANOVA) was conducted wherein item-wise data entailed one row per trial. Results revealed a significant main effect of accent, $F(1, 9022) = 355.600$, $p < 0.001$, partial $\eta^2 = 0.038$, a significant main effect of spectral shifting, $F(6, 9022) = 3094.720$, $p < 0.001$, partial $\eta^2 = 0.673$, a significant effect of talker nested in accent, $F(8, 9022) = 27.010$, $p < 0.001$, partial $\eta^2 = 0.023$, and a significant interaction between accent and spectral shifting, $F(6, 9022) = 14.050$, $p < 0.001$, partial $\eta^2 = 0.009$. As expected, a large effect was observed for spectral shifting, with accent only having a small effect on results. Analyses were performed in R 3.4.0 using the stats package for the ANOVA and the heplots package for measuring effect sizes (Fox et al., 2018).

Figures 4(a) and 4(b) show the mean naturalness ratings across shift factors for each foreign-accented and native-accented talker, respectively. Figure 5 displays the mean ratings for each accent condition. Figures 4(a), 4(b) and 5 show that, for extreme downward spectral shifting (-3), male voices were perceived as extremely unnatural regardless of the accent of the talker. This was expected, since this extreme shift (which is outside of the normal human speech range) was meant to serve as an anchor point for "extremely unnatural" voices. Extreme upward shifts (3) were still perceived as less natural for male and female speech, though not to the extent that extreme downward shifts had on male voices. Only male voices were perceived as somewhat unnatural when spectrally shifted down two steps. These results replicate previous findings from Fu and Shannon (1999) (who shifted only the spectral envelope in a noise vocoder) and Assmann and Nearey (2008) (who simultaneously shifted both the spectral envelope and F0 using STRAIGHT). They found a general pattern that, for males, upward shifts lead to formant patterns that fall within the natural range of female and child voices, while downward shifts produce formant patterns that are atypically low with respect to the distribution. For females, downward shifts lead to formant patterns that fall within the natural range of male voices, while upward shifts fall within the natural range of child voices. They also found that downward shifts had a greater impact on men's voices compared to women's and children's voices. Specifically, they reported a greater impairment in vowel identification when listeners heard male voices that were spectrally shifted down compared to when listening to women's or children's voices shifted down. Also similar to the aforementioned findings, we found that an upward scaling by a factor of two for male and female talkers did not severely impact naturalness ratings in either accent condition. The above ANOVA revealed a significant effect of accentedness on the naturalness ratings, indicating a bias in the naturalness judgments based on the talkers' accentedness. As shown in Figure 5, listeners rated foreign-accented talkers as less natural than native-accented talkers in all conditions including when speech was unshifted. This was an unexpected finding considering that listeners were explicitly instructed not to rate naturalness based on how intelligible or how foreign-accented a talker was perceived to be.

5 Experiment 3: Intelligibility

5.1 Listeners

Participants consisted of 64 monolingual, native speakers of English (age range: 18–43 years, mean: 24.2 years) who had only ever resided in Texas and reported variable exposure to foreign-accented speech ($n = 16$ per condition). These participants did not participate in either of the first two control experiments. All participants reported normal hearing and passed a hearing screening at 20 dB HL at octave frequencies from 250 Hz to 8000 Hz in both ears. Participants were students at the University of Texas at Dallas and were compensated with course credit. The experiment took

