単音節知覚および話者同定における話者の親近性の効果
— 家族の声と他人の声の比較 –

ジェームス マグナソン [1]、 山田玲子 [1]、 ハワード ナスバウム [2]

[1]619-02, 京都府相楽郡精華町光台 2-2, ATR 人間情報通信研究所
(magnuson@hip.atr.co.jp, yamada@hip.atr.co.jp)

[2] シカゴ大学心理学部
(h-nusbaum@uchicago.edu)

あらまし　　話者適応の文脈へのチューニング理論(contextual tuning theory)では作動記憶に蓄えられた話者性情報を利用することにより音声の認識が促進されると述べている [9]。本論文では長期記憶に蓄えられた親近性の高い話者の声の情報が利用されているかどうか [12, 10] という点について検討を行った。具体的には、話者同定課題、単音節同定課題、単音節検出課題における家族の声と他人の声の比較を行うことにより、親近性の高い話者の声を聞く際に適応過程を必要とするかどうか調べた。その結果、話者同定課題および品質を劣化させた単音節同定課題において話者に対する親近性が成績を促進することが明らかになった。しかし、単音節検出課題では一貫して単一話者が呈示される方が複数話者が呈示される場合より反応時間が短かく、話者の親近性に関わらず典型的な話者適応 [9] が観察された。これらの結果は、長期記憶に蓄えられた話者情報は話者や単音節の同定を促進するが、話者適応の初期過程が完結するまでは利用されないことを示唆する。本結果の話者適応および話者同定に関する理論への寄与について述べる。

和文キーワード　　音声知覚、話者変動、話者同定

Variability in Familiar and Novel Talkers: Effects on Mora Perception and Talker Identification

James S. Magnuson[1], Reiko A. Yamada[1], and Howard C. Nusbaum[2]

[1]ATR Human Information Processing Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan
(magnuson@hip.atr.co.jp, yamada@hip.atr.co.jp)

[2]University of Chicago Department of Psychology, 5848 S. University, Chicago, IL 60637, USA
(h-nusbaum@uchicago.edu)

Abstract　　Contextual tuning theories of talker normalization state that listeners can use information about a talker's vocal characteristics stored in working memory to facilitate recognition of that talker's speech [9]. We investigated whether people can use information about a familiar talker's voice, stored in long-term memory [12, 10], in the same way. That is, can subjects circumvent normalization processes when listening to highly familiar talkers, such as familiy members? We found that familiarity with a talker's voice facilitated performance in a talker identification task, and also in a mora identification task with degraded stimuli. However, in a monitoring paradigm that typically results in faster performance in single-talker than multiple-talker conditions[9], we found the typical normalization effect for both familiar and unfamiliar talkers. Thus, while information about talkers that listeners have in long-term memory can be used strategically to facilitate, e.g., segment identification, that information is not available until the initial processes of talker normalization are complete. We discuss the implications of the results for theories of talker normalization and talker identification.

英文 key words　　speech perception, talker variability, talker identification

1 Introduction

Much of the work on perceptual normalization of talker differences and talker identification has proceeded in mutual isolation. A recent exception is Johnson's theory of talker-dependent, exemplar-based systems for talker identification and vowel identification [1]. Theories which relate talker identification and speech perception may be more parsimonious than post-hoc attempts to integrate separate theories developed in isolation.

However, the cues used to recognize voices may vary from talker to talker, and in some cases the best cues to talker identity are contained in higher-level structure than the information most relevant for segment identification. By playing samples of famous voices backwards, Van Lancker et al. [12] demonstrated that the effects of distorting information about syllable structure, temporal relations, and phonetic cues on the ability of listeners to identify talkers are different for different talkers; for some, the effect is negligible, but for others identification accuracy falls dramatically.

Thus, there is reason to doubt that listeners use the same information for identifying talkers and recognizing the utterances produced by those talkers. However, Nygaard, Sommers and Pisoni [10] have shown that familiarity with a talker can facilitate performance in a word identification task. Nygaard et al. trained subjects to identify a set of 10 talkers. At the end of nine days of training, subjects who had reached an accuracy criterion of 70% in the talker-identification task were more accurate at transcribing speech produced by the trained-on talkers than speech produced by talkers they had not heard before when the speech was presented in noise. However, fewer than 50% of the subjects reached the accuracy criterion, and subjects who had not reached the criterion did not show a facilitation effect for trained-on talkers. This suggests that a very high degree of familiarity is required before the representation of a talker can be used to facilitate speech perception.

In this paper, we report the results of three experiments designed to determine whether listeners can use their knowledge of highly-familiar talkers' vocal characteristics to circumvent talker normalization processes. In the first experiment, we tested whether subjects can tune to highly familiar talkers (family members) more quickly than to unfamiliar talkers. In the second experiment, we verified that subjects could identify their family members' voices more accurately than voices they were trained to identify in the experimental context. In the final experiment, in order to compare the effects of experimental training and long-term experience with voices on identification, we asked subjects to transcribe moras presented in noise that were produced by talkers that were highly familiar (family members), that subjects had been trained to identify, or that subjects had heard but not been trained to identify.

