Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-020-02203-y

Talker familiarity and the accommodation of talker variability

James S. Magnuson [1] · Howard C. Nusbaum [2] · Reiko Akahane-Yamada [3] · David Saltzman [1]

Accepted: 5 November 2020

Abstract
A fundamental problem in speech perception is how (or whether) listeners accommodate variability in the way talkers produce speech. One view of the way listeners cope with this variability is that talker differences are *normalized* – a mapping between talker-specific characteristics and phonetic categories is computed such that speech is recognized in the context of the talker's vocal characteristics. Consistent with this view, listeners process speech more slowly when the talker changes randomly than when the talker remains constant. An alternative view is that speech perception is based on talker-specific auditory exemplars in memory clustered around linguistic categories that allow talker-independent perception. Consistent with this view, listeners become more efficient at talker-specific phonetic processing after voice identification training. We asked whether phonetic efficiency would increase with talker familiarity by testing listeners with extremely familiar talkers (family members), newly familiar talkers (based on laboratory training), and unfamiliar talkers. We also asked whether familiarity would reduce the need for normalization. As predicted, phonetic efficiency (word recognition in noise) increased with familiarity (unfamiliar < trained-on < family). However, we observed a constant processing cost for talker changes even for pairs of family members. We discuss how normalization and exemplar theories might account for these results, and constraints the results impose on theoretical accounts of phonetic constancy.

Keywords Psycholinguistics · Speech perception

Introduction

One of the most remarkable aspects of spoken language understanding is *phonetic constancy* despite multiple sources of acoustic variability, such as differences in acoustic environment, speaking rate, and talker characteristics. A talker's productions of a particular phoneme may vary depending on segmental context (Liberman, Delattre, & Cooper, 1952), prosodic context (Cutler, Dahan, & Donselaar, 1997; Fougeron & Keating, 1997), or discourse context (e.g., whether information is new or old; Fowler & Housum, 1987; Fowler, Levy, & Brown, 1997; Nooteboom & Kruyt, 1987). The problem is compounded by differences between talkers (e.g., in size or other physical aspects of the vocal tract, or in dialect), yielding a many-to-many mapping between acoustic patterns and percepts (Peterson & Barney, 1952). Most research on talker differences in speech perception has focused on how listeners might overcome the contribution of talker variability to the lack of invariance problem. For decades, theoretical accounts assumed that talker differences must be *normalized* – that is, that listeners must compute a mapping in real time using talker-specific characteristics and the relationship of acoustic patterns to phonetic categories (Gerstman, 1968; Ladefoged & Broadbent, 1957; Miller, 1989; Nearey, 1989; Nusbaum & Morin, 1992; Potter & Steinberg, 1950; Rakerd & Vebrugge, 1987; Strange, 1989; Syrdal & Gopal, 1986; Traunmuller, 1981).

The normalization hypothesis has been supported by numerous studies documenting a performance cost in speech perception associated with talker variability (e.g., slower and/or more errorful performance when stimuli are mixed across test trials from multiple talkers rather than when stimuli are blocked by talker: Choi, Hu, & Perrachione, 2018; Magnuson & Nusbaum, 2007; Martin, Mullennix, Pisoni, &

✉ James S. Magnuson
james.magnuson@uconn.edu

[1] Department of Psychological Sciences, and CT Institute for the Brain and Cognitive Sciences, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269-1020, USA

[2] Department of Psychology, University of Chicago, Chicago, IL, USA

[3] Advanced Telecommunications Research Institute International, Kyoto, Japan

Published online: 04 January 2021

Summers, 1989; Mullennix, Pisoni, & Martin, 1989; Nusbaum & Morin, 1992). This cost is often interpreted as reflecting the operation of a normalization mechanism that uses a talker's vocal characteristics to compute a scaled mapping between acoustic patterns and phonological categories, an interpretation supported by evidence showing increased demands on attention processes with talker changes (e.g., Wong et al., 2004). Moreover, some results are consistent with the idea (cf. Joos, 1948) that normalization yields a mapping that is applied to subsequent utterances in the absence of evidence for a talker change. For example, Ladefoged and Broadbent (1957; see Ladefoged, 1989, for a replication) found that identification of a target /bVt/ syllable frame (where V stands for *vowel*) could be shifted reliably depending on the talker characteristics in an immediately preceding carrier phrase (**"**Please say what this word is"): an item identified as *bit* following one synthetic talker's carrier phrase was identified as *bet* following another's, depending on the heights of the first and second formants (F1 and F2). The fact that a carrier phrase changes perception of a following target word suggests that listeners do not automatically encode the talker characteristics of each segment or syllable independently as has been suggested by some views of talker normalization. These theories (so-called *intrinsic normalization* or *structural estimation* theories) argue that each segment of speech contains sufficient acoustic information (e.g., from F0 and F3) to calibrate phonetic recognition to the speaker's vocal characteristics (e.g., Miller & Liberman, 1979; Shankweiler, Strange, & Verbrugge, 1977; Syrdal & Gopal, 1986; see Neary, 1989, or Nusbaum & Magnuson, 1997, for reviews).

Two related alternatives to talker normalization theories have been proposed. Proponents of *exemplar* models of speech perception (e.g., Johnson, 1994, 1997, 2005; Pierrehumbert, 2002) and proponents of *nonanalytic episodic* theories (Goldinger, 1998; Pisoni, 1997; Nygaard & Pisoni, 1998) have argued that similarity to stored traces of speech that statistically sample the space of talker characteristics would provide a sufficient basis for phonetic constancy without explicit normalization. Instead, on such views, an incoming speech sample activates acoustically similar episodes in memory, and the incoming speech is recognized based upon the set of activated traces. Recognition efficiency is predicted to be proportional to the number of similar stored exemplars. From this perspective, there is no difference (for the perceptual system) between acoustic information corresponding to phonetic structure and acoustic information corresponding to vocal characteristics, and both are encoded automatically as an integral whole as part of recognition. While exemplar theories allow for or posit explicitly analytic mechanisms (e.g., exemplars are based on the results of formant or feature extraction), nonanalytic theories posit that episodes are stored in memory as intact, unanalyzed wholes. Note that while there are some subtle differences between episodic and nonanalytic theories, they share the key assumption that phonetic constancy does not require normalization or accommodation of talker variability. We will group them together throughout this paper based on this salient similarity.

Episodic/nonanalytic theories reject normalization in light of evidence that listeners are sensitive to (presumably) phonetically irrelevant surface details of utterances. For example, several studies suggest there is a contingent relationship between phonetic and indexical information (e.g., talker identity, dialect, etc.) carried in speech, as listeners have difficulty ignoring irrelevant variability in either dimension (Mullennix & Pisoni, 1990), and training on talker identification facilitates phonetic perception of speech produced by trained-on talkers (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994), suggesting perceptual learning occurs for both types of information. There is also ample evidence that listeners encode talker-specific surface details and other "non-linguistic" characteristics of speech. Consistency in talker characteristics facilitates memory (e.g., Church & Schacter, 1994; Craik & Kirsner, 1974; Creelman, 1957; Goldinger, 1996; Goldinger, Pisoni, & Logan, 1991; Palmeri, Goldinger, & Pisoni, 1993; Schacter & Church, 1992; Sheffert & Fowler, 1995), as well as performance in speech-processing tasks such as shadowing (Goldinger, 1998; Pufahl & Samuel, 2014).

On such views, the problem of phonetic constancy has been classically misconstrued, and in fact may be simplified by considering regularities introduced by indexical and phonetic variation; that is, the acoustic correlates of talker variability are not random, and may *constrain*, per the notion of "lawful variability" (Elman & McClelland, 1986), rather than *complicate,* the problem of phonetic constancy (and vice versa; cf. Remez et al., 1997). In part, this rejection of normalization follows from an overgeneralization about normalization theories; some proponents of exemplar and nonanalytic theories claim that normalization must entail recoding in a talker-invariant, abstract phonetic code, and discarding of any non-phonetic surface detail. While some have made explicit claims of this sort, normalization does not logically depend on reductive abstraction that explicitly discards nonphonetic details (e.g., Joos, 1948, in perhaps the first detailed consideration of talker accommodation, proposed that the signal and internal categories must be brought into registration by warping either or both; see footnote 3 of Magnuson & Nusbaum, 2007). But how could exemplar/nonanalytic theories explain costs associated with talker changes?

On these views, if recent instances carry more weight than older ones, talker consistency will lead to faster performance, on the assumption that successive samples of speech will be more similar to each other than samples from different talkers (although this overlooks an important issue in the lack of invariance problem; two successive speech samples from one talker can be dramatically different acoustically). In

contrast, a talker change weakens activation of speech samples in long-term memory because of the reduced similarity between successive samples.

The emphasis in the exemplar view on contingent encoding of phonetic and indexical characteristics (Nygaard & Pisoni, 1998) and the importance of the number of exemplars in memory activated by any particular sample of speech (Goldinger, 1998) suggests an important issue for both the exemplar and normalization views: the impact of talker familiarity. Talkers become familiar because we hear them often. Presumably, this leads to many opportunities for learning (encoding and storing in memory) their phonetic and indexical characteristics and increases the number and/or strength of stored exemplars in memory. This predicts that speech perception should be generally more efficient for familiar talkers than unfamiliar talkers. Indeed, experience identifying talkers' identities provides sufficient familiarity with their productions that subsequently listeners are significantly better at identifying speech produced by those talkers than by unfamiliar talkers (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). High familiarity acquired outside the laboratory also provides a significant boost in one's ability to separate speech produced concurrently by two talkers (e.g., Johnsrude, Mackey, Hakyemez, Alexander, Trang, & Carlyon, 2013, found a significant advantage in a cocktail-party paradigm when one talker was the listener's spouse), and in identifying speech in adverse conditions (e.g., Souza, Gehani, Wright, & McCloy, 2013, found advantages for "frequent communication partners" – spouses or friends).

However, a normalization theory like *contextual tuning* (Joos, 1948; Ladefoged & Broadbent, 1957; Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997; Zhang & Chen, 2016) can make similar predictions given that a sample of a talker's speech provides information about the talker's vocal characteristics. If the operations required to map a talker's characteristics to phonetic categories are subject to perceptual learning, the operations for adjusting to a familiar talker's characteristics might become automatized. This would predict that adjusting to a change from one familiar talker to another ought to be easier than a change between two unfamiliar (and therefore unautomatized) talkers (Logan, 1988, 2002; Nygaard, Sommers, & Pisoni, 1994).

