Structured phonetic variation facilitates talker identification

Divya Ganugapati[1] & Rachel M. Theodore[1,2]

[1]Department of Speech, Language, and Hearing Sciences

University of Connecticut

850 Bolton Road, Unit 1085

Storrs, CT 06269-1085

divya.ganugapati@uconn.edu, rachel.theodore@uconn.edu

[2]Connecticut Institute for the Brain and Cognitive Sciences

University of Connecticut

337 Mansfield Road, Unit 1272

Storrs, CT 06269-1272

Abbreviated title: Phonetic variation facilitates talker identification

Author to whom correspondence should be addressed:

Rachel M. Theodore, Ph.D.

rachel.theodore@uconn.edu

Abstract

Listeners use talker-specific phonetic structure to facilitate language comprehension. This study tests whether sensitivity to talker-specific phonetic variation also facilitates talker identification. During training, two listener groups learned to associate talkers' voices with cartoon pseudo-faces. For one group, each talker produced characteristically different voice-onset-time values; for the other group, no talker-specific phonetic structure was present. After training, listeners were tested on talker identification for trained and novel words, which was improved for those who heard structured phonetic variation compared to those who did not. These findings suggest an additive benefit of talker-specific phonetic variation for talker identification beyond traditional indexical cues.

1. Introduction

In order to map the acoustic signal to meaning, listeners must solve the lack of invariance problem for speech, which can arise, for example, because multiple acoustic forms are produced for a given speech sound, or because one or more phonemes of the canonical form may be omitted in a given word. There is a rich literature demonstrating that some variability in speech acoustics is highly structured, including variability associated with talkers' idiolects. For example, talkers show differences in their production of formant frequencies for vowels (Hillenbrand, Getty, Clark, & Wheeler, 1995), spectral center of gravity for fricatives (Newman, Clouse, & Burnham, 2001), and voice-onset-time (VOT) for word-initial stop consonants (Allen, Miller, & DeSteno, 2003; Hullebus, Tobin, & Gafos, 2018 (German); Theodore, Miller, & DeSteno, 2009). In other words, talkers have characteristic idiolectal patterns for acoustic-phonetic properties of speech, including VOT. Listeners can track talkers' characteristic productions (Theodore & Miller, 2010) and dynamically modify the mapping to speech sounds to reflect talker-specific phonetic distributions (e.g., Clayards, Tanenhaus, Aslin, & Jacobs, 2008; Theodore, Myers, & Lomibao, 2015). Listeners also show increased word transcription accuracy for familiar compared to unfamiliar talkers (Nygaard & Pisoni, 1998). Collectively, these findings demonstrate that listeners derive talker-specific mappings to speech sounds that serve to facilitate language comprehension.[1]

The interplay between talker processing and linguistic processing is also observed in the domain of voice processing. Listeners show increased talker identification for talkers speaking a

[1] Unless otherwise indicated in parentheses following each citation, the examined language in cited studies was American English. In English, there is a two-way phonological voicing contrast between short-lag VOTs that cue voiced stops and long-lag VOTS that cue voiceless stops (Lisker & Abramson, 1964).

familiar compared to an unfamiliar language [e.g., Goggin, Thompson, Strube, & Simental, 1991 (English, German, Spanish)]. There is some evidence to suggest that experience with the linguistic sound structure plays an important role in talker identification, consistent with frameworks that outline *a priori* computational expectations that talker-specific phonetic variation should facilitate voice processing (Kleinschmidt & Jaeger, 2015). For example, listeners who have regular exposure to a nonnative language show increased talker identification for that language compared to listeners without regular exposure (Orena, Theodore, & Polka, 2015). Other studies have shown that listeners can identify native-language voices from sine-wave speech analogs (Remez, Fellowes, & Rubin, 1997), a signal manipulation that removes traditional indexical properties (e.g., fundamental frequency) but preserves some idiosyncratic phonetic variation, and that listeners can learn to use VOT as a cue to talker identity for voices that are otherwise identical (Francis & Driscoll, 2006).

