Attention, Perception, & Psychophysics (2020) 82:2237–2243
https://doi.org/10.3758/s13414-020-01971-x

SHORT REPORT

Talker normalization is mediated by structured indexical information

Christian E. Stilp[1] · Rachel M. Theodore[2,3]

Published online: 19 February 2020

Abstract

Speech perception is challenged by indexical variability. A litany of studies on talker normalization have demonstrated that hearing multiple talkers incurs processing costs (e.g., lower accuracy, increased response time) compared to hearing a single talker. However, when reframing these studies in terms of stimulus structure, it is evident that past tests of multiple-talker (i.e., low structure) and single-talker (i.e., high structure) conditions are not representative of the graded nature of indexical variation in the environment. Here we tested the hypothesis that processing costs incurred by multiple-talker conditions would abate given increased stimulus structure. We tested this hypothesis by manipulating the degree to which talkers' voices differed acoustically (Experiment 1) and also the frequency with which talkers' voices changed (Experiment 2) in multiple-talker conditions. Listeners performed a speeded classification task for words containing vowels that varied in acoustic-phonemic ambiguity. In Experiment 1, response times progressively decreased as acoustic variability among talkers' voices decreased. In Experiment 2, blocking talkers within mixed-talker conditions led to more similar response times among single-talker and multiple-talker conditions. Neither result interacted with acoustic-phonemic ambiguity of the target vowels. Thus, the results showed that indexical structure mediated the processing costs incurred by hearing different talkers. This is consistent with the Efficient Coding Hypothesis, which proposes that sensory and perceptual processing are facilitated by stimulus structure. Defining the roles and limits of stimulus structure on speech perception is an important direction for future research.

Keywords Speech perception · Categorization · Perceptual categorization and identification

Introduction

The world is far from random; instead, objects and events in the environment are highly structured. According to the Efficient Coding Hypothesis (Attneave, 1954; Barlow, 1961), sensory systems detect and exploit this structure in order to facilitate processing. Efficient coding has been extremely productive for understanding visual processing and perception (Field, 1987; Geisler, 2008; Olshausen & Field, 1996; Simoncelli, 2003), and recent applications to speech perception have been equally promising (Gervain & Geffen, 2019; Kluender, Stilp, & Kiefte, 2013; Kluender, Stilp, & Llanos, 2019; Stilp & Kluender, 2010). For example, the statistical structure of sentence contexts influences subsequent vowel categorization (Stilp & Assgari, 2019), as predicted by efficient coding.

While not originally conceived as such, studies of talker normalization (e.g., Bradlow, Nygaard, & Pisoni, 1999; Nygaard, Sommers, & Pisoni, 1995; Pisoni, 1997; Sommers, Nygaard, & Pisoni, 1994) reflect perceptual sensitivity to stimulus structure. In these studies, listeners perform a task (e.g., phoneme categorization, word identification, recognition memory) with stimuli spoken by a single talker or by multiple talkers. Listeners generally show higher accuracy and/or faster response time in single-talker compared to multiple-talker conditions. Hearing one talker (i.e., highly structured stimuli) facilitates speech perception, whereas hearing multiple talkers (i.e., minimally structured stimuli) incurs processing costs.

✉ Rachel M. Theodore
rachel.theodore@uconn.edu

[1] Department of Psychological and Brain Sciences, University of Louisville, 317 Life Sciences Building, Louisville, KY 40292, USA

[2] Department of Speech, Language, and Hearing Sciences, University of Connecticut, 2 Alethia Drive, Unit 1085, Storrs, CT 06269-1085, USA

[3] Connecticut Institute for the Brain and Cognitive Sciences, University of Connecticut, 337 Mansfield Road, Unit 1272, Storrs, CT 06269-1272, USA

