Talker information influences spectral contrast effects in speech categorization[a)]

Ashley A. Assgari[b)] and Christian E. Stilp
Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292, USA

(Received 6 May 2015; revised 9 October 2015; accepted 13 October 2015; published online 13 November 2015)

Spectral contrast effects, the perceptual magnification of spectral differences between sounds, have been widely shown to influence speech categorization. However, whether talker information alters spectral contrast effects was recently debated [Laing, Liu, Lotto, and Holt, Front. Psychol. **3**, 1–9 (2012)]. Here, contributions of reliable spectral properties, between-talker and within-talker variability to spectral contrast effects in vowel categorization were investigated. Listeners heard sentences in three conditions (One Talker/One Sentence, One Talker/200 Sentences, 200 Talkers/200 Sentences) followed by a target vowel (varying from /ɪ/-/ɛ/ in F_1, spoken by a single talker). Low-F_1 or high-F_1 frequency regions in the sentences were amplified to encourage /ɛ/ or /ɪ/ responses, respectively. When sentences contained large reliable spectral peaks (+20 dB; experiment 1), all contrast effect magnitudes were comparable. Talker information did not alter contrast effects following large spectral peaks, which were likely attributed to an external source (e.g., communication channel) rather than talkers. When sentences contained modest reliable spectral peaks (+5 dB; experiment 2), contrast effects were smaller following 200 Talkers/200 Sentences compared to single-talker conditions. Constant recalibration to new talkers reduced listeners' sensitivity to modest spectral peaks, diminishing contrast effects. Results bridge conflicting reports of whether talker information influences spectral contrast effects in speech categorization. © *2015 Acoustical Society of America*.
[http://dx.doi.org/10.1121/1.4934559]

[ICB] Pages: 3023–3032

I. INTRODUCTION

The auditory system is remarkably sensitive to changes in the acoustic input. When spectra of a preceding acoustic context and a subsequent target sound differ, the auditory system perceptually magnifies this difference. This is known as a spectral contrast effect, where perception of the target sound is biased away from the spectrum of the preceding context. Spectral contrast effects have been widely shown to influence speech categorization. One of the earliest demonstrations of this is the seminal findings of Ladefoged and Broadbent (1957). When the range of the first formant (F_1) in the preceding sentence was shifted down to lower frequencies (more closely resembling the low F_1 in /ɪ/), listeners labeled the subsequent target vowel as the higher-F_1 /ɛ/. When the range of F_1 in the preceding sentence was shifted up to higher frequencies (more closely resembling the high F_1 in /ɛ/), listeners labeled the target vowel as /ɪ/.

While such effects have long been known to occur, the source of spectral contrast effects has been debated. Ladefoged and Broadbent (1957) interpreted their results as the auditory system normalizing to acoustic properties of a talker's voice. They proposed that listeners learned the formant structure of the talker's sentence and used that information to interpret the target vowel, consistent with the suggestions of Joos (1948). Others have suggested that these effects are rooted in speech production and compensate for coarticulation (Fowler, 2006). Several studies during the last few decades focused less on talkers (or speech) as the basis of these effects and more on simple acoustics (e.g., Watkins, 1991; Watkins and Makin, 1994; Lotto and Kluender, 1998; Holt, 2005; Stilp *et al.*, 2010; Sjerps *et al.*, 2011; Huang and Holt, 2012; Laing *et al.*, 2012; Stilp *et al.*, 2015).

This focus on acoustics has been advanced most strongly by Holt and colleagues. Instead of having a speech context precede the target speech sound, they presented a series of sine tones ("tone histories") that sampled the frequency region predicted to produce spectral contrast effects. Tone histories successfully produced spectral contrast effects when they preceded /ga/-/da/ and /ɛ/-/ʌ/ continua (Holt, 2005, 2006; Huang and Holt, 2012). In a similar experiment, Laing and colleagues (2012) manipulated F_1 or F_3 regions in a speech context to induce perception of different talkers. Following these contexts, listeners identified targets from a /ga/-/da/ continuum varying in their F_2 and F_3 transitions. When manipulations of talker identity occurred in the F_1 region (which is spectrally remote from the F_3 region that is key for the /g/-/d/ distinction), no spectral contrast effect was observed. However, both tone histories and speech with manipulations in the F_3 region were successful in producing spectral contrast effects. They concluded that talker information is neither necessary nor sufficient to produce spectral contrast effects, which are instead dictated by the acoustics of the preceding context.

[a)]Results presented at the 169th Meeting of the Acoustical Society of America, Pittsburgh, PA.
[b)]Electronic mail: ashley.assgari@louisville.edu

However, investigations of spectral contrast effects have been ill-designed to measure contributions of talker information. Studies utilizing tone histories as acoustic contexts mimicked some key acoustic properties of speech (i.e., long-term average spectrum in a particular frequency region), but this does not approximate the acoustic complexity and extreme variability of natural speech. The amplitude envelope of a tone history is far more consistent than that of speech. In addition, tone histories sample the frequency region of interest more often and more regularly than speech. Finally, sine tones are rarely if ever encountered in natural acoustic environments, bringing ecological validity into question. When studies of spectral contrast effects did use speech as the acoustic context, it was a single talker producing one sentence that listeners heard dozens if not hundreds of times during the experiment, limiting within- and between-talker acoustic variability. Watkins (1991) examined spectral contrast effects using a variety of acoustic contexts across experiments, yet within an experiment, only a single context was used. When Laing *et al.* (2012) tested whether talker information influenced spectral contrast effects, they manipulated a single formant frequency (in a single sentence) to produce contexts that sounded like they were spoken by different talkers. This approach restricted acoustic variability to a narrow frequency region, but acoustic differences between talkers span broad frequency regions if not the entire spectrum (Peterson and Barney, 1952).

A long line of literature demonstrates high perceptual sensitivity to talker information, often termed talker normalization. Talker normalization is the perceptual adjustment or adaptation to talker-specific properties of the speech signal. Several studies demonstrated that speech perception was faster and/or more accurate when hearing a single talker than when hearing multiple talkers (Creelman, 1957; Verbrugge *et al.*, 1976; Assmann *et al.*, 1982; Martin *et al.*, 1989; Mullennix *et al.*, 1989). This pattern of results is consistent across phoneme recognition and word recognition tasks and holds for as few as 2 and as many as 30 talkers. Yet, the influence of talker information on spectral contrast effects has never been adequately tested.

There is reason to believe that talker information does influence spectral contrast effects. Each has been described as a means of normalizing or accounting for extreme acoustic variability in the speech signal. Talker normalization and spectral contrast effects both entail perceptual adjustment to stable acoustic properties of a listening context. Mullenix and colleagues (1989) suggested "talker normalization processes may be related to other low-level sensory encoding processes which are also sensitive to the changes and variability in acoustic information in the speech signal" (pp. 375–376). Spectral contrast effects stem from low-level sensory processes that emphasize changes in the acoustic signal. Thus, spectral contrast and talker normalization may be complementary interpretations of how listeners adjust to acoustic characteristics that convey talker identity.