Figure 1 presents a schematic of contextual tuning based on Magnuson and Nusbaum (2007). As speech is processed continuously, samples are monitored for evidence of a talker change. When a change is detected, a process is triggered that computes the acoustic-perceptual mapping for the talker. Achieving a robust mapping may take many milliseconds of input. As processing continues, the mapping is continuously tuned based on bottom-up fit and top-down error checking (e.g., a parse of "fress orange" would suggest the need to adjust the /s/-/ʃ/ boundary). Given the recent report from Choi et al. (2018) that mixed-talker processing costs are

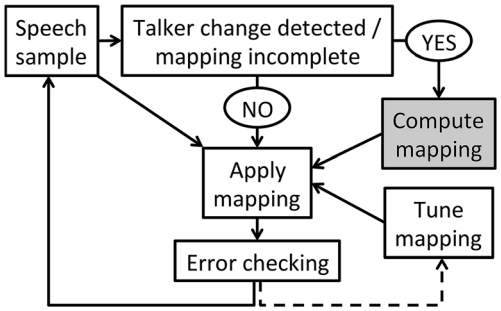

Fig. 1 Schematic of contextual tuning theory based on Nusbaum and Magnuson (1997). The schematic is intended to depict processes that happen in continuous cascade, not in strict serial order. Note the arrow from "Speech sample" to "Apply mapping," indicating that as tuning is attempted, the system continues to operate with the current estimate by default. The dashed arrow indicates the top-down (closed-loop) aspect of contextual tuning: error checking (e.g., difficulty in lexical or supralexical parsing) drives constant tuning. The gray box ("Compute mapping") is the resource-demanding component hypothesized to create relative slowdowns in mixed- versus blocked-talker conditions. Source: Magnuson (2018a)

observed even when stimuli are not phonetically ambiguous, we propose that detecting a talker change would also trigger reanalysis, even in the absence of any linguistic ambiguity. Although the diagram may imply serial stages, consider, for example, the arrow from "Speech sample" to "Apply mapping": even as a talker change is detected, the speech stream continues, and the system must apply in parallel the current mapping even as it is in the process of adjusting to a change in talker. Crucially, on this view, the process of *computing* the mapping after a talker change is particularly attentionally demanding, leading to the consistent slowing observed in mixed- versus blocked-talker conditions in simple tasks like word or syllable monitoring (Nusbaum & Morin, 1992).

If the mapping for familiar talkers could be stored in memory (whether in the form of a mapping per se or the *operations* [cf. Kolers, 1976; Kolers & Ostry, 1974] to appropriately warp inputs to align them with representations), perhaps the resource-demanding mapping computation could be bypassed, as schematized in Fig. 2. In the elaborated

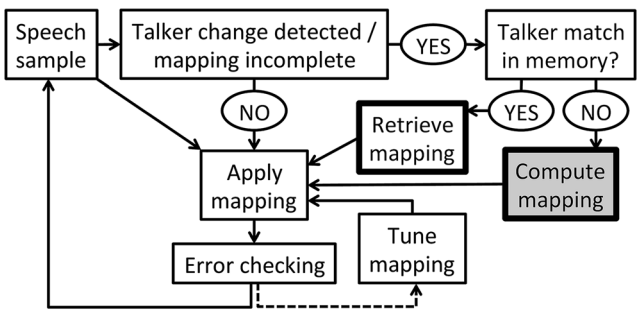

Fig. 2 The contextual tuning diagram elaborated to include a potential way for familiarity to avoid computing talker-percept mappings. If a familiar talker is (implicitly) detected, perhaps the mapping for that talker can be retrieved for memory, allowing the resource-demanding "Compute mapping" step to be skipped. Source: Magnuson (2018b)

schematic, detecting a change in characteristics triggers a search for a mapping for a known talker (or perhaps for a talker highly similar to the new talker; Zhang & Chen, 2016). If retrieving a mapping is less demanding than computing one, this would predict a reduced cost for a change to a familiar talker.

Thus, contextual tuning, exemplar, and episodic/nonanalytic accounts all potentially predict that talker-change costs may not be observed for highly familiar talkers (albeit for different reasons). However, if the processing cost typically found after a talker change reflects an obligatory process that monitors the acoustic-phonetic mapping from speech to perceptual categories, it may be that familiarity cannot affect this cost; before operations specific to familiar vocal characteristics can be brought to bear, the speech signal may need to be analyzed sufficiently to allow the familiar talker characteristics to be detected.

We examined in three experiments whether perceptual processing advantages found for recognizing the speech of familiar talkers (e.g., Nygaard & Pisoni, 1998) extend to adjusting to talker changes. In Experiment 1, we compared the costs of talker changes between familiar talkers (family members) with costs for changes between unfamiliar talkers. In Experiment 2, we tested the familiarity of family-member voices by comparing voice identification for family members' voices with voices learned in a single experimental session. In Experiment 3, in order to compare the effects of experimental training and long-term experience with voices on speech- (rather than voice-) identification performance, we asked subjects to transcribe morae presented in noise that were produced by talkers that were very familiar (family members), newly familiar talkers that subjects had been trained to identify, or talkers that subjects had heard but not been trained to identify.

Experiment 1

We designed Experiment 1 to test whether processing costs that typically accompany changes between unfamiliar talkers would be observed for highly familiar talkers. We used a speeded target-monitoring procedure (Nusbaum & Morin, 1992) in which listeners are shown a visual target (orthographic representation of phoneme or word) and monitor a sequence of utterances, pressing a key whenever they hear the target (with multiple instances of the target positioned within a pseudo-random sequence of spoken distractors). In Experiment 1, the speech consisted of morae (Japanese syllables) produced by familiar talkers (subjects' family members) and unfamiliar talkers. In a series of *blocked-talker* blocks of trials, all morae were produced by a single talker. In a series of *mixed-talker* blocks, the morae were produced by two talkers. In the mixed-talker condition, the two talkers could both be familiar, or one or both could be unfamiliar. This allowed us to examine whether a change in talker slows processing compared to blocked-talker conditions, even when the two talkers are highly familiar. Note that many studies use laboratory training to familiarize participants with talkers (with notable exceptions such as the long-time co-workers' voices used by Remez et al., 1997). By using extremely familiar voices – a participant's family members – we can exploit a level of familiarity much greater than could be achieved via brief laboratory training (although we compare family members to trained-on talkers in Experiment 2). Furthermore, we expect the minimal level of familiarity for family members to be higher than for co-workers or other potentially familiar adult voices. This should result in better speech recognition for these highly familiar family members.

Method

Materials We recorded two parents and one child from nine Japanese families reading lists of Japanese morae (consonant-vowel sequences, in the case of each of our items). The families were recruited from Kyoto and Nara prefectures in the area near ATR (Advanced Telecommunications Research Institute International, where the experiments were conducted). Children ranged in age from 7 to 12 years old. Adults and older children read a list of 100 morae. Younger children read a 45-item subset of the full list, but this included all items ultimately used for this experiment. The morae were recorded and simultaneously digitized at a sampling rate of 44.1 kHz and 16-bit resolution, and were later down-sampled to 22.05 kHz. Items were hand-edited to remove silence at the beginning and end of each utterance, and RMS amplitude was digitally normalized. Average mora duration was approximately 180 ms.

Participants Both adults from seven of the nine families we recorded participated in Experiment 1, as well as one adult from another family (the mother from family 3, "fam3-mom"). The father from one family (fam3-dad) and both adults from one family (family 5) were unable to participate in the experiment. One participant (fam1-mom) was excluded due to data recording errors. Thus, a total of 15 adults participated in Experiment 1, with data from 14 included for analyses. All of the subjects were native speakers of Japanese, and all reported having normal hearing and normal or corrected-to-normal vision, and no history of hearing or speech disorders. Sample size was similar to those in previous studies using speeded monitoring tasks to examine multi-talker processing costs (e.g., Magnuson & Nusbaum, 2007; Nusbaum & Morin, 1992).

Procedure We used a speeded target-monitoring task based on that described by Nusbaum and Morin (1992), and measured response times and accuracy. Subjects were presented with an

orthographic (hiragana) representation of a target mora on a computer display and were instructed to press a response button whenever they heard the mora they saw on the screen. Morae were presented on-line to subjects seated at workstations over STAX Lambda-SR-Signature headphones. See Fig. 3 for a schematic of the procedure.

On each trial, subjects heard a sequence of 16 morae with a stimulus-onset asynchrony between mora onsets of 830 ms. Trials were separated by 3,000 ms of silence, during which a message appeared on the screen to alert subjects that the target mora was changing. Four target morae were randomly positioned among 12 distractors, with the following constraints: targets could not be first or last in a trial, and targets had to be separated by at least one distractor. Four morae served as targets (/bo/, /gu/, /ki/, and /pa/) and 16 as distractors (/be/, /bu/, /ga/, /go/, /ji/, /ka/, /ko/, /me/, /mu/, /na/, /ni/, /pe/, /pi/, /ri/, /ro/, and /zo/). Target morae served as distractors when they were not the target.

Each subject listened to four talkers in the blocked-talker condition, in which all targets and distractors in each trial were produced by a single talker. The four talkers were a familiar adult (*Fa*, the subject's spouse), a familiar child (*Fc*, the subject's child), an unfamiliar adult (*Ua*), and an unfamiliar child (*Uc*). Half the subjects were assigned male unfamiliar talkers from one of the families, and half were assigned female unfamiliar talkers from another family. The same pair of unfamiliar talkers was assigned to husbands and wives from the same family. Therefore, there were equal numbers of female and male subjects listening to male and female unfamiliar talkers. Each subject also listened to six pairs of talkers in the mixed-talker condition, where half the targets and distractors were produced by each of two talkers and randomly ordered. The talker pairs were: *FaFc* (familiar), *UaUc* (unfamiliar), *FaUa, FaUc, FcUa*, and *FcUc* (crossed). Presentation order of blocked-talker and mixed-talker trials across subjects was controlled with a Latin square design. All manipulations were within subjects. There were eight trials per talker in the blocked-talker condition (with four targets per trial) and 16 trials per talker pair in the mixed-talker condition (such that there was the same number of trials [8] per talker in the blocked condition and within each mixed-condition pair).

Results

All analyses were conducted using R version 3.6.2 (R Core Team, 2019). Analyses of reaction time and accuracy for Experiment 1 were performed with linear mixed-effects models (with sum-coded categorical fixed effects), implemented with the R packages *afex* (Singmann, Bolker, Westfall, Aust, & Ben-Shachar, 2020) and *lme4* (Bates, Mächler, Bolker, & Walker, 2015). We first created all possible permutations of random effects structures (which *afex* treats as sum-coded by default), and then applied a backwards-stepping selection procedure to the models that converged, and selected the model with the maximal random effects structure that accounted for significantly more variance than the next-most maximal model (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). All fixed effects were set as factors and were sum coded. Mixed-effects models were fit using the *afex* command "*mixed,*" and the values from the resulting ANOVA table are reported for each model. P-values were estimated using the Satterthwaite (1946) method or a likelihood-ratio test (for logistic models with binomial outcomes).