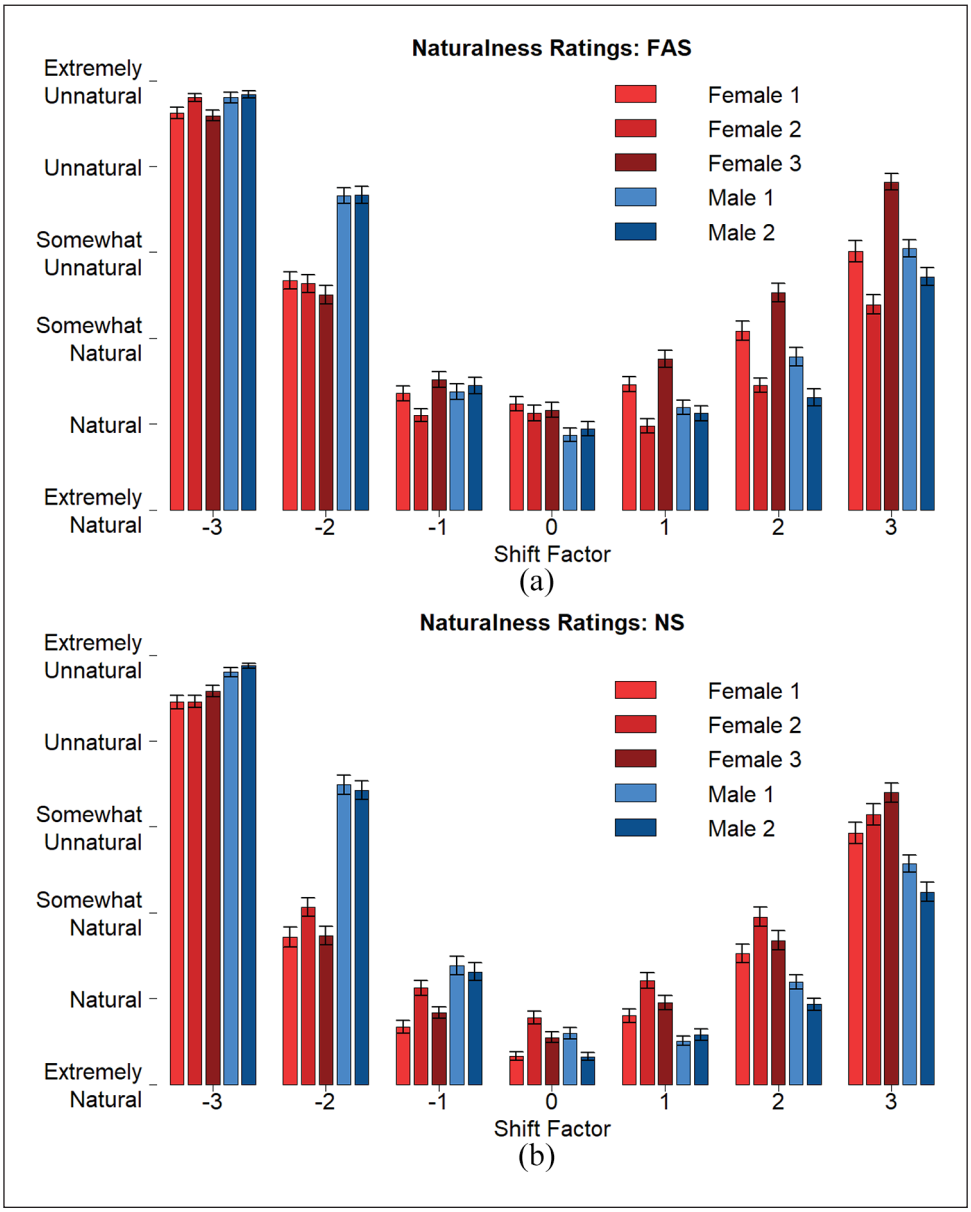

Figure 4. (a) Mean naturalness ratings across shift factors for each talker in the foreign-accented speech condition. A shift of 0 entails that no shifts were made to the speech stimuli. Positive shift factors indicate upward scaling of the spectral envelope and fundamental frequency (F0) for each sentence, and negative shift factors indicate downward scaling of the spectral envelope and F0 for each sentence. Standard error of the mean (*SEM*) bars are shown; and (b) mean naturalness ratings across shift factors for each talker in the native-accented speech condition. A shift of 0 entails that no shifts were made to the speech stimuli. Positive shift factors indicate upward scaling of the spectral envelope and F0 for each sentence, and negative shift factors indicate downward scaling of the spectral envelope and F0 for each sentence. SEM bars are also shown.

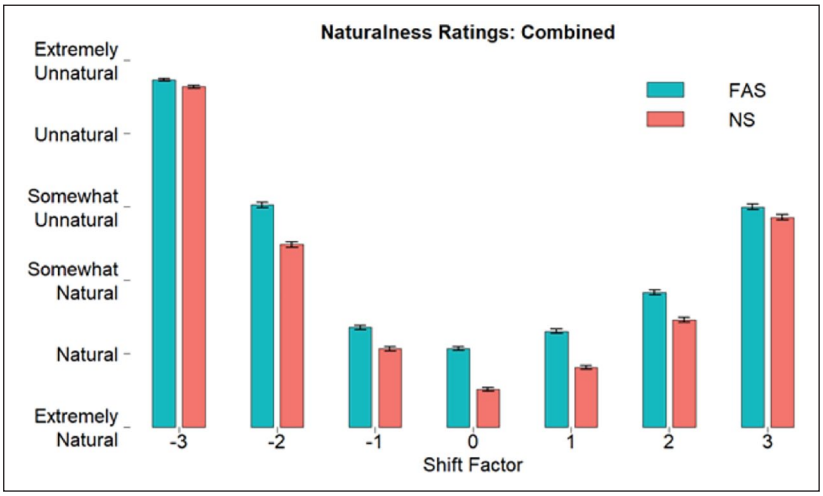

Figure 5. Mean naturalness ratings for each accent condition across shift factors. A shift factor of 0 denotes responses to unshifted stimuli. Positive shift factors indicate upward scaling of the spectral envelope and fundamental frequency for each sentence; negative shift factors indicate downward scaling. Error bars indicate the standard error of the mean.

place in a sound-attenuating booth, where stimuli were played at a comfortable level (around 65 dB SPL) through Sennheiser HD-598 headphones using Tucker-Davis System 3 and RP2.1 hardware. Stimuli and conditions were randomized and presented using custom scripts created in MATLAB (The MathWorks, Inc.). All experimental procedures were reviewed and approved by the University of Texas at Dallas Institutional Review Board.

5.2 Design and procedure

The intelligibility experiment was a 4 × 4 mixed design: 4 listening conditions (between-subjects factor) by 4 blocks (within-subjects factor). The four listening conditions are described as follows: (a) for the *Foreign-Accented Single Talker Unshifted* condition, listeners heard a single foreign-accented talker across 40 sentences, and this talker was randomly selected from the five foreign-accented talkers presented in the multiple-talker condition; (b) for the *Foreign-Accented Multiple Talkers Unshifted* condition, listeners heard five different foreign-accented talkers across 40 sentences. Presentation of these five talkers was randomly interspersed for each block, and listeners heard each talker twice per block (additional details regarding blocks are described below). Data from both unshifted foreign-accented conditions were reported in our previous work (Kapolowicz et al., 2018), and again, presented here to allow for direct comparisons between shifted and unshifted speech. Since listeners were different for each condition, and all listeners shared the same relevant demographical information, statistical comparisons across conditions are allowed; (c) for the *Foreign-Accented Single Talker Shifted* condition, listeners heard unshifted speech as well as spectrally shifted speech from the same foreign-accented talker (randomly selected for each listener from one of the five talkers in the Foreign-Accented Multiple Talkers Unshifted condition) across 40 sentences to simulate five different talkers. These shift factors ($\pm 2, \pm 1, 0$) were randomly interspersed within each block, and listeners heard each shift factor twice per block; and (d) the *Native-Accented Single Talker Shifted* condition was the same

as the Foreign-Accented Single Talker Shifted condition, only listeners heard native-accented stimuli rather than foreign-accented stimuli.