2 Experiment 1: Normalization

Nusbaum and Morin [9] presented subjects with vowels, CV and CVC syllables, and words in a speeded-target monitoring task. Subjects saw an orthographic representation of a target, and were instructed to hit a key whenever they heard that target among a set of distractors played through headphones. Nusbaum and Morin used two talker-variability conditions: in the blocked-talker condition, all stimuli were produced by a single talker; in the mixed-talker condition, utterances from at least two talkers were presented in random order. Subjects were consistently slower (by approximately 25 ms) to respond in the mixed-talker condition than in the blocked-talker condition for each sort of stimulus. This "normalization effect" (which also interacts with cognitive load) is thought to result from the time it takes to compute a representation of talker characteristics which enables appropriate mappings from acoustics to percepts. When the talker does not change, the representation is held in working memory and can be referenced more efficiently than talker characteristics could be recomputed for every sample of speech, which results in a performance advantage in the blocked-talker condition. In other words, given a constant context of talker characteristics, listeners can "tune" to a talker and constrain the amount of processing necessary for recognition.

If the representations of talkers stored in long-term memory for talker identification are compatible with the (hypothesized) process of contextual tuning, we might expect that those representations could be referenced in less time than it takes to compute a representation for talker normalization. A listener might be able to avoid recomputing talker characteristics when the talker changes from one highly familiar talker to another.

We followed the procedure developed by Nusbaum and Morin [9] for speeded-target monitoring, using familiar talkers (family members) and unfamiliar talkers to determine whether or not long-term memory representations of familiar talkers can be referenced in time to avoid computing talker characteristics after a talker change.

2.1 Method

2.1.1 Stimuli

We recorded two parents and one or two children from seven Japanese families reading lists of Japanese moras (consonant-vowel sequences). Adults and older children read a list of 100 moras. Younger children read a 45 item subset of the full list. Stimuli 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. Each stimulus was hand-edited so that there was a minimum of silence at the beginning and end of each utterance, and average RMS amplitude was digitally normalized.

2.1.2 Subjects

Both adults from the six of the seven families recorded participated in Experiment 1. All of the subjects were native speakers of Japanese with no history of hearing or speech disorders.

2.1.3 Procedure

We used the monitoring paradigm described by Nusbaum and Morin (1992). A speeded-target monitoring task was used and hit rate, false alarm rate, and response times were calculated. 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. Stimuli were presented on-line to subjects seated at NeXT workstations over STAX Lambda headphones.

In each trial, subjects heard a sequence of sixteen moras. Zeroes were added to the end of each stimulus so that there was 830 ms between the onsets of moras. Trials were separated by 3000 ms of silence, during which a message appeared on the screen to alert subjects that the target mora was changing. Four target moras were randomly positioned among twelve distractors, with these constraints: targets could not be first in a trial, targets could not be last in a trial, and targets had to be separated by at least one distractor.

Four of the moras served as targets (*bo, gu, ki,* and *pa*) and sixteen as distractors (*be, bu, ga, go, ji, ka, ko, me, mu, na, ni, pe, pi, ri, ro,* and *zo*). The target moras *bo, gu, ki,* and *pa* were also used as distractors when they were not chosen as the target.

Each subject listened to four talkers in 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, UaUc, FaUa, FaUc, FcUa and FcUc. Presentation order of blocked-talker and mixed-talker trials across subjects was controlled with a Latin square design.

2.2 Results and discussion

We performed analyses of variance on two forms of the data. First, hit rate, false alarm rate and response time were organized by talker pair for blocked- and mixed-talker conditions. Although there were no reliable differences in hit or false-alarm rates (hit rates were above 94% for all talker pairs in both blocked and mixed conditions; false alarm rates were below .05%), subjects were reliably faster to respond to targets in the blocked-talker condition than in the mixed-talker condition, for both familiar and unfamiliar talkers ($F(1,9)=22.822$, $p=.001$; see Figure 1). The size of this effect is consistent with the results of previous uses of this paradigm with native speakers of American English (e.g., [9], [6]). The interaction of talker pair by talker condition was nearly significant ($F(5,45)=2.333$, $p=.058$), due to the lack of any difference between blocked and mixed conditions for the FcUc talker pair (see Figure 2).