The current study investigates the influence of talker information on spectral contrast effects in vowel categorization. Within- and between-talker acoustic variability are manipulated by varying the number of talkers and number of unique sentences spoken in three conditions. In one condition, a single sentence spoken by one talker is repeated 200 times in an experimental block (One Talker/One Sentence). A second condition presents a unique talker and a unique sentence on every trial (200 Talkers/200 Sentences) to maximize between-talker acoustic variability. The third condition presents 200 unique sentences spoken by a single talker (One Talker/200 Sentences) to test the role of within-talker acoustic variability while controlling for between-talker variability. Spectral contrast effects in vowel categorization are tested following sentences that have had reliable spectral peaks added by bandpass filters at one of two filter gains: one predicted to produce large contrast effects (+20 dB spectral peak added to preceding sentence, experiment 1) and one predicted to produce smaller but still significant contrast effects (+5 dB spectral peak, following Stilp *et al.*, 2015; experiment 2).

These manipulations introduce competing predictions for the present experiments. Arguments that talker information does not influence spectral contrast effects are based on experiments that added a large spectral peak to the preceding acoustic context (Laing *et al.*, 2012). While large spectral peaks are common in short-term speech spectra (e.g., vowel formant peaks), it is far less common to observe such dramatic peaks in the long-term average spectrum of speech.[1] Listeners might be reluctant to treat large reliable spectral peaks as being produced by a talker and instead attribute them to an external source like the communication channel (much as listeners do for complex systematic spectral distortions; Watkins, 1991). Attributing reliable spectral peaks to the communication channel dissociates them from talker information, diminishing the role of talker information in the task. We predict that when sentences feature a large reliable spectral peak (+20 dB) in experiment 1, spectral contrast magnitudes will be comparable irrespective of the number of talkers.

Modest spectral peaks (+5 dB), on the other hand, are much more common in long-term average speech spectra. Listeners are expected to attribute modest reliable spectral peaks to talkers rather than the communication channel. Thus, acoustic information and talker information would be emanating from the same source, providing a strong test of the role of talker information in spectral contrast effects. If talker information does not influence spectral contrast effects, then as in experiment 1, spectral contrast effect magnitudes should be comparable irrespective of the number of talkers. If talker information does influence spectral contrast effects, then categorization performance should be sensitive to the number of talkers. Constant recalibration to new talkers should interrupt listeners' adaptation to talker characteristics including these modest reliable spectral peaks, making them less effective in producing spectral contrast effects. Hearing the same talker does not require such recalibration and is conducive to adaptation to talker characteristics, thus maintaining the presence and magnitudes of spectral contrast effects. In experiment 2, changing the talker on each trial (200 Talkers/200 Sentences) is predicted to induce repeated recalibration to changes in talker identity, resulting in diminished spectral contrast effects relative to single-talker

conditions (One Talker/One Sentence, One Talker/200 Sentences).

II. EXPERIMENT 1

A. Methods

1. Participants

Sixteen undergraduates at the University of Louisville participated in exchange for course credit. All participants were native English speakers and reported normal hearing.

2. Stimuli

a. Sentences. Sentence stimuli were drawn from three sources: (1) a recording of "Please say what this vowel is" (2174 ms), spoken by the second author and used in Stilp *et al.* (2015); (2) the Hearing In Noise Test (HINT; Nilsson *et al.*, 1994); and (3) the Texas Instruments/Massachusetts Institute of Technology (TIMIT) database (Garofolo *et al.*, 1990). Sentences from the HINT and TIMIT corpora were selected according to two criteria. First, sentence duration was constrained to be within one second of the single sentence tested by Stilp *et al.* (2015). Second, sentence spectra were comparable to those in the sentence used in Stilp *et al.* in either the 100–400 Hz (low F_1) or 550–850 Hz (high F_1) region, as amplifying these frequency regions produced spectral contrast effects in identification of target vowels in an /ɪ/-/ɛ/ continuum in Stilp *et al.* (2015). One hundred unique sentences from each corpus with reasonably flat spectra from 100 to 400 Hz were selected (HINT sentence mean duration = 1728 ms; TIMIT sentence mean duration = 2271 ms). The process was repeated for 100 different sentences with reasonably flat spectra in the 550–850 Hz region (HINT sentence mean duration = 1750 ms; TIMIT sentence mean duration = 2225 ms). In all, 200 HINT sentences (adult list) spoken by a single male talker and 200 TIMIT sentences spoken by 200 different talkers (138 males, sampled from all dialect regions) were selected.

b. Filters. Sentences were processed by the same band-pass filters used by Stilp *et al.* (2015). Filter bandwidths were fixed at 300 Hz, and peak gain was set to +20 dB. The low-F_1 filter had a center frequency of 250 Hz and the high-F_1 filter had a 700 Hz center frequency. Filters were created using the fir2 function in MATLAB with 1200 coefficients. The single sentence tested by Stilp *et al.* (2015) was processed by each filter. TIMIT and HINT sentences that were comparable to the single sentence in low-F_1 regions were processed by the low-F_1 filter, and likewise for high-F_1 sentences and the high-F_1 filter (Fig. 1).

c. Vowels. Vowel targets were the same as those previously used in Stilp *et al.* (2015). For a detailed description of the generation procedures, see Winn and Litovsky (2015). Briefly, tokens of /ɪ/ and /ɛ/ were recorded by the second author. Formant contours were extracted using PRAAT (Boersma and Weenink, 2014). In the /ɪ/ endpoint, F_1 linearly increased from 400 to 430 Hz while F_2 linearly decreased from 2000 to 1800 Hz. In the /ɛ/ endpoint, F_1 linearly decreased from 580 to 550 Hz while F_2 linearly decreased from 1800 to 1700 Hz. Formant trajectories were linearly interpolated to create a ten-step continuum of formant tracks. These formant tracks were then used as filters applied to single voice source extracted from the /ɪ/ endpoint, producing the ten-step continuum of vowel tokens. Energy above 2500 Hz was replaced with the energy high-passed-filtered from the original /ɪ/ token for all vowels. Spectra from the continuum endpoints are depicted in Fig. 1(d). Final vowel stimuli were 246 ms in duration with fundamental frequency set to 100 Hz throughout the vowel.

All sentences and vowel targets were equated in RMS amplitude.[2] A vowel target was appended to each sentence with a 50 ms inter-stimulus interval. In the One Talker/200 Sentences and 200 Talkers/200 Sentences conditions, each vowel target was appended to ten different randomly selected sentences, resulting in 200 unique sentence/vowel combinations. In the One Talker/One Sentence condition, each vowel target was appended to each filtered sentence, producing 20 unique sentence/vowel combinations. Stimuli were upsampled to 44 100 Hz.