To gain familiarity with the procedure, listeners in all conditions heard unshifted sentences from the same native-accented talker for a 10-sentence practice block (corresponding to block 1). Listeners' responses for block 1 were not included in statistical analyses. The remaining 40 sentences were divided into four 10-sentence blocks (blocks 2–5). The procedure was blocked using 10-sentence increments to consider if adaptation to stimuli occurred earlier than in the final exposure block. Here, "adaptation" is meant to capture on-line processing that may occur in situations such as over the course of a single conversation. As such, "adaptation" is considered in the current experiment as a statistically significant improvement in intelligibility scores when comparing block 2 with block 5. In each trial, the target sentence was randomly selected (without replacement) from previously recorded/processed sentences. The experiment was completed in a single session lasting approximately 30 minutes. The intelligibility scores were calculated using an automated algorithm written in MATLAB. Intelligibility scores were calculated as the ratio of correctly identified keywords to the total number of presented keywords. The scoring routine also considered homophones (e.g., see vs. sea) as correct responses. Scores were rationalized arcsine unit (RAU) transformed for statistical comparisons to mitigate potential ceiling and floor effects (Studebaker, 1985).

5.3 Results

An ANOVA on the RAU transformed intelligibility scores was performed in R 3.4.0 using the ez package (Lawrence, 2011) and revealed a significant main effect of listening condition, $F(3, 60) = 80.779$, $p < 0.001$, partial $\eta^2 = 0.657$, a non-significant main effect of block, $F(3, 180) = 2.564$, $p = 0.056$, partial $\eta^2 = 0.020$, and a non-significant interaction between block and listening condition, $F(9, 180) = 0.911$, $p = 0.517$, partial $\eta^2 = 0.023$. Planned comparisons were made using the "emmeans" package (Lenth, 2019) in R 3.6.1. As hypothesized, the difference between the scores in blocks 2 and 5 in the Foreign-Accented Single Talker Unshifted condition was statistically significant, $t(180) = -2.123$, $p = 0.0351$, indicating that talker-specific rapid adaptation occurred when F0 and the spectral envelope were not shifted. This difference, however, did not reach significance in the Foreign-Accented Single Talker Shifted condition, $t(180) = -0.948$, $p = 0.3446$, nor for the Foreign-Accented Multiple Talkers Unshifted condition, $t(180) = -0.754$, $p = 0.4521$. For the Foreign-Accented Single Talker Shifted condition, listeners were presented with only one talker for the duration of the experiment, revealing that shifting F0 and the spectral envelope was enough to prevent the rapid adaptation observed in the unshifted condition (see Figure 6 for a graphical representation of results).

The results from the intelligibility experiment reveal that spectrally shifting speech from a single foreign-accented talker leads to perceptual patterns that were similar to those observed when listeners actually heard multiple foreign-accented talkers. Intelligibility scores in the Foreign-Accented Single Talker Shifted condition and in the Foreign-Accented Multiple Talkers Unshifted condition were much lower than intelligibility scores for perception in the Foreign-Accented Single Talker Unshifted condition. Unlike in the Foreign-Accented Single Talker Unshifted condition, increased exposure did not result in improvements in intelligibility scores for the Foreign-Accented Multiple Talkers Unshifted condition nor for the Foreign-Accented Single Talker Shifted condition. Also, unlike for foreign-accented speech, near ceiling performance was observed for spectrally shifted native-accented speech. This finding indicates that spectral shifting, in itself, is not a likely explanation to explain the drop in performance that was observed for the shifted foreign-accented speech condition.

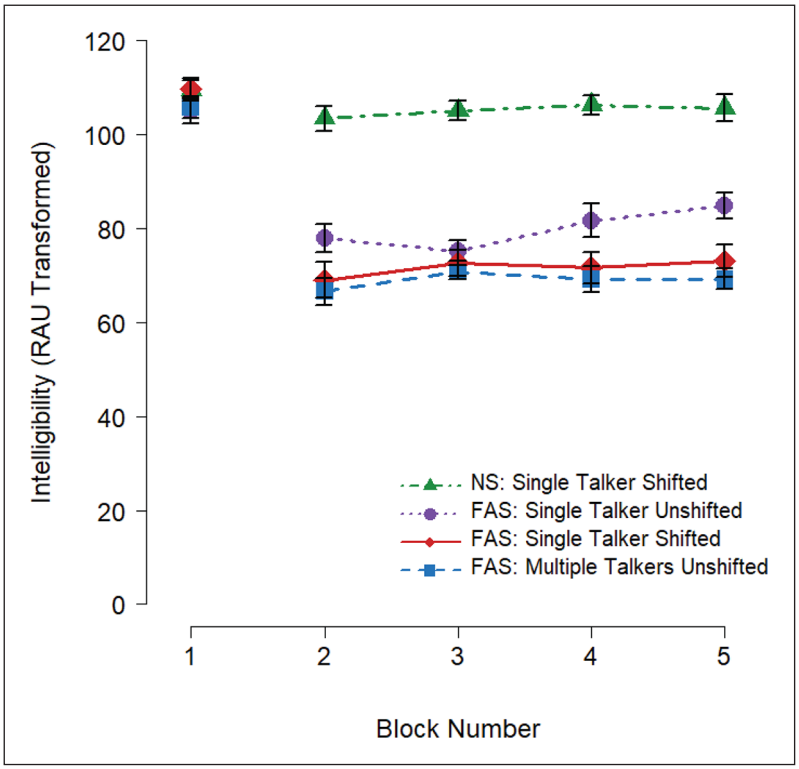

Figure 6. Mean intelligibility performance with increased exposure. Each block indicates exposure to 10 sentences. Block 1 corresponds to the practice session where unshifted speech from a single native-accented female talker was heard. Scores for this block were not analyzed. Unshifted conditions denote unprocessed speech. Shifted conditions comprise speech from a single talker but simulated to sound like speech from five different talkers. Error bars indicate the standard error of the mean.