Throughout this literature, talker normalization studies have provided fairly uniform tests of sensitivity to stimulus structure. Multiple-talker stimuli often consist of speech from a small number of men and women, with minimal acoustic details provided regarding the specific degree of indexical variation among talkers. This approach is sufficient to induce processing costs for multiple-talker compared to single-talker stimulus sets, but does not represent the graded nature of indexical structure present in the environment. Efficient coding makes a novel prediction for talker normalization: as stimulus structure increases, processing costs associated with multiple talkers should abate. Stimulus structure is a broad term that has been used in the psycholinguistic literature to refer to, for example, hierarchical variability within a phonetic category (Kleinschmidt & Jaeger, 2015) or within-talker acoustic predictability across phonetic categories (Chodroff & Wilson, 2017). Here we use this term, as it is used in the efficient coding literature, to encapsulate any type of predictability that may be encoded and subsequently used by listeners to facilitate perception. We tested this prediction by manipulating the degree to which talkers' voices differed acoustically (Experiment 1) and also how often talkers' voices changed (Experiment 2) in multiple-talker environments. Listeners' response times for word identification were used to quantify processing cost.

Methods

Participants

The participants were 72 monolingual speakers of American English (53 women, 19 men; mean age = 20 ± 2 years[1]; n = 36 in each experiment, no one completed both experiments). None had a history of speech, language, or hearing disorders according to self-report; all passed a hearing screen on the day of testing. Two additional participants were tested but excluded from the study due to failure to meet the accuracy criterion described below.

Stimuli and procedure

Tokens of the words *he'd*, *hoed*, and *who'd*, each produced by ten talkers, were drawn from the Hillenbrand corpus (Hillenbrand, Getty, Clark, & Wheeler, 1995). Words were selected to achieve low (/i/ - /o/) and high (/o/ - /u/) acoustic-phonemic ambiguity in the vowel contrasts, as in Choi, Hu, and Perrachione (2018). The talkers formed three levels of talker variability: single talker (one man or one woman; each heard by half of the participants), mixed talker with low fundamental frequency (F_0) variability (two men, two women), and mixed talker with high F_0 variability (two men, two women). The selected talkers showed consistent F_0 across tokens (i.e., within 30 Hz). In multiple-talker conditions, four talkers were selected so that F_0 was either minimally (low variability) or maximally (high variability) different across talkers, as shown in Table 1. Mean word duration (625 ± 81 ms) did not vary across talker variability conditions ($F(2, 27) = 0.271$, $p = 0.765$) or target vowels ($F(2, 27) = 0.016$, $p = 0.985$). Intensity of the tokens was equated using Praat.

Talker variability (single, mixed low, mixed high) and acoustic-phonemic ambiguity (low, high) were both manipulated within-subjects, forming six blocks. Block order was counterbalanced across participants in each experiment. Each block consisted of 20 trials for each vowel (i.e., /i/ and /o/ for low ambiguity blocks, /o/ and /u/ for high ambiguity blocks). Following Choi et al. (2018), single-talker blocks tested 20 repetitions of each vowel; mixed-talker blocks tested five repetitions of each vowel from each of the four talkers. In Experiment 1, stimulus presentation within each block was randomized separately for each participant. In Experiment 2, mixed-talker trials were blocked to present ten consecutive trials from each of the four talkers. Talker order, held constant across ambiguity conditions and participants, was determined by maximizing the change in F_0 each time the talker (and, concomitantly, talker gender) changed. Within each talker blocking, order of the ten trials remained randomized. To illustrate this manipulation, Fig. 1 shows trial-by-trial F_0 in each block for one participant in each experiment; F_0 is highly structured for the mixed-talker conditions in Experiment 2 compared to Experiment 1.

Table 1 Mean fundamental frequency (Hz) of the tokens for the three talker variability conditions and the two acoustic-phonemic ambiguity conditions; talker identifiers correspond to those used in the Hillenbrand corpus (Hillenbrand et al., 1995)

Talker variability		Acoustic-phonemic ambiguity			
		Low		High	
Condition	Talker	/i/	/o/	/o/	/u/
Single	w16	237	228	228	238
	m18	133	135	135	139
Mixed Low	m01	173	175	175	181
	m45	215	202	202	212
	w26	194	191	191	189
	w49	206	189	189	201
Mixed High	m03	94	98	98	102
	m11	160	158	158	159
	w11	216	206	206	237
	w33	287	264	264	291

[1] Here and throughout, variability values reported in the main text indicate standard deviation.