Accuracy A generalized linear mixed-effects model with a logit link was fit to assess participants' accuracy in the word-monitoring task stimulus-by-stimulus. For a non-target, inhibiting a response was a correct rejection (scored as 1), whereas a response was a false alarm (scored as 0). The opposite was true for targets (responses were hits and coded as 1,

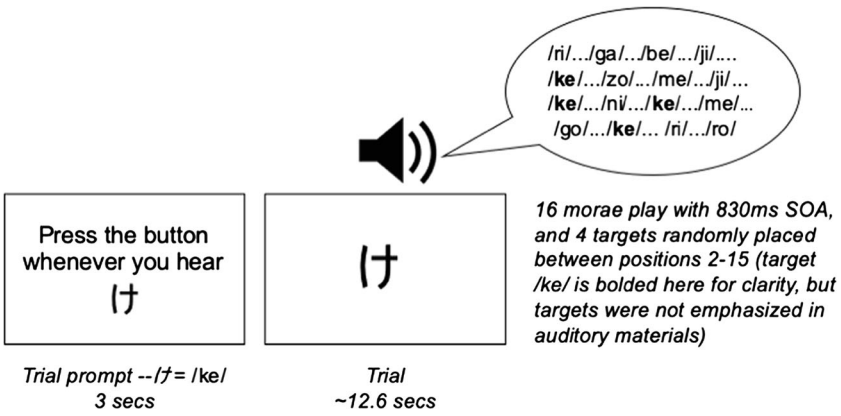

Fig. 3 Experimental procedure for Experiment 1. At the beginning of each trial, a message appeared for 3 s informing participants what mora to monitor for (messages were in Japanese). Then the target mora was displayed continuously throughout the trial as participants heard a sequence of 16 morae with four instances of the target (but never in first or last position). Source: Magnuson (2020)

while failures to respond were misses and coded as 0). The selected model contained fixed effects of Blocking (items Blocked by talker, or items from two talkers Mixed), talker Familiarity (Familiar vs. Unfamiliar), and talker Age (Adult vs. Child), and their interactions, as well as by-subject random slopes and intercepts for the three-way interaction. Accuracy on the talker-monitoring task was near ceiling (mean proportion correct = 0.990, *SD* = 0.10; see Table 1 for a summary by talker Familiarity, talker Age, and Blocking). No main effects or interactions were significant.

Reaction time For the monitoring task from Experiment 1, we began by selecting only correct responses and then removed any responses with reaction times (RTs) more than 3 standard deviations from the mean (which removed only 20 responses, or 0.3%). Summary data is presented in Table 2. We first fit generalized linear-mixed effects models to the reaction time data, and applied the model selection procedure described above. The selected model included fixed effects of (talker) Familiarity (Familiar vs. Unfamiliar), (talker) Age (Adult vs. Child), Blocking (Blocked or Mixed talker trials), and their interactions, along with by-subject random slopes and intercepts for the three-way interaction term. The model was applied to raw RTs with an inverse Gaussian distribution with identity link, as recommended by Lo and Andrews (2015). Central tendencies for RTs as a function of talker Age, talker Familiarity, and Blocking, along with distributions and individual data points, are presented in Fig. 4. Means for individual participants are shown in Fig. 5.

The main effect of Blocking was significant ($\chi^2 = 6.30$, $p = 0.012$), with slower responses overall for Mixed ($M = 361$ ms, $SD = 99$) versus Blocked trials ($M = 340$, $SD = 94$). The main effect of Age was also significant ($\chi^2 = 12.40$, $p < 0.001$), with faster responses to targets produced by Adult talkers ($M = 345$ ms, $SD = 94$) than Child talkers ($M = 366$ ms, $SD = 101$). Crucially, the main effect of talker Familiarity was not significant ($\chi^2 = 0.84$, $p = 0.360$), with a trend for *slower* responses for Familiar talkers ($M = 360$ ms, $SD = 101$) than for Unfamiliar talkers ($M = 351$ ms, $SD = 96$). There was a significant interaction of talker Familiarity and talker Age ($\chi^2 = 5.74$, $p = 0.017$). We used the R package *emmeans* (Lenth, 2020) to investigate the source of this interaction via post hoc analyses, which yielded the following conclusions: There was a significant effect of talker Age for Unfamiliar talkers ($p < 0.0001$), as participants were slower to respond for Unfamiliar Child talkers ($M = 371$, $SD = 104$) than Unfamiliar Adult talkers ($M = 331$, $SD = 84$), but response latencies were similar ($p = 0.110$) for Familiar Child ($M = 361$, $SD = 99$) and Familiar Adult ($M = 359$, $SD = 102$) talkers. No other interactions were significant. We also confirmed that the effect of

Table 1 Mean (standard deviation) accuracy (proportion correct) by talker Familiarity, Age, and Blocking in Experiment 1

Familiarity	Age	Accuracy	
		Blocked	Mixed
Familiar	Adult	0.996 (0.06)	0.994 (0.05)
Familiar	Child	0.997 (0.05)	0.986 (0.12)
Unfamiliar	Adult	0.996 (0.07)	0.989 (0.10)
Unfamiliar	Child	0.990 (0.10)	0.986 (0.12)

Table 2 Mean (standard deviation) response times (RTs) by talker Familiarity, Age, and Blocking in Experiment 1 (all mixed combinations included)

Familiarity	Age	RT (ms)		RT (log ms)	
		Blocked	Mixed	Blocked	Mixed
Familiar	Adult	330 (97)	369 (102)	5.76 (0.28)	5.87 (0.28)
Familiar	Child	353 (96)	364 (100)	5.83 (0.27)	5.86 (0.27)
Unfamiliar	Adult	313 (77)	337 (86)	5.72 (0.23)	5.79 (0.25)
Unfamiliar	Child	363 (99)	374 (105)	5.86 (0.27)	5.89 (0.28)

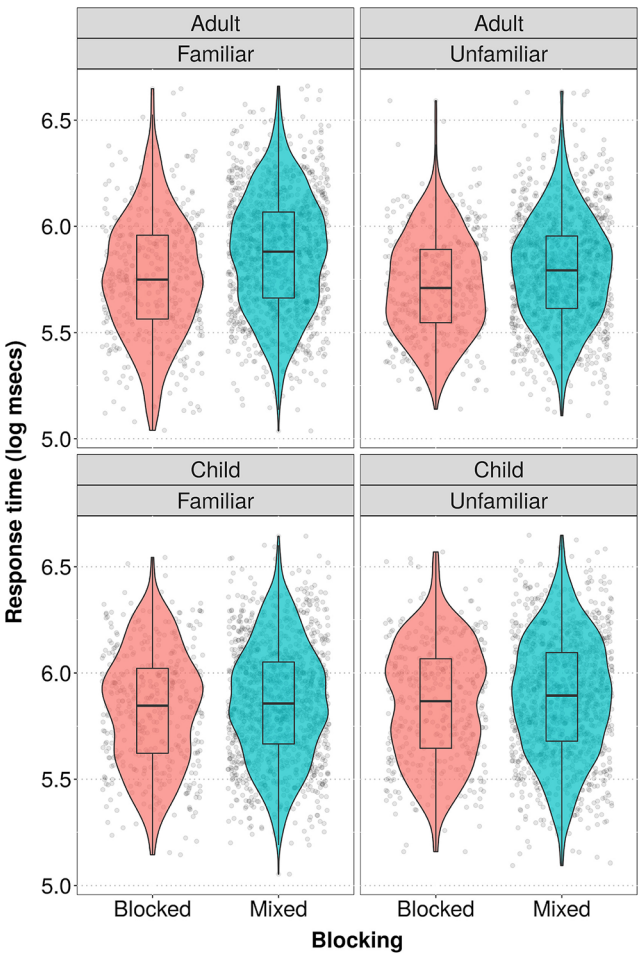

Fig. 4 Violin plots of log RT in Experiment 1 (superimposed over jittered individual trial data points)

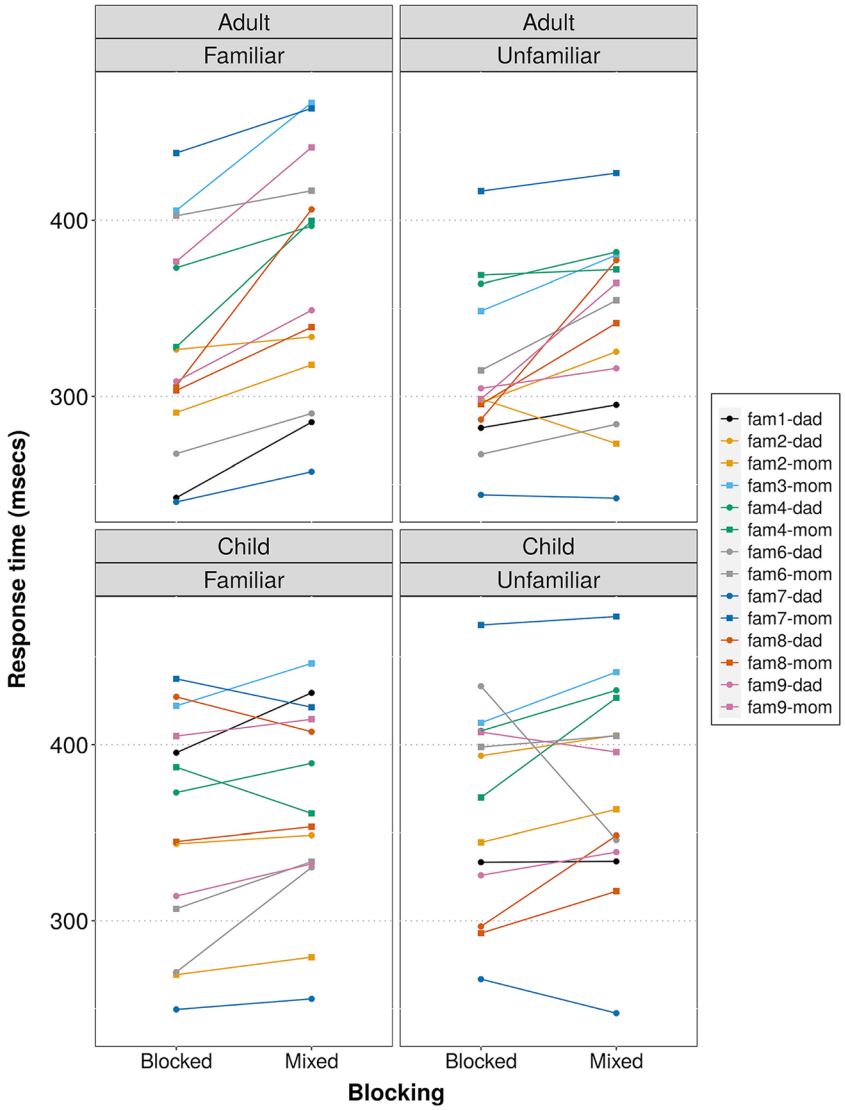

Fig. 5 Individual subject response times for Experiment 1 by talker Familiarity, talker Age, and Blocking

Blocking was significant for both Unfamiliar talkers (Blocked: $M = 337$, $SD = 92$; Mixed: $M = 355$, $SD = 97$; $p = 0.005$) and Familiar talkers (Blocked: $M = 342$, $SD = 97$; Mixed: $M = 366$, $SD = 101$; $p < 0.0001$).

However, note that this analysis does not isolate the conditions where Familiarity could have maximal impact. When we consider Mixed trials for Familiar Adult and Child talkers in this analysis, we are including cases where those Familiar talkers' productions were mixed with each other, but also when they were mixed with Unfamiliar talkers' productions. It is possible that Familiarity may be able to mitigate the impact of talker changes when *both* talkers are familiar (e.g., if talker adjustments can be avoided for familiar talkers but not for unfamiliar ones, the adjustment to unfamiliar talkers may be masking the benefit of familiarity). To assess this, we conducted a follow-up analysis restricting the data to Mixed trials that only included either the Familiar Adult and Familiar Child paired together *or* the Unfamiliar Adult and Unfamiliar Child paired together. The logic is that these combinations represent maximal conditions of Familiarity (both talkers Familiar or both talkers Unfamiliar). Summary data for this restricted analysis are presented in Table 3. We used

Table 3 Mean (standard deviation) response times (RTs) by talker Familiarity and talker Age in Mixed-talker trials restricted to cases where both talkers were Familiar or both were Unfamiliar in Experiment 1

Familiarity	Age	RT (ms) Mixed	RT (log ms) Mixed
Familiar	Adult	376 (108)	5.89 (0.28)
Familiar	Child	367 (102)	5.87 (0.28)
Unfamiliar	Adult	340 (84)	5.80 (0.24)
Unfamiliar	Child	376 (102)	5.89 (0.27)

the same model selection procedure. The selected model included fixed effects of talker Familiarity (Familiar vs. Unfamiliar), talker Age (Adult vs. Child), Blocking (Blocked or Mixed talker trials), and their interactions, along with by-subject random slopes and intercepts for the three-way interaction term.