The second analysis of variance was performed with the data organized by familiarity (familiar or unfamiliar), talker age (adult or child), and talker condition (blocked or mixed). Again, there were no effects on accuracy or false alarm rates. While there was not a main effect of familiarity ($F(1,10)=.006$, $p=.939$), there was an effect of talker age (with subjects faster to respond to targets produced by adult talkers; $F(1,10)=15.270$, $p=.003$) and interactions between talker age and condition (the difference between RT on children and adults is larger in blocked than in mixed condition; $F(1,10)=8.236$, $p=.017$), and talker age and familiarity ($F(1,10)=6.350$, $p=.030$). It appears that the effect of talker age is due to a large difference in the time it takes to respond to unfamiliar adults and unfamiliar children (but leaves us with the question of why subjects should be able to respond so much faster to unfamiliar adults than familiar adults and children). Figure 3 shows that the effect of condition is largest for familiar and unfamiliar adults and that the effect of condition on familiar and unfamiliar children is quite small. A possible explanation for the small effect of talker condition on child talkers (as well as the lack of an effect for talker pair FcUc) is that the vocal characteristics of the familiar and unfamiliar children may be much more similar than the vocal characteristics of the familiar and unfamiliar adults (see [6] for a discussion of when small differences between talkers do and do not result in normalization effects). Some of the children also tended to prevoice voiced consonants relatively longer than adults, and some also varied their speech rate more than adults, which could be confounding factors.

This experiment replicates previous results with native speakers of another language (American English), and extends them to address the question of whether or not familiar and unfamiliar talkers require the same processing time attributed to a process of talker normalization. There is no observable advantage in normalization for familiar talkers (e.g., there is no advantage of the FaFc condition over any of the others). It seems that listeners are still computing the talkers' vo-

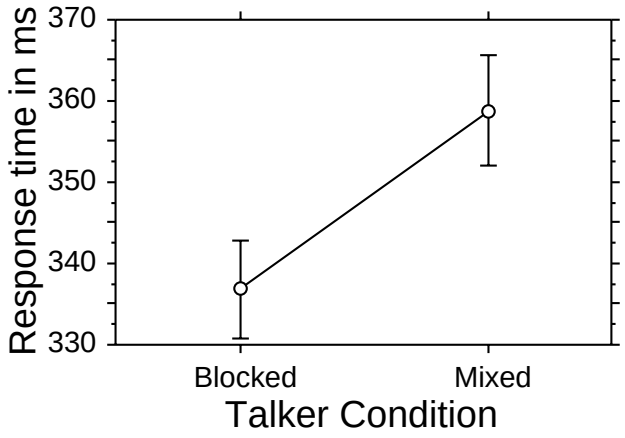

Figure 1: Effect of talker condition in Experiment 1 (bars represent standard error.)

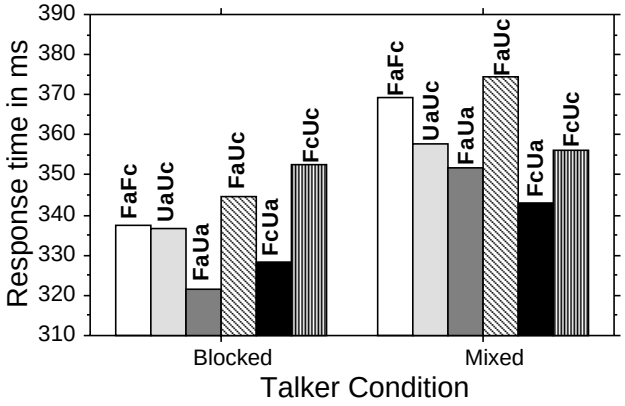

Figure 2: Interaction of talker pair and talker condition in Experiment 1.

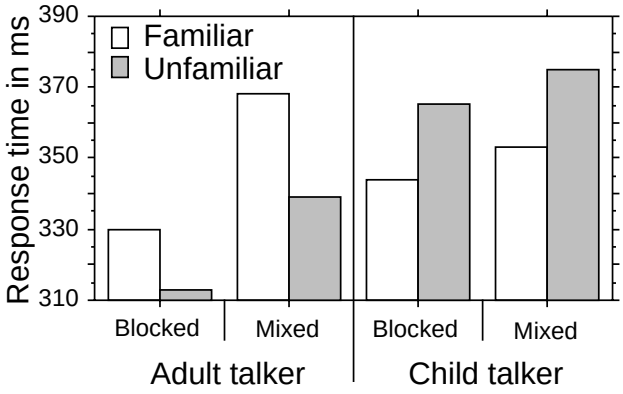

Figure 3: Interaction of familiarity, talker age and condition in Experiment 1.

cal characteristics even when the talkers are highly familiar. Thus, it appears that familiarity with a talker's voice does not change the initial processes of talker normalization.

3 Experiment 2: Talker Identification

Most of the previous perceptual studies of talker identification (or discrimination) have used much longer stimuli than those we used in Experiment 1 (e.g., 2-4 s [12], 6-120 s [4]). The lack of an advantage for familiar versus unfamiliar talkers, and the typical normalization effect for a monitoring task (slower RT in mixed than blocked condition) for unfamiliar and familiar talkers may be due to the fact that the stimuli were so short (on the order of a few hundred ms) that subjects would not have been able to identify the familiar talkers. It is also possible that subjects were able to develop representations of the unfamiliar talkers during the course of the experiment. Recent research indicates that fairly detailed representations of talker characteristics are encoded without conscious effort, even during a lexical-decision task, and are available for later cued recall of spoken words [11, 5].

