The Effects of Familiarity with a Voice on Speech Perception. [*]

James S. Magnuson, Reiko A. Yamada (ATR Human Information Processing Laboratories)
Howard C. Nusbaum (University of Chicago, Department of Psychology)

概要

We review progress in an on-going investigation into the relationship between mechanisms of voice identificiation and word recognition. In the first experiment, we found that talker normalization effects occur even when subjects listen to highly familiar talkers (family members). 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. We found that familiarity with a voice predicted accuracy: the more experience subjects had with a talker's voice, the easier that talker's words were to transcribe when presented in noise.

1 Introduction

In this report, we review progress in an on-going investigation into the relationship between mechanisms of voice identificiation, talker normalization, and word recognition.

The variability that exists between talkers presents a substantial problem to theories of speech perception, as well as speech recognition by machine. Because of physiological differences (e.g., vocal tract length, head size, age, sex), social differences (e.g., differences in accent), the way different talkers produce the same speech sound may differ acoustically, and the way they produce different speech sounds may be very similar acoustically. Despite this variability, people seem to effortlessly match the acoustic signals to the proper linguistic percepts.

However, careful experimental designs reveal that people are more efficient (faster and/or more accurate) at speech perception tasks when they listen to only one talker than when they listen to two or more talkers in random alternation. For example, Kato and Kakehi [1] have demonstrated that accuracy gradually increases in transcribing speech when the talker is kept constant, and drops off sharply when the talker changes. Nusbaum and Morin [2] presented subjects with vowels, CV and CVC syllables, and words in a speeded-target monitoring task, in 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" 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. This possibility motivated Experiment 1.

2 Experiment 1: Normalization

The stimuli for all three experiments were drawn from the same database. 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.

Both adults from 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.

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.

Each subject listened to four talkers in a 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 (the subject's spouse), a familiar child (the subject's child), and an unfamiliar adult and an unfamiliar child. Each subject also listened to every possible pairing of the four talkers in a mixed-talker condition, where half the targets and distractors were produced by each of two talkers and randomly ordered.

2.1 Results

Although there were no reliable differences in hit rates (above 94% in all conditions) or false-alarm rates (below

[*] 日本人による米語r/l/ 音知覚における話者変動の効果: 適応か基準設定か？ジェームス マグナソン、山田玲子(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).
Email: magnuson@hip.atr.co.jp, yamada@hip.atr.co.jp.

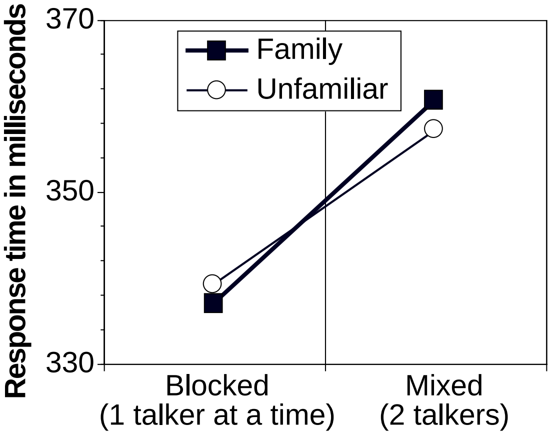

Figure 1. Effect of talker condition in Experiment 1.

.05% in all conditions), 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¡.01; see Figure 1). It seems that listeners are still computing the talkers' vocal 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

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.

In Experiment 2, ten of the subjects who participated in Experiment 1 were first familiarized with the voices of two new unfamiliar adults and two new unfamiliar children. This familiarization was followed by training on the four unfamiliar talkers, practice at identifying all six talkers (familiar and unfamiliar), and a talker-identification test.

Subjects learned to identify the new unfamiliar talkers fairly well based on training with relatively few (30) mora tokens (M = 80%). Performance for familiar talkers was also high (M = 92%). This suggests that the use of relatively short stimuli should not have been the cause of the lack of familiarity effects in Experiment 1.

4 Experiment 3: Talker identification and mora transcription

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. Because several weeks elapsed between Experiments 2 and 3, subjects were refamiliarized with the unfamiliar talkers they had been trained to identify in Experiment 2. Subjects were then tested in their ability to identify the six talkers they had been tested on in Experiment 2.

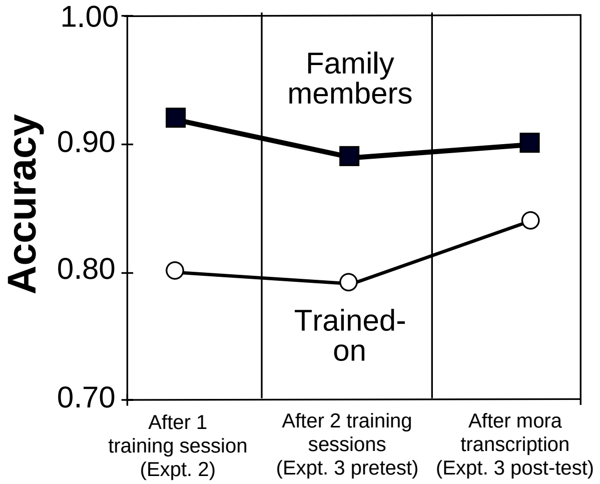

Figure 2. Voice identification accuracy as a function of familiarity in Experiments 2 and 3.

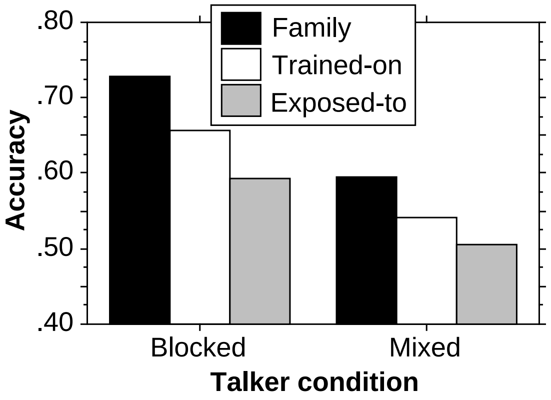

Figure 3. Mora transcription accuracy as a function of talker familiarity in Experiment 3.

Following the identification task, subjects were asked to transcribe moras produced by three different pairs of talkers: *highly-familiar talkers* (the familiar adult and child from Experiments 1 and 2); *trained-on talkers* (one pair of unfamiliar talkers they had been trained to identify in 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*.

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.

Following the transcription task, subjects' ability to identify the six talkers used in Experiment 2 was tested again.

4.1 Results

In figure 2, accuracy in the three voice identification tests are compared. 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 on the *unfamiliar* talker pairs was approaching that on the familiar talkers.

An ANOVA revealed that subjects were significantly more accurate in the mora transcription task when stimuli were *blocked* by talker (M=66%) than when the talker changed randomly from trial to trial (M=55%; F(1,9)=5.42, p¡.05; see Figure 3.

The effect of familiarity was also significant (F(2,18)=3.64; p=¡.05). Subjects were more accurate at identifying moras produced by their family members (M=66%) 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 3).

5 Summary

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.

Acknowledgments

We thank Dr. Hideki Kawahara and Inge-Marie Eigsti for comments which greatly improved this paper.

REFERENCES

[1] 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.

[2] 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.