3. Procedure

After acquisition of informed consent, participants were led into a sound attenuating booth (Acoustic Systems, Inc., Austin, TX). A custom MATLAB script led the participants through the experiment. Stimuli were D/A converted by RME HDSPe AIO sound cards (Audio AG, Haimhausen, Germany) on personal computers and passed through a programmable attenuator (TDT PA4, Tucker-Davis Technologies, Alachua, FL) and headphone buffer (TDT HB6). Stimuli were presented diotically at 70 dB sound pressure level (SPL) over circumaural headphones (Beyerdynamic DT-150, Beyerdynamic Inc. USA, Farmingdale, NY). At the end of each trial, participants clicked the mouse to indicate whether the target vowel sounded more like "ih (as in 'bit')" or "eh (as in 'bet')". The experiment was self-paced and allowed the participants the opportunity to take breaks between each of the three testing blocks. Each block tested 200 trials in a single condition in random orders and lasted approximately 12 min. Block orders were counterbalanced across participants.

B. Results

Listeners were required to meet a performance criterion of 80% mean accuracy for identifying vowel continuum endpoints in each experimental block. Three listeners failed to meet this criterion, so all results for these listeners were removed from subsequent analysis. Figures 2(a)–2(c) show mean responses and logistic regression fits for the remaining 13 listeners. Logistic regressions were fit to each listener's data for high-F_1 and low-F_1 precursor conditions. Logistic regression midpoints were defined as the stimulus step number at which listeners would label the vowel target /ɛ/ 50% of the time. Vowel continuum steps were numbered 1–10 [see x axes of Figs. 2(a)–2(c)], and midpoints were interpolated between these steps as needed. Midpoints were calculated for each regression fit (low F_1, high F_1) for each listener. Spectral contrast effect magnitude was defined as

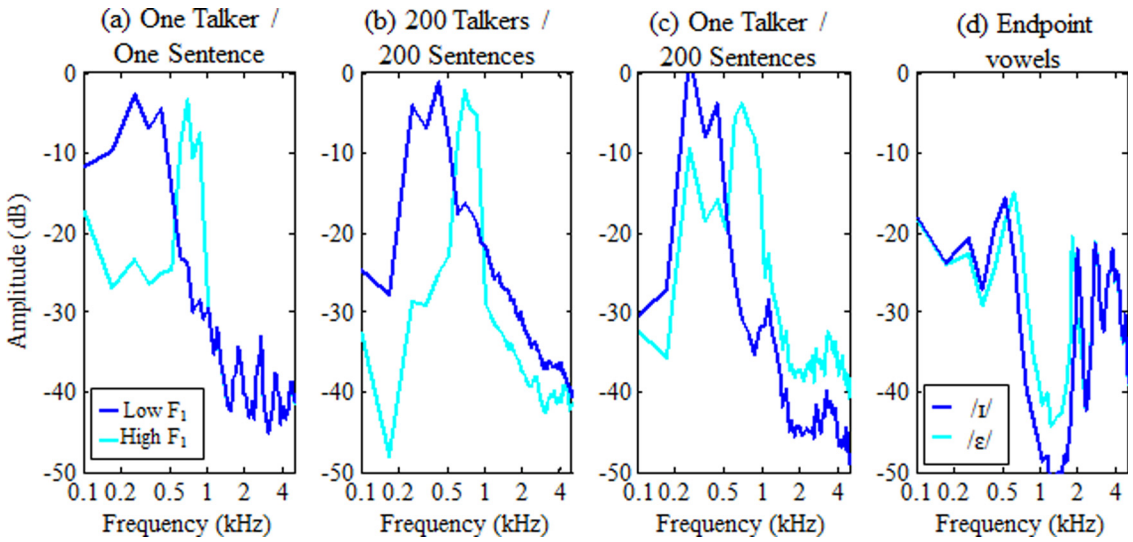

FIG. 1. (Color online) Long-term average spectra for stimuli presented in experiment 1. Separate lines indicate spectra for sentences with +20 dB low-F_1 or high-F_1 peaks added by filtering in (a) One Talker/One Sentence, (b) 200 Talkers/200 Sentences, and (c) One Talker/200 Sentences conditions. (d) Long-term average spectra for endpoints of the continuum of vowel targets. Vowels principally differ in F_1 center frequency (/ɪ/ = 415 Hz, /ɛ/ = 565 Hz).

the difference in midpoints between the high-F_1 and low-F_1 regression fits, measured in the number of stimulus steps. Mean contrast effect magnitudes were extremely similar across conditions [One Talker/One Sentence: $M = 1.77$, SE = 0.31; 200 Talkers/200 Sentences: $M = 1.74$, SE = 0.23; One Talker/200 Sentences: $M = 1.52$, SE = 0.22; Fig. 2(d)].[3] To retain sensitivity to paired comparisons between conditions, results were analyzed using paired-sample t-tests. Spectral contrast effect magnitudes did not differ between any conditions (all $t_{12} < 1.24$, $p > 0.22$, Bonferroni-corrected for multiple comparisons).

C. Discussion

Spectral contrast effect magnitudes were comparable in the face of within-talker acoustic variability (comparing One Talker/One Sentence to One Talker/200 Sentences) and between-talker acoustic variability (comparing One Talker/One Sentence and One Talker/200 Sentences to 200 Talkers/200 Sentences). Thus, talker information and acoustic variability did not influence spectral contrast effects in vowel categorization when reliable spectral peaks in the preceding acoustic context were large (+20 dB). Results are consistent