6 Discussion

To investigate the hypothesis that stable spectral envelope and F0 cues stemming from a single talker aid listeners with rapidly adapting to foreign-accented speech in a talker-dependent manner, we covaried the spectral envelope and F0 of individual talkers up or down in increments of 8% and 30%, respectively, and assessed how this manipulation would affect listeners' intelligibility scores initially and after increased exposure. Two control experiments were first conducted to ensure that, when different spectral shifts were introduced to each of a sequence of sentences from a single talker, these sentences were perceived as coming from different talkers (Experiment 1) and perceived as being within the natural human range of speech (Experiment 2). We then compared conditions where listeners heard spectrally shifted speech from a single talker with conditions where listeners heard unprocessed sentences from a single talker and when listeners heard unprocessed sentences from multiple talkers. When these spectral shifts were implemented, the adaptation effect that was observed for unprocessed sentences spoken by a single foreign-accented talker was eliminated, and the performance drop was similar to when listeners heard unprocessed speech from multiple talkers (Experiment 3). This outcome supports the hypothesis that listeners utilize stable spectral envelope and F0 cues during early online processing of adapting to foreign-accented

speech in a talker-dependent manner (i.e., talker-dependent adaptation) to a greater extent than relying on other stable cues, such as similarities in pronunciation across different talkers who share the same native language (i.e., accent-dependent adaptation).

It has been consistently shown that listeners can adapt to foreign-accented speech (e.g., Baese-Berk et al., 2013; Bradlow & Bent, 2008; Sidaras et al., 2009), even with as little as one minute of exposure (Clarke & Garrett, 2004). Although research suggests that listeners readily adapt to foreign-accented speech after minimal exposure, listeners still often report difficulty understanding foreign accents, and the time-course and specificity of adaptation remain unclear (Trude et al., 2013), especially regarding speech produced by either a single talker or multiple talkers. These questions are of particular importance in speech communication given that foreign-accented talkers outnumber native talkers of English, and communication between these two groups is increasing (Graddol, 2006; Jenkins, 2000). Several other factors influencing adaptation to foreign-accented speech have been investigated, including the effect of baseline intelligibility scores (e.g., Bradlow & Bent, 2008), the effect of accent (e.g., Bent & Holt, 2013), the role of feedback (e.g., Cooper & Bradlow, 2016), and the role of semantic predictability (e.g., Baese-Berk et al., 2021). In the present work, we addressed the potential role of two talker-specific spectral cues in regard to the relationship between the perception of foreign-accented speech and talker variability. We compared conditions where listeners heard the same talker repeatedly as opposed to a succession of different talkers. Clarifying perceptual differences between single- and multiple-talker conditions is an important step toward elucidating the underlying mechanisms involved in the perceptual process of adapting to foreign-accented speech.

Talker identification is a relatively easy, even subconscious, perceptual process for native speech, but perhaps this process becomes more critical when perceiving foreign-accented speech. The perceived divergences from native speech norms across segmental and suprasegmental dimensions can make non-native speech more difficult to understand (Lane, 1963). In this case, we speculate that the talker-specific cues provided by the spectral envelope and F0 could aid with perceptual recalibration to foreign-accented sounds that depart from listeners' preconceived category boundaries for the target signal. This notion is consistent with the idea that talker-specific spectral information is important (perhaps necessary) when perceiving foreign-accented speech by helping listeners to infer indexical cues to the appropriate talker/gender/age-specific representations of the speech signal.

This perceptual process involving accurate talker representation could work in concert with other aspects of the second language (L2) speech signal which are more similar to the listener's own production patterns, such as when there is overlap in production for a foreign-accented talker's first language and L2 categories that assimilates to the correct target in the L2 (Best, 1995; Flege, 1995; Thomson et al., 2009). Also, when L2 production deviates enough from other L2 categories, misclassification resulting from two distinct tokens assimilating into the same category can be avoided (Best, 1995; Flege, 1995; Thomson et al., 2009). Other production patterns which vary from native production norms have also been shown to aid listeners, such as foreign-accented talkers having slower speaking rates (Anderson Hsieh & Koehler, 1988). Despite production of foreign-accented speech being slower than for native-accented speech, there is also greater variability in speaking rate across utterances for foreign-accented speech (Baese-Berk & Morrill, 2015), perhaps rendering this cue as less reliable for perceiving and adapting to foreign-accented speech.

For the present study, we focused on the role of the spectral envelope and F0 to assess listeners' reliance on talker-specific cues for perceptual adaptation to foreign-accented speech. The spectral scale factors and signal processing implemented in this study preserved the temporal structure of the speech stimuli, including the speaking rate, relative segment durations, and overall amplitude contour. This allowed for an indirect assessment of the contribution of such cues to perception of

foreign-accented speech. The findings suggest that these other cues do not provide a basis for adaptation to foreign-accented speech, while stable spectral envelope and F0 cues appear to make a substantial contribution.