Neuroimaging findings have shown that brain regions responsible for mapping sound to meaning are sensitive to speaker information in addition to lexical information (Chandrasekaran, Chan, & Wong, 2011). Listeners show sensitivity to voice information at early, pre-attentive stages of processing, challenging the view that cues to voice identity are discarded in the process of mapping speech to meaning (Knösche, Lattner, Maess, Schauer, & Friederici, 2002 (German); Tuninetti, Chládková, Peter, Schiller, & Escudero, 2017 (Dutch, Australian English)). Moreover, brain regions associated with voice processing are also sensitive to talker-specific phonetic variation (Knösche et al., 2002; Myers & Theodore, 2017). In Myers and Theodore (2017), listeners heard two talkers produce characteristically different VOTs for word-initial voiceless stops during a brief exposure phase. Following exposure, neural activation was measured using fMRI while listeners completed a phonetic categorization task for VOTs that were either consistent or inconsistent with their exposure. Of interest to the current work, right temporoparietal regions implicated in voice processing showed sensitivity to the consistency between VOT variant and talker exposure as reflected by increased activation for VOTs that were atypical compared to typical of the speaker based on previous exposure. The observed sensitivity to talker-specific VOT in voice processing neural regions is striking because the talkers' voices differed on a host of traditional indexical properties (e.g., fundamental frequency) in addition to their characteristic difference in VOT production, suggesting that talker-specific phonetic structure can be exploited for voice processing.

Here we test this hypothesis directly. In two experiments, two groups of listeners completed a training phase where they heard /g/- and /k/-initial words produced by three female speakers and learned to associate each voice with a cartoon pseudo-face. For one group, there was a structured relationship between VOT and talker, but for the other group, no talker-specific structure was provided. For both groups, the talkers' voices differed with respect to traditional indexical properties and thus sensitivity to phonetic variation was not required to perform the talker identification task (cf. Francis & Driscoll, 2006). After training, both groups completed a talker identification test phase for trained and novel words. The duration of the training phase was very brief (Experiment 1) or relatively longer (Experiment 2). If listeners can in principle use structured phonetic variation to facilitate voice processing over and above the benefit of traditional indexical cues, then we would expect to observe heightened talker identification at test for listeners in the structured compared to the unstructured training group.

2. Experiment 1

2.1 Participants and stimuli

Forty monolingual speakers of American English (mean = 20 years, SD = 2 years, 28 women, 12 men) were recruited from the University of Connecticut community. No participant had a history of speech, language, or hearing disorder per self-report. All participants passed a hearing screen administered at 25 dB for octave frequencies between 500 and 4000 Hz. Listeners received partial course credit or monetary compensation ($5) for their participation and were randomly assigned to either the structured (n = 20) or unstructured (n = 20) exposure condition.

Stimuli consisted of single-word utterances produced by three female speakers of American English with perceptually distinct voices. Stimuli were drawn from four VOT continua (*goal*-*coal*, *gain*-*cane*, *bowl*-*pole*, *bane*-*pain*) that were created for each talker following methods outlined in Allen and Miller (2004); word duration was equivalent across continua and talkers (ranging between 501 and 511 ms). For each talker and each voiced endpoint (i.e., *goal*, *gain*, *bowl*, *bane*), a VOT continuum was created based on the voiced endpoint by successively changing voiced cycles to voiceless cycles using a speech synthesizer (ASL, KayPENTAX, Montvale, NJ), increasing VOT by 4-5 ms with each iteration of the synthesis procedure. This procedure yielded continua that perceptually ranged from voiced to voiceless minimal pairs (e.g., *goal*-*coal*), with many VOT variants cueing each member of the pair.

As shown in Fig. 1, tokens from these continua were selected to form three sets, two for use during training (i.e., structured and unstructured exposure groups) and one for use during test. Both the structured and unstructured training sets contained 72 tokens drawn from the *goal*-*coal* and *gain*-*cane* continua that included six repetitions of each voiced-initial word (6 repetitions X 2 voiced-initial words X 3 talkers = 36 voiced-initial items) in addition to 36 voiceless-initial items. The same voiced-initial items were used in both the structured and unstructured sets, and consisted the voiced endpoints of each continuum; VOTs were equivalent across talker and word (ranging between 35 and 40 ms). For the structured set, the voiceless-initial items consisted of three repetitions of two VOT variants for each word and each talker (3 repetitions X 2 VOT variants X 2 words X 3 talkers = 36 voiceless-initial items). The VOT variants were selected so that each talker had a characteristic VOT, with talker 1 producing VOTs of 75 and 85 ms, talker 2 producing VOTs of 115 and 125 ms, and talker 3 producing VOTs of 155 and 165 ms. These ranges span the range of VOTs observed in the literature for American English stops (e.g., Theodore et al., 2009). For the unstructured set, the voiceless-initial items consisted of one repetition of six VOT variants for each talker, corresponding to the VOTs of 75, 85, 115, 125, 155, and 165 ms (1 repetition X 6 VOT variants X 2 words X 3 talkers = 36 voiceless-initial items). Accordingly, both the structured and unstructured training sets contained equal numbers of voiced- and voiceless-initial items, and there were equal numbers of each voiceless-initial VOT variant. The critical difference between the two training sets is that a talker-specific structure for voiceless-initial VOTs was present in the structured but not the unstructured training sets.