Experiment 2 was designed to verify that subjects were able to identify the familiar talkers, and examine how well subjects could identify new voices after relatively small amounts of training. Subjects were trained to identify two new unfamiliar adults and two new unfamiliar children. Then they were tested on how well they could identify the familiar and unfamiliar talkers.

3.1 Method

3.1.1 Subjects

The same subjects who participated in Experiment 1 participated in Experiment 2.

3.1.2 Stimuli

Three subsets of the mora set recorded for Experiment 1 were used. 5 moras were used for familiarization, 10 for training, and 20 for testing. For each subject, stimuli were produced by the familiar adult and familiar child they heard in Experiment 1, as well as two new unfamiliar adults and two new unfamiliar children. The unfamiliar talkers were of the same sex as the familiar talkers for each subject, and were chosen to have a measured average fundamental frequency within approximately 10 Hz of the appropriate familiar talker.

3.1.3 Procedure

Stimuli were presented on-line to subjects seated at NeXT workstations over STAX 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 moras. For each trial, subjects had to choose between 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 stimulus was repeated and subjects answered again. This was repeated for each stimulus until subjects answered correctly. Subjects heard two repetitions of six stimuli from each of the talkers.

The next three blocks were for training. First, subjects had 20 trials from each of the two unfamiliar adults only (2 repetitions of 10 items), and then from the two unfamiliar children only (2 repetitions of 10 items). Stimuli were presented randomly so that the talker also varied randomly from trial to trial. The stimuli 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 stimuli from all four unfamiliar talkers presented in random order (2 repetitions of 10 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, 1 repetition of 2 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, and feedback was eliminated. The practice block consisted of two stimuli from each talker, chosen randomly from the list of items used in the familiarization block and presented in random order. The test block used 2 repetitions of 20 new items produced by each of the six talkers presented in random order.

3.2 Results and Discussion

Subjects learned to identify the new unfamiliar talkers fairly well based on training with relatively few (30) mora tokens (M = 75% for unfamiliar adults in testing, M = 84% for unfamiliar children). Performance for familiar talkers was also high (M = 92% for familiar adults, M = 83% for familiar children). This suggests that the use of relatively short stimuli should not have been the cause of the lack of familiarity effects in Experiment 1 (despite the the similarity in accuracy for familiar and unfamiliar children, which we will discuss shortly). A comparison of these results to previous results for 5 talkers in a discrimination task (familiar or unfamiliar) [4], where accuracy was only around 70% for 6 s stimuli, suggests that our feedback method was effective (or our task, featuring two highly familiar talkers, was much easier).

We performed analyses of variance with data organized by familiarity and talker age, with accuracy and response time as dependent measures. While there were no reliable effects of familiarity or talker age on accuracy – although on average subjects were more accurate on familiar talkers (M = 88%) than unfamiliar talkers (M = 80%) – the interaction between familiarity

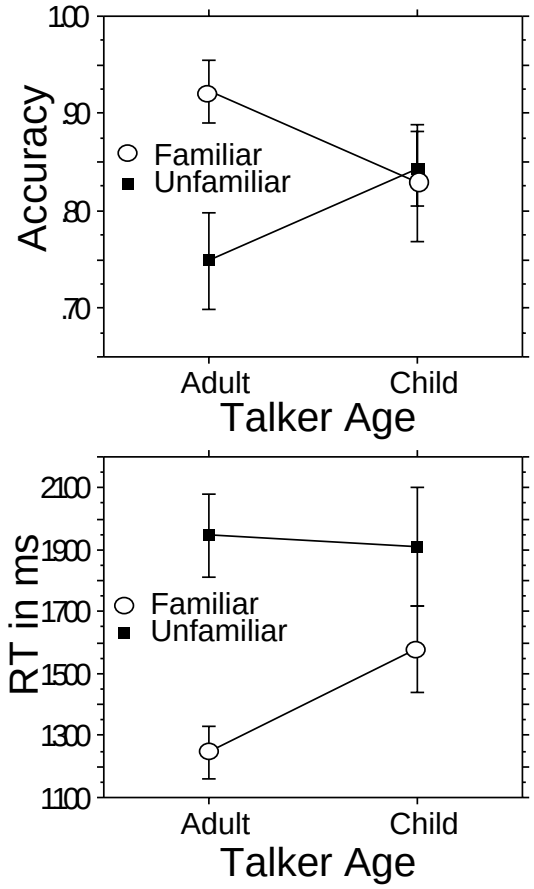

Figure 4: Interaction of familiarity and talker age on accuracy (top panel) and response time (bottom panel) in Experiment 2 (bars represent standard error).

and talker age was significant ($F(1,10)=6.186$, $p=.032$). In the top panel of Figure 4 you can see that subjects were much better at identifying familiar adults than unfamiliar adults, but there was not much difference between familiar and unfamiliar children.