6.1 Talker discrimination

Speech sounds include acoustic information related to unique talker characteristics in addition to linguistic content (van Dommelen, 1990). When a change in talker is detected, the normalization process is disrupted (Barreda, 2012). This can be more detrimental when perceiving speech in difficult conditions, such as when perceiving foreign-accented speech. Acoustic information implicated in the normalization process includes the spectral envelope and the average F0, properties that differ across talkers due to physical constraints such as age and sex (Titze, 1989) and size of the talker (Smith et al., 2005). Spectral envelope and F0 cues play an important role in the perception of talker age and sex in children and adults (Bachorowski & Owren, 1999; Barreda & Assmann, 2018; Hillenbrand & Clark, 2009). It has also been shown that vocal-tract length (related to the spectral envelope) is more important than source characteristics (related to F0) for talker identity. For example, Kuwabara and Takagi (1991) manipulated formant frequency and F0 characteristics of speech and found that listener identification of individual talkers was relatively undisturbed by changes in F0 but was impaired when formant characteristics were altered by as little as 5%. Gaudrain et al. (2009) found that listeners perceived talkers as different when F0 varied by 45% and when the spectral envelope varied by 25%. Overall, these findings show a larger contribution of vocal-tract length characteristics compared to source characteristics when identifying different talkers. However, when performing signal processing manipulations on speech from a single talker with the goal of simulating a change in talker, as was observed in Experiment 1, then covarying these characteristics should produce a more natural voice quality (Assmann et al., 2006), indicating that both source and filter components of the vocal apparatus contribute to talker perception (Baumann & Belin, 2010). Such an assessment was conducted in Experiment 2.

6.2 Naturalness ratings

Experiment 2 assessed the perceived naturalness of spectrally shifted native-accented and foreign-accented speech. Our results were consistent with previous studies reporting that downward scaling of speech from male talkers results in judgments perceived as being less natural than downward scaling of speech from female talkers, and spectrally scaling speech from male and female talkers downward beyond the normal human range results in speech being perceived as extremely unnatural (Assmann & Nearey, 2008). An unanticipated finding was that foreign-accented speech was judged as less natural than native-accented speech even when the spectral information was unshifted. In this experiment, listeners were explicitly instructed *not* to rate the level of naturalness for each talker based on perceived intelligibility or on perceived level of foreign-accentedness. One possible explanation for these results is that listeners were unable to entirely separate naturalness from accentedness in their judgments. A related finding was reported by Mackey et al. (1997). In their study, listeners used a 9-point Likert scale to rate perceived naturalness of speech from talkers with foreign accents and regional accents that varied from their own as well as from talkers who shared their same regional dialect. They found that listeners rated speech from talkers who did not share their same regional dialect as sounding less natural (Mackey et al., 1997). Nevertheless, for the shifts implemented in the Intelligibility experiment, stimuli from both accents were rated as being relatively natural.

6.3 Intelligibility

In Experiment 3, brief exposure to spectrally shifted foreign-accented speech from a single talker produced similar patterns to unshifted speech from five different foreign-accented talkers who share the same native language. We also found that intelligibility scores only improved with increased exposure when listeners perceived unshifted speech from a single foreign-accented talker. These results indicate that intelligibility of foreign-accented speech is impaired and adaptation is disrupted when listeners perceive that they are hearing speech from different talkers. Specifically, the finding that rapid adaptation to a single foreign-accented talker was disrupted when spectral envelope and F0 cues were shifted to simulate different talkers, combined with the finding of a lack of adaptation occurring in the multiple-talker condition, support a theory of talker normalization for rapid perceptual adaptation to foreign-accented speech.

These results underscore the importance of talker-specific spectral envelope and F0 cues with regards to rapidly adapting to foreign-accented speech. The spectral manipulations utilized in the present work gave listeners the impression that they were perceiving different talkers, as tested in Experiment 1. Independent support for this finding also comes from subjective reports of the listeners in Experiment 3, who were different from the listeners in Experiment 2. After completing the experiment, every listener in Experiment 3 reported hearing different talkers in the shifted and multiple-talker conditions, but consistently heard a single talker in the single-talker unshifted condition. A plausible interpretation is that the impression of different talkers disrupted the adaptation process, resulting in performance that was nearly identical to that observed when listeners were presented with actual different talkers. This finding supports a theory of extrinsic normalization, whereby listeners utilize information (i.e., spectral envelope and F0 cues) from preceding sounds and apply that information to successive sounds from that same talker to adapt to foreign-accented speech.

The role of talker variability may be explained by considering the results reported from Barreda (2012), who found improvements in conditions where listeners perceived that they were hearing vowels produced by a single talker, provided that formant spacing cues remained similar across exposure. In conditions where listeners perceived that they were hearing multiple talkers, performance remained stable over time (no improvement or decrement occurred), similar to our present findings (Barreda, 2012). Interestingly, when listeners perceived that they were hearing different talkers due to a shifting of source cues, the decrement in performance that was observed in perceived single talker conditions when formant spacing cues were incongruent between voices was reduced (Barreda, 2012). Barreda's (2012) key conclusion was that the process of detecting a change in talker is what determines how listeners utilize extrinsic information, such as the spectral envelope and F0, to guide perception. Barreda (2012) described these findings to suggest that when listeners perceive utterances as coming from the same talker, they are able to utilize extrinsic information, wherein an estimate of the formant space is refined from the estimate used in the preceding utterance and applied to current and preceding utterances. But when listeners perceive a change in talker across utterances, they utilize an intrinsic process, relying solely on the information provided in the current utterance (Barreda, 2012).

The work presented by Barreda (2012) showcases the specific relevance of a perceived change in talker. However, even small shifts in the spectral envelope and F0 can produce an outcome similar to the one reported here. Magnuson and Nusbaum (2007) exposed listeners to synthetic voices with minimally different F0s, and informed them that they were either hearing the same voice or different voices. The listeners who were told that they were hearing different voices responded with longer reaction times than those who were told that they were hearing the same voice, despite both groups being exposed to the same stimuli (Magnuson & Nusbaum, 2007). This finding supports the notion that the extrinsic normalization/talker adaptation process can be disrupted by instructions to the

listeners that they are hearing different talkers. The lack of adaptation to spectrally shifted foreign-accented speech observed in our present findings may be a consequence of listeners perceiving that they are hearing different talkers, rather than (or in addition to) a failure to recalibrate to changes in F0 and spectral envelope cues. Future studies could ascertain the smallest spectral shifts needed to replicate our results and investigate whether listeners associate these shifts with a change in talker identity. This interpretation leads to a further prediction, namely that providing instructions to listeners in the shifted conditions that they were actually hearing the same talker could change the pattern of results to be more similar to those in the single-talker unshifted condition.