The test set was identical for the two exposure groups and contained the four words used during training (*goal*, *gain*, *coal*, *cane*) and four novel words (*bowl*, *bane*, *pole*, *pain*) for each talker (3 talkers X 2 repetitions X 8 words = 48 test tokens). The voiced-initial tokens (*goal*, *gain*, *bowl*, *bane*) were the voiced endpoints of each continuum; as for the *goal* and *gain* tokens, VOTs for the *bowl* and *bane* tokens were equivalent across talker and word (ranging between 15 and 20 ms). The voiceless-initial tokens (*coal*, *cane*, *pole*, *pain*) included the VOTs intermediate to those used in the structured exposure set (talker 1 = 80 ms, talker 2 = 120 ms, talker 3 = 160 ms) and corresponding VOT tokens from the *bowl-pole* and *bane-pain* continua (talker 1 = 60 ms, talker 2 = 100 ms, talker 3 = 140 ms). The shorter VOTs of the labial compared to the velar tokens are consistent with how place of articulation influences VOT (Lisker & Abramson, 1964).

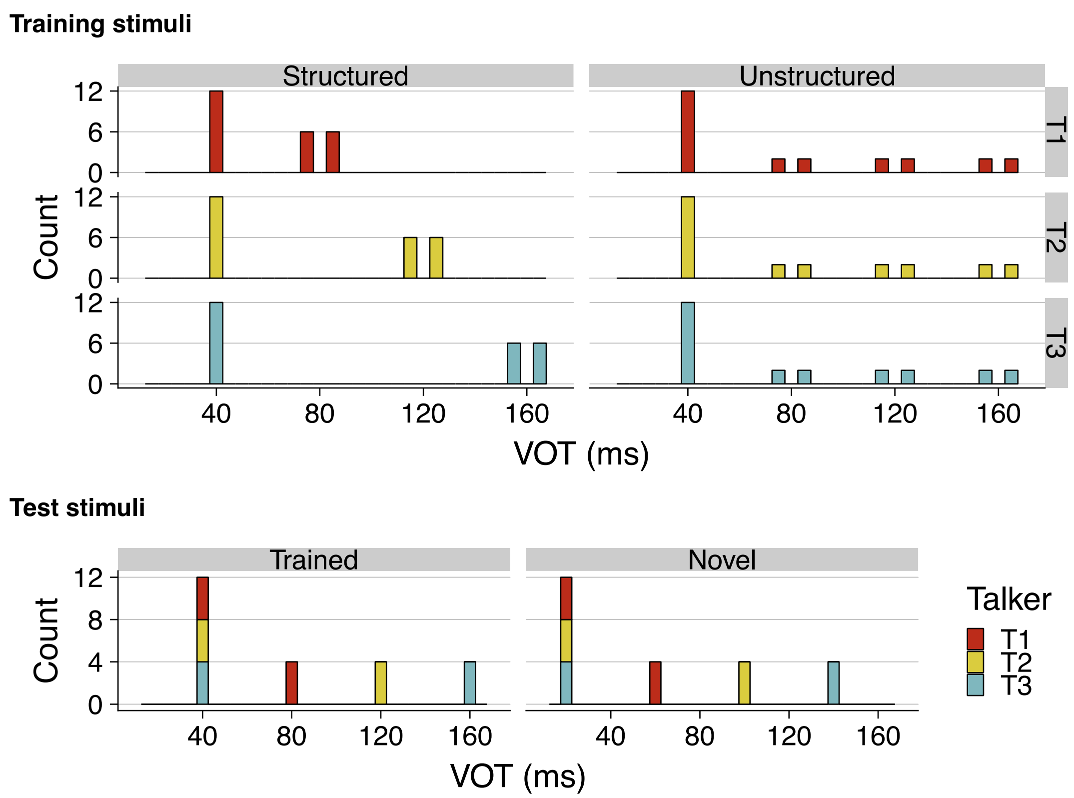

Fig. 1 (Color online) The top panel shows histograms of VOTs presented during training for the structured and unstructured exposure conditions. For the structured exposure condition, each talker (i.e., T1, T2, T3) shows a characteristic VOT production. For the unstructured exposure condition, there is no characteristic relationship between talker and VOT. The bottom panel shows histograms of VOTs presented at test for the trained and novel words; the same test stimuli were used for both exposure groups. For illustration purposes, voiced tokens are plotted as 40 ms VOT (the trained, velar-initial words) or 20 ms VOT (the novel, labial-initial words); as described in the main text, the exact VOTs of the voiced-initial words were within 5 ms of these values.

2.2 Procedure

All testing was completed in a sound-attenuated booth. Auditory stimuli were presented via headphones at a comfortable listening level held constant across participants. Participants completed three phases: familiarization, training, and test. Familiarization consisted of 12 trials (2 repetitions X 2 words X 3 talkers) using the (voiced-initial) *goal* and *gain* tokens that were selected for the training (and test) phases. On a single trial, the auditory stimulus was presented along with the cartoon pseudo-face. Participants were told, “Your job is to listen, look, and try to remember what that voice sounds like." No responses were collected during familiarization. The training phase