The analysis of response time revealed a strong effect of familiarity. Subjects were faster to respond to stimuli produced by familiar talkers than unfamiliar talkers ($F(1,10)=17.686$, $p=.002$; see Figure 4, bottom panel). Subjects were faster to respond to adults (M = 1650 ms) than children (M = 1764 ms), but not significantly so ($F(1,10)=3.214$, $p=.103$). The interaction of familiarity and talker age was nearly significant ($F(1,10)=4.846$, $p=.052$); see Figure 4, bottom panel).

The interaction of familiarity and talker age demonstrates that although subjects are not more accurate at recognizing familiar children than unfamiliar children, when they do recognize them, they are faster to respond – perhaps because they are more confident of their response. This could be due to larger variability in the children's utterances; it is sometimes difficult to elicit constant prosodic patterns when recording children. Or it could be that identifying familiar and unfamiliar talkers in this task required different numbers of steps. First, subjects must decide whether the talker is an adult or a child. Then subjects may decide whether the talker is familiar or not. For familiar talkers, the process ends here. For unfamiliar talkers, an additional discrimination is required, which may explain the constant latency between 1900 and 2000 ms required for unfamiliar adults and children.

4 Experiment 3: Talker identification and Mora Identification

It is possible that the lack of familiarity effects in Experiment 1 is due simply to subjects being unable to retrieve information about talker identity (which we know they have, from the results of Experiment 2) from memory quickly enough. In that case, advantages of long-term memory representations of talker characteristics may only appear in higher-level tasks. For example, recognizing the voice of a familiar talker with an odd accent from a short initial sample of speech may aid recognition of characteristic productions. Distinctive structural characteristics could also aid recognition in degraded conditions. Kato and Kakehi [3, 2] has demonstrated that trained transcribers are able to tune to the voices of talkers when listening to degraded speech over the course of approximately 3 to 5 mora samples.

In Experiment 3, we tested the possibility that subjects could use knowledge about talkers in a higher-level task than the one we used for Experiment 1. We presented degraded speech 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, that subjects had never been asked to identify). In addition, stimuli were presented in two talker conditions, as in Experiment 1: *blocked* and *mixed*. With this manipulation, we attempted to replicate the "tuning" phenomena reported by Kato and Kakehi [3, 2].

4.1 Method

4.1.1 Subjects

At the time of this writing, three of the six families that participated in Experiments 1 and 2 had participated in Experiment 3 (for a total of six subjects).

4.1.2 Stimuli

For talker identification training and testing, the stimuli were the same as those used for Experiment 2. The same talkers (Fa, Fc, Ua1, Uc1, Ua2, and Uc2) assigned to subjects for Experiment 2 were assigned in Experiment 3.

For each subject, the stimuli for mora identification were produced by each of 6 talkers: Fa, Fc, Ua1, Uc1, Ea and Ec. "Ea" and "Ec" were a pair of talkers subjects had been *exposed* to in an earlier experiment: the unfamiliar talker pair from Experiment 1 (Ua, Uc). All subjects heard the unfamiliar pair they had heard in Experiment 1 as the pair of *exposed-to* talkers. This allowed us to compare the effects of simple exposure to talkers in the experimental setting with the effects of explicit talker identification training.

For mora identification in noise, we used 15 of the stimuli used for talker identification testing and 15 stimuli that had not been used in Experiments 1 and 2. This allowed us to compare performance on *old* and *new* stimuli. If we observed a performance advantage for *familiar* and *trained-on* talkers, it would be possible that the advantage was due to instance-specific memories for particular stimuli. If there was no performance

advantage for *old* stimuli, we could be sure that other effects are due to experience with talkers rather than particular stimuli.

In order to avoid ceiling effects on accuracy, we made the stimuli for mora identification noisy by randomly selecting 10% of the samples of each stimulus and changing the signs of the values of these samples. This resulted in a sufficient level of degradation that the stimuli were moderately difficult to identify. In addition, such degradation preserves the amplitude envelope of the stimuli and is not as fatiguing to listen to as, e.g., broad-band white noise.

4.1.3 Procedure

There were seven parts to the experiment. First, subjects were refamiliarized to the same four unfamiliar talkers they had heard in Experiment 2 (Ua1, Uc1, Ua2, and Uc2). The refamiliarization block was identical to the familiarization block used in Experiment 2, except that only 2 stimuli per talker were used. Second, subjects were retrained on the four unfamiliar talkers. This retraining was identical to the training session used in Experiment 2, although the stimuli were in new, randomly-generated orders. Third, subjects 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, subjects were given a talker identification pretest. The pretest was identical to the test used in Experiment 2, except that only 15 stimuli per talker were used. Then subjects transcribed moras presented in noise. We will explain this in greater detail in the next paragraph. Finally, subjects were given a talker identification posttest.