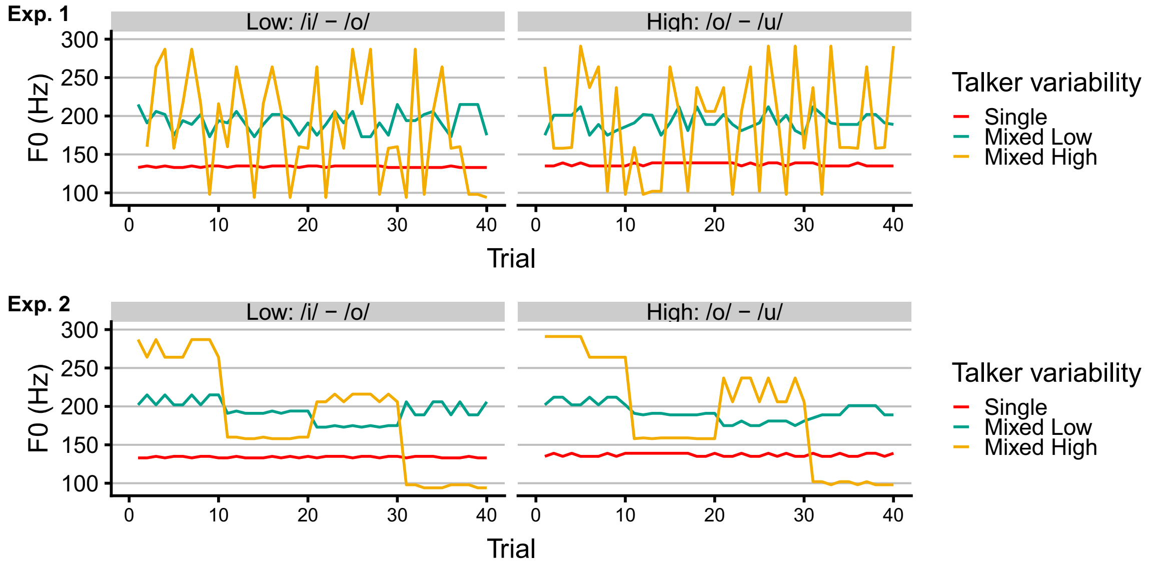

Fig. 1 Trial-by-trial F_0 in each condition for a representative participant in Experiment 1 (**top**, subject E1.001) and Experiment 2 (**bottom**, subject E2.001). Talker order was m45, w26, m01, w49 for mixed-low talker variability and w33, m11, w11, m03 for mixed-high talker variability. Trial-level F_0 variability for mixed-talker conditions in Experiment 1, where trials were completely randomized within conditions, is increased relative to Experiment 2, where trials for mixed-talker conditions were blocked by talker

On each trial, participants were instructed to identify the word as quickly and accurately as possible. Participants responded by pressing a labeled button on a response box (Cedrus RB-740). A visual stimulus assigning a number to each word was displayed throughout each block. Trials were separated by 2,000 ms, timed from the participant's response. All testing was completed in a sound-attenuated booth. Auditory stimuli were presented via headphones (Sony MDR-7506) at a comfortable listening level that was held constant across participants. Stimulus presentation and response collection were controlled using SuperLab (version 4.5) running on a Mac OS X operating system.

Results: Experiment 1

High accuracy for word identification (≥ 0.95 proportion correct) was an inclusion criterion for this study; accordingly, accuracy across participants was near ceiling (mean = 0.99 ± 0.01). Incorrect trials were excluded from analysis. Response times (RTs, in milliseconds) were log-transformed, and trials exceeding three standard deviations from each participant's mean log RT were excluded (< 1% of trials). Figure 2 shows the mean RT in each condition for each participant, in addition to boxplots aggregating across participants. Trial-level log RTs were submitted to a linear mixed effects model using lme4 (Bates, Maechler, Bolker, & Walker, 2015) in R (R Development Core Team, 2016). The Satterthwaite approximation of degrees of freedom was used to evaluate statistical significance using the t distribution as implemented in lmerTest (Kuznetsova, Brockhoff, & Christensen, 2017). The model included fixed effects of talker variability, acoustic-phonemic ambiguity, and their interaction. Talker variability was treatment-coded with mixed-low variability as the reference level. Ambiguity was sum-coded (low = -0.5, high = 0.5). The random effects structure consisted of random intercepts by subject and random slopes by subject for talker variability, ambiguity, and their interaction. Estimated marginal means from the model are shown in Table 2.