was of fixed length, consisting of one randomization of the 72 items appropriate for the specific exposure group (Fig. 1). On each trial, an auditory stimulus was presented simultaneously with a visual array of three cartoon pseudo-faces. Participants were directed to select the cartoon associated with the talker's voice by pressing an appropriately labeled button on the response box. Feedback was provided in the form of "Yes!" for correct responses and "No." for incorrect responses. Trials were separated by an ISI of 2000 ms. The test phase consisted of one randomization of the 48 test stimuli. The procedure was identical to that during training except that no feedback was provided during test. Participants were given a brief break between the training and test phases, and the entire session lasted approximately 15 minutes.

2.3 Results

The raw data and analysis script for all results presented in this manuscript can be retrieved at https://osf.io/jt37x/?view_only=d682f75915cb4ad4960688d695abcc35. Mean proportion correct talker identification responses for training and test is shown in Fig 2(a). It appears that both groups learned to identify the talkers, given above chance performance at both training and test, and that the magnitude of learning is comparable between conditions. For the training phase, trial-level responses (0 = incorrect, 1 = correct) were analyzed using a generalized linear mixed-effects model (GLMM) with the binomial response family specifying exposure as a fixed effect (structured = 1, unstructured = -1) and random intercepts by subject and talker, implemented using the lme4 package (Bates et al., 2014). The model showed no relationship between exposure condition and talker identification accuracy during training ($\hat{\beta} = -0.154$, $SE = 0.146$, $z = -1.052$, $p = 0.293$). For the test phase, trial-level responses (0 = incorrect, 1 = correct) were analyzed using a GLMM with the fixed effects of exposure group (structured = 1, unstructured =

-1), item type (trained = 1, novel = -1), and their interaction, in addition to random slopes by subject for item type and random intercepts by subject and talker. Accuracy was higher for trained compared to novel words ($\hat{\beta}$ = 0.210, SE = 0.066, z = 3.186, p = 0.001). There was no main effect of exposure condition ($\hat{\beta}$ = -0.023, SE = 0.154, z = -0.148, p = 0.883), nor an interaction between item type and exposure condition ($\hat{\beta}$ = 0.023, SE = 0.064, z = 0.358, p = 0.720).