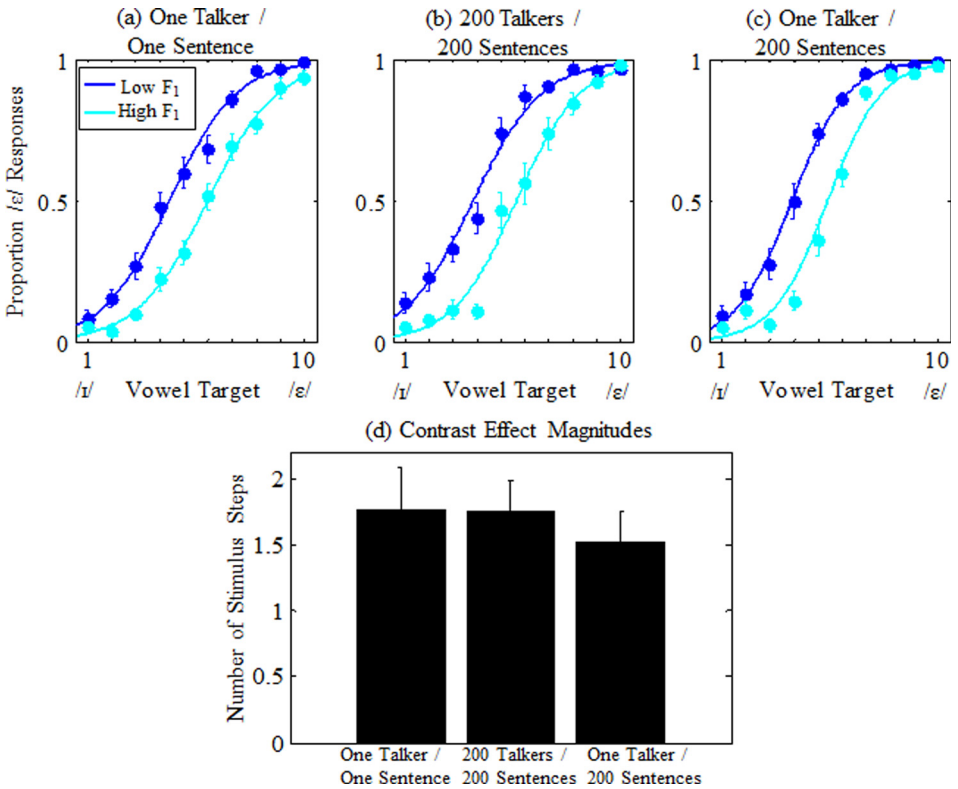

FIG. 2. (Color online) Results from experiment 1 where sentence contexts featured large reliable spectral peaks (+20 dB). Panels 2(a)–2(c) depict the mean proportions of /ɛ/ responses as a function of vowel target from the ten-step vowel continuum. Logistic regression functions are fit to mean responses for (a) One Talker/One Sentence, (b) 200 Talkers/200 Sentences, and (c) One Talker/200 Sentences conditions. Circles indicate mean proportions of /ɛ/ responses; error bars indicate 1 standard error. (d) Mean spectral contrast effect magnitudes in experiment 1. Contrast effect magnitudes were defined as the number of stimulus steps separating midpoints of logistic regression functions, calculated for each listener and then averaged. Error bars indicate 1 standard error.

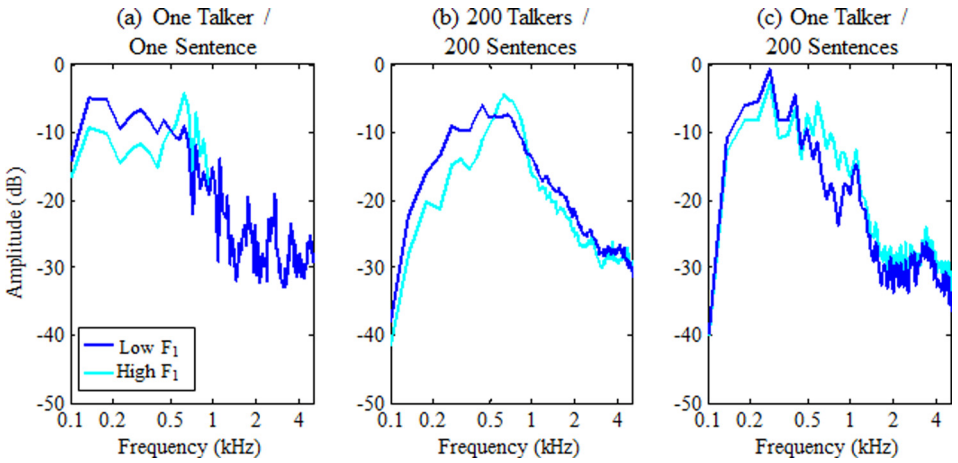

FIG. 3. (Color online) Long-term average spectra for stimuli presented in experiment 2. Separate lines indicate spectra for sentences with +5 dB low-F_1 or high-F_1 peaks added by filtering in (a) One Talker/One Sentence, (b) 200 Talkers/200 Sentences, and (c) One Talker/200 Sentences conditions. Sentences are the same as those presented in experiment 1 and depicted in Fig. 1.

with those of Holt and colleagues (Huang and Holt, 2012; Laing *et al.*, 2012), who reported comparable spectral contrast effects when the preceding acoustic context was a talker or a series of sine tones.

Earlier research framed spectral contrast effects as the result of listeners compensating for systematic distortion of the communication channel (Watkins, 1991). Watkins and colleagues manipulated their preceding acoustic contexts using filters that captured the difference between two vowel spectra (endpoints of the target vowel continuum). These spectral envelope differences were often dramatic, adding spectral peaks to the preceding context that were often +15 dB and sometimes up to +30 dB (Watkins, 1991; Watkins and Makin, 1994, 1996a, 1996b). Many investigations of spectral contrast effects, including experiment 1, presented acoustic contexts with similarly dramatic spectral peaks (see Stilp *et al.*, 2015 for review). It is quite possible that these spectral properties were not perceived as being produced by each talker's speech but instead by the communication channel, as every sound the listeners heard featured these fairly extreme spectral peaks. As such, the acoustic information responsible for producing the contrast effect was functionally separated from the talkers' speech. This might explain the apparent insensitivity to talker information in experiment 1, as similar results were observed whether 1 or 200 talkers were heard.

Experiment 2 added very modest reliable spectral peaks (+5 dB) to sentence spectra. These modest spectral peaks are expected to be attributed to the talkers' speech rather than the communication channel, providing a more sensitive test of talker influences on spectral contrast effects than experiment 1. Repeated recalibration to a new talker on each trial (200 Talkers/200 Sentences) is predicted to interfere with listeners' adjustment to these stable spectral properties, thus diminishing the size of spectral contrast effects. Experiment 2 also tests whether this interference manifests amidst extreme acoustic variability within a single talker (One Talker/200 Sentences) compared to minimal between- or within-talker variability (One Talker/One Sentence). If the results of experiment 1 were not circumstantial and talker information truly plays no role in spectral contrast effects, then experiment 2 will simply replicate this result for modest reliable spectral peaks.

III. EXPERIMENT 2

A. Methods

1. Participants

Fourteen undergraduates from the University of Louisville participated in exchange for course credit. All participants were native English speakers and reported normal hearing. None participated in experiment 1.

2. Stimuli

Stimuli were the same as those reported in experiment 1 with the exception of filtering. Sentences were processed with a +5 dB bandpass filter with the same bandwidths and center frequencies (Fig. 3). Filtered sentences were low-pass filtered at 5 kHz to match the spectral bandwidth of the target vowels.

3. Procedure

The procedure was the same as in experiment 1.

B. Results

Due to the performance criterion of 80% accuracy on labeling vowel continuum endpoints, two participants' complete data sets were removed from further analysis. Figures 4(a)–4(c) show mean responses and logistic regression fits across the remaining 12 listeners. As expected, spectral contrast effect magnitudes were smaller than those observed in experiment 1 owing to more modest spectral peaks being added to the sentences. Three one-way *t*-tests against 0 indicated that statistically significant contrast effects occurred in all conditions [all t's > 3.02, p's < 0.01; Fig. 4(d)]. Unlike experiment 1, contrast effect magnitudes differed by condition.[4] Paired-sample *t*-tests indicate that contrast effects in the 200 Talkers/200 Sentences condition ($M = 0.26$, SE $= 0.09$) were significantly smaller than those in the One Talker/200 Sentences condition ($M = 0.52$, SE $= 0.09$) ($t_{11} = 2.71$, $p = 0.02$). Contrast effects in the 200 Talkers/200 Sentences condition were marginally smaller than those in the One Talker/One Sentence condition ($M = 0.51$, SE $= 0.12$) ($t_{11} = 1.60$, $p = 0.14$). This trend is heavily influenced by one listener in the One Talker/One Sentence condition exhibiting

