THE EFFECTS OF SPEECH PERCEPTION AND SPEECH COMPREHENSION ON TALKER IDENTIFICATION

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

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

ABSTRACT

Listeners identify talkers more accurately in their native language than an unknown, foreign language in a phenomenon called the *language-familiarity effect.* However, the psychological basis for this effect remains unknown. Some have suggested that the linguistic processes involved in speech perception and comprehension facilitate native-language talker identification. Others have argued that talker identification is independent of linguistic processing and that increased familiarity with statistical properties of speech acoustics is sufficient to produce this effect. We report two experiments investigating whether linguistic processes facilitate talker identification. Experiment 1 reveals that there is no native-language advantage for time-reversed speech, suggesting that acoustic factors do not explain the language-familiarity effect. Experiment 2 reveals that talker identification accuracy improves as a function of the linguistic content in speech, suggesting that both speech perception and comprehension contribute to talker identification. Together, these results demonstrate a true linguistic basis for the language-familiarity effect in talker identification.

Keywords: talker identification, speech perception, reversed speech, nonsense speech, comprehension

1. INTRODUCTION

The language-familiarity effect, first described nearly three decades ago [9], has recently been the focus of considerable research interest in order to understand why native-language voices are identified more accurately than foreign-language ones. Originally described in terms of language-based schemata for voices [4], authors have more recently appealed to models of talker variability in speech perception to explain how speech and talker identification processes may be integrated. A number of studies provide evidence suggesting that core linguistic abilities contribute to the language familiarity effect: Mere repeated exposure to foreign-language talkers without speech comprehension does not attenuate the language-familiarity effect [7], but greater linguistic experience in a foreign-language does [2]. Interestingly, individuals with phonologically-based reading impairment do not show the language-familiarity effect [8], further implicating phonological processing as integral to talker identification. However, recent reports challenge the role of linguistic processes in talker identification [3], suggesting instead that voices are better understood using psychological models of face perception than linguistic models of speech perception. Here, we describe two experiments that assess competing claims in these two views of talker identification, ultimately demonstrating that linguistic processes in speech perception and lexical access work together to enhance talker identification by human listeners.

2. EXPERIMENT 1

Based on subjective judgments of voice similarity from reversed speech, a recently reported study has suggested that speech comprehension is not necessary for the language-familiarity effect [3]. We sought to verify this assertion by objectively assessing listeners' abilities to learn talkers' identity from forward and time-reversed speech in their native language and an unfamiliar foreign one.

2.1. Methods

2.1.1. Participants

Two groups of participants successfully completed this study: native speakers of American English and native speakers of Mandarin Chinese. The English group (N=16) consisted of young adults with no prior exposure to Mandarin. The Mandarin group (N=14) consisted of young adults born in China and who had been in the United States for less than four years. Inclusion criteria required participants to have a self-reported history free from speech, language, or hearing problems and to perform above chance (20%) in all conditions. Because we are interested in the basis of the language-familiarity effect, we included only participants exhibiting this effect (*i.e.*, who perform better in their native language than the other language) in the forward speech condition. Additional participants who were recruited but failed to meet the inclusion criteria were excluded (2 English, 12 Mandarin). Participants gave written informed consent and were paid for participating.

2.1.2. Stimuli

Participants learned to identify talkers from hearing them say short sentences. Examples of the stimuli appear in Table 1. English stimuli came from List 13 of the phonetically-balanced “Harvard Sentences” [6]. Mandarin stimuli came from List 1 of the “Mandarin Speech Perception Test,” a published corpus of phonetically balanced sentences [5].

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. Talkers had regionally homogeneous accents in each language. None of the talkers were recorded in both languages or participated in the experiment. Recordings were made at 44.1 kHz in a sound attenuated recording booth and normalized to 65 dB SPL RMS amplitude. Sentence durations (mean ± s.d.) were 1.83 ± 0.25s in English and 1.58 ± 0.21s in Mandarin. Each recording was additionally time-reversed to produce an incomprehensible but acoustically matched stimulus.