Compared to the mixed-talker low variability condition (mean = 675 ± 144 ms)[2], RTs were faster in the single-talker condition (mean = 613 ± 131 ms; $\widehat{\beta}$ = -0.096, SE = 0.014, t = -6.699, $p < 0.001$) and slower in the mixed-talker high variability condition (mean = 703 ± 134 ms; $\widehat{\beta}$ = 0.045, SE = 0.014, t = 3.152, p = 0.003). The pairwise comparison between the mixed-talker high variability and single-talker conditions was tested for this model using the emmeans package (Lenth, 2019), which showed that RTs were significantly slower in the former compared to the latter ($\widehat{\beta}$ = -0.141, SE = 0.016, t = -9.124, $p < 0.001$). The model also showed a main effect of ambiguity ($\widehat{\beta}$ = 0.141, SE = 0.024, t = 6.006, $p < 0.001$), with faster RTs for

[2] In the main text, empirical means and their corresponding standard deviations are reported to describe the raw data. To calculate the empirical means, we first calculated by-subject means in the condition(s) of interest; thus, empirical means reflect grand means. Estimated marginal means for each of the six conditions of each experiment as derived from the linear mixed-effects models are shown in Table 2.

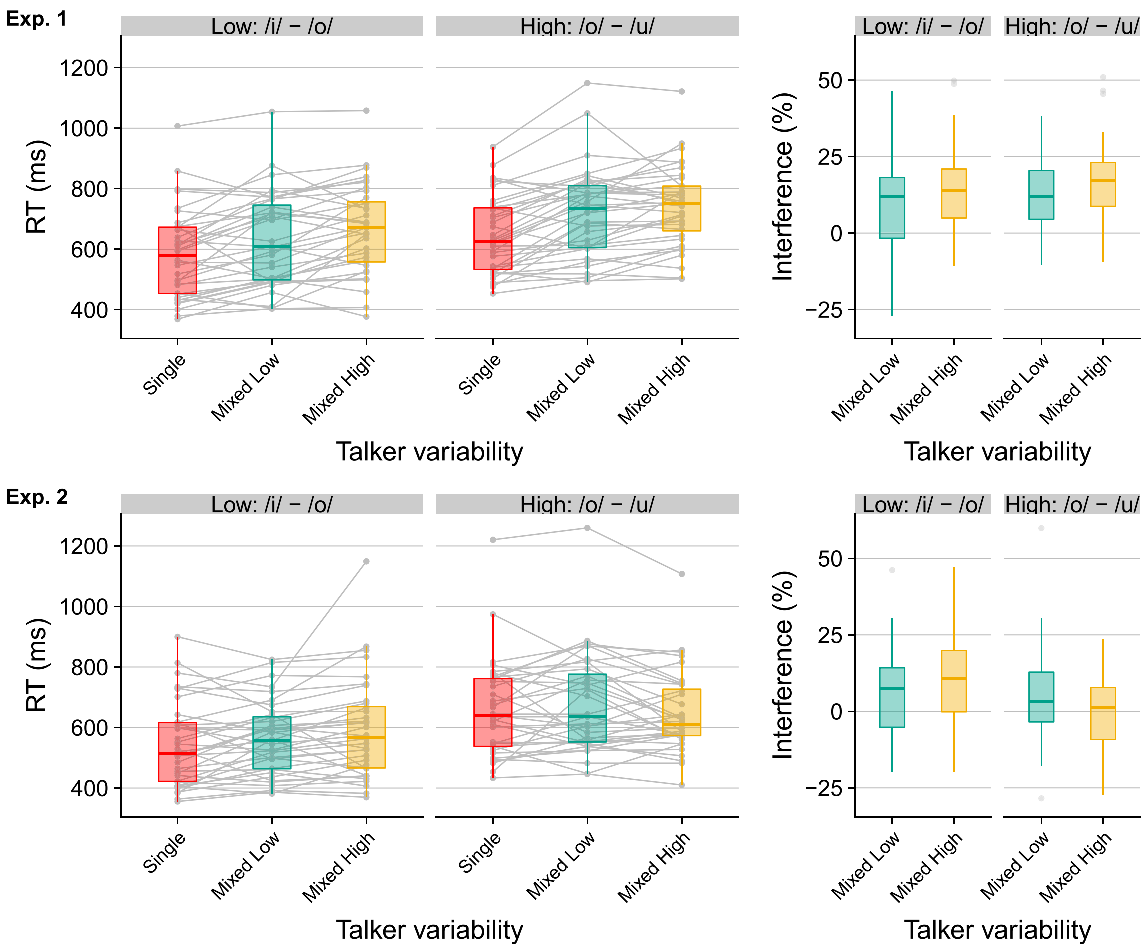

Fig. 2 Results from Experiment 1 (top) and Experiment 2 (bottom). At left is empirical mean response time (RT, in milliseconds) for each participant and boxplots aggregated across participants. At right are the empirical interference distributions across participants; interference was calculated as the difference between the mixed-talker conditions and the single-talker conditions, scaled to each participant's mean RT in the single-talker condition as follows: [(mixed – single / single) × 100]

the low- (mean = 626 ± 142 ms) compared to the high-ambiguity condition (mean = 702 ± 127 ms). There was no interaction between ambiguity and talker variability for either contrast ($\widehat{\beta}$ = -0.025, *SE* = 0.029, *t* = -0.864, *p* = 0.393 and $\widehat{\beta}$ = -0.028, *SE* = 0.027, *t* = -1.050, *p* = 0.301, respectively).[3]