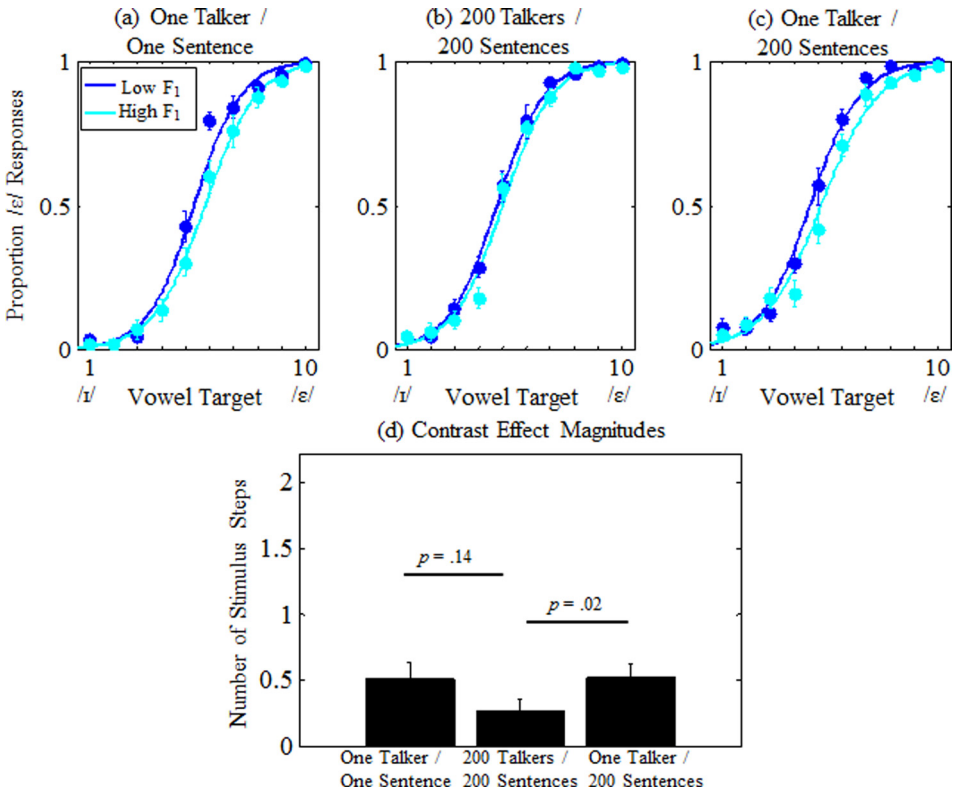

FIG. 4. (Color online) Results from experiment 2, where sentence contexts featured small reliable spectral peaks (+5 dB). Notation is identical to Fig. 2. Logistic regression functions are fit to mean responses for (a) One Talker/One Sentence, (b) 200 Talkers/200 Sentences, and (c) One Talker/200 Sentences conditions. Circles indicate mean proportions of /ɛ/ responses; error bars indicate 1 standard error. (d) Mean spectral contrast effect magnitudes in experiment 2. Error bars indicate 1 standard error.

a contrast effect in the opposite direction (magnitude = −0.40; mean of remaining listeners = 0.63) even though s/he met the 80% accuracy criterion on vowel endpoints and exhibited contrast effects that were highly consistent with group means in the other two conditions (One Talker/200 Sentences: listener contrast effect = 0.53, group mean with that listener excluded = 0.52; 200 Talker/200 Sentences: listener contrast effect = 0.26, group mean with that listener excluded = 0.26). When the comparison is conducted without this participant, contrast effects are significantly different from one another ($t_{10} = 2.41$, $p = 0.04$). Means for the One Talker/One Sentence and One Talker/200 Sentences conditions were extremely similar to one another, rendering *t*-tests for group differences unnecessary.

C. Discussion

The magnitudes of spectral contrast effects in experiment 2 depended on the number of talkers. When the acoustic context was spoken by a single talker, contrast effect magnitudes were larger than when the acoustic context featured sentences spoken by 200 different talkers. This finding reveals that spectral contrast effects in vowel categorization are sensitive to talker information, consistent with past research on talker normalization (Creelman, 1957; Verbrugge, 1976; Mullenix *et al.*, 1989; Martin *et al.*, 1989) and contrary to claims of insensitivity to talker information (Laing *et al.*, 2012).

Similar to experiment 1, contrast effect magnitudes were comparable regardless of the number of sentences produced by a single talker (One Talker/One Sentence versus One Talker/200 Sentences). Thus, between-talker acoustic variability is responsible for attenuating contrast effects in the 200 Talkers/200 Sentences condition and not variability due to different sentences being spoken. Results also extend Holt (2005, 2006), where frequencies in tone histories were presented in different randomized orders from trial to trial. Here, contrast effect magnitudes maintained over substantially greater acoustic variability when they were spoken by a single talker.

In the One Talker/One Sentence condition, the same talker produced the precursor sentence and the target vowels, presenting a straightforward case of talker normalization. In the One Talker/200 Sentences condition, a different talker produced the precursor sentences but the same target vowels were used. Despite this difference in talkers, contrast effect magnitudes were similar across conditions. This raises the question as to whether this result embodies talker normalization or some broader form of acoustic normalization. Many acoustic properties provide cues to talker identity, but two of the primary cues are fundamental frequency and vowel space. Barreda (2012) reported that, of these two cues, fundamental frequency was far more effective in indicating a change in talker identity. Mean fundamental frequency was highly comparable across single-talker conditions (One Talker/One Sentence mean $f0 = 100$ Hz; One Talker/200 Sentences mean $f0 = 110$ Hz, SE = 1), suggesting a change in talker was not sufficiently cued by mean pitch. Magnuson and Nusbaum (2007) also reported that when listeners were not given explicit instructions about the number of talkers (similar to listeners in the present experiments), they treated a 10 Hz difference in $f0$ as coming from the same talker. Thus, contrast effect magnitudes were comparable due at least in part to the acoustic similarity in talker's voices (as indicated by mean pitch). This is consistent with Goldinger (1996), who reported greater facilitation of phoneme and

word recognition when the (different) talker's voice was more perceptually similar to the target voice (same gender, similar relative pitches).