The results were quite similar. The main effect of Blocking was significant ($\chi^2 = 6.37$, $p = 0.012$), with slower responses overall for Mixed ($M = 365$ ms, $SD = 100$) versus Blocked trials ($M = 340$, $SD = 94$). The main effect of Age was also significant ($\chi^2 = 9.14$, $p = 0.003$), with faster responses to targets produced by Adult talkers ($M = 340$ ms, $SD = 95$) than Child talkers ($M = 365$ ms, $SD = 100$). Crucially, as in the original analysis, the main effect of talker Familiarity was not significant ($\chi^2 = 1.20$, $p = 0.273$), with the same weak trend for *slower* responses for Familiar talkers ($M = 357$ ms, $SD = 102$) than for Unfamiliar talkers ($M = 348$ ms, $SD = 94$). We again observed a significant interaction of talker Familiarity and talker Age ($\chi^2 = 5.87$, $p = 0.015$). Post hoc analyses using the R package *emmeans* (Lenth, 2020) showed the same pattern as in the original analysis: there was a significant effect of talker Age for Unfamiliar talkers ($p < 0.0001$), with participants slower to respond for Unfamiliar Child talkers ($M = 370$, $SD = 100$) than Unfamiliar Adult talkers ($M = 327$, $SD = 82$), but response latencies were more similar ($p = 0.385$) for Familiar Child ($M = 360$, $SD = 99$) and Familiar Adult ($M = 353$, $SD = 105$) talkers. No other interactions were significant. We again confirmed that the effect of Blocking was significant for both Unfamiliar talkers (Blocked: $M = 337$, $SD = 92$; Mixed: $M = 358$, $SD = 95$; $p = 0.014$) and Familiar talkers (Blocked: $M = 342$, $SD = 97$; Mixed: $M = 372$, $SD = 105$; $p = 0.0001$). Thus, we observed significantly slower responses for Mixed-talker conditions even when the mixed talkers were both highly familiar.

Discussion

In Experiment 1, we observed a constant cost of mixing talkers (with a main effect of Blocking on accuracy, as well as effects of Blocking on RT for both Familiar and Unfamiliar talker pairs). Overall, the RT effects had a similar magnitude to what has been observed in previous studies using this paradigm with native speakers of American English (e.g., Nusbaum & Morin, 1992; Magnuson & Nusbaum, 2007). We found no advantage for our extremely familiar talkers (family members) compared to strangers: it appears that even familiar pairs of talkers require accommodation, given reliably slower responses and poorer accuracy in the mixed-talker condition compared to the blocked-talker condition. Although effects appear to be smaller for Child talkers (Fig. 4), note that nearly all participants showed the pattern of slower RTs for Mixed-talker than Blocked-talker trials for all talkers (see Fig. 5, where 11 of 14 participants showed the predicted trend for Familiar and Unfamiliar Child talkers, and 13 or 14 of 14 participants showed the predicted trend for Adult talkers).

This result is counter to the nonanalytic, exemplar, and normalization predictions we discussed earlier: Each account provides a logical basis to expect better performance following a change between familiar talkers compared to a change between unfamiliar talkers (contingent learning of indexical and phonetic characteristics [Nygaard & Pisoni, 1998], activation of more exemplars for familiar talkers in an episodic lexicon model [Goldinger, 1998], or perceptual learning on the operations of a normalization mechanism, cf. Fig. 2).

The consistent slowing of responses to speech produced by a different talker, regardless of talker familiarity, suggests that listeners are processing the change in talker characteristics, possibly in service of computing the mapping between talker characteristics and phonetic categories (i.e., normalization or accomodation, as in the *contextual tuning* hypothesis schematized in Fig. 1). On a normalization account, while information in memory about the talker may facilitate phonetic recognition, the change in talker must first be detected from the acoustic signal, and internal reference frames for segmental interpretation must be adjusted accordingly. However, an alternative interpretation of our results is that the manipulation of familiarity was insufficient to support recognition of talker characteristics. That is, the lack of a main effect of familiarity suggests the possibility that although we used speech from talkers who *should* have been familiar, perhaps the listeners were unable to recognize talkers from the short speech samples we used, preventing any familiarity advantage. We address this concern in Experiments 2 and 3.

Experiment 2

Although it has been conjectured that talker-change costs result from the simultaneous encoding of talker identity and linguistic message (e.g., Mullennix & Pisoni, 1990), it is not entirely clear that the acoustic properties that support talker identification and (putative) talker normalization are necessarily the same. Nonetheless, if listeners were unable to recognize a talker from the speech samples we used, it might also be the case that the speech samples contained insufficient detail to allow episodic benefits to emerge. Most previous perceptual studies of talker identification (or discrimination) have used longer duration materials than those we used in Experiment 1 (e.g., 2–4 s, Van Lancker et al., 1985; 6–120 s, Legge, Grosmann, & Pieper, 1984). In Experiment 2, we used a talker identification task to test whether the morae we used provided an adequate basis for recognizing familiar talkers, as well as a sufficient basis for learning to recognize unfamiliar talkers.

Method

Subjects The same subjects who participated in Experiment 1 participated in Experiment 2, approximately 1 month after Experiment 1. This included the mother from Family 1 (whose data for Experiment 1 were lost due to recording errors). The father from Family 3 also participated, for a total of 16 participants (results do not change materially if the two additional participants who did not participate in Experiment 1 are excluded).

Materials For each participant, three adult-child pairs of talkers were used in Experiment 2: the familiar talkers (a subject's spouse and child), and two unfamiliar talker pairs (one adult and one child) that the participant had not heard in Experiment 1. The unfamiliar talkers (whom participants would be trained to identify) were of the same sex as the familiar talkers for each subject. In order to ensure a fairly constant level of difficulty for all subjects, the unfamiliar talkers were chosen to have a measured average fundamental frequency within approximately 10 Hz of the appropriate familiar talker. Three subsets of the morae recorded for Experiment 1 were used. Five morae were used for familiarization, 10 were used for training, and 20 were used for testing.

Procedure Morae were presented on-line to subjects seated at workstations over STAX Lambda-SR-Signature headphones. There were six blocks in Experiment 2. The first block provided familiarization with the novel talkers. Subjects heard the four unfamiliar talkers in a fixed order. The talker order was cycled through five times with different morae. For each trial, subjects had to make a four-alternative choice among keys labeled (in Japanese): *unfamiliar adult 1, unfamiliar adult 2, unfamiliar child 1,* and *unfamiliar child 2*. When subjects answered correctly, they heard a chime. When they answered incorrectly, they heard a buzzer and then the item was repeated and they answered again.

The next three blocks were for training. First, subjects were presented with 20 trials from each of the two unfamiliar adults only (two repetitions of ten items), and then from the two unfamiliar children only (two repetitions of ten items). Although there were still four choices, each part was effectively a two-alternative forced choice, as adults and children were not confusable. Morae were presented in random order so that the talker varied from trial to trial. The items used for these two blocks were the same ones used for the familiarization block. After training separately on the adults and children, subjects had a final training block with new items produced by all four unfamiliar talkers presented in random order (two repetitions of ten new items per talker). Feedback was given for all training blocks in the same form as for the familiarization block.

Training was followed by a practice block with all six talkers (familiar and unfamiliar, one presentation of two items per talker) and a test block with all six talkers. *Familiar adult* and *familiar child* were added to the response keys for the practice and test blocks (making this a four-alternative forced choice; though again, adults and children were not confusable, which effectively narrowed the task to three choices), and feedback was eliminated. The practice block consisted of two items produced by each talker, chosen randomly from the list of items used in the familiarization block and presented in random order. The test block used two repetitions of 20 new items produced by each of the six talkers presented in random order. All manipulations were within subjects.

Results

Accuracy Participants learned to identify the new unfamiliar talkers (which we will label "trained" talkers for the sake of clarity when we reach Experiment 3) fairly well based on relatively little training with just 30 mora tokens. Figure 6 shows mean accuracy by talker Familiarity and Age for individual participants. There is substantial variability, but also an apparent advantage for familiar talkers (adults in particular), which is confirmed in Table 4. A generalized linear mixed-effects model with a logit link was fit to assess participants' accuracy in the talker identification task (with misidentifications coded as 0 and correct identifications coded as 1). We used the same selection procedure described for Experiment 1 to select among possible random effect structures. The selected model contained fixed effects of talker Familiarity (Familiar vs. Trained-on), and talker Age (Adult vs. Child), and their interaction, as well as by-subject random slopes and intercepts for the two-way interaction and the main effect of talker Age. There was a significant effect of talker Familiarity ($\chi^2 = 7.48$, $p = 0.006$), resulting from greater proportion correct on Familiar talkers ($M = 0.87$, $SD = 0.33$) than Trained-on talkers ($M = 0.80$, $SD = 0.40$). The main effect of talker Age was not significant ($\chi^2 = 0.23$, $p = 0.632$), but the interaction of Age and Familiarity was significant ($\chi^2 = 7.01$, $p = 0.008$). The pattern in Table 4 gives a sense of the source of the interaction, with higher accuracy for Familiar Adult than Familiar Child talkers, and the opposite pattern for Unfamiliar talkers. Post hoc tests indicated that the effect of Age was not significant for either Familiar talkers ($p = 0.082$) or Trained-on talkers ($p = 0.131$). However, the effect of Familiarity was significant for Adult talkers ($p = 0.001$) but not for Child talkers ($p = 0.957$). This may follow from either (or both) the somewhat lower accuracy for Familiar Child talkers compared to Familiar Adult talkers, or the better performance on Trained-on Child talkers than Trained-on Adult talkers. This suggests that for our participants, their own child's voice did not stand out as much as their spouse's, and/or that the unfamiliar Trained-on Child talkers were more distinctive than the unfamiliar Trained-on Adult talkers.