Following Choi et al. (2018), interference effects of talker variability were calculated in each ambiguity condition (Fig. 2, right). Interference was calculated as the difference in mean RT between each mixed-talker condition and the single-talker condition, scaled to each participant's mean RT in the single-talker condition: [(mixed – single / single) × 100]. Consistent with the main effect of talker variability in the model, interference values were higher for the mixed-high variability condition compared to the mixed-low variability condition. The null interaction between variability and ambiguity reflects similar displacement of interference distributions across ambiguity conditions.

Table 2 For each experiment, estimated marginal means (in milliseconds) and corresponding 95% confidence interval (in parentheses) for the single, mixed low, and mixed high talker variability conditions for each of the acoustic-phonemic ambiguity conditions. Estimated marginal means were derived for the linear mixed effects models described in the main text using the emmeans package in R

Experiment	Talker variability	Acoustic-phonemic ambiguity	
		Low	High
1	Single	550 (506–598)	618 (579–660)
	Mixed Low	598 (551–649)	689 (644–737)
	Mixed High	634 (588–685)	710 (670–754)
2	Single	512 (471–557)	621 (576–669)
	Mixed Low	535 (498–575)	637 (590–688)
	Mixed High	556 (510–606)	613 (574–653)

[3] The analysis deviated from the preregistration in one way. Specifically, the mixed-talker low variability condition was used as the reference level instead of the single-talker condition. This is because we were wrong in the preregistration; in order to test for a monotonic change across talker variability, the reference level needs to be set to the intermediate condition.

Results: Experiment 2

Mean accuracy across participants was near ceiling (mean = 0.99 ± 0.01). Incorrect trials were excluded, RTs were log-transformed, and trials exceeding three standard deviations from each participant's mean log RT were excluded (1% of trials). Figure 2 (bottom) shows mean RT in each condition for each participant and boxplots aggregating across participants. Trial-level log RTs were submitted to a linear mixed effects model as described for Experiment 1; estimated marginal means from the model are shown in Table 2.