It is unsurprising that mean sentence pitch was both higher and more variable in the 200 Talkers/200 Sentences condition (mean $f0 = 149$ Hz, SE = 3). This distribution was essentially comprised of two smaller distributions, one for male talkers ($n = 138$, mean $f0 = 122$ Hz, SE = 1) and one for female talkers ($n = 62$, mean $f0 = 210$ Hz, SE = 3). If similarity in mean pitch helped maintain contrast effects across One Talker/One Sentence and One Talker/200 Sentences conditions, then differences in mean pitch are expected to attenuate contrast effects in the 200 Talkers/200 Sentences conditions. To examine this possibility, mean sentence pitch was calculated in PRAAT by extracting pitch contours from 50 to 500 Hz. Pitch contours were visually inspected and manually edited to remove any erroneous pitch values, then the mean was calculated. These mean pitches were then averaged across sentences that preceded the same target vowel in a given filtering condition (e.g., mean pitch across the 10 low-F_1 sentences that preceded vowel target 1; mean pitch for the 10 high-F_1 sentences that preceded vowel target 1, etc.). Figure 5(a) plots the absolute difference in mean pitch across filtering conditions for each vowel target (e.g., absolute difference between mean pitch of the 10 low-F_1 sentences and that of the 10 high-F_1 sentences, all of which preceded vowel target 1, etc.). As expected, high talker variability produced a wide range of mean pitch differences, from very modest (vowel targets 7, 10) to relatively large (vowel targets 5, 9). Large pitch differences for sentences preceding vowel target 5 were due to 10 male talkers being randomly selected to precede the target vowel in the low-F_1 condition (mean $f0 = 123$ Hz) while four female and six male talkers were selected in the high-F_1 condition (mean $f0 = 169$ Hz). Similarly, large pitch differences for sentences preceding vowel target 9 were due to three male and seven female talkers preceding the target vowel in the low-F_1 condition (mean $f0 = 187$ Hz) and six male and four female talkers being selected in the high-F_1 condition (mean $f0 = 149$ Hz). The same analyses were conducted on sentences in the One Talker/200 Sentences condition for comparison, and most pitch differences were on the order of a few hertz.

The relationship between mean pitch differences and contrast effect magnitudes was investigated as follows. Contrast effect magnitudes were approximated by calculating the difference in mean proportions of /ɛ/ responses for each vowel target [i.e., the vertical distances between circles in Fig. 4(b), and the same in Fig. 4(c)]. In Fig. 5(b), /ɛ/ response probabilities following the high-F_1-filtered precursor were subtracted from those following the low-F_1-filtered precursor; positive differences were consistent with spectral contrast. For mid-continuum vowels (which are generally influenced the most by contextual factors), response differences corresponded well to absolute differences in mean pitch. In the One Talker/200 Sentences condition, mean pitch differences were small (1–8 Hz) and response differences were comparatively large (6%–15%), consistent with spectral contrast. In the 200 Talkers/200 Sentences

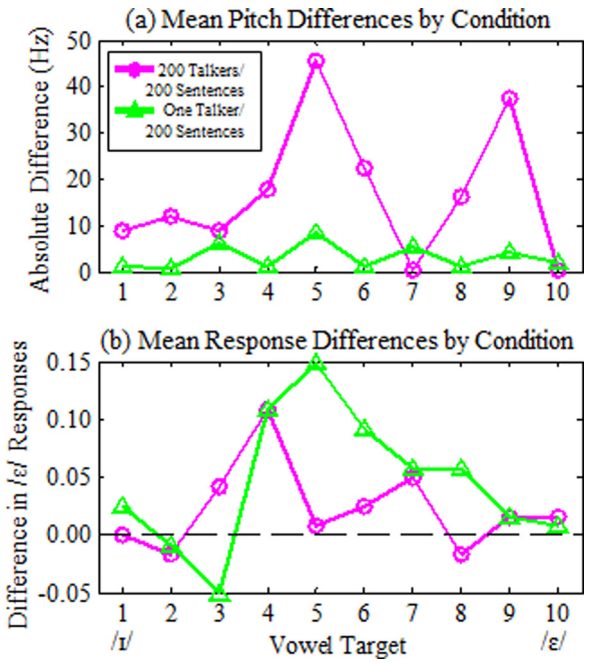

FIG. 5. (Color online) (a) Absolute differences in mean sentence pitch for 200 Talkers/200 Sentences (circles) and One Talker/200 Sentences (triangles) conditions in experiment 2. Mean sentence pitch was calculated for each of the ten sentences that preceded each vowel target (10 in low-F_1-amplified condition, 10 in high-F_1-amplified condition), and grand means were calculated. Symbols in (a) indicate the absolute differences in these grand means. (b) Differences in the mean proportions of /ɛ/ responses for each vowel target in the 200 Talkers/200 Sentences (circles) and One Talker/200 Sentences (triangles) conditions in experiment 2. These differences are equal to the vertical distance between symbols in Fig. 4(b) as well as Fig. 4(c). Positive differences are consistent with spectral contrast effects.

condition, for vowel target 5, mean pitch differences were very large (46 Hz) and response differences were extinguished (1%). While mean pitch differences were not as extreme for vowel target 6 (22 Hz), response differences were still very small (2.5%). When mean pitch differences were small for vowel target 7 (<1 Hz), response differences (5%) were on par with those observed in the One Talker/200 Sentences condition (6%). Across these stimuli, small differences in mean pitch (low talker variability) were conducive to contrast effects while large differences in mean pitch (high talker variability) corresponded to diminished contrast effects.

While a potential relationship between talker pitch and spectral contrast effect magnitude is intriguing, caution is warranted for several reasons. First, apparent exceptions to this relationship are also observed.[5] Vowel targets 4 and 8 exhibited similar patterns of mean pitch differences within each condition (18 and 16 Hz for 200 Talkers/200 Sentences; 1 and 1 Hz for One Talker/200 Sentences) but very different behavioral results. For vowel target 8, response differences were 6% in the One Talker/200 Sentences condition and −2% in the 200 Talkers/200 Sentences condition, consistent with pitch differences influencing behavioral results. For vowel target 4, response differences were 11% in both conditions, or comparable spectral contrast effects despite talker variability. Second, analyses examined broad trends in pitch differences across low-F_1 and high-F_1 conditions, but this approach averaged pitch measurements within tokens and

among tokens in the same filtering condition. Finally, talkers were randomly assigned to different vowel targets in order to maximize between-talker variability from trial to trial. Talker pitch was not experimentally controlled in any way, calling for targeted manipulations that formally test this relationship to determine whether these results are suggestive or serendipitous.

IV. GENERAL DISCUSSION

Spectral contrast effects have been widely shown to influence speech categorization, but contributions of talker information to this process have been debated. The present experiments address this debate through two key manipulations: acoustic variability (both between- and within-talker) and the magnitudes of spectral peaks that produce spectral contrast effects (large or modest). In experiment 1, comparable contrast effects were observed regardless of whether the preceding sentence context was spoken by 1 or 200 different talkers. Results appear to be consistent with studies showing talker information is not necessary for producing spectral contrast effects (Holt, 2005, 2006; Huang and Holt, 2012; Laing *et al.*, 2012). However, all of these experiments presented acoustic contexts with large spectral peaks (on par with +20 dB tested here). Listeners might treat such an extreme and pervasive spectral property as originating from the communication channel rather than each talker's speech. In experiment 2, preceding sentences featured more modest reliable spectral peaks (+5 dB) that were predicted to be attributed to talkers and not the communication channel. Listeners exhibited smaller spectral contrast effects following 200 sentences spoken by 200 different talkers compared to 1 or 200 sentences produced by a single talker. Repeated recalibration to new talkers interrupted listeners' adaptation to talker characteristics including these modest reliable spectral peaks, thereby diminishing the magnitudes of spectral contrast effects. No such interference was observed when sentences were all spoken by the same talker. Thus, when reliable spectral peaks in the listening context are large, spectral contrast effects are not influenced by talker information (experiment 1), but when reliable spectral peaks are modest, spectral contrast effects are influenced by talker information (experiment 2). These results bridge conflicting reports of talker information influencing speech perception (i.e., talker normalization; Creelman, 1957; Verbrugge *et al.*, 1976; Assmann *et al.*, 1982; Martin *et al.*, 1989; Mullennix *et al.*, 1989) with reports that it does not (Laing *et al.*, 2012). Results also shed important light on how listeners respond to stable spectral properties in speech by attributing them to talkers or to other sources such as the communication channel.