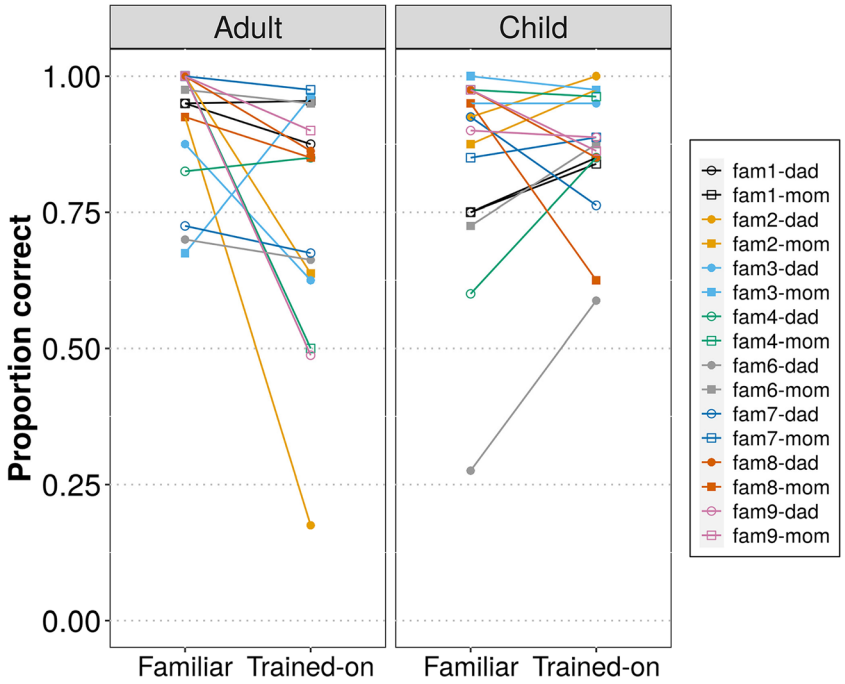

Fig. 6 Proportion correct for individual subjects in the talker identification task in Experiment 2 by talker Familiarity and talker Age3

Reaction time

We also examined RT, although accuracy was our primary concern. The RT pattern complemented that of accuracy (see Table 4). As in Experiment 1, we began by selecting only correct responses and then removed any responses with RTs more than 3 standard deviations from the mean (which removed 1.4% of responses). We first fit generalized linear-mixed effects models to the reaction time data, and applied the model selection procedure described for Experiment 1. The selected model included fixed effects of talker Familiarity (Familiar vs. Trained), talker Age (Adult vs. Child), and their interaction, along with by-subject random slopes and intercepts for the two-way interaction term. As in Experiment 1, the model was applied to raw RTs with an inverse Gaussian distribution with identity link, as recommended by Lo and Andrews (2015).

As in the accuracy analysis, the main effect of Familiarity was significant ($\chi^2 = 5.38$, $p = 0.020$), with faster responses overall for Familiar talkers ($M = 1397$ ms, $SD = 712$) than Trained talkers ($M = 1785$, $SD = 830$). The main effect of Age was not significant ($\chi^2 = 1.87$, $p = 0.171$). There was a significant interaction of talker Familiarity and talker Age ($\chi^2 = 5.14$, $p = 0.023$). Post hoc analyses using the R package *emmeans* (Lenth, 2020) showed that the effect of Familiarity was significant for Adult talkers ($p < 0.0001$) but not Child talkers ($p < 0.644$), despite the trend for faster responses for Familiar Child talkers compared to Trained-on Child talkers (note that variability was quite high).

Table 4 Means (and standard deviations in parentheses) for accuracy (proportion correct) and reaction time (RT; for correct responses) by talker Familiarity and talker Age in Experiment 2

Familiarity	Age	Accuracy	RT (ms)
Familiar	Adult	0.905 (0.293)	1,296 (563)
Familiar	Child	0.843 (0.364)	1,507 (831)
Trained-on	Adult	0.740 (0.440)	1,826 (820)
Trained-on	Child	0.859 (0.348)	1,748 (838)

Discussion

We observed advantages for familiar talkers in both accuracy and reaction time. Clearly, listeners can reliably identify talkers from a single mora, and there is sufficient talker-specific information contained in a mora to confer performance advantages for familiar talker identification. The present results suggest that the lack of a familiarity effect in Experiment 1 was not a consequence of the morae being too short to provide an adequate basis for retrieving talker characteristics (at least not those necessary for identification). The reliable effects of familiarity on accuracy and RT demonstrate that family members were indeed significantly more familiar to the participants than the trained-on talkers. However, the stimulus properties relevant for voice identification may not overlap completely with those relevant for phonetic processing; Experiment 3 tests whether our mora stimuli are sufficiently long to enable talker-specific benefits from phonetic processing under challenging conditions with degraded

stimuli (as Nygaard et al., 1994, found for phonetic identification of speech in noise).

Experiment 3

Speech perception is more accurate for familiar talkers than for unfamiliar talkers (Nygaard et al., 1994; Nygaard & Pisoni, 1998). However, the results of Experiment 1 suggest that this recognition advantage may not generalize to the perceptual costs imposed by talker variability, as we observed talker change costs for even potentially maximally familiar talkers – participants' own family members. Experiment 2 tested whether the speech samples we used might be too short to provide sufficient information about talker characteristics. While the results of Experiment 2 demonstrate that morae convey enough indexical information to support learning talker vocal characteristics and to support talker identification, it is not clear that the speech perception advantage reported by Nygaard and colleagues (Nygaard et al., 1994; Nygaard & Pisoni, 1998) would be obtained with the same materials. We designed Experiment 3 to test whether our materials provide sufficient information about talker characteristics to enable both learning of talker identity (as we observed in Experiment 2) and subsequent facilitation of speech perception under difficult conditions, as Nygaard and colleagues (Nygaard et al., 1994; Nygaard & Pisoni, 1998) found with training on words and sentences presented in noise.

Experiment 3 was conducted with 12 subjects from Experiment 2 about 8 weeks later. We first reinforced each participant's talker identification training with a short bout of the training regimen from Experiment 2. Then we presented participants with degraded speech ("sample-degraded", with the sign changed for 10% of waveform samples selected at random) produced by three different pairs of talkers: highly *Familiar talkers* (the familiar adult and child from Experiments 1 and 2), *Trained-on talkers* (unfamiliar adult 1 and unfamiliar child 1 from Experiment 2), and *Exposed-to talkers* (the unfamiliar adult and child from Experiment 1, whom participants had heard in the mora monitoring task in Experiment 1, but whom subjects had never been asked to identify). The task was to transcribe each mora. In addition, the morae were presented in two conditions, as in Experiment 1: blocked by talker, and items from multiple talkers mixed within a block.

The primary prediction for Experiment 3 is that since the morae provide sufficient cues for talker identification (given the results of Experiment 2), we ought to find that talker identity training facilitates identification of degraded morae (as predicted by the results of Nygaard & Pisoni, 1998). If we fail to observe such an advantage, this would leave doubt about the relevance of Experiment 1 to the question of whether familiarity mitigates the contribution of talker variability to the lack of invariance problem.

Method

Participants Twelve participants who participated in Experiment 2 participated in Experiment 3 approximately 8 weeks later. These participants were the mothers and fathers from families 2, 4, 6, 7, 8, and 9. Other participants declined to return.

Materials For *talker identification training and testing*, the items were the same as those used for Experiment 2. The same unfamiliar talkers that participants were trained to identify in Experiment 2 (with labels "unfamiliar adult 1," "unfamiliar adult 2," "unfamiliar child 1," and "unfamiliar child 2") were assigned to subjects in Experiment 3.

For each subject, the items for *mora identification* were produced by each of six talkers: the Familiar adult and child (the participant's spouse and child), a Trained-on adult and child (the talkers assigned for each participant as unfamiliar adult 1 and unfamiliar child 1 for Experiment 2) and an "Exposed-to" pair of talkers that the participant had heard in the mora-monitoring task in Experiment 1, but that had not been included in talker identification training in Experiment 2. Including *Trained-on* and *Exposed-to* talkers allowed us to compare the effects of simple exposure to talkers in the experimental setting with the effects of explicit talker identification training. There were 30 morae for each talker (/bi/, /ba/, /bo/, /gi/ /ge/, /gu/, /ki/, /ke/, /ku/, /mi/, /ma/, /mo/, /ni/, /ne/, /no/, /nu/, /pi/, /pe/, /pa/, /po/, /pu/, /ri/, /re/, /ra/, /ro/, /ru/, /ze/, /za/, /zo/, and /zu/). Each appeared once per talker in the Blocked condition and once per talker in the Mixed condition, in random order.

In order to avoid ceiling levels of accuracy, we degraded the items for mora identification by randomly selecting 10% of the samples of each item and changing the signs of the values of these samples. This resulted in a sufficient level of degradation that the morae were moderately difficult to identify, while preserving the amplitude envelope of the items (Horii, House & Hughes, 1971; O'Malley & Peterson, 1966).

Procedure There were five parts to the experiment. First, participants were re-familiarized to the same four unfamiliar talkers they had heard in Experiment 2 (*Ua1, Uc1, Ua2, and Uc2*), though only one pair would be used later as the "trained-on" talkers. The re-familiarization block was identical to the familiarization block used in Experiment 2, except that only two morae per talker were used.

Second, participants were retrained to identify the four unfamiliar talkers. This retraining was identical to the training session used in Experiment 2, although the items were in new,

randomly generated orders. The purpose of this retraining was to ensure participants would return at least to the level of performance they exhibited by the end of Experiment 2, prior to examining the effect of phonetic identification on later talker identification.

Third, participants practiced identifying the four unfamiliar talkers along with the two familiar talkers. This practice block was identical to the one used in Experiment 2. Fourth, participants were given a talker identification test identical to the test used in Experiment 2, except that only 15 morae per talker were used.

Then subjects transcribed sample-degraded morae produced by the familiar talkers, the exposed-to talkers, and the trained-on talkers (one pair of the "unfamiliar" talkers' subjects had been trained to identify in Experiments 2 and 3). The mora identification in noise phase consisted of two blocks. In one block, 30 morae produced by each of the six talkers were presented consecutively (blocked-talker condition). After 30 items from one talker, the 30 items from the next talker followed immediately. In the other block, the same set of 30 items per talker (180 total) was randomly ordered (mixed-talker condition). The order of mixed and blocked conditions was counterbalanced across participants.

Participants were seated at workstations. At the beginning of each trial, the trial number appeared on the screen, and the mora was played simultaneously. There was a 2-s inter-trial-interval during which participants were to transcribe what they had heard onto a numbered answer sheet. As we described earlier, zeroes were added to the end of each mora such that each was 830 ms long. Thus, the interval between mora onsets was 2,830 ms. Finally, we note that all manipulations were within subjects.

Results

Talker identification Results following the talker-training post-test are shown for individual subjects in Fig. 7. The additional talker identification training resulted in approximately equal numbers of participants showing modest advantages for Familiar or Trained-on talkers. Only two participants failed to identify all talkers above chance levels (fam4-mom failed to identify the unfamiliar child above chance, and fam6-dad failed to identify his own child above chance). We first examined the accuracy data with all participants included (see Table 5 for summaries).