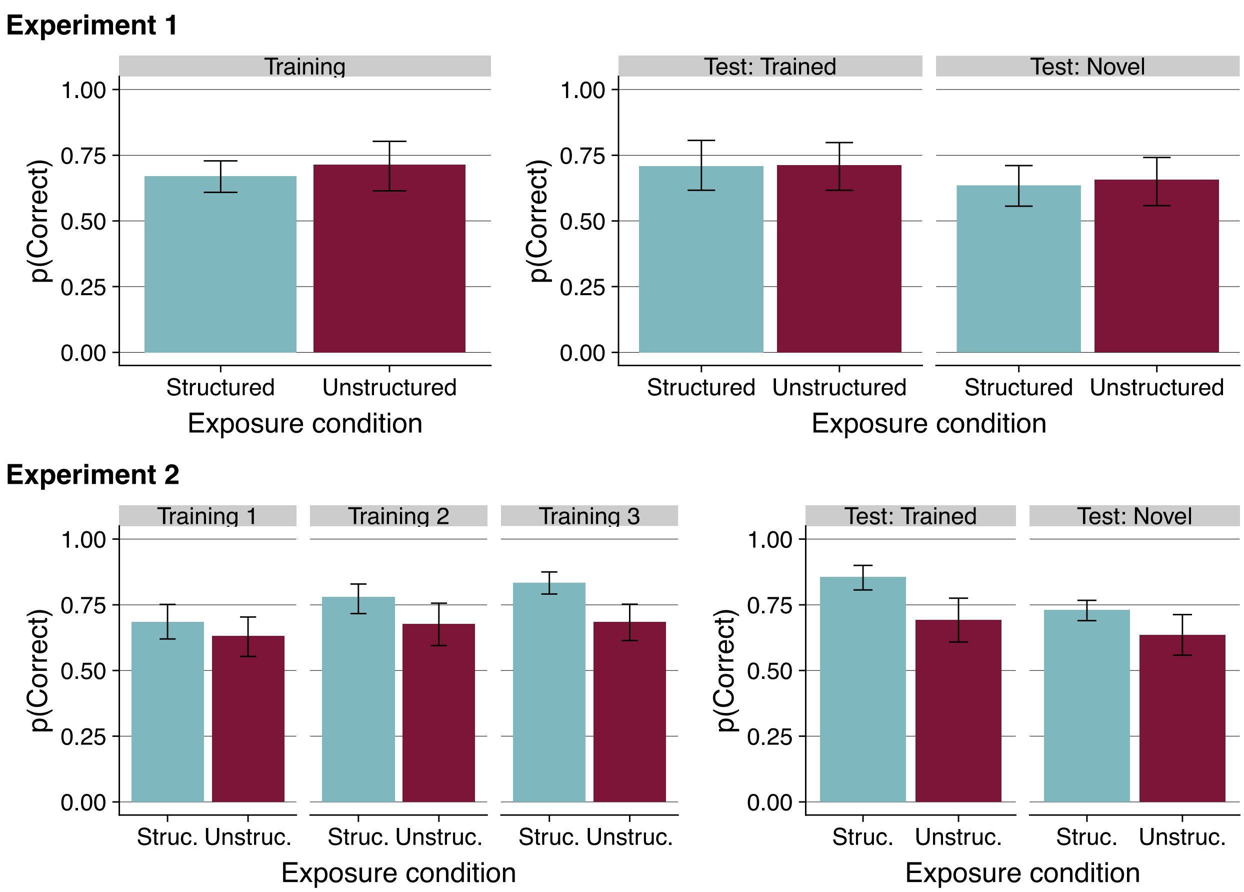

Fig. 2 (Color online) The top panel shows mean proportion correct talker identification for the structured and unstructured exposure groups during training (left) and test (right) for Experiment 1. The bottom panel shows mean proportion correct talker identification during training (left) and test (right) for the two exposure conditions in Experiment 2. Error bars indicate bootstrapped 95% confidence intervals calculated over by-subject means.

3. Experiment 2

In experiment 1, listeners successfully learned to identify voices with brief exposure to single-word productions; however, there was no additional benefit given exposure to structured versus unstructured phonetic variation. Experiment 2 tests whether a facilitative effect of structured phonetic variation on talker identification would emerge given a longer exposure period.

3.1 Methods

The participants were 40 monolingual speakers of American English (mean = 20 years, SD = 1 years, 26 women, 14 men) who did not participate in experiment 1 following the criteria outlined previously. Participants were compensated with partial course credit or $10. Listeners were randomly assigned to either the structured (n = 20) or unstructured (n = 20) exposure condition. The stimuli and procedure for experiment 2 were identical to those used in experiment 1 with one critical exception; instead of one block of training (72 trials), listeners completed exactly three blocks of training (216 trials). Each of the three training blocks was a unique randomization of the 72 training items appropriate for each exposure condition, as described for experiment 1. The entire procedure lasted approximately 30 minutes.

