EPISODIC MEMORY FOR WORDS ENHANCES THE LANGUAGE FAMILIARITY EFFECT IN TALKER IDENTIFICATION

Deirdre E. McLaughlin, Sara C. Dougherty, Rebecca A. Lember, & Tyler K. Perrachione

Department of Speech, Language, and Hearing Sciences, Boston University, USA
tkp@bu.edu

ABSTRACT

Speech perception and talker identification are intertwined. Speech from a single talker is recognized more efficiently than speech from multiple talkers; likewise, in the "language familiarity effect", talker identification is more accurate in one's native language than a foreign one. Models of speech perception implicate episodic memory to explain effects of phonetic variability. We investigated whether these models can also account for the language familiarity effect. Listeners learned to identify voices speaking English and Mandarin in conditions differentially favoring episodic memory: (1) all talkers repeated the same sentences and (2) each talker said completely unique sentences with no repeated words. The language familiarity effect was stronger when talkers' speech had identical content, suggesting that episodic lexical access enhances talker identification in a native language. Foreign-language talker identification did not differ between the conditions, suggesting that episodic memory for voices is filtered by lexical abstraction possible only in a familiar language.

Keywords: talker identification, episodic memory, speech perception, language familiarity

1. INTRODUCTION

Studies of speech perception have established that indexical variability affects speech processing. Talker and rate variation, two sources of variability that affect the phonetic realization of speech, have been shown to significantly affect listeners' reaction times in word or speech sound identification tasks [8,9]. Listeners also have better memory for spoken words when indexical features, such as those associated with different talkers, are consistent between encoding and recognition [2]. Similarly, linguistic processing affects our perception of indexical features, such as our ability to identify talkers by voice. Previous studies of talker identification have demonstrated that listeners are better at identifying voices in their native language than in an unfamiliar foreign language – a phenomenon known as the *language familiarity effect* [6,11]. Although a variety of studies using different methods have consistently demonstrated the language familiarity effect [3,6,10,11], there is not yet a good model of the exact linguistic processes that make native-language talker identification easier. Some authors have argued that the source of this effect is the ability to notice idiosyncratic differences in talker phonetics when compared to abstract phonological representations of words [10], whereas others have argued that mere familiarity with the statistical properties of the acoustic features of one's native language is sufficient to give listeners an advantage [4].

In this study, we designed an experiment to test whether linguistic processing – specifically, lexical access – contributes to the language familiarity effect. Episodic theories of lexical access suggest that listeners' representations of spoken words include information about the specific phonetic realization whenever a word was perceived, including the indexical features related to talker identity [7]. Although sometimes contrasted with talker normalization models of speech perception [7], there is ample evidence for the psychological reality of both processes during typical speech perception [2,9]. Because speech perception and talker identification processes are linked, information stored during one process is likely accessible by the other. If, as episodic models of speech perception predict, listeners form memories of both the abstract word and its specific phonetic realization, then recognizing similarities or differences between the idiosyncratic properties of speakers' voices during talker identification should be facilitated when different talkers say the same words as each other. However, this should only be true in a native language, when there are abstract word forms against which to store these memories.

Three hypotheses follow from the idea that accessing words is important for talker identification: First, listeners will identify voices in their native language more accurately when they are able to compare different talkers' productions of the same words. Second, listeners will not differ in their ability to identify foreign-language talkers when they say the same vs. different words, because they do not have the abstract lexical representations against which they can form and compare memories

of talkers' voices. Third, listeners should demonstrate a larger language familiarity effect (an even greater discrepancy between native- versus foreign-language talker identification) when the content of different talkers' speech is the same than when the content of their speech is unique.

2. METHODS

2.1. Participants

Speakers of American English (N=16, age 18-29, M=20.5 years) completed this study. Inclusion criteria required participants to have a self-reported history free from speech, language, or hearing problems; to perform above chance (20%) in all conditions, and to demonstrate the language familiarity effect (perform better in English than Mandarin). Recruited participants who failed to meet the inclusion criteria were excluded from the study (N=6). Participants provided written informed consent and received monetary compensation for their participation.

Table 1: Examples of sentence stimuli for each language condition.

English Sentences
Granola is best in yogurt.
Policemen chase criminals.
Ships sail to distant shore.
Trolleys roll down busy streets.
Mandarin Sentences
今天的阳光真好 jīn tiān de yáng guāng zhēn hǎo *It's a nice sunny day.*
节假日不用门票 jié jiǎ rì bù yòng mén piào *No ticket is needed during holiday.*
晚上一块去跳舞 wǎn shàng yī kuài qù tiào wǔ *Let's go dancing together tonight.*
对面有两所高中 duì miàn yǒu liǎng suǒ gāo zhōng *There are two high schools across the street.*

2.2. Stimuli

Participants learned to identify talkers from hearing them say short sentences. Examples of the sentence stimuli are shown in Table 1. We generated a corpus of 100 English sentences in which no word was ever repeated within or between sentences. The English sentences were 6-8 syllables long (M=7), and were phonotactically balanced by controlling positional probability of phonemes and biphones [13]. The Mandarin sentences came from a published corpus of 100 phonetically balanced sentences [5], each 7 syllables long, and in which characters were repeated only extremely rarely.