Previous studies of spectral contrast effects routinely utilized a single token as the preceding acoustic context, presenting it to listeners dozens if not hundreds of times during the experiment. The current study added acoustic variability to the context by manipulating the number of sentences produced by each talker. Within-talker acoustic variability did not affect the magnitudes of spectral contrast effects, as performance in One Talker/One Sentence and One Talker/200 Sentences conditions was comparable in each experiment. This extends spectral contrast effects to situations where the acoustic context is highly uncertain. Stilp *et al.* (2015) first reported spectral contrast effects following +5 dB reliable spectral peaks, concluding that spectral contrast effects might influence everyday speech perception more than previously considered. Experiment 2 builds on this suggestion as contrast effects influenced vowel categorization when the preceding context and talker were unpredictable from trial to trial.

A wide range of studies demonstrate higher-level influences on lower-level acoustic processing in speech perception. Listeners' expectations of talker gender (Johnson *et al.*, 1999), dialect region (Hay *et al.*, 2006a), age and social status (Hay *et al.*, 2006b), or the number of talkers (Magnuson and Nusbaum, 2007) can alter phoneme perception and/or reaction times. Directing listeners' attention to different aspects of the speech signal, such as linguistic versus talker information, can also affect word recognition (Theodore *et al.*, 2015). In addition, perception of an ambiguous phoneme is biased toward forming a valid word (Norris *et al.*, 2003). The present results may provide another example of higher-level information influencing lower-level processing. In experiment 2, changing talkers and sentences on each trial resulted in diminished spectral contrast effects. Changes in talker identity were cued by both high-level talker characteristics (accent, dialect region, gender, etc.) and lower-level acoustic properties (pitch, duration, prosodic details, etc.). Changing low-level acoustic properties alone due to sentence variability did not diminish the magnitudes of spectral contrast effects (comparable results across One Talker/200 Sentences and One Talker/One Sentence conditions), but such changes are smaller within-talker than between-talkers. While between-talker variability inherently introduces acoustic variability across sentences, changes in talker appear to be responsible for influencing lower-level acoustic processing (diminishing the size of spectral contrast effects).

Listeners' ability to separate phonemic information from talker information is debated (see Pisoni, 1993 for a review). In a study by Mullenix and Pisoni (1990), listeners attended to either voice or word information in a modified Garner interference paradigm. When voice and word both varied, listeners could not ignore irrelevant variation in the unattended property, suggesting that talker and phoneme information are processed simultaneously. The present results seem to provide differing accounts of this interpretation. In experiment 1, contrast effect magnitudes were comparable, suggesting that listeners processed phonemic information independent of talker information. In experiment 2, contrast effects were smaller in the multiple-talker condition than single-talker conditions, suggesting phonemic and talker information were processed simultaneously. However, sentences in experiment 2 were processed to add very modest reliable spectral peaks (+5 dB, as opposed to +20 dB in experiment 1). It is important to note that these stimuli more closely resemble unfiltered speech used in talker normalization experiments, including Mullenix and Pisoni (1990). Thus, listeners' ability to separate phonemic information from talker information can be influenced by specific

acoustic properties of the preceding context, such as the magnitude of reliable spectral peaks.

Perceptual benefits are observed when hearing familiar talkers rather than novel talkers (Nygaard *et al.*, 1994; Nygaard and Pisoni, 1998). In these experiments, listeners were taught to associate different voices with common names, then, at test, identified words spoken either by familiar or novel talkers. Listeners were faster and more accurate when identifying words spoken by familiar talkers. It is conceivable that listeners implicitly became familiar with and perhaps "learned" the single talker following 200 presentations of the same or unique sentences. This familiarity appears to be conducive to spectral contrast effects in experiment 2 as constant recalibration to new talkers in the multi-talker condition resulted in diminished effects. However, spectral contrast effects are a unique dependent variable for measuring influences of talker information and familiarity. For example, spectral contrast effects alter perception of a target sound based on the acoustics of the preceding context. This is not well captured by accuracy, especially for ambiguous mid-continuum stimuli. Additionally, the magnitudes of spectral contrast effects do not reflect the quality of performance. Larger spectral contrast effects do not necessarily imply the listener is performing "better." Despite these differences, the relationship between talker information and spectral contrast effects might aid understanding of low-level processes that contribute to talker normalization (Mullenix *et al.*, 1989).

In conclusion, the present results illuminate how talker information and reliable spectral properties of the listening context influence speech categorization. When the preceding acoustic context possesses large reliable spectral peaks, listeners exhibit comparable spectral contrast effects irrespective of the number of talkers. This is consistent with attributing the large spectral peaks to the communication channel and not to the talkers' speech. When the preceding context possesses modest reliable spectral peaks, contrast effects are smaller when the context is spoken by 200 different talkers versus a single talker speaking 1 or 200 different sentences. This is consistent with attributing the modest spectral peaks to the talkers' speech, and sensitivity to this spectral peak was disrupted by constant recalibration to new talkers. Thus, listeners' sensitivity to talker information depends on key acoustic properties of the preceding context, such as the magnitudes of reliable spectral peaks.

ACKNOWLEDGMENTS

The authors thank Rachel Theodore, Santiago Barreda, and two anonymous reviewers for their very helpful comments. The authors also thank Kara Hendrix, Almina Klanco, Emily Nations, and Caitlyn Stromatt for assistance with data collection.

[1]This argument pertains to long-term average speech spectra measured at the source. It is acknowledged that other factors influence the speech signal before it reaches the listener's cochlea and that in some cases, these factors can add fairly substantial spectral peaks to the signal (e.g., room acoustics, head-related transfer function), but these peaks primarily occur at higher frequencies than those tested in the present experiments (F_1 peaks below 1 kHz).

[2]Due to a programming error, sentences were not low-pass filtered at 5 kHz to match the spectral bandwidth of target vowels as was done in experiment 2. However, frequencies above 5 kHz are spectrally remote relative to the F_1 regions of interest (below 1 kHz) and were at least -40 dB relative to peak spectral amplitude in the precursor. Thus frequencies above 5 kHz likely did not affect performance in experiment 1. Spectral contrast effect magnitudes in the One Talker/One Sentence condition did not statistically differ from those tested in Stilp *et al.* (2015) (independent-samples *t*-test: $t_{24} = 0.41$, $p = 0.68$).