The mora identification in noise phase consisted of two blocks. In one block, 30 stimuli from each talker were presented consecutively (*blocked*-talker condition). After 30 stimuli from one talker, the 30 stimuli from the next talker followed immediately. In the other block, the same set of stimuli were completely randomly-ordered (*mixed*-talker condition). Subjects were seated at NeXT workstations. At the beginning of each trial, the trial number appeared on the screen. Then the stimulus was played. There was a two-second inter-trial-interval during which subjects were to transcribe what they had heard onto a numbered answer sheet. As mentioned in Section 2.1.3, zeroes were added to the end of each stimulus such that each was 830 ms long. Thus, the interval between stimulus onsets was 2830 ms.

4.2 Results

4.2.1 Talker Identification: Testing

We conducted an ANOVA comparing performance on the test from Experiment 2, the pretest in Experiment 3, and the posttest in Experiment 3. For the six subjects tested in Experiment 3, accuracy increased with each additional test ($F(2,10)=5.412$; $p=.026$; see Figure 5). As can be seen in Figure 5, accuracy on *family* talker pairs did not improve from test-to-test since it was initially very high. By the time of the posttest in Experiment 3, accuracy was nearly as high on the *unfamiliar* talker pairs, although performance varied much more for those pairs (standard deviation and standard error were approximately double those for familiar talkers).

There was also a significant effect of familiarity ($F(2,10)=4.191$; $p=.048$). Subjects were much more

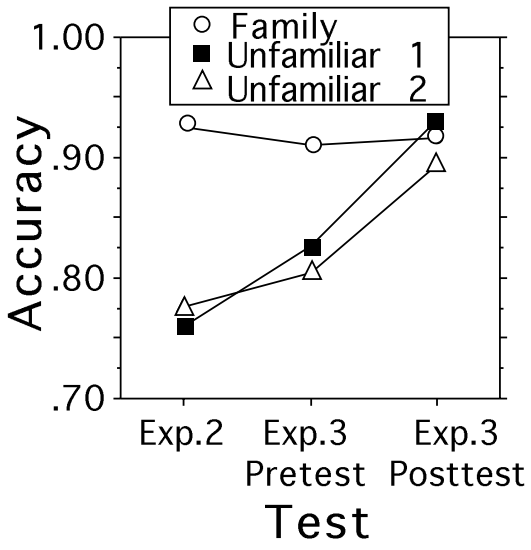

Figure 5: Interaction of familiarity and test in talker identification. The effect of test was significant, but the interaction of familiarity and test was not. Familiarity is included to illustrate the differences between talker pairs. The pair "Unfamiliar 1" differs from "Unfamiliar 2" in that Unfamiliar 1 was also heard in the mora identification task.

accurate at identifying the voices of their family members (M=92%) than the voices of the unfamiliar talkers they also heard in the mora identification (Ua1 and Uc1, M=84%) and the voices of the unfamiliar talkers they only heard in the identification training (Ua2 and Uc2, M=83%).

Subjects were significantly more accurate at identifying the voices of children (M=93%) than adults (M=80%; $F(1,5)=7.164$; $p=.044$). An interaction of familiarity and age reveals that, as in Experiment 2, this effect was due to poor performance on unfamiliar adults ($F(2,10)=4.709$; $p=.036$; see Figure 6).

4.2.2 Mora Identification

Talker condition (*blocked* vs. *mixed*) An ANOVA revealed that subjects were significantly more accurate when stimuli were *blocked* by talker (M=67%) than when the talker changed randomly from trial to trial (M=48%; $F(1,5)=9.102$; $p=.03$).

Familiarity (*family* vs. *trained-on* vs. *exposed-to*) The effect of familiarity was nearly significant ($F(2,10)=3.74$; $p=.061$). Subjects were more accurate at identifying moras produced by their family members (M=64%) than unfamiliar adults they had been trained to identify (M=58%) and talkers they had heard before but had not been trained to identify (M=49%; see Figure 7.

Age (*adult* vs. *child*) In contrast to the talker identification results, subjects were significantly more accurate at identifying *adults* (M=62%) than *children* (M=53%; $F(1,5)=8.682$; $p=.032$).

Stimulus condition (*old* vs. *new*) There was no advantage of *old* items over *new* items.

Trial-by-trial tuning We did not observe the sort of trial-by-trial tuning Kato and Kakehi [3, 2] reported, although accuracy did tend to increase from the beginning to the end of *blocked* sessions. The lack of a clear trend was probably due to the small number of subjects

