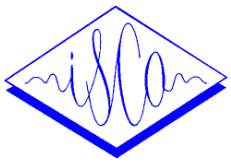

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

ARE REPRESENTATIONS USED FOR TALKER IDENTIFICATION AVAILABLE FOR TALKER NORMALIZATION?

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

[1]ATR Human Information Processing Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan
[2]University of Chicago Department of Psychology, 5848 S. University, Chicago, IL 60637, USA

ABSTRACT

Contextual tuning theories of talker normalization state that listeners can use information about a talker's vocal characteristics stored in working memory to recognize that talker's speech [8]. We investigated whether people can use information about a familiar talker's voice, stored in long-term memory [10], in the same way. That is, whether people can circumvent talker normalization processes when listening to familiar talkers by referencing the representations they use for talker identification. We presented subjects with stimuli produced by familiar and unfamiliar talkers in a monitoring paradigm that typically results in faster performance in a single-talker condition than a multiple-talker condition. We found the typical normalization effect for both familiar and unfamiliar talkers, suggesting that even if talker representations used for identification are compatible with those used for normalization, they cannot be retrieved more quickly than the representations used for normalization can be computed. We verified subjects' ability to identify familiar talkers in a second experiment, and found that familiarity facilitated both accuracy and response time in the identification task. We discuss the implications of the results for theories of talker normalization and 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. Van Lancker et al. [10] demonstrated that the effects of distorting information about syllable structure, temporal relations and phonetic cues by playing samples of famous voices backwards apply differentially to 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. In this paper, we report the results of two experiments designed to determine whether one consistent effect of talker variability holds for familiar as well as unfamiliar talkers.

Nusbaum and Morin [8] 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.

2. EXPERIMENT 1: NORMALIZATION

We followed the procedure developed by Nusbaum and Morin [8] 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 SR-Signature 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. Husbands and wives were assigned the same unfamiliar talkers. 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., [8], [5]). 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).

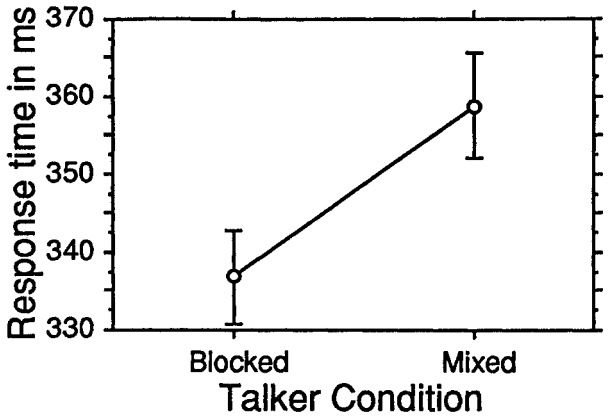

Figure 1. Effect of talker condition in Experiment 1 (bars represent standard error.)

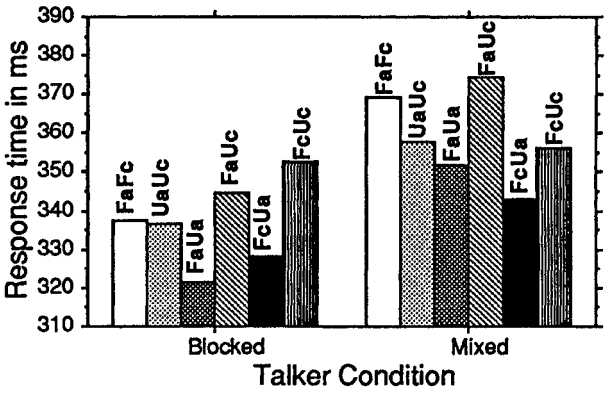

Figure 2. Interaction of talker pair and talker condition in Experiment 1.

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 puzzling 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 unfamil-

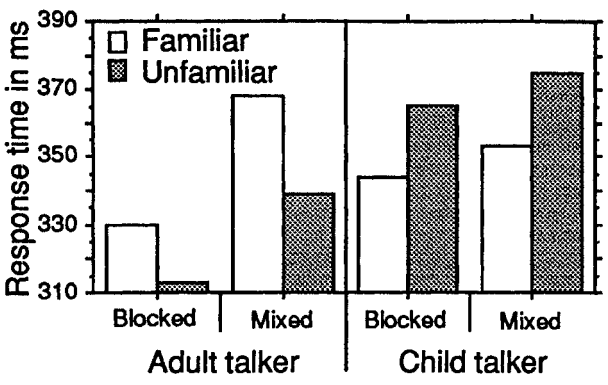

Figure 3. Interaction of familiarity, talker age and condition in Experiment 1.

iar adults (see [5] 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, which could be a confounding factor.

This experiment replicated previous results with native speakers of another language (American English), and extends them to address the question of whether or not familiar 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' vocal characteristics even when the talkers are highly familiar. Thus, familiarity with a talker's voice does not change the initial processes of talker normalization.

3. EXPERIMENT 2: 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 [10], 6-120 s [3]). The lack of an advantage for familiar vs. 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 [9], [4].

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 new subsets of the mora set recorded for Experiment 1 were used. 20 moras were used for familiarization, 20 for training and 40 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.

The next three blocks were for training. First, subjects heard the 20 trials from each of the two unfamiliar adults only, and then from the two unfamiliar children only. 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 20 new stimuli from all four unfamiliar talkers presented in random order. 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) 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 40 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 relatively few (45) 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) [3], 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 and talker age was significant ($F(1,10)=6.186$, $p=.032$). In the left panel of Figure 4 you can see that subjects were much better at identifying familiar adults than unfamiliar adults, but there was not

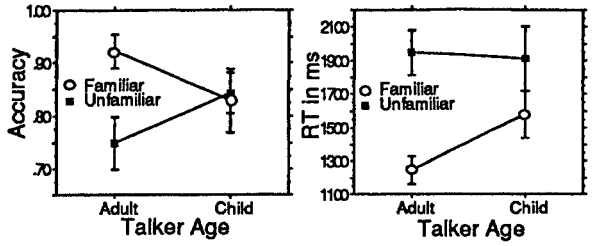

Figure 4. Interaction of familiarity and talker age on accuracy (left panel) and response time (right panel) in Experiment 2 (bars represent standard error).

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, right 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, right 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. GENERAL DISCUSSION

The two experiments discussed here show that, although representations of highly-familiar talkers in long-term memory facilitate accuracy and speed of talker identification, those representations cannot be referenced in order to circumvent the response-time effect resulting from talker variability examined in Experiment 1. Subjects are slower to respond when the speech of even highly-familiar talkers is mixed than when speech is 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 is not an accuracy advantage for familiar children in the identification task, although there is 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 [8], [5] and [6] 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 present results 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, [7]), the effect of mixing talkers on recognition speed should have been reduced for the familiar talkers because, as demonstrated in the Experiment 2, familiar talkers are 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.

It is possible that the advantages of long-term memory representations of talker characteristics may only apply 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. Kakehi [2] has demonstrated that tuning to talkers in degraded speech takes place over the course of approximately 3 or 4 mora samples. The next step for this research is to examine the intersection between talker familiarity and contextual tuning in a context less constrained by time.

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

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.