3.2 Results

Performance during the training and test phases is shown in Fig. 2. Visual inspection suggests that compared to the unstructured group, the structured group showed (1) greater improvement over the three blocks of training and (2) improved talker recognition at test. Separate GLMMs were constructed for the training and test data, with trial-level accuracy (0 = incorrect, = correct) as the predicted value in each model. The training model contained fixed effects of condition (structured = 1, unstructured = -1) and block (treatment-coded with two contrasts; block 1 as the reference level in each), random slopes by subject for block, and random intercepts by subject and talker. The results showed a main effect of block for both the block 2 vs. block 1 contrast ($\hat{\beta}$ = 0.410, *SE* = 0.080, z = 5.139, $p < 0.001$) and the block 3 vs. block 1 contrast ($\hat{\beta}$ = 0.583, *SE* =

0.087, $z = 6.735$, $p < 0.001$), indicating that talker identification accuracy improved across the training blocks. There was no main effect of condition ($\hat{\beta} = -0.128$, $SE = 0.128$, $z = -1.005$, $p = 0.315$), nor an interaction between condition and block for the block 2 vs. block 1 contrast ($\hat{\beta} = 0.143$, $SE = 0.078$, $z = 1.841$, $p = 0.066$). However, a robust interaction was observed between condition and block for the block 3 vs. block 1 contrast ($\hat{\beta} = 0.308$, $SE = 0.085$, $z = 3.614$, $p < 0.001$), indicating that those in the structured exposure group improved to a greater degree in block three compared to block one than those in the unstructured exposure group.

The test model contained the fixed effects of exposure condition (structured = 1, unstructured = -1), item type (trained = 1, novel = -1), and their interaction. Random effects included random slopes by subject for exposure and item type, and random intercepts by subject and talker. There was a main effect of exposure ($\hat{\beta} = 0.354$, $SE = 0.121$, $z = 2.932$, $p = 0.003$), with increased accuracy for the structured compared to the unstructured exposure group, a main effect of item type ($\hat{\beta} = 0.311$, $SE = 0.062$, $z = 5.044$, $p < 0.001$), with increased accuracy for trained compared to novel items, and an interaction between exposure and item type ($\hat{\beta} = 0.138$, $SE = 0.060$, $z = 2.320$, $p = 0.020$). Simple effects analyses showed that the item type effect was reliable for both the structured ($\hat{\beta} = 0.449$, $SE = 0.091$, $z = 4.921$, $p < 0.001$) and unstructured exposure groups ($\hat{\beta} = 0.173$, $SE = 0.080$, $z = 2.128$, $p = 0.030$), and that the exposure effect was robust for the trained words ($\hat{\beta} = 0.492$, $SE = 0.153$, $z = 3.210$, $p = 0.001$) but not for the novel words ($\hat{\beta} = 0.216$, $SE = 0.113$, $z = 1.917$, $p = 0.055$). Thus, the interaction observed in the full model can be attributed a greater difference between the structured and unstructured exposure groups for the trained compared to the novel items.

4. Conclusions

Here we examined whether listeners can use structured phonetic variation to facilitate voice processing. Given brief exposure to talkers' voices, access to structured phonetic variation did not show any additional benefit to talker identification beyond the traditional indexical cues (e.g., fundamental frequency) available to both exposure groups. However, given a more extended period of exposure, listeners who heard talkers produce characteristic VOTs showed improved talker identification compared to listeners who were not exposed to talker-specific phonetic variation. The facilitative effect of talker-specific phonetic variation resulted in an increased rate of learning across the exposure period and increased talker identification accuracy at test primarily for trained words, given the marginal influence of exposure condition on talker identification for novel words. Generalization of talker-specific VOT patterns to a novel place of articulation for talker identification would parallel patterns observed for phonetic processing (Theodore & Miller, 2010) and be consistent with findings showing that talker differences in VOT production are stable across place of articulation (Theodore et al., 2009); however, no robust evidence in support of generalization was observed in the current work.