Table 1: Example stimuli used in Experiment 1

English Sentences [6]
Type out three lists of orders.
The harder he tried the less he got done.
The cup cracked and spilled its contents.
Mandarin Sentences [5]
今天的阳光真好 jīn tiān de yáng guāng zhēn hǎo *It's a nice sunny day.*
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.*

2.1.3. 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 comprehensibility of the speech (forward vs. reversed). In each condition, 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). Participants indicated which talker they heard by pressing the corresponding number on a keypad. Talkers were counterbalanced between the forward and reversed conditions in each language to control for differences in the distinctiveness of any voice or combination of voices.

Participants first learned to identify talkers in the training phase, in which they heard one sentence spoken by each of the talkers while the corresponding avatar appeared on the screen. After hearing all the talkers in turn, listeners heard one of them say the sentence again while all five avatars appeared on the screen and indicated which of the talkers was speaking by pressing the corresponding button. Listeners received feedback as to whether they had selected correctly, or who the correct response should have been. This active practice was repeated until all the talkers had said the sentence twice. The training phase was then repeated for the next sentence, and so on in this way until listeners had practiced identifying talkers from five different sentences. The training phase consisted in total of 50 passive exposure trials and 50 trials with feedback.

Listeners were then tested on their ability to correctly identify the talkers in a test phase: they heard each talker say the five training sentences plus five new sentences, and chose which of the talkers they thought was speaking. The test consisted of 50 trials with no feedback given. Stimuli were presented with PsychoPy (v1.8.0) at a comfortable listening level via Sennheiser HD 380 Pro circumaural headphones in a sound attenuated booth. The order of the four conditions was counterbalanced across participants.

2.2 Results

Accuracy on the test phase of each condition was analyzed in R using generalized linear mixed effects models for binomial data, with condition as the fixed factor and a maximal random effects structure including random intercepts and slopes by participant, and random intercepts by talker [1].

Despite a reliable language-familiarity effect for both groups of listeners when identifying talkers from typical, forward speech, neither the English listeners nor the Mandarin listeners exhibited a native-language advantage when identifying talkers from time-reversed speech (Fig. 1).

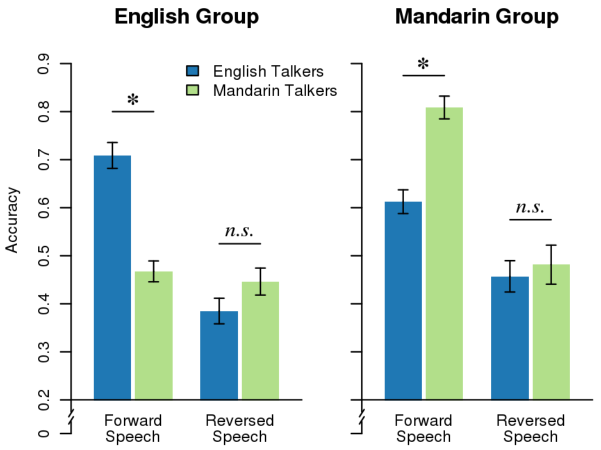
English listeners learned to identify English-speaking talkers significantly more accurately from forward than time-reversed speech (71% vs. 39%; z=10.06, $p<2\times10^{-16}$). In contrast, they identified Mandarin-speaking voices equally accurately for forward and time-reversed speech (47% vs. 45%; z=0.88, p=0.38). Importantly, English listeners were no more accurate learning to identify time-reversed English voices than they were time-reversed Mandarin voices (39% vs. 45%; z=1.22, p=0.22).

Mandarin listeners also learned to identify talkers speaking their native language significantly more accurately from forward than time-reversed speech (81% vs. 48%; z=9.57, $p<2\times10^{-16}$). The Mandarin listeners also showed this effect when identifying

English talkers (61% vs. 46%; z=4.62, $p<4\times10^{-6}$). Importantly, the Mandarin listeners also showed no native language advantage when identifying time-reversed Mandarin versus English voices (48% vs. 46% z=0.33, p=0.74).