[3]When contrast effects are evaluated with response probabilities rather than number of stimulus steps, on average listeners responded "eh" 15.00% (One Talker/One Sentence), 14.38% (One Talker/200 Sentences), and 16.77% (200 Talkers/200 Sentences) more often following the low-F_1-emphasized precursor compared to the high-F_1-emphasized precursor.

[4]When contrast effects are evaluated with response probabilities rather than number of stimulus steps, on average listeners responded "eh" 4.50% (One Talker/One Sentence), 4.58% (One Talker/200 Sentences), and 2.33% (200 Talkers/200 Sentences) more often following the low-F_1-emphasized precursor compared to the high-F_1-emphasized precursor.

[5]For vowel target 9, small pitch differences occurred in the One Talker/200 Sentences condition and large pitch differences occurred in the 200 Talkers/200 Sentences condition, yet similar patterns of responses were observed (Fig. 5). Vowel continuum members near or at series endpoints are far less sensitive to contextual factors than mid-continuum members. This is evident by the negligible effects of filtering on responses to vowel continuum endpoints [see Figs. 4(a)–4(c)]. It is difficult to determine whether this result suggests a null relationship between talker pitch and vowel identifications, unambiguous formant cues in the target vowel that diminish the influence of preceding context, or both.

Assmann, P. F., Nearey, T. M., and Hogan, J. T. (**1982**). "Vowel identification: Orthographic, perceptual, and acoustic aspects," J. Acoust. Soc. Am. **71**(4), 975–989.

Barreda, S. (**2012**). "Vowel normalization and the perception of speaker changes: An exploration of the contextual tuning hypothesis," J. Acoust. Soc. Am. **132**(5), 3453–3464.

Boersma, P., and Weenink, D. (**2014**). "PRAAT: Doing phonetics by computer (version 5.3.61) [computer program]," http://www.praat.org (Last viewed January 1, 2014).

Creelman, C. D. (**1957**). "Case of the unknown talker," J. Acoust. Soc. Am. **29**, 655.

Fowler, C. A. (**2006**). "Compensation for coarticulation reflects gesture perception, not spectral contrast," Percept. Psychophys. **68**(2), 161–177.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N. (**1990**). *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus* (CDROM) (National Institute of Standards and Technology), NIST Order No. PB91-505065.

Goldinger, S. D. (**1996**). "Words and voices: Episodic traces in spoken word identification and recognition memory," J. Exp. Psychol. Learn. Mem. Cogn. **22**(5), 1166–1183.

Hay, J., Nolan, A., and Drager. K. (**2006a**). "From *fush* to *feesh*: Exemplar priming in speech perception," Ling. Rev. **23**(3), 351–379.

Hay, J., Warren, P., and Drager. K. (**2006b**). "Factors influencing speech perception in the context of a merger in progress," J. Phon. **34**, 458–484.

Holt, L. L. (**2005**). "Temporally nonadjacent nonlinguistic sounds effect speech categorization," Psych. Sci. **16**(4), 305–312.

Holt, L. L. (**2006**). "The mean matters: Effects of statistically defined nonspeech spectral distributions on speech categorization," J. Acoust. Soc. Am. **120**(5), 2801–2817.

Huang, J., and Holt, L. L. (**2012**). "Listening for the norm: Adaptive coding in speech categorization," Front. Psychol. **3**(10), 1–6.

Johnson, K., Strand., E. A., and D'Imperio, M. (**1999**). "Auditory-visual integration of talker gender in vowel perception," J. Phon. **27**, 359–384.

Joos, M. (**1948**). "Acoustic phonetics," Language **24**, 5–136.

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**(1), 98–104.

Laing, E. J. C., Liu, R., Lotto, A. J., and Holt, L. L. (**2012**). "Tuned with a tune: Talker normalization via general auditory processes," Front. Psychol. **3**, 1–9.

Lotto, A. J., and Kluender, K. R. (**1998**). "General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification," Percept. Psychophys. **60**(4), 602–619.
Magnuson, J. S., and Nusbaum, H. C. (**2007**). "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," J. Exp. Psychol. Hum. Percept. Perform. **33**(2), 391–409.
Martin, C. S., Mullennix, J. W., Pisoni, D. B., and Summers W. V. (**1989**). "Effects of talker variability on recall of spoken word lists," J. Exp. Psychol. Learn. Mem. Cogn. **15**(4), 676–684.
Mullennix, J. W., and Pisoni, D. B. (**1990**). "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**(4), 379–390.
Mullennix, J. W., Pisoni, D. B., and Martin C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.
Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**(2), 1085–1099.
Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cog. Psychol. **47**, 204–238.
Nygaard, L. C., and Pisoni D. B. (**1998**). "Talker-specific learning in speech perception," Percept. Psychophys. **60**(3), 355–376.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psych. Sci. **5**(1), 42–46.
Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.
Pisoni, D. B. (**1993**). "Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning," Speech. Commun. **13**(1–2), 109–125.
Sjerps, M. J., McQueen, J. M., and Mitterer, H. (**2011**). "Constraints on the processes responsible for the extrinsic normalization of vowels," Atten. Percept. Psychophys. **73**(4), 1195–1215.
Stilp, C. E., Alexander, J. M., Kiefte, M., and Kluender, K. R. (**2010**). "Auditory color constancy: Calibration to reliable spectral properties across nonspeech context and targets," Atten. Percept. Psychophys. **72**(2), 470–480.
Stilp, C. E., Anderson, P. W., and Winn, M. B. (**2015**). "Predicting contrast effects following reliable spectral properties in speech perception," J. Acoust. Soc. Am. **137**(6), 3466–3476.
Theodore, R. M., Blumstein, S. E., and Luthra, S. (**2015**). "Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis," Atten. Percept. Psychophys. **77**(5), 1674–1684.
Verbrugge, R. R., Strange, W., Shankweiler, D. P., and Edman, T. R. (**1976**). "What information enables a listener to map a targets vowel space?," J. Acoust. Soc. Am. **60**(1), 198–212.
Watkins, A. J. (**1991**). "Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **90**(6), 2942–2955.
Watkins, A. J., and Makin, S. J. (**1994**). "Perceptual compensation for speaker differences and spectral-envelope distortion," J. Acoust. Soc. Am. **96**(3), 1263–1282.
Watkins, A. J., and Makin, S. J. (**1996a**). "Effects of spectral contrast on perceptual compensation for spectral-envelope distortion," J. Acoust. Soc. Am. **99**(6), 3749–3757.
Watkins, A. J., and Makin, S. J. (**1996b**). "Some effects of filtered contexts on the perception of vowels and fricatives," J. Acoust. Soc. Am. **99**(1), 588–594.
Winn, M. B., and Litovsky, R. Y. (**2015**). "Using speech sounds to test functional spectral resolution in listeners with cochlear implants," J. Acoust. Soc. Am. **137**(3), 1430–1442.