Ten female native speakers of American English (age 20-29, M=23 years) recorded the English sentences, and ten female native speakers of standard Mandarin (age 18-36, M=26 years) recorded the Mandarin sentences. In each language condition, talkers had regionally homogeneous accents. None of the talkers recorded sentences in both language conditions or participated in the experiment. The sentences were recorded at 44.1 kHz in a sound attenuated recording booth and normalized to 65 dB SPL RMS amplitude using Praat. The recorded English sentences were an average duration of 1.66 ± 0.18s and the recorded Mandarin sentences were an average duration of 1.58 ± 0.09s (mean ± s.d).

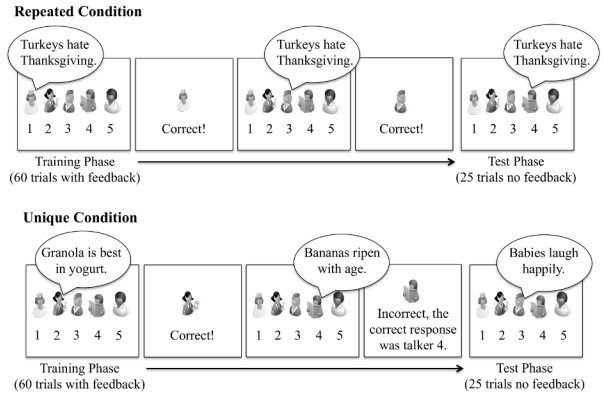
2.3. Procedure

Stimuli were presented to participants in a 2×2 factorial design in which we varied the language being spoken (English vs. Mandarin) and the sentence content (repeated vs. unique). In the repeated condition, the same sentences were spoken by all talkers. In the unique condition, each talker spoke a unique set of sentences, and no words were ever repeated between sentences, whether spoken by the same talker or by a different talker. Sentences were counterbalanced across conditions so that, for a given listener, a sentence was not used in both conditions, but across listeners, all sentences were used in both conditions. In all condition × language combinations, participants learned to identify five different talkers by the sound of their voice. Talkers were represented by both a cartoon avatar and a number (1-5). The procedure is depicted in Figure 1. Participants indicated which talker they heard speaking by pressing the corresponding number on a keypad. Talkers were counterbalanced between the repeated and unique conditions to control for differences in distinctiveness of any voice or combination of voices.

Participants first learned to identify the voices in the *training phase*, in which they heard a talker say a sentence, chose who among the five talkers they thought was speaking, and received corrective feedback indicating whether they had selected correctly or who the correct response should have been. They were tested on their ability to correctly identify the talkers in the *test phase*, in which they

heard a talker say a sentence and chose which of the five talkers they thought was talking, but did not receive any feedback. In the training phase of the repeated condition, listeners heard the same 12 sentences spoken by all talkers. In the test phase of the repeated condition, listeners heard a subset of five sentences randomly selected from the 12 training sentences and spoken by every talker. In the training phase of the unique condition, listeners heard 12 unique sentences spoken by each of the five talkers, such that no talker repeated any words, and the words in each talker's sentences were not in any of the sentences spoken by any of the other talkers. In the test phase of the repeated condition, listeners heard each talker say five new sentences, for a total of 25 completely unique test sentences. Stimulus delivery was controlled with PsychoPy (v1.8.0). All stimuli were delivered to the participants at a comfortable listening level using Sennheiser HD 380 Pro circumaural headphones in a sound attenuated booth.

Figure 1: Schematic representation of the training and test phase in each condition.



2.4. Data analysis

Participants' accuracy on the test phase of each condition was analyzed in R using the generalized linear mixed effects model for binomial data implemented in the package 'lme4'. Fixed factors in the model included Language (English vs. Mandarin) and Condition (Repeated vs. Unique); a maximal random effects structure included within-participant intercepts and slopes and within-talker intercepts [1]. Other inferential statistics were conducted by calculating participants' mean accuracy in each condition (the number of trials in which the talker was identified correctly out of the total number of trials); proportional data were arcsine transformed prior to parametric statistics [12].