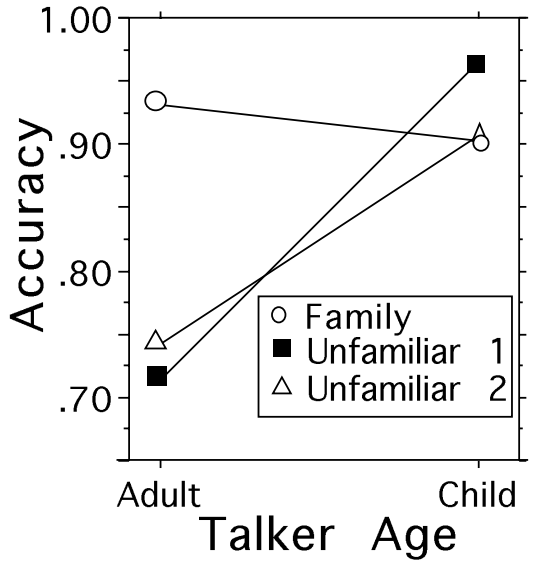

Figure 6: Interaction of familiarity and talker age in the talker identification portion of Experiment 3. The pair "Unfamiliar 1" differs from "Unfamiliar 2" in that Unfamiliar 1 was also heard in the mora identification task.

we have run so far. However, it is clear from the effect of talker condition that subjects were able to exploit stable talker characteristics in the *blocked* condition.

4.3 Discussion

Even after a break of 8 or more weeks between Experiments 2 and 3, subjects quickly reached and exceeded the level of talker identification accuracy of Experiment 2 in Experiment 3. There was also a considerable increase in accuracy from pretest to posttest in Experiment 3, suggesting that subjects' representations of the individual talkers' characteristics were reinforced even during talker identification and mora identification sessions. Familiarity was also correlated with accuracy in the mora identification phase of the experiment. However, we cannot be sure that the advantage of *trained-on* talkers over *exposed-to* talkers was due to training rather than simple exposure. We are planning experiments designed to control for exposure versus training effects.

It is clear, though, that in the talker identification training subjects were attending to phonetic detail at a level sufficient to aid them in the subsequent mora identification task. It is also important to note that the information subjects were using for talker identification were not purely phonetic: although subjects were much more accurate at identifying the voices of children than those of adults, they were less accurate at identifying moras produced by children. This confirms the results of Van Lancker et al. [12], who found that the qualities that made talkers distinctive varied from talker to talker. To borrow an example from Terrence Nearey, the voices of Popeye (with low F0 and high formants) and Julia Child (with high F0 and low formants) are distinctive, but the qualities that make them highly-distinctive talkers do not necessarily make them highly-intelligible talkers. While what we know about a familiar voice may enable us to identify the source, it may not always be useful for acoustic-to-phonetic processing.

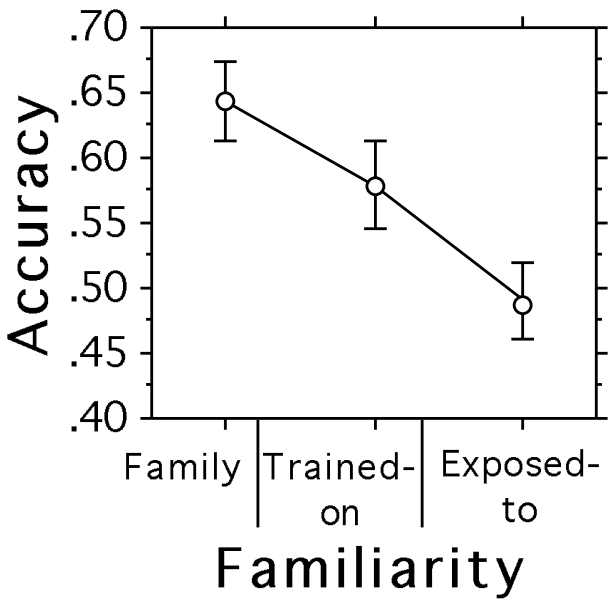

Figure 7: The nearly-significant effect of familiarity in the mora identification portion of Experiment 3.

In fact, the significant effect of talker condition (*blocked* versus *mixed*) in mora identification suggests that using knowledge about a familiar talker in a difficult perception context is a strategic process which requires attention. In accordance with a contextual tuning theory of talker normalization, subjects were much more accurate when stimuli were blocked by talker than when stimuli from different talkers were mixed. It seems that as long as the talker remained constant, subjects were able to focus their attention on the phonetic details most relevant for identifying moras. That is, they were able to exploit talker stability to constrain the amount of attention they allotted to analyzing vocal characteristics (as opposed to phonetic details which were not necessarily correlated with the qualities used to identify talkers). We plan to conduct further studies using cognitive load manipulations to test this hypothesis further.

Thus, the results of Experiment 3 show that the perceptual processes of speech perception can be quickly adapted to exploit temporary stability. Also, it appears that there are multiple processes used for talker identification, since distinctiveness does not always predict intelligibility, but does in some cases (e.g., as a function of familiarity). That is, qualities that make some talkers distinctive may facilitate acoustic-phonetic processing, but this is not always true.

5 General discussion

The three experiments discussed here show that, although representations of highly-familiar talkers in long-term memory facilitate accuracy and speed of talker identification, as well as accuracy at identifying speech in noise, those representations cannot be referenced in order to circumvent the response-time effect resulting from talker variability examined in Experiment 1.