Mandarin listeners identified voices more accurately than English listeners for forward speech (z=2.06, p<0.04) – an effect resulting from greater overall accuracy, not a different language-familiarity effect magnitude (no nativeness × group interaction: z=0.05, p=0.96). However, English and Mandarin listeners did not differ in overall ability to identify voices from time-reversed speech (z=1.10, p=0.27).

Figure 1: Listeners who exhibit a reliable language-familiarity effect when identifying talkers from natural speech demonstrate no such advantage when learning to identify talkers from incomprehensible time-reversed speech.



2.3 Discussion

In stark contrast to the large and reliable language-familiarity effect for forward speech, we found no evidence of a native-language advantage when identifying talkers from time-reversed speech. Neither English nor Mandarin listeners were better at identifying talkers in their native language from time-reversed speech. This result conflicts with a recent experiment using subjective judgments of talker similarity, from which it was concluded that the language-familiarity effect is independent of speech comprehension [3]. Conversely, we found that the language-familiarity effect exists *only* when native-language speech is comprehensible.

We also found that, while English listeners did not differ for forward vs. time-reversed Mandarin voices, Mandarin listeners performed better on forward than time-reversed English speech. This may reflect differences in the groups' familiarity with their respective non-native language. Likewise, the Mandarin group appeared slightly more accurate when identifying forward voices, akin to the recent report that pitch perception abilities of Mandarin speakers facilitate talker identification [11].

3. EXPERIMENT 2

Having implicated speech perception processes as important to the language-familiarity effect, we next asked whether talkers were identified more accurately from meaningful speech than from phonologically legal but meaningless speech.

3.1. Methods

3.1.1. Participants

A new sample of native speakers of American English (N=24) completed Experiment 2. Inclusion criteria were the same as Experiment 1. Six additional participants were recruited but, failing to meet inclusion criteria, were excluded from the analysis. Participants gave written informed consent and were paid for participating. No talkers or participants from Experiment 1 also participated in Experiment 2.

Table 2: Example stimuli used in Experiment 2

Phonologically balanced nonword sentences
ðit staʊ hɛsət ri rugəl daɪpəts
ðid əgar ri staɪnɚ ladi ðet
waɪbəl waθ rɪz sonəd hæð bæðl̩
krʌkt kɪpənd θɪtʃ panəst tældi

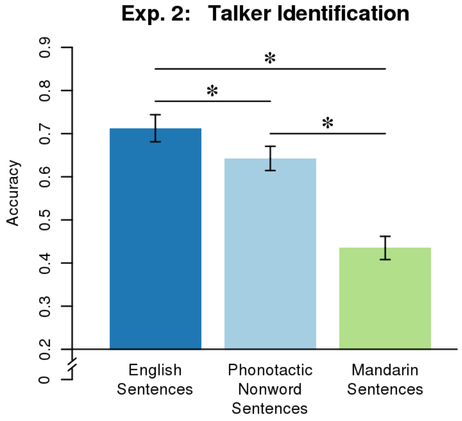
3.1.2. Stimuli

The English and Mandarin sentences from Experiment 1 were used again in Experiment 2. In addition, we developed a set of nonword sentences that were phonetically and phonotactically balanced against the real English sentences (Table 2). We generated these sentences from a broad phonetic transcription of the Harvard Sentences used in Experiment 1. This transcription was then scrambled to produce sentences that contained no real English words, but which were otherwise identical to the English sentences in every way: The English and nonword sentences did not differ in number of phonemes or number of syllables (both p=1.0). All combinations of phonemes in the nonword sentences complied with the rules of English phonology and phonotactics. We controlled the nonword sentences' probabilistic phontactics by matching the mean phoneme and biphone positional probabilities [10] of the nonwords in these sentences with the words in the real English sentences (phones: $F_{1,18}$=1.78, p=0.20; biphones: $F_{1,18}$=0.88, p=0.36). Nonword sentences were produced by the same ten English-speaking talkers as in Experiment 1, all of whom had training in phonetics, were extensively familiarized with the nonword sentences, and were recorded several times to ensure fluency. Nonword sentence recordings were 1.97 ± 0.19s.