In the present study, lower intelligibility scores for spectrally shifted speech were only observed for foreign-accented speech. Intelligibility scores were near ceiling performance for spectrally shifted native-accented speech. Similar results for spectrally shifted native-accented speech were reported by Holmes et al. (2018), who found that, when listeners were exposed to spectrally-shifted sentences from familiar talkers, intelligibility scores were higher than when exposed to unfamiliar voices, even though listeners thought that they were hearing different talkers in both cases. Their results, combined with the present results, suggest that, for native-accented speech, the impression of hearing different talkers is not as detrimental as it is for foreign-accented speech, which is inconsistent with reports from Magnuson and Nusbaum (2007) and Barreda (2012). Instead, listeners are still able to utilize other talker-specific cues to obtain an intelligibility benefit from exposure to familiar native-accented talkers (Holmes et al., 2018). The present results for native-accented speech are in line with their findings; however, the results for foreign-accented speech imply that listeners are less able to benefit from other talker-specific cues to aid with rapid adaptation.

The spectrally shifted native-accented speech condition was included in the present study as a control to confirm that the process of spectrally shifting speech, in itself, produces little-or-no decline in intelligibility. Although it is beyond the scope of the present work, it should be considered that, because performance for native-accented speech is near ceiling performance, even upon initial exposure, no effect of adaptation can be detected. However, this result does not necessarily entail that these cues are less important for perception of native-accented speech. In the present work, stimuli were presented in quiet. If the listeners were, instead, exposed to native-accented speech in a noisy condition, then we would expect that they would adapt more quickly in conditions where they were exposed to only a single talker as opposed to different talkers. Previous work has shown that increased exposure/familiarity to a single talker aids listeners with perception of speech in noisy conditions (e.g., Johnsrude et al., 2013). The present results suggest that a similar talker-dependent listening strategy seems to be employed when perceiving foreign-accented speech.

Although listeners in the present study did not show adaptation to foreign-accented speech spoken by multiple talkers, previous studies have shown that such adaptation can occur when a longer exposure period is provided (e.g., Bradlow & Bent, 2008). Moreover, previous results have also reported that listeners can adapt not only to novel talkers but also to novel accents (Baese-Berk et al., 2013). In both of these cases, however, listeners were exposed to training over a two-day period as opposed to having only a single training session, as was the case in the present study. Additional exposure may, therefore, be necessary to provide listeners with other potentially reliable cues, such as systematic variation that can be observed in speech patterns across different foreign-accented talkers (Baese-Berk et al., 2013). In line with our present findings, however, Xie et al. (2018) found that listeners who were trained on Mandarin-accented speech and tested the same day only retained their improvement in performance for the same talker; they found a reduction in performance for a novel Mandarin-accented talker if tested on the same day as training had occurred. Interestingly, and in support of the results from Bradlow and Bent (2008) and Baese-Berk et al. (2013), Xie et al. (2018) also found that listeners showed improvements for the novel talker condition when tested on the second day, suggesting that sleep may have facilitated generalization to a novel talker. The focus of the present work was

limited to on-line processing mechanisms when listeners are perceiving foreign-accented speech, but the results of Xie et al. (2018) reveal interesting implications for the role of memory consolidation to aid with this perceptual adaptation process.

6.4 Implications for CI users

These results may have direct implications for CI users. CIs deliver degraded speech signals which limit access to talker-specific voice cues such as the spectral envelope and F0 (Fuller et al., 2014; Başkent et al., 2016; Gaudrain & Başkent, 2018). Previous work has shown that CI users have difficulty with talker variability (Chang & Fu, 2006; Kaiser et al., 2003) and with perception of foreign-accented speech (Ji et al., 2014). CI users also struggle to adapt to foreign-accented speech, even when there is only a single talker (Kapolowicz et al., 2020). The results from the present work, which underscore the importance of talker-specific voice cues when adapting to foreign-accented speech, may help to explain the added difficulties experienced by CI users when they are perceiving foreign-accented speech. It is unclear whether CI users can adapt to foreign-accented speech over a longer exposure period by utilizing different listening strategies that do not depend on talker-specific voice cues, such as by learning to rely on accent-dependent cues (i.e., production patterns that are specific to a particular accent) that are more easily transmitted through their devices. In support of this possibility, Kapolowicz et al. (2020) found that a few CI users' intelligibility scores improved with increased exposure when perceiving Mandarin-accented English from several different talkers whereas this was not the case with increased exposure to only a single talker. Future studies could attempt to uncover why certain CI users are able to use a more effective listening strategy while others are not, or if those who initially struggle can eventually adopt the more effective listening strategy if given a longer exposure period.

Acknowledgements

We thank those who participated in these experiments as well as Christina Mai and Danni Yang for assistance with data collection. We also thank Dr. Rachel Theodore and our anonymous reviewers for providing insightful comments to improve the manuscript.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD


Michelle R. Kapolowicz https://orcid.org/0000-0003-4404-5270

Note

1. Note that talker is not independent from accent (i.e., it is nested in accent). Nonetheless, we considered accent as a separate factor to investigate the effects of talkers' accentedness on the naturalness ratings.

References

Abdi, H., Edelman, B., Valentin, D., & Dowling, W. J. (2009). *Experimental design and analysis for psychology*. Oxford University Press.