Because the current paradigm provided feedback during training, it may have encouraged explicit learning of the mapping between VOT and talker; as a consequence, the incentive for learning this relationship (for trained items) might be exaggerated compared to more implicit learning paradigms. Though feedback was provided during training, the talkers' voices differed in traditional indexical properties (e.g., fundamental frequency) in addition to the phonetic manipulation, and thus sensitivity to talker-specific phonetic cues was not required in order to learn to identify the talkers' voices. This manipulation is in contrast to Francis and Driscoll (2006), where a difference in within-category VOT was the only cue available to distinguish talkers' voices. Thus, listeners can use talker-specific phonetic variation to facilitate talker identification not only when it is the only cue available (Francis & Driscoll, 2006; Remez et al., 1997), but also when it co-occurs with variation in fundamental frequency and vocal quality. In the current work, the facilitative influence of talker-specific phonetic variation on talker identification was only observed given the longer exposure period provided in experiment 2, suggesting that (1) listeners may require exposure in order to learn talker-specific phonetic structure on a time course that was present in experiment 2 but not in experiment 1, and/or (2) traditional indexical cues to voice identity may be weighted more heavily during initial exposure compared to phonetic cues. One avenue for future research is to examine whether nonnative listeners receive the same benefit for structured phonetic variation as the native listeners tested here; doing so would shed light on potential mechanisms that contribute to the native language benefit for talker identification. Specifically, it may be the case that when perceiving speech in the native language, listeners can use their knowledge of the linguistic sound structure to help parse phonetic variation in the input as a language-general cue versus a talker-specific cue (Kleinschmidt & Jaeger, 2015), but in the absence of expertise with linguistic sound structure, the listener may not be able to determine which aspects of the phonetic stream are licensed by the phonological system versus being attributable to a talker's idiolect (Perrachione & Wong, 2007).

To conclude, it is well established that there are tight, bi-directional influences between the phonetic processing and indexical processing mechanisms, which are observed behaviorally (Creel & Bregman, 2011; Nygaard & Pisoni, 1998; Theodore & Miller, 2010) and in the neural response to speech input (e.g., Chandrasekaran et al., 2011; Knösche et al., 2002; Myers & Theodore, 2017; Tuninetti et al., 2017). The current results further demonstrate that listeners' sensitivity to talker differences in phonetic properties of speech is one aspect of representational knowledge that mediates the relationship between speech perception and voice processing.

Acknowledgments

Research was supported by a seed grant to R.M.T. from the Connecticut Institute for the Brain and Cognitive Sciences, a SHARE award to D.G. from the Office of Undergraduate Research at the University of Connecticut, and by National Science Foundation (BCS) grant 1827591 to R.M.T. Portions of these data were submitted for presentation at the 176th meeting of the Acoustical Society of America, Victoria, British Columbia.

References and links

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *The Journal of the Acoustical Society of America*, *113*(1), 544–552.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Grothendieck, G. (2014). Package 'lme4.' *R Foundation for Statistical Computing, Vienna*, *12*.

Chandrasekaran, B., Chan, A. H. D., & Wong, P. C. M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*(10), 2690–2700. https://doi.org/10.1162/jocn.2011.21631

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809.

Creel, S. C., & Bregman, M. R. (2011). How talker identity relates to language processing. *Language and Linguistics Compass*, *5*(5), 190–204.

Francis, A. L., & Driscoll, C. (2006). Training to use voice onset time as a cue to talker identification induces a left-ear/right-hemisphere processing advantage. *Brain and Language*, *98*(3), 310–318.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, *19*(5), 448–458.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111.

Hullebus, M. A., Tobin, S., & Gafos, A. (2018). Speaker-specific structure in German voiceless stop voice onset times. In *Proceedings of Interspeech* (pp. 1403–1407).

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Knösche, T. R., Lattner, S., Maess, B., Schauer, M., & Friederici, A. D. (2002). Early parallel processing of auditory word and voice information. *NeuroImage*, *17*(3), 1493–1503.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, *165*, 33–44. https://doi.org/10.1016/j.bandl.2016.11.001

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *The Journal of the Acoustical Society of America*, *109*(3), 1181–1196.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Attention, Perception, & Psychophysics*, *60*(3), 355–376.

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, *143*, 36–40.

Perrachione, T. K., & Wong, P. C. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia*, *45*(8), 1899–1910.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *The Journal of the Acoustical Society of America*, *128*(4), 2090–2099.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982. https://doi.org/10.1121/1.3106131

Theodore, R. M., Myers, E. B., & Lomibao, J. A. (2015). Talker-specific influences on phonetic category structure. *The Journal of the Acoustical Society of America*, *138*(2), 1068–1078.

Tuninetti, A., Chládková, K., Peter, V., Schiller, N. O., & Escudero, P. (2017). When speaker identity is unavoidable: neural processing of speaker identity cues in natural speech. *Brain and Language*, *174*, 42–49.