Compared to the mixed-talker low variability condition (mean = 619 ± 135 ms), RT was numerically but not statistically faster in the single-talker condition (mean = 600 ± 140 ms; $\widehat{\beta}$ = -0.034, SE = 0.018, t = -1.988, p = 0.055) and comparable to the mixed-talker high¯variability condition (mean = 618 ± 143 ms; $\widehat{\beta}$ = -0.001, SE = 0.015, t = -0.042, p = 0.9671). Pairwise comparison of the estimated marginal means for the model showed slower RTs in the mixed-talker high variability condition compared to the single-talker condition ($\widehat{\beta}$ = -0.034, SE = 0.014, t = -2.503, p = 0.044). The model showed a main effect of ambiguity ($\widehat{\beta}$ = 0.175, SE = 0.021, t = 8.260, p < 0.0001), confirming faster RTs for the low- (mean = 566 ± 134 ms) compared to the high-ambiguity contrast (mean = 659 ± 144 ms). Ambiguity did not interact with talker variability for the single versus mixed-low contrast ($\widehat{\beta}$ = 0.018, SE = 0.032, t = 0.547, p = 0.588), but did for the mixed-low versus mixed-high contrast ($\widehat{\beta}$ = -0.078, SE = 0.029, t = -2.682, p = 0.011).

To examine the nature of the interaction, the emmeans package was used to test pairwise comparisons in the model using the Tukey method to adjust for multiple comparisons. RTs were slower in the high- compared to the low-ambiguity condition for the single-talker ($\widehat{\beta}$ = -0.193, SE = 0.025, t = -7.807, p < 0.0001), mixed-low talker variability ($\widehat{\beta}$ = -0.175, SE = 0.021, t = -8.260, p < 0.0001), and mixed-high talker variability conditions ($\widehat{\beta}$ = -0.097, SE = 0.026, t = -3.734, p < 0.001). In the low acoustic-phonemic ambiguity condition, there was no reliable difference in RT between the single-talker and mixed-low talker variability conditions ($\widehat{\beta}$ = 0.044, SE = 0.023, t = 1.890, p = 0.157), nor between the mixed-low and mixed-high talker variability conditions ($\widehat{\beta}$ = -0.038, SE = 0.021, t = -1.858, p =0.166); however, RTs were slower in the mixed-high talker variability condition compared to the single-talker condition ($\widehat{\beta}$ = -0.082, SE = 0.024, t = -3.361, p = 0.005). In the high acoustic-phonemic ambiguity condition, there was no reliable difference between any of the talker variability conditions (single-talker vs. mixed-low talker variability: $\widehat{\beta}$ = 0.026, SE = 0.024, t = 1.063, p = 0.543; single-talker vs. mixed-high talker variability: $\widehat{\beta}$ = 0.014, SE = 0.021, t = 0.653, p = 0.792; mixed-low talker variability vs. mixed-high talker variability: $\widehat{\beta}$ = 0.040, SE = 0.022, t = 1.818, p = 0.179). Thus, the interaction observed in the full model reflects slower RTs for the mixed-talker high variability condition in the low but not the high acoustic-phonemic ambiguity condition.