Anderson Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, *38*(4), 561–613. https://doi.org/10.1111/j.1467-1770.1988.tb00167.x

Assmann, P. F., & Nearey, T. M. (2007). Effects of frequency shifts on the identification of vowels and words in sentences. *Journal of the Acoustical Society of America*, *122*(5), 3064–3065. https://doi.org/10.1121/1.2942936

Assmann, P. F., & Nearey, T. M. (2008). Identification of frequency-shifted vowels. *Journal of the Acoustical Society of America*, *124*(5), 3203–3212. https://doi.org/10.1121/1.2980456

Assmann, P. F., Dembling, S., & Nearey, T. M. (2006). Effects of frequency shifts on perceived naturalness and gender information in speech. *Proceedings of the Ninth International Conference on Spoken Language Processing*, *2006*, 889–892. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.119.3904&rep=rep1&type=pdf

Assmann, P. F., Nearey, T. M., & Bharadwaj, S. (2008). Analysis and classification of a vowel database. *Canadian Acoustics*, *36*(3), 148–149.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *Journal of the Acoustical Society of America*, *106*(2), 1054–1063. https://doi.org/10.1121/1.427115

Baese-Berk, M. M., & Morrill, T.H. (2015). Speaking rate consistency in native and non-native speakers of English. *Journal of the Acoustical Society of America: Express Letters, 138*(3), EL223–228. https://doi.org/10.1121/1.4929622

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America: Express Letters*, *133*(3), EL174–180. https://doi.org/10.1121/1.4789864

Baese-Berk, M. M., Bent, T., & Walker, K. (2021). Semantic predictability and adaptation to nonnative speech. *Journal of the Acoustical Society of America: Express Letters*, *1*, 015207. Advance online publication 21 January 2021. https://doi.org/10.1121/10.0003326

Barreda, S. (2012). Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis. *Journal of the Acoustical Society of America*, *132*(5), 3453–3464. https://doi.org/10.1121/1.4747011

Barreda, S. (2020). Vowel normalization as perceptual constancy. *Language*, *96*(2), 224–254. https://doi.org/10.1353/lan.2020.0018

Barreda, S., & Assmann, P.F. (2018). Modeling the perception of children's age from speech acoustics. *Journal of the Acoustical Society of America: Express Letters*, *143*(5), EL361. https://doi.org/10.1121/1.5037614

Başkent, D., Gaudrain, E., Tamati, T.N., & Wagner, A. (2016). Perception and psychoacoustics of speech in cochlear implant users. In A. T. Cacace, E. de Kleine, A. Genene-Holt, & P. van Dijk (Eds.), *Scientific foundations of audiology: Perspectives from physics, biology, modeling, and medicine* (pp. 285–319). Plural Publishing, Inc.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. https://doi.org/10.1007/s00426-008-0185-z

Bent, T., & Holt, R. F. (2013). The influence of talker and foreign-accent variability on spoken word identification. *Journal of the Acoustical Society of America*, *133*(3), 1677–1686. https://doi.org/10.1121/1.4776212

Best, C. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed.), *Speech perception and linguistic experience: Theoretical and methodological issues* (pp. 171–204). York Press.

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. https://doi.org/10.1016/j.cognition.2007.04.005

Chang, Y., & Fu, Q.-J. (2006). Effects of talker variability on vowel recognition in cochlear implants. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1331–1341. https://doi.org/10.1044/1092-4388(2006/095)

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented speech. *Journal of the Acoustical Society of America*, *116*(6), 3647–3658. https://doi.org/10.1121/1.1815131

Cooper, A., & Bradlow, A. R. (2016). Linguistically guided adaptation to foreign-accented speech. *Journal of the Acoustical Society of America: Express Letters*, *140*(5), EL378–EL384. https://doi.org/10.1121/1.4966585

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues* (pp. 229–273). York Press.

Fox, J., Friendly, M., & Monette, G. (2018). heplots: Visualizing Tests in Multivariate Linear Models. R package version 1.3-5. URL https://CRAN.R-project.org/package=heplots

Fu, Q. J., & Shannon, R. V. (1999). Effects of electrode configuration and frequency allocation on vowel recognition with the nucleus-22 cochlear implant. *Ear and Hearing*, *20*(4), 332–344. https://doi.org/10.1097/00003446-199908000-00006

Fuller, C. D., Gaudrain, E., Clarke, J. N., Galvin, J. J., Fu, Q.-J., Free, R. H., & Başkent, D. (2014). Gender categorization is abnormal in cochlear implant users. *Journal of the Association for Research in Otolaryngology*, *15*(6), 1037–1048. https://doi.org/10.1007/s10162-014-0483-7

Gaudrain, E., & Başkent, D. (2018). Discrimination of voice pitch and vocal-tract length in cochlear implant users. *Ear and Hearing*, *39*(2), 226–237. https://doi.org/10.1097/AUD.0000000000000480

Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. (2009). The role of glottal pulse rate and vocal tract length in the perception of speaker identity. INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6–10, 2009. https://hal.archives-ouvertes.fr/hal-02144510/document

Graddol, D. (2006). *English next*. British Council.

Hillenbrand, J. M., & Clark, M. J. (2009). The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, *71*(5), 1150–1166. https://doi.org/10.3758/APP.71.5.1150

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognized as familiar. *Psychological Sciences*, *29*(10), 1575–1583. https://doi.org/10.1177/0956797618779083

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

IEEE Subcommittee (1969). IEEE recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics*, *17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Jenkins, J. (2000). *The phonology of English as an international language*. Oxford University Press.