We fit a generalized linear mixed-effects model with a logit link to assess participants' accuracy in the talker identification task (with misidentifications coded as 0 and correct identifications coded as 1). We used the same selection procedure described for Experiment 1 to select among all permutations of possible random effect structures. The model selected contained fixed effects of talker Familiarity (Familiar vs. Trained), and talker Age (Adult vs. Child), and their interaction, as well as by-subject random slopes and intercepts for the two-way interaction and the main effect of talker Age. There was a significant effect of talker Familiarity ($\chi^2 = 7.03$, $p = 0.008$), resulting from greater proportion correct on Familiar talkers ($M = 0.872$, $SD = 0.334$) than Trained-on talkers ($M = 0.825$, $SD = 0.380$). The main effect of talker Age was not

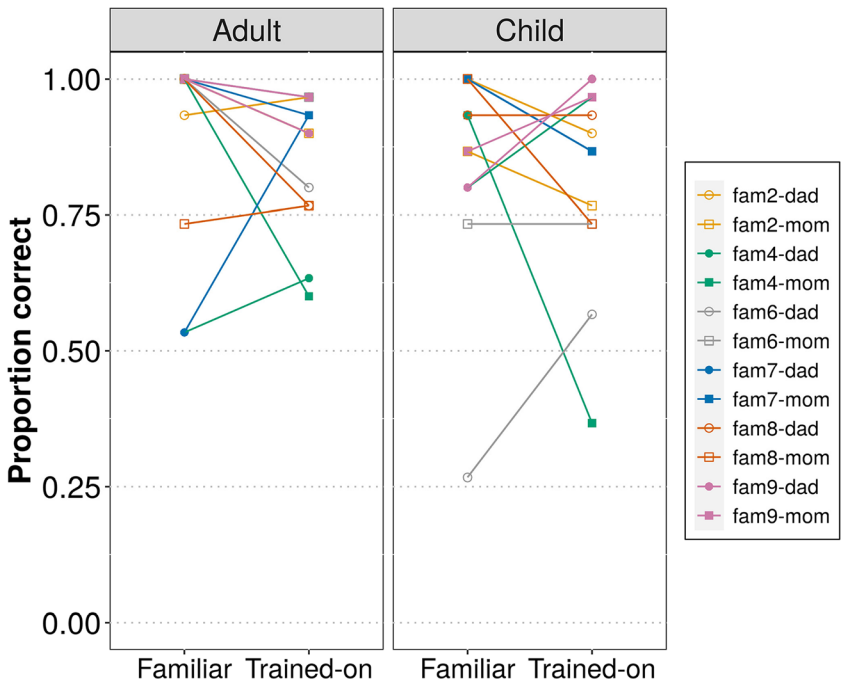

Fig. 7 Post-test talker identification accuracy by familiarity for individual subjects in Experiment 3

Table 5 Mean (standard deviation) for accuracy (proportion correct) and reaction time (RT) by talker Familiarity and talker Age in the talker identification task in Experiment 3

Familiarity	Age	Accuracy	RT (ms)
Familiar	Adult	0.894 (0.308)	1,296 (563)
Familiar	Child	0.850 (0.358)	1,507 (831)
Trained-on	Adult	0.844 (0.363)	1,826 (820)
Trained-on	Child	0.806 (0.396)	1,748 (838)

significant ($\chi^2 = 2.85$, $p = 0.091$), nor was the interaction of Age and Familiarity ($\chi^2 = 2.46$, $p = 0.117$). We repeated the analysis with the poor-performing participants (fam4-mom and fam6-dad), but this did not alter the pattern of significance.

Speech-in-noise (mora transcription) task Mora transcription responses were considered correct only if the participant produced the expected mora in full (i.e., we did not give partial credit if the participant supplied the correct consonant but incorrect vowel, or vice versa). Chance performance could be considered somewhere between approximately 1/100 (based on approximately 100 CV syllables in Japanese) or 1/30 (given that we used 30 morae, although participants did not know that). In either case, participants performed well above chance, with a mean proportion correct of 0.63. Proportion correct by Blocking, talker Age, and Talker is shown in Fig. 8.

We fit a generalized linear mixed-effects model with a logit link to assess participants' accuracy in the mora transcription in noise task (with incorrect transcriptions coded as 0 and correct transcriptions coded as 1). We used the same selection procedure described for Experiment 1 to select among all permutations of possible random effect structures. The selected model contained fixed effects of Blocking (Blocked or Mixed), Talker (Familiar, Trained-on, or Exposed-to), and talker Age (Adult vs. Child), and their interactions, as well as by-subject random slopes and intercepts for the three-way interaction and the main effects of Blocking and Talker. There was a significant effect of Blocking ($\chi^2 = 4.45$, $p = 0.035$), resulting from greater proportion correct when items were Blocked by talker ($M = 0.648$, $SD = 0.478$) rather than Mixed ($M = 0.610$, $SD = 0.488$). The main effect of talker Age was also significant ($\chi^2 = 34.2$, $p < 0.001$), resulting from greater proportion correct for Adult ($M = 0.686$, $SD = 0.464$) than Child talkers ($M = 0.572$, $SD = 0.495$). Despite trends for accuracy to increase with talker familiarity in Fig. 8, the main effect of Talker was not significant ($\chi^2 = 2.39$, $p = 0.302$). No interactions were significant.

While we did not see the predicted clear impact of talker familiarity, to further investigate the apparent trend in Fig. 8 for accuracy to be lower for Exposed-to talkers, we conducted post hoc tests comparing each level of Talker. Accuracy was highest for Familiar talkers ($M = 0.661$, $SD = 0.47$), then for Trained-on talkers ($M = 0.639$, $SD = 0.48$), and lowest for Exposed-to talkers ($M = 0.590$, $SD = 0.49$). While the difference between Familiar and Trained-on talkers was

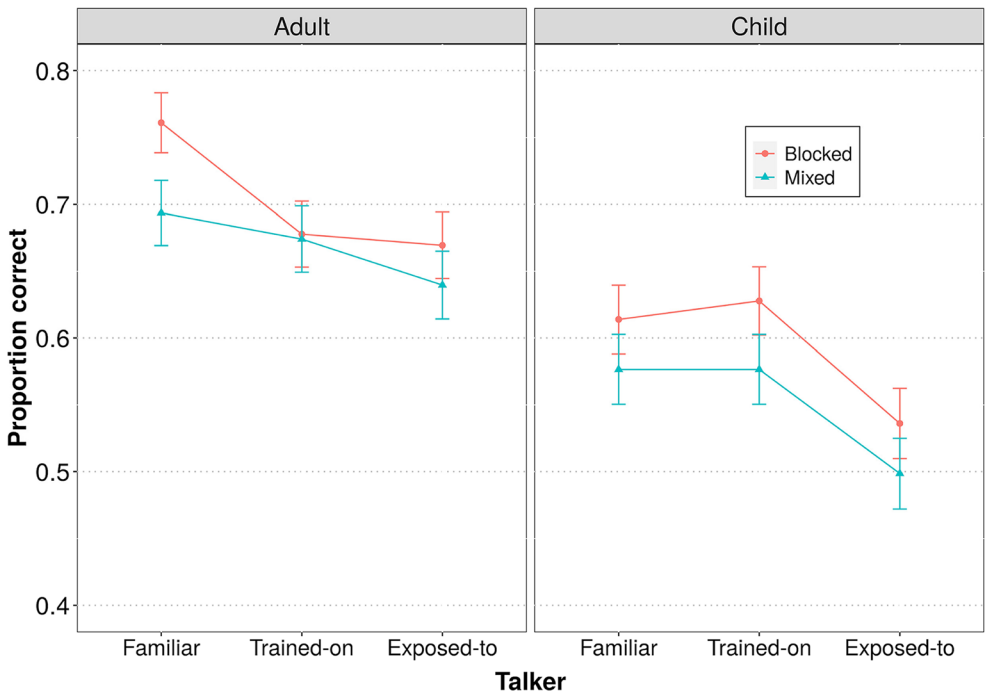

Fig. 8 Transcription accuracy for morae presented in noise for Adult and Child talkers by talker Familiarity and Blocking in Experiment 3. Error bars indicate standard error

not significant ($p = 0.649$, demonstrating the impact of talker identification training), the difference between Familiar and Exposed-to talkers was ($p = 0.008$; there was also a strong trend for Trained-on vs. Exposed-to: $p = 0.069$). Thus, we see evidence for benefits of talker familiarity (through daily, real-life experience or laboratory training) for this speech-in-noise task.

Discussion

Even after a break of 8 or more weeks between Experiments 2 and 3, most participants quickly recovered the level of talker identification accuracy observed after training in Experiment 2. Accuracy in the mora identification task was also correlated with familiarity, both as a function of extensive exposure outside the lab (family members) and talker identification training in the lab (trained-on vs. exposed-to talkers). This replicates the generalization effect reported by Nygaard et al. (1994) and Nygaard and Pisoni (1998) – training on talker identity benefits linguistic identification.

Thus, the short CV items used in the current experiments provide a sufficient basis for retrieving talker characteristics that facilitate both talker and linguistic identification. The results of Experiments 2 and 3 demonstrate that the morae used in Experiment 1 provide sufficient talker information to lead to reliable effects of talker familiarity in other tasks. This means that the performance deficits observed for talker variability in mora monitoring in Experiment 1 cannot be attributed to the qualities of the materials used. Thus, even when listening to highly familiar talkers such as family members, talker variability exacts a performance cost.

General discussion

As we discussed in the *Introduction*, quite different theoretical perspectives on talker differences (talker normalization theories vs. nonanalytic and exemplar theories) predict a reduction of talker variability effects with increased talker familiarity. On the normalization view, repeated adjustments to familiar talkers' voices might automatize the talker-specific adjustments reducing processing costs (Nygaard et al., 1994). On an exemplar view, an advantage for familiar talkers should follow from greater representation of familiar talkers among stored exemplars in memory (Goldinger, 1998; Johnson, 1990, 1997) or from contingent/integral encoding of indexical and phonetic properties of voices, leading to general benefits in any processing task as a result of perceptual learning (Nygaard & Pisoni, 1998). Increased talker familiarity therefore should change the way listeners process talker variability. However, the results of Experiment 1 show that this prediction does not hold: a change in talkers results in a consistent processing cost, even when the talkers are highly familiar.

Experiments 2 and 3 showed that this result cannot be attributed to inadequate cues to talker characteristics in the materials we used, since the items provided an adequate basis for identifying familiar talkers and learning to identify unfamiliar ones. Experiment 3 also showed that, as has been reported previously (Nygaard et al., 1994; Nygaard & Pisoni, 1998), familiarity with a talker as a result of laboratory training in talker identification facilitates speech perception under difficult conditions, and furthermore, that extensive familiarity based on exposure to talkers outside the lab (e.g., years of experience with family members) has similar effects.

If talker familiarity improves recognition accuracy under noisy and degraded listening conditions, why doesn't familiarity improve recognition performance when there is talker variability? In Experiment 1, similar slowing was observed for mora recognition given changes between familiar talkers and changes between unfamiliar talkers. We suggest that the cost results from prerequisite processing of a talker's voice that must occur before any familiarity benefit can occur. That is, there is a *parallel-contingent* relation (Turvey, 1973) between voice characteristics and phonetic identification, as has previously been discussed by Mullennix and Pisoni (1990). That is, the implications of a talker's acoustic-phonetic characteristics for mapping to perceptual categories must be taken into account in some way before familiarity can be exploited. This could be explicit talker recognition, or simply sufficient analysis of speech samples to trigger talker-specific procedural memory. We hypothesize that it is the latter, and that this analysis corresponds to attunement to talker characteristics as in Figs. 1 and 2. This analysis apparently must occur for familiar or unfamiliar talkers, leading to the constant cost observed in Experiment 1. That is, even if procedural memory for mapping a talkers' productions to phonological categories boosts subsequent processing (e.g., identification in noise, as in Experiment 3 and Nygaard et al., 1994), knowledge about a talker, procedural or otherwise, can only benefit processing once talker characteristics have been detected. This process appears to be fast and efficient, although it imposes significant, detectable demands on speech processing, resulting in a constant cost of about 20–30 ms that cannot be improved upon by extensive experience with individual talkers.