Subjects were slower to respond when the speech of even highly-familiar talkers was mixed than when speech was blocked by talker. The exception of the FcUc (familiar child – unfamiliar child) talker pair in

Experiment 1 may be due to greater overall vocal similarity of the children used in the study. Indeed, there was not an accuracy advantage for familiar children in the identification task, although there was a response time advantage. This suggests that larger subsets of the familiar and unfamiliar talkers' utterances were confused when the talkers were children. However, even when the familiar and unfamiliar children were discriminable, sufficient similarity between the talkers could explain the lack of an effect for mixing the talkers from the talker pair FcUc – see [9, 6, 7] for evidence that some highly-discriminable talker pairs are similar enough in vowel space and average F0 that they do not require separate calibration.

The results of Experiment 1 suggest that the long-term representations of familiar talkers' vocal characteristics do not appear to be useful in reducing the time it takes to recognize speech when that speech is produced by a mix of talkers. If the increase in time were due to competition between talker identification and speech recognition (as suggested by Mullennix and Pisoni, [8]), the effect of mixing talkers on recognition speed should have been reduced for the familiar talkers because, as Experiment 2 demonstrated, familiar talkers aree identified substantially faster than unfamiliar talkers. The lack of an effect or interaction between familiarity and recognition processing in the mixed-talker case strongly suggests that the increased recognition time is due to the process of normalizing for the differences between talkers rather than talker identification.

The results of Experiment 3 show that familiarity with the voice of a talker (if not necessarily the ability to identify the talker) facilitates segment identification of degraded speech. The significant accuracy improvement when stimuli were blocked by talker indicates that the effect results from strategic deployment of attention to phonetic detail (rather than assessing vocal characteristics) when the talker remains constant.

The accuracy advantage for *trained-on* versus *exposed-to* talkers indicates that subjects were able to develop representations of talker characteristics that could be used to facilitate segment identification after relatively small amounts of training. The increase in accuracy on the talker-identification tests – even after a several-week interval between training sessions – demonstrated that the representations developed in a one-hour experiment, based on relatively few tokens from the *trained-on* talkers, were available from long-term memory. In addition, the absence of an accuracy advantage for *old* versus *new* stimulus items in mora identification suggests that the representations developed in training were abstract, rather than instance-specific. Finally, the lack of a strong correlation between talker distinctiveness and talker intelligibility suggests that, as Johnson [1] hypothesized, there may be multi-modal, talker-specific representations of vocal characteristics in long-term memory. Once the talker has been identified, if the representation includes predictions about phonetically-relevant details, it can be used strategically to facilitate segment identification.

To summarize, the results of the experiments reported here indicate that: (1) representations of talker characteristics in long-term memory cannot be accessed quickly enough to circumvent the initial processes of talker normalization (assessing vocal characteristics), (2) however, stability in source identity can be determined without a complete analysis of talker characteristics (since there is a response-time advantage when stimuli are blocked by talker rather than mixed), supporting a contextual theory; (3) humans can develop long-lasting representations of vocal- characteristics from limited talker-identification training, which can be used strategically in a mora-identification task, (4) although the qualities that make a voice distinctive do not necessarily make it more intelligible.

Acknowledgments

We thank Dr. Hideki Kawahara and Inge-Marie Eigsti for comments which greatly improved this paper, and Rie Morita for help running subjects.

References

[1] Johnson, K. (1994). Memory for vowel exemplars. J. Acoust. Soc. Am., 95, 2977.

[2] Kakehi, K. (1992). Adaptability to differences between talkers in Japanese monosyllabic perception.. In Y. Tohkura, Y. Sagisaka, and E. Vatikiotis-Bateson (Eds.), Speech Perception, Speech Production, and Linguistic Structure, pp. 135-142. Tokyo: OHM.

[3] 加藤和美、 筧一彦 (1988). 音声知覚における話者への適応性の検討. 日本音響学会誌, 44, 180-186.

[4] Legge, G. E., Grosmann, C., and Pieper, C. M. (1984). Learning unfamiliar voices. J. Experimental Psychology: Learning, Memory, and Cognition, 10, 298-303.

[5] Luce, P. A., and Lyons, E. A. (1994). The representation of voice information in spoken word recognition: Differential effects of repetition in lexical decision and recognition. J. Acoust. Soc. Am., 95, 2872.

[6] Magnuson, J. S. and Nusbaum, H. C. (1993). Talker differences and perceptual normalization. J. Acoust. Soc. Am., 93, 2371.

[7] Magnuson, J. S. and Nusbaum, H. C. (1994). Some acoustic and non-acoustic conditions that produce talker normalization. Proceedings of the 1994 Spring Meeting of the Acoust. Soc. Japan, 637-638.

[8] Mullenix, J. W. and Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics, 47, 379-390.

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

[10] Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (1994). Speech perception as a talker-contingent process. Psychological Science, 5, 42-46.

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

[12] Van Lancker, D., Kreiman, J., and Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters, part I: Recognition of backward voices. J. Phonetics, 13, 19-38.