Compared to Experiment 1, adding trial-level indexical structure in Experiment 2 attenuated the processing cost associated with talker variability. As shown in Fig. 2 (right), the interquartile range for three of the four interference distributions in Experiment 1 does not include zero (which would indicate no interference compared to the single-talker condition). In contrast, the interquartile range for three of the four interference distributions in Experiment 2 does include zero. This interaction between talker variability and experiment was directly tested in a linear mixed effects model following the structure outlined previously with the addition of experiment as a fixed effect (sum-coded, experiment 1 = -0.5, experiment 2 = 0.5). RTs were numerically but not significantly slower in Experiment 1 (mean = 663 ± 132 ms) compared to Experiment 2 (mean = 612 ± 136 ms; $\widehat{\beta}$ = -0.095, SE = 0.050, t = -1.902, p = 0.061). There was no interaction between experiment and acoustic-phonemic ambiguity ($\widehat{\beta}$ = 0.034, SE = 0.032, t = 1.069, p = 0.289). However, the interaction between experiment and talker variability was reliable for both the mixed-talker low versus single-talker contrast ($\widehat{\beta}$ = 0.061, SE = 0.023, t = 2.716, p = 0.008) and the mixed-talker low versus mixed-talker high contrast ($\widehat{\beta}$ = -0.046, SE = 0.021, t = -2.166, p = 0.034). Simple slope analyses showed no change in RTs across experiments for the single-talker condition ($\widehat{\beta}$ = -0.033, SE = 0.051, t = -0.656, p = 0.514), a numerical but not significant decrease for mixed-talker low variability ($\widehat{\beta}$ = -0.095, SE = 0.050, t = -1.904, p = 0.061), and significant decrease for mixed-talker high variability ($\widehat{\beta}$ = -0.140, SE = 0.048, t = -2.909, p = 0.005).

Discussion

Stimulus structure can mediate the processing costs incurred when hearing multiple talkers. In Experiment 1, processing time for mixed-talker blocks decreased as variability in F_0 decreased, decreasing further still for single-talker blocks. Past studies of talker normalization have generally focused on the presence/absence of such processing costs; here we reveal that these costs are graded. In Experiment 2, blocking talkers within mixed-talker conditions further attenuated processing costs, leading to performance that was more similar among talker variability conditions. Across experiments, the perceptual benefits of stimulus structure interacted with each other. Blocking talkers within mixed-talker conditions only made responses faster for stimuli with the greatest indexical variability. These results support an efficient coding approach to talker normalization, as speech perception amidst indexical variability was increasingly facilitated by trial-level stimulus structure.

Talker normalization depends on acoustic characteristics of talkers' voices. This point was first raised by Goldinger (1996), who reported a correlation between similarity ratings and perception of words spoken by different talkers. As acoustic differences across talkers (defined in large part by F_0) increased, RT increased and word recall accuracy decreased. Talker acoustics show graded influence for other aspects of speech processing, including spectral context effects (Ladefoged & Broadbent, 1957). Context effects are attenuated when talkers' fundamental frequencies are highly variable, and are in some cases equivalent to single-talker conditions when F_0 is minimally variable (Assgari & Stilp, 2015; Assgari, Theodore, & Stilp, 2019). Acoustic similarity also affects perception of both phonetic and indexical properties of the speech signal, which are interdependent (Mullennix & Pisoni, 1990). Choi et al. (2018) found that the processing costs incurred by hearing multiple talkers increased when phonetic properties were more similar. This pattern was not observed in the current work, suggesting that structured indexical variation can diminish potentially additive influences of phonetic and indexical variability.

Here, talker normalization is revealed to be a graded process, but is it obligatory? Previous work has suggested that this is in fact the case, given that talker variability challenged categorization of even acoustically unambiguous phonemes (Choi et al., 2018). In Experiment 1, processing costs were observed for talkers who minimally differed in F_0; however, the interference distributions observed in Experiment 2 suggest that stimulus structure may be sufficiently great to eliminate these costs entirely. Indeed, the results of Experiment 2 showed no difference in processing time among the three talker variability conditions for the high ambiguity contrast, nor between the single and mixed-talker low variability conditions for the low-ambiguity contrast. These results provide an existence proof that some types of trial-level stimulus structure *can* sufficiently eliminate the processing cost associated with mixed-talker input, at least when measured at the block level as is standard in the talker normalization literature. Importantly, the current manipulations of structured indexical variation reflect only two of the many ways that structure may be provided, and not all are expected to benefit perception equally. For example, context phrases rich in spectro-temporal information provided no more resilience to talker variability than a single neutral vowel matched in duration (Choi & Perrachione, 2019). Defining the limits of perceptual facilitation resulting from stimulus structure is an important direction for future research.