Indeed, Nusbaum and Morin (1992) made a compelling case that this cost is due to increased working memory load (based on their results showing that talker variability interacts with phonological working memory; e.g., RT increased significantly as load increased in mixed-talker conditions, but not in blocked talker conditions). Wong, Nusbaum, and Small (2004) found that in a mixed-talker condition, cortical activity increases in areas associated with speech (posterior superior temporal gyrus), consistent with increased capacity demands in processing other spoken language materials (Just et al., 1996), but also

areas associated with shifts of attentional processing (superior parietal cortex, e.g., see Posner, 2003; Yantis et al., 2002).

These studies were used to argue that a change in talker imposes a measurable load on working memory due to the increased uncertainty about phonetic recognition (see Nusbaum & Morin, 1992), even though there is no difference in the explicit memory aspects of the recognition task in the blocked- and mixed-talker conditions. The observed working memory load was interpreted as reflecting the need to test among alternative phonetic interpretations of the acoustic pattern of a stimulus. When there is a talker change, listeners appear to shift attention to a different set of acoustic properties (including F0 and F3; see Nusbaum & Morin, 1992) that provide information about talker vocal characteristics that can reduce this uncertainty. As proposed by Choi et al. (2018), on the basis of their finding that a mixed-talker cost is observed even when talker variability does not result in task-relevant phonetic ambiguity (e.g., deciding whether one has heard *sigh* or *buy*), this also appears to be an obligatory process under typical circumstances. However, consistent with active attentional control, Magnuson and Nusbaum (2007) found that talker normalization may be modulated by listeners' expectations. They presented listeners with speech produced by two synthetic "talkers" differing in fundamental frequency by 10 Hz (with only very minor corollary variation in other parameters), and told different groups of subjects they would be hearing two talkers, or one talker, and a third group received no instructions about talkers. Only the group expecting to hear two talkers showed the typical talker-change cost (approximately 20 ms slowing in mixed- vs. blocked-talker condition). While this effect may only be observable under extremely specific conditions (a case where there is discernible variation that is not automatically attributed to talker variability), it is consistent with an active control mechanism that can be modulated by attention (Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997).

With respect to exemplar and episodic/nonanalytic theories, the familiarity benefits we observed in talker identification (Experiment 2) and phonetic processing (Experiment 3) are quite consistent with previous findings. However, it is not apparent how current exemplar models would account for the attentional and expectation effects we have just reviewed. A constant performance cost for talker changes might be consistent with an exemplar model, if we assume that a processing benefit should accrue when successive stimuli are highly similar in respect to the overall characteristics of the talker that produced them (although why similarity in talker characteristics should produce a measurable effect given potentially larger phonetic variability within a talker [e.g., one talkers' productions of "ball" and "done"] than between talkers [e.g., one talkers' productions of "ball" and "done"].

The implications for the contingent/integral encoding of phonetic and indexical characteristics account are more complex. The results of Nygaard and Pisoni (1998), among others reviewed earlier, call for a rethinking of the conventional separation of linguistic and indexical properties of speech. As we discussed, we agree with Nygaard et al. (1994) and Mullennix and Pisoni (1990) that the perception of these dimensions is contingent. It is important to note that change deafness studies indicate that talker changes are not always detected (e.g., Vitevitch, 2003 reports such evidence, and that processing advantages associated with preserving talker characteristics depend on whether or not talker changes are noticed) and detection of a talker change may depend on listener expectations to monitor for the change (Fenn et al., 2011; Theodore, Blumstein, & Luthra, 2015).

We differ with Nygaard and Pisoni (1998) as to the degree and nature of the contingency. Nygaard and Pisoni argue these aspects of the signal are wholly inseparable, which is the basis for their argument against normalization: if the two dimensions are not separable perceptually or in memory, a normalization mechanism that purports to isolate and operate on the phonetic dimension is illogical. Further, by appeal to an instance-based mechanism like that posited in Goldinger's (1998) episodic lexicon theory, normalization is unnecessary (although in Goldinger's simulations, separate elements were used to code indexical and phonetic characteristics, whereas in real speech, indexical and phonetic characteristics are not distinct; see Magnuson & Nusbaum, 2007, pp. 404-405, for a more detailed critique).

Nygaard and Pisoni cite the results of Mullennix and Pisoni (1990) as evidence of the inseparability of the dimensions, given that Mullennix and Pisoni found evidence of integral processing using a Garner (1974) interference task. However, Mullennix and Pisoni emphasized the asymmetries in the interference patterns they observed. When talker and phonetic variability were increased together, there were linear increases in the amount of interference observed, but phonetic variation interfered less with talker processing than talker variability interfered with phonetic identification. There was also greater interference for the phonetic task from task-irrelevant phonetic variability than with talker variability, and vice-versa. When talker and phonetic variability were increased separately (Mullennix & Pisoni's Experiment 2), interference increased linearly as phonetic variability increased, while a nearly constant cost was found for talker variability, whether the number of talkers was four or 16. This suggests a more complex relationship between these aspects of the signal than all-or-none integrality. Indeed, many authors, including Garner (1974), have proposed that there is a continuum of separability of stimulus dimensions, rather than discrete categories of integrality or separability (e.g., Ashby & Maddox, 1994; Potts, Melara, & Marks, 1998).

While integrality of talker and phonetic information is consistent with the contingency view of Nygaard and colleagues, partial, asymmetric contingency is consistent with the contextual tuning account, which closely resembles the original

explanation of Mullennix and Pisoni (1990). They interpreted the absence of a talker set-size effect as indicating that listeners *can* selectively ignore phonetic variability while attending to talkers' voices (even though the two dimensions are closely related), and that "two qualitatively different types of processes are utilized" for processing phonetic and talker information. They proposed that operations that decode talker and phonetic identity work in parallel, but in a contingent fashion (cf. Turvey, 1973). Because talker characteristics condition phonetic realizations of speech sounds, phonetic identification is hierarchically dependent on an analysis of talker characteristics. Varying talker characteristics simultaneously varies phonetic characteristics (for many or most words), even if the word produced is constant. In contrast, when the talker is constant, the talker characteristics available to a listener do not change dramatically between most words.

Indeed, whether effects of surface specificity are observed depends on the task used and how quickly it can be performed. Luce and Lyons (1998) found specificity effects in an explicit recognition task, but not in an implicit task (priming in lexical decision). McLennan and Luce (2005) have found a temporal dissociation between processing of lexical information and surface specificity (speaking rate and talker identity). When processing was fast (because the materials were relatively easy to process), they observed equivalent priming for repeated words whether rate or talker identity was constant or varied between presentations. When processing was slowed by the relative difficulty of the materials, greater priming was found when surface specificity was preserved. This difference in time course for phonetic information and non-phonetic surface variability is consistent with the interpretation that priming effects in these experiments depend upon reactivation of linguistic representations and episodic event memories that are distinct (but not completely independent, in the sense that there are redundancies and associations between them).

Our view of the larger picture is that speech perception emerges from multiple processes working in parallel on different, but not necessarily independent, aspects of the signal. This is consistent with evidence from multiple neurophysiological maps in auditory cortex (e.g., Dick et al., 2012; Hackett, 2007; Woods et al., 2009), indicating that there are parallel auditory representations subserving different aspects of perception. We expect that these include (but are not limited to) the recovery of phonetic categories, indexical information about the talker, and episodic traces of the speech event and its context. As Andruski, Blumstein, and Burton (1994) showed, processes recovering the linguistic message preserve subcategorical detail at least until initial contact is made with the lexicon. It is also clear that episodic traces of speech events that include a high degree of surface detail are simultaneously laid down (Goldinger, 1996, 1998; Palmeri et al., 1993), although these specificity effects may depend on the salience of talker changes (Fenn et al., 2011; Theodore et al., 2015; Vitevitch, 2003). While representations of prior speech episodes may also be activated as speech is heard and potentially influence lexical processing (Goldinger, 1998), there seems to be a distinction between phonetic representations and other aspects of speech, as the nature of the task determines to what degree non-phonetic surface details of speech are activated (Luce & Lyons, 1998), and these representations also appear to operate on slightly different time scales, with priority for phonetic representations (McLennan & Luce, 2005). Similar observations have led some to the conclusion that listeners *both* encode instances and derive abstract (talker-invariant) encodings (Magnuson & Nusbaum, 2007; Pierrehumbert, 2016; Pisoni & Levi, 2007). Work by Myers and colleagues has begun providing evidence for parallel neural representations that are talker-invariant versus talker-specific (e.g., Myers & Theodore, 2017; Salvata, Blumstein, & Myers, 2012).

Much of the debate about how listeners achieve phonetic constancy despite talker variability in the past few years has been shaped by the assertion that normalizing talker differences necessarily implies that surface detail is stripped from the speech signal and is forever lost. This is not the only view of normalization. Although preservation of surface detail has been argued to be evidence against normalization (e.g., Goldinger, 1996, 1998; Palmeri et al., 1993; Pisoni, 1997), it is not dispositive. We agree that a strong abstractionist view is held by some (e.g., Andruski et al., 1994, assumed nonphonetic variation is eventually discarded, but focused on determining how late in lexical access subcategorical information is available – and indeed found that the effects of subcategorical perturbations of VOT became undetectable in their procedure after 50–250 ms). However, this is not a requirement of normalization. Perhaps the earliest specific description of a normalization mechanism was provided by Joos (1948), and on his account, either internal representations or the acoustic signal are warped to bring the two into registration, with no information discarded. Evidence for contingencies of "linguistic" and "non-linguistic" characteristics of utterances in memory and processing definitively rules out a model of speech perception in which there is a single representation of the signal from which phonetically irrelevant information is cast off as quickly as possible. But this evidence does not remove the need to explain how listeners achieve phonetic constancy despite talker differences.

Conclusions

We presented three main results. First, as has been found in several other studies (e.g., Magnuson & Nusbaum, 2007; Nusbaum & Morin, 1992), stability in talker characteristics improves performance in linguistic tasks, which is consistent with both normalization and exemplar theories. Second, in accord with several previous reports (Johnsrude et al., 2013;

Nygaard et al., 1994; Nygaard & Pisoni, 1998; Souza et al., 2013), we found that familiarity with a talker's characteristics (acquired in the lab or outside the lab) facilitates linguistic processing. Third, and most importantly, the results of Experiment 1 indicate that there is a constant cost associated with talker changes independent of talker familiarity. Even the talkers we might expect to be the most familiar possible (one's own spouse and child) appear to require adjustments in acoustic-perceptual mapping when there is a talker change. This result is also consistent with normalization and exemplar accounts, but imposes constraints on both.

If statistical sampling of variability using exemplar mechanisms is the basis for phonetic constancy, the lack of facilitation in changes from one familiar talker to another should constrain the relative weighting of short-term versus long-term exemplar traces – short-term similarity appears to swamp long-term similarity, at least so far as the impact of talker changes is concerned. Or, as we discussed above, it may be even more likely that any benefit for familiarity depends on first having access to familiar talker characteristics through analytic processing of speech after a talker change.