Ji, C., Galvin, J. J., Chang, Y., Xu, A., & Fu, Q. J. (2014). Perception of speech produced by native and nonnative talkers by listeners with normal hearing and listeners with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *57*(2), 532–542. https://doi.org/10.1044/2014_JSLHR-H-12-0404

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Joos, M. A. (1948). Acoustic phonetics. *Language, 24*, Suppl. *2*, 1–136.

Kaiser, A. R., Kirk, K. I., & Lachs Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46*(2), 390–404. https://doi.org/10.1044/1092-4388(2003/032)

Kapolowicz, M. R., Montazeri, V., & Assmann, P. F. (2016). The role of spectral resolution in foreign-accented speech perception. *Proceedings of Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, 3289–3293. https://www.isca-speech.org/archive/Interspeech_2016/pdfs/1585.PDF

Kapolowicz, M. R., Montazeri, V., & Assmann, P. F. (2018). Perceiving foreign-accented speech with decreased spectral resolution in single- and multiple-talker conditions. *Journal of the Acoustical Society of America: Express Letters*, *143*(2), EL99. https://doi.org/10.1121/1.5023594

Kapolowicz, M. R., Montazeri, V., Baese-Berk, M. M., Zeng, F.-G., & Assmann, P. F. (2020). Rapid adaptation to non-native speech is impaired in cochlear implant users. *Journal of the Acoustical Society of America: Express Letters*, *148*(3), EL267–EL272. https://doi.org/10.1121/10.0001941

Kawahara, H., Masuda-Katsuse, I., & de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, *27*, 187–207. https://doi.org/10.1016/S0167-6393(98)00085-5

Kawahara, H., de Cheveigné, A., Banno, H., Takahashi, T., & Irino, T. (2005). Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT. *INTERSPEECH-2005*, 537–540. https://www.isca-speech.org/archive/archive_papers/interspeech_2005/i05_0537.pdf

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Wiley-Blackwell.

Kuwabara, H., & Takagi, T. (1991). Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method. *Speech Communication*, *10*(5–6), 491–495. https://doi.org/10.1016/0167-6393(91)90052-U

Kuwabara, H., & Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: Control and conversion. *Speech Communication*, *16*(2), 165–173. https://doi.org/10.1016/0167-6393(94)00053-D

Lane, H. (1963). Foreign accent and speech distortion. *Journal of the Acoustical Society of America*, *35*(4), 451–453. https://doi.org/10.1121/1.1918501

Lawrence, M. A. (2011). ez: Easy analysis and visualization of factorial experiments. R package version 3.0-0. URL http://CRAN.R-project.org/package=ez

Lenth, R. (2019). emmeans: Estimated Marginal Means, aka Least-Squares

Means. R package version 1.4.1. https://CRAN.R-project.org/package=emmeans

Mackey, L. S., Finn, P., & Ingham, R. J. (1997). Effect of speech dialect on speech naturalness ratings: A systematic replication of Martin, Haroldson, and Triden (1984). *Journal of Speech, Language, and Hearing Research*, *40*(2), 349–360. https://doi.org/10.1044/jslhr.4002.349

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088–2113. https://doi.org/10.1121/1.397861

Nearey, T. M., & Assmann, P. F. (2007). Probabilistic 'sliding template' models for indirect vowel normalization. In M. J. Solé, P. S. Beddor, & M. Ohala (eds.), *Experimental approaches to phonology* (pp. 246–269). Oxford University Press.

Nygaard, L. C., & Pisoni, D. P. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376. https://doi.org/10.3758/bf03206860

Sidaras, S. K., Alexander, J. E. D., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *Journal of the Acoustical Society of America*, *125*(5), 3306–3316. https://doi.org/10.1121/1.3101452

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia*, *49*(14), 3831–3846. https://doi.org/10.1016/j.neuropsychologia.2011.09.044

Smith, D. R. R., Patterson, R. D., Turner, R, Kawahara, H., & Irino, T. (2005). The processing and perception of size information in speech sounds. *Journal of the Acoustical Society of America*, *117*(1), 305–318. https://doi.org/10.1121/1.1828637

Studebaker, G. (1985). A "rationalized" arcsine transform. *Journal of Speech and Hearing Research*, *28*(3), 455–462. https://doi.org/10.1044/jshr.2803.455

Tamati, T. N., Sijp, L., & Başkent, D. (2020). Talker variability in word recognition under cochlear implant simulation: Does talker gender matter? *Journal of the Acoustical Society of America: Express Letters*, *147*(4), EL370–EL376. https://doi.org/10.1121/10.0001097

Tamati, T. N., Pisoni, D. B., & Moberly, A. C. (2021). The perception of regional dialects and foreign accents by cochlear implant users. *Journal of Speech, Language, and Hearing Research*, *64*, 683–690. https://doi.org/10.1044/2020_JSLHR-20-00496

Thomson, R. I., Nearey, T. M., & Derwing, T. M. (2009). A modified statistical pattern recognition approach to measuring the crosslinguistic similarity of Mandarin and English vowels. *Journal of the Acoustical Society of America*, *126*(3), 1447–1460. https://doi.org/10.1121/1.3177260

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, *85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Trude, A. M., Tremblay, A., & Brown-Schmidt, S. (2013). Limitations of adaptation to foreign accents. *Journal of Memory and Language*, *69*(3), 349–367. https://doi.org/10.1016/j.jml.2013.05.002

van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language and Speech*, *33*(3), 259–272. https://doi.org/10.1177/002383099003300302

Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter- and intratalker variation on Cantonese level tones. *Journal of Speech, Language, and Hearing Research*, *46*(2), 413–421. https://doi.org/10.1044/1092-4388(2003/034)

Xie, X., Sayako Earle, F., & Myers, E. B. (2018). Sleep facilitates generalization of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210. https://doi.org/10.1080/23273798.2017.1369551