Bottom-up and top-down influences combine to shape speech perception (Davis & Johnsrude, 2007; McClelland, Mirman, & Holt, 2006; Sohoglu, Peelle, Carlyon, & Davis, 2012). Here, indexical structure was an important bottom-up influence when hearing different talkers. Previous studies have reported various higher-level influences on talker normalization including instructions and/or expectations (Johnson, Strand, & D'Imperio, 1999; Magnuson & Nusbaum, 2007), previous experience (Nygaard, Sommers, & Pisoni, 1994), and attention (Nusbaum & Morin, 1992). Defining the relative contributions of lower-level stimulus structure and higher-level factors for perception of different talkers' speech will be highly illuminating.

Acknowledgements This research was supported by a seed grant to R.M.T. from the Connecticut Institute for the Brain and Cognitive Sciences and by National Science Foundation (BCS) grant 1827591 to R.M.T.

Open Practices Statement The data and materials for all experiments are available at https://osf.io/mcfg4/. Experiment 1 was preregistered (https://osf.io/2ew68).

References

Assgari, A. A., & Stilp, C. E. (2015). Talker information influences spectral contrast effects in speech categorization. *Journal of the Acoustical Society of America*, *138*(5), 3023–3032.

Assgari, A. A., Theodore, R. M., & Stilp, C. E. (2019). Variability in talkers' fundamental frequencies shapes context effects in speech perception. *Journal of the Acoustical Society of America*, *145*(3), 1443–1454.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183–193.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith (Ed.), *Sensory communication* (pp. 53–85). Cambridge, Mass.: MIT Press.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Attention, Perception, & Psychophysics*, *61*(2), 206–219.

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: Covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47.

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention Perception & Psychophysics*, *80*(3), 784–797.

Choi, J. Y., & Perrachione, T. K. (2019). Time and information in perceptual adaptation to speech. *Cognition*, *192*, 1–14.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research*, *229*(1–2), 132–147.

Field, D. J. (1987). Relations between the statistics of natural images and the response properties of cortical cells. *Journal of the Optical Society of America A*, *4*(12), 2379–2394.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Reviews in Psychology*, *59*, 167–192.

Gervain, J., & Geffen, M. N. (2019). Efficient neural coding in auditory and speech perception. *Trends in Neurosciences*, *42*(1), 56–65.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1166–1183.

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

Johnson, K, Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*(4), 359–384.

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), 148–203.

Kluender, K. R., Stilp, C. E., & Kiefte, M. (2013). Perception of vowel sounds within a biologically realistic model of efficient coding. In G. S. Morrison & P. F. Assmann (Eds.), *Vowel Inherent Spectral Change* (pp. 117–151). Springer Berlin.

Kluender, K. R., Stilp, C. E., & Llanos, F. (2019). Longstanding problems in speech perception dissolve within an information-theoretic perspective. *Attention, Perception, & Psychophysics*.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. 10.18637/jss.v082.i13

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104.

Lenth, R. (2019). emmeans: Estimated marginal means, aka least-squares means. (Version R package version 1.3.4). Retrieved from https://CRAN.R-project.org/package=emmeans

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.

McClelland, J. L., Mirman, D., & Holt, L. L. (2006). Are there interactive processes in speech perception? *Trends in Cognitive Sciences*, *10*(8), 363–369.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.), *Speech Perception, Speech Production, and Linguistic Structure* (pp. 113–134). Tokyo: OHM.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Attention, Perception, & Psychophysics*, *57*(7), 989–1001.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In Keith Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (pp. 9–32). Burlington, MA: Morgan Kaufmann Publishers.

R Development Core Team. (2016). *"R: A language and environment for statistical computing."* Vienna, Austria: R Foundation for Statistical Computing.

Simoncelli, E. P. (2003). Vision and the statistics of the visual environment. *Current Opinion in Neurobiology*, *13*(2), 144–149.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience*, *32*(25), 8443–8453.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America*, *96*(3), 1314–1324.

Stilp, C.E., & Kluender, K. R. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(27), 12387–12392.

Stilp, C. E., & Assgari, A. A. (2019). Natural speech statistics shift phoneme categorization. *Attention, Perception, & Psychophysics, 81(6),* 2037–2052.