3. RESULTS

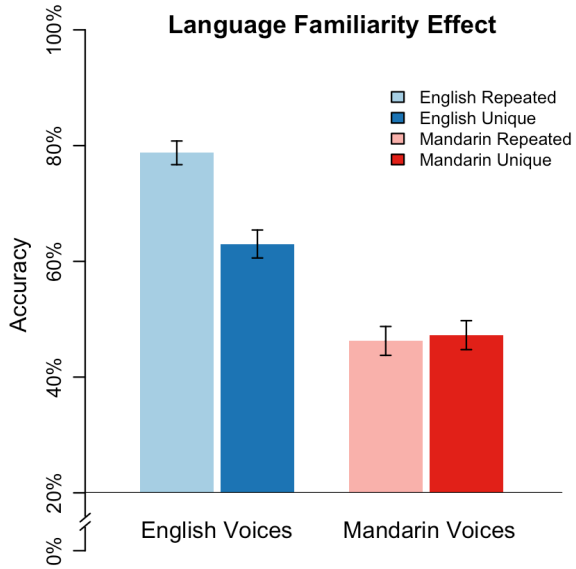
3.1. Effects of lexical repetition on talker identification

There was a significant effect of condition (z=4.22, p=2.5×10^{-5}), such that participants were more accurate at identifying talkers when they heard all talkers say the same sentences, and less accurate when the content of each talker's speech was totally unique.

However, there was also an interaction effect between language and condition (z=3.26, p=0.0011), such that the effect of condition was greater in English than in Mandarin. Therefore, in order to fully understand the effect of lexical repetition on talker identification in a native vs. foreign language, we repeated the generalized linear mixed effects model, but on the data from each language separately.

In English, listeners identified talkers more accurately when they all said the same sentences (z=4.23, p=2.4×10^{-5}). However, the Mandarin talkers were not identified more accurately regardless of whether they all said the same sentences or all said unique sentences (z=0.23, p=0.82).

Figure 2: Participant accuracy at voice identification. Error bars represent the standard error of the mean.



3.2. Effects of lexical repetition on the language familiarity effect

There was both an overall effect of language in the main model (z=2.48, p=0.013), as well as for each

condition separately (Repeated: z=6.28, p<3.5×10^{-10}; Unique: z=2.20, p=0.028), such that listeners always identified the voices speaking English better than those speaking Mandarin. In order to evaluate the extent to which the language familiarity effect depends on comparing word-level pronunciations across talkers, we calculated the magnitude of each participant's language familiarity effect in each condition as the difference between their performance in the English and Mandarin voices. The effect size (Cohen's d) of the language familiarity effect was 2.61 when talkers all said the same sentences, but only 1.12 when they said unique sentences. Although both of these effect sizes are nominally large, the magnitude of the language familiarity effect (difference between English and Mandarin) was significantly greater when talkers said the same words than when no words were repeated across talkers (t_{15}=3.12, p=0.007).

4. DISCUSSION

In this study, listeners learned to identify talkers from their voice in four conditions that manipulated whether listeners understood the linguistic content of talker's speech and whether repeated words were available for forming and comparing episodic memories. We found that participants were more accurate at identifying talkers in their native language when the talkers all said the same words than when they all said different words. This suggests that listeners' talker identification abilities are enhanced when they can use episodic memories for words in their native language to highlight the distinguishing phonetic idiosyncrasies of talkers.

In contrast, participants performed no differently when identifying talkers speaking in an unfamiliar foreign language, regardless of whether those talkers all said the same words or different words. That is, there was no benefit of hearing repeated speech on talker identification accuracy when the content of that speech was incomprehensible to listeners. This suggests that the enhancement of talker identification abilities made possible by episodic memories for speech requires listeners to be able to access meaningful words.

Finally, the magnitude of the language familiarity effect in talker identification was greater when the content of talkers' sentences was the same than when there were no repeated words between talkers. This indicates that a principal reason for superior talker identification abilities in one's native language is the availability of memory traces of the specific words one is hearing, against which to compare the speech of an attended talker. In short, speech perception and comprehension play a facilitatory role in talker identification from speech in a familiar language, but not in an unknown, foreign one.

5. REFERENCES

[1] Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang*. 68, 255-278.

[2] Bradlow, A. R., Nygaard, L. C., Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Percept. Psychophys*. 61, 206-219.

[3] Bregman, M. R., Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition.* 130, 85-95.

[4] Fleming, D., Giordano, B. L., Caldara, R., Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *P. Natl. Acad. Sci. USA.*, 111, 13795-13798.

[5] Fu, Q. J., Zhu, M., & Wang, X. (2011). Development and validation of the Mandarin speech perception test. *J. Acoust. Soc. Am.* 129, EL267-EL273.

[6] Goggin, J. P., et al. 1991. The role of language familiarity in voice identification." *Mem. Cognition***.** 19, 448-458.

[7] Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychol. Rev.* 105, 251.

[8] Green, K. P., Tomiak, G. R., Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Percept. Psychophys.* 59, 675-692.

[9] Mullennix, J. W., Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Percept. Psychophys*. 47, 379-390.

[10] Perrachione, T. K., Del Tufo, S. N., Gabrieli, J. D. E. (2011). Human voice recognition depends on language ability. *Science.* 333, 595.

[11] Perrachione, T. K., Wong, P. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia.* 45, 1899-1910.

[12] Studebaker, G. A. (1985). A rationalized arcsine transform. *J. Speech Lang. Hear. R.* , 28, 455-462.

[13] Vitevitch, M.S., Luce, P.A. (2004) A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behav. Res. Meth. Instr.* 36, 481-487.