3.1.3. Procedure

The procedures of Experiment 1 were repeated for Experiment 2 with the following exceptions: Instead of four factorial conditions, participants learned to identify talkers in three conditions that parametrically varied the similarity between the talkers' speech and listeners' native language: (i) English sentences, (ii) Phonologically balanced nonword sentences, and (iii) Mandarin sentences. Talkers in the English and Nonwords conditions were counterbalanced to control for individual differences in vocal distinctiveness. Likewise, the talkers used in the Mandarin condition were permuted from among the ten recorded. Participants in Experiment 2 learned to identify talkers in each condition during a training phase (50 exposure trials and 50 practice trials with feedback), and their talker identification performance was assessed in a test phase (50 trials with no feedback). The order of conditions was counterbalanced across participants. Statistical analysis procedures were the same as Experiment 1.

Figure 2: Talker identification performance improves with greater linguistic processing



3.2 Results

Participants learned to identify talkers most accurately in the English condition (71.3 ± 15%), followed by the Phonologically Balanced Nonwords condition (64.3 ± 14%), and least accurately in the Mandarin condition (43.5 ± 13%) (Fig. 2**)**. The difference between performance in each of these conditions was significant (English vs. Mandarin: z=4.79, $p<2\times10^{-6}$, Cohen's d=1.98; English vs. Nonwords: z=2.71, p<0.007, d=0.49; Nonwords vs. Mandarin: z=3.75, p<0.0002, d=1.57).

3.3 Discussion

Talker identification improves as a function of the number of linguistic processes that can be brought to bear. Talkers are identified least accurately when listeners recognize no words and are unfamiliar with the sound system (Mandarin speech). Performance improves when listeners are familiar with the types of speech sounds and their distribution, but in which there are no familiar words (nonword sentences). However, talker identification is best from meaningful native-language speech in which listeners are familiar with both the phonetics and lexical content.

4. CONCLUSIONS

These results suggest that the language-familiarity effect in talker identification arises from linguistic processes underlying speech perception and comprehension. When speech is incomprehensible or has unusual phonetic properties (time-reversed speech and foreign-language speech) talker identification is poorest. Familiar sounds in meaningless words facilitate talker identification, but not as much as meaningful speech. That speech perception and lexical access both contribute to the language-familiarity effect affirms an integrated system for speech and voice perception in human talker identification [7,8].

5. REFERENCES

[1] Barr, D., *et al.* (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 255-278.

[2] Creel, S., Bregman, M. (2014). Gradient language dominance affects talker learning. *Cognition*. 130, 85.

[3] Fleming, D., *et al.* (2014). A language-familiarity effect for speaker discrimination without comprehension. *P. Natl. Acad. Sci.* 111, 13795-13798.

[4] Goggin, J., *et al.* (1991). The role of language familiarity in voice identification. *Mem. Cogn*. 19, 448.

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

[6] IEEE. (1969). IEEE recommended practices for speech quality measure-ments. *IEEE Trans. Audio Electroacoust.* 17, 225-246.

[7] Perrachione, T., 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.

[8] Perrachione, T., *et al.* (2011). Human voice recognition depends on language ability. *Science*. 333, 595.

[9] Thompson, C. (1987). A language effect in voice identification. *Appl. Cognitive Psych.* 1, 121-131.

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

[11] Xie, X., Myers, E. (2015). The impact of musical training and tone language experience on talker identification. *J. Acoust. Soc. Am.* 137, 419-432.