While the current results by themselves do not provide a dispositive basis for preferring one account over another, the normalization explanation accounts for a broader range of data. Normalization is consistent with the processing cost associated with talker changes, as well as with the context sensitivity of talker-specific phonetic categorization (Ladefoged & Broadbent, 1957). In particular, the contextual tuning theory of normalization (Magnuson & Nusbaum, 2007; Nusbaum & Morin, 1992; Nusbaum & Magnuson, 1997) also provides an account for effects of attention (Nusbaum & Morin, 1992; Wong et al., 2004) including top-down expectations (Magnuson & Nusbaum, 2007). It does not simultaneously account for specificity effects (Goldinger, 1996, 1998; Martin et al., 1989; Mullennix et al., 1989; Palmeri et al., 1993), and thus this view requires other processes operating in parallel to provide a comprehensive account of speech perception.

However, while exemplar/episodic approaches promise simultaneous accounts of specificity effects and phonetic constancy despite talker variability, they have yet to propose a plausible account of the latter. Goldinger's (1998) simulations that simultaneously accounted for specificity effects and phonetic constancy were a crucial first step, but depended on the unrealistic assumption that indexical and phonetic characteristics could be coded independently (different units in the input vectors represented the two types of information), while in real speech, talker differences condition the realization of phonetic productions (and vice-versa).

We suggest that benefits of learning about talkers actually depend upon processes that simultaneously support normalization and access to any memory for a talker – to apply memory for a talker's characteristics (and/or the operations to apply to them), enough speech must be processed to afford retrieval (indeed, the analysis of talker characteristics relevant for phonetic mapping may be a fundamental component of talker recognition, given the utility of talker-specific phonetic patterns for talker identification; Remez et al., 1997). That is, consistent with the conclusions of Mullennix and Pisoni (1990), we assume there are multiple processes that operate on the speech signal, hierarchically organized in a parallel-contingent manner (Turvey, 1973). Determining the mapping between characteristics of the current talker and internal representations must be among the first and most important processes that operate, as suggested by our finding that even the most familiar talkers appear to require attention-demanding perceptual accommodation.

Author Note We thank Inge-Marie Eigsti, Jennifer Pardo, Hideki Kawahara, and Tsuneo Yamada for comments that greatly improved this paper. Preparation of the manuscript was facilitated by the following grants: NSF 1754284 (PI: JSM), NSF IGERT 1144399 (PI: JSM), and NSF NRT 1747486 (PI: JSM).

Open Practices Statement Data and R scripts are available at https://osf.io/9xmnc.

References

Andruski, J. E., Blumstein, S. E., & Burton, M. (1994). The effect of subphonetic differences on lexical access. *Cognition, 52*, 163–187.

Ashby, F. G., & Maddox, W. T. (1994). A response time theory of separability and integrality in speeded classification. *Journal of Mathematical Psychology, 38*, 423–466.

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

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*, 784–797.

Church, B.A., & Schacter, D.L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 521–533.

Craik, F.I.M., & Kirsner, K., (1974). The effects of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology, 26,* 274-284.

Creelman, C.D. (1957). The case of the unknown talker. *Journal of the Acoustical Society of America 29*, 655.

Cutler, A., Dahan, D., & Donselaar, W. van (1997). Prosody in the comprehension of spoken language: a literature review. *Language & Speech, 40*, 141–201.

Dick, F., Tierney, A.T., Lutti, A., Josephs, O. Sereno, M.I., & Weiskopf, N. (2012). In vivo functional and myeloarchitectonic mapping of human primary auditory areas. *Journal of Neuroscience*, *32,* 16095–16105.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech wave. In J. S. Perkell & D. H. Klatt (Eds.), Invariance and Variability in Speech Processes (pp. 360-380). Lawrence Erlbaum Associates: Hillsdale, NJ.

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *Quarterly Journal of Experimental Psychology*, 64, 1442–1456.

Fougeron, C. A., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America, 101,* 3728 – 3740.

Fowler, C. A., & Housum, J. (1987). Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, 489–450.

Fowler, C. A., Levy, E. T., & Brown, J. M. (1997). Reductions of spoken words in certain discourse contexts. *Journal of Memory and Language, 37,* 24–40.

Garner, W. R. (1974). *The Processing of Information and Structure*. Potomac, Maryland: Lawrence Erlbaum.

Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio Electroacoustics, AU-16*, 78–80.

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

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 17*, 152–162.

Hackett, T.A. (2007). Organization and correspondence of the auditory cortex of humans and nonhuman primates. In J.H. Kass (Ed.), *Evolution of the nervous system*, (pp 109 –119). Oxford, UK: Elsevier.

Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*. https://doi.org/10.3389/fnsys.2014.00035

Horii, Y., House, A.S., & Hughes, G.W. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America, 49*, 1849–1856.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88*, 642–654.

Johnson, K. (1994). Memory for vowel exemplars. *Journal of the Acoustical Society of America, 95*, 2977.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 145–166). San Diego: Academic Press.

Johnson, K. (2005). Speaker normalization in speech perception. In D.B. Pisoni & R. Remez (Eds.), *The Handbook of Speech Perception*. Oxford: Blackwell Publishers. pp. 363–389.

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

Joos, M. (1948). *Acoustic phonetics*. Baltimore: Linguistic Society of America.

Just, M. A., Carpenter, P. A., Keller, T. A., Eddy, W. F., Rep, M., van Dijl, J. M., Suda, K., Schatz, G., et al. (1996). Brain activation modulated by sentence comprehension. *Science, 274(5284)*, 114–116.

Kolers, P. A. (1976). Reading a year later. Journal of Experimental Psychology: Human Learning and Memory, 2, 554–565.

Kolers, P. A. and Ostry, D. J. (1974). Time course of loss of information regarding pattern analyzing operations. *Journal of Verbal Learning and Verbal Behavior, 13,* 599–612.

Ladefoged, P. (1989). A note on "Information conveyed by vowels" *Journal of the Acoustical Society of America, 85*, 2223–2224.

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

Legge, G. E., Grosmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 298–303.

Lenth, R. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. R package version 1.4.6. https://CRAN.R-project.org/package=emmeans

Liberman, A. M., DeLattre, P. D., & Cooper, F. S. (1952). The role of selected stimulus variables in the perdcetion of unvoiced stop consonants. *American Journal of Psychology, 65*, 497–516.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2015.01171

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Logan, G. D. (2002). An instance theory of attention and memory. *Psychological Review, 109*, 376–400.

Luce, P. A., & Lyons, E. A. (1998). Specificity of memory representations for spoken words. *Memory & Cognition, 26*, 708–715.

Magnuson, J. S. (2018a). Contextual tuning theory without memory. *Figshare* https://doi.org/10.6084/m9.figshare.5977387.v1

Magnuson, J. S. (2018b). Contextual tuning with memory. *Figshare* https://doi.org/10.6084/m9.figshare.5977444.v1

Magnuson, J. (2020). Mora monitoring procedure. *Figshare* https://doi.org/10.6084/m9.figshare.12560294.v1

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*, 391–409.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology*: *Learning, Memory, & Cognition*, *15*, 676–684.

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 306–321.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America, 85,* 2114–2134.

Miller, J. L., & Liberman, A. M. (1979). Some effects of later-occurring information on the perception of stop consonant and semivowel. *Perception & Psychophysics, 25*, 457–465.

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

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85*, 365–378.

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. Brain and Language, 165, 33–44.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Nooteboom, S. G., & Kruyt, J. G. (1987). Accent, focus distribution, and the perceived distribution of given and new information: An experiment. *Journal of the Acoustical Society of America, 82*, 1512 – 1524.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 109–132). San Diego: Academic Press.

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

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

O'Malley, M.H., & Peterson, G.E. (1966). An experimental method for prosodic analysis. *Phonetica, 15*, 1 – 13.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 19, 309–328.

Peterson, G. E. and Barney, H. L. (1952). Control methods used in a study of vowels. *Journal of the Acoustical Society of America, 24*, 175–184.

Pierrehumbert, J. (2002) Word-specific phonetics. In C. Gussenhoven and N. Warner (Eds.), *Laboratory Phonology 7*, pp. 101–139. Berlin: Mouton de Gruyter.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2*, 33–52.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullennix (Eds.), *Talker Variability in Speech Processing* (pp. 9–32). San Diego: Academic Press.

Pisoni, D.B. & Levi, S.V. (2007). Representations and representational specificity in speech perception and spoken word recognition. In M.G. Gaskell (Ed.), The Oxford Handbook of Psycholinguistics, pp. 3–18. Oxford University Press: UK.

Posner, Michael I. (2003). Imaging a science of mind. *Trends in Cognitive Sciences, 7(10)*, 450–453.

Potter, R., & Steinberg, J. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America, 22*, 807–820.

Potts, B.C., Melara, R. D., & Marks, L. E. (1998). Circle size and diameter tilt: A new look at integrality and separability. Perception & Psychophysics, 60, 101–112.

Pufahl, A. & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. Cognitive Psychology, 70, 1–30.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rakerd, B. & Verbrugge, R. R. (1987). Evidence that the dynamics information for vowels is talker independent in form. *Journal of Memory and Language, 26*,558–563.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception & Performance, 23,* 651–666.

Salvata, C, Blumstein, S.E., Myers, E. B. (2012). Speaker Invariance for Phonetic Information: an FMRI Investigation. Language and Cognitive Processes, 27(2), 210–230.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2*, 110–114. https://doi.org/10.2307/3002019

Schacter, D. L., & Church, B. A. (1992). Auditory priming and explicit memory for words and voices. Journal of Experimental Psychology: Learning, Memory, & Cognition, 18, 915–930.

Shankweiler, D., Strange, W., & Verbrugge, R. (1977). Speech and the problem of perceptual constancy. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing* (pp. 315–345). Hillsdale, NJ: Erlbaum.

Sheffert, S. M. & Fowler, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. Journal of Memory and Language, 34, 665–685.

Singmann, H., Bolker, B., Westfall, J., Aust, F. & Ben-Shachar, M. S. (2020). afex: Analysis of Factorial Experiments. R package version 0.27–2. https://CRAN.R-project.org/package=afex

Souza, P. E., Gehani, N., Wright, R. A., & McCloy, D. R. (2013). The advantage of knowing the talker. Journal of the American Academy of Audiology, 24(8), 689–700.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America, 85,* 2135–2153.

Syrdal, A. K. and Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79*, 1086–1100.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics, 77*, 1674–1684.

Traunmuller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America, 69*, 1465–1475.

Turvey, M. T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review, 80*, 1–52.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters, part I: Recognition of backward voices. *Journal of Phonetics, 13*, 19–38.

Vitevitch, M.S. (2003). Change deafness: The inability to detect changes in a talker's voice. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 333–342.

Wong, P.C.M., Nusbaum, H.C., & Small, S.L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*, 1173–1184.

Woods, D. L, Stecker, G.C., Rinne T,. Herron T.J., Cate, A.D., Yund, E.W., Liao, I., & Kang, X. (2009). Functional maps of human auditory cortex: Effects of acoustic features and attention. PLoS One 4:e5183.

Yantis, S., Schwarzbach, J., Serences, J. T., Carlson, R. L., Steinmetz, M. A., Pekar, J. J., Courtney, S. M. (2002). Transient neural activity in human parietal cortex during spatial attention shifts. *Nature Neuroscience, 5(10)*, 995–1002.

Zhang, C. & Chen, S. (2016). Towards an integrative model of talker normalization. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 1252–1268.