Matching Speaking to Singing Voices and the Influence of Content

Zehra F. Peynircioğlu, Brian E. Rabinovitz, and Juliana Repice, *Washington, DC*

Summary: Objectives and Hypothesis. We tested whether speaking voices of unfamiliar people could be matched to their singing voices, and, if so, whether the content of the utterances would influence this matching performance. Our hypothesis was that enough acoustic features would remain the same between speaking and singing voices such that their identification as belonging to the same or different individuals would be possible even upon a single hearing. We also hypothesized that the contents of the utterances would influence this identification process such that voices uttering words would be easier to match than those uttering vowels.

Study Design. We used a within-participant design with blocked stimuli that were counterbalanced using a Latin square design. In one block, mode (speaking vs singing) was manipulated while content was held constant; in another block, content (word vs syllable) was manipulated while mode was held constant, and in the control block, both mode and content were held constant.

Method. Participants indicated whether the voices in any given pair of utterances belonged to the same person or to different people.

Results. Cross-mode matching was above chance level, although mode-congruent performance was better. Further, only speaking voices were easier to match when uttering words.

Conclusions. We can identify speaking and singing voices as the same or different even on just a single hearing. However, content interacts with mode such that words benefit matching of speaking voices but not of singing voices. Results are discussed within an attentional framework.

Key Words: Singing voices–Speaking voices–Linguistic content–Voice perception–Voice identification.

Discrimination and recognition of speaking voices is a well-researched area.[1–3] The prosodic features, intonation, accent, and phonetics of the contents of the utterances are important,[4] and so are the acoustic features comprising the timbre of the voice.[5] These acoustic features are shaped by the physiology of the sound source and the vocal tract, and individual differences in vocal fold size and vocal tracts result in unique voices.[6]

Singing and speaking voices have different acoustic features. But singing voices can still be identified, although the task is harder when compared with the identification of speaking voices.[7] One interesting question that remains unanswered in this realm is whether one can tell the identity of the same voice speaking and singing. In disguising voices, alterations obtained by physical changes in the vocal tract, such as lowering of the larynx and changes to lip and tongue positions, are similar to those that occur during singing.[8] Thus, one may expect a relatively poor performance in matching a person's singing voice to his or her speaking voice. Nevertheless, everyday observations with friends, family members, or famous people whose voices we may have heard many times suggest that we can often recognize their voices whether we hear them speaking or singing. For instance, we have showed in informal demonstrations that many individuals who have watched South Park episodes can recognize that it is Cartman singing even if they have never heard him sing before. Thus, enough acoustic features must remain constant and are also likely to be aided by paralinguistic components for us to identify voices, at least of people that we know.

But can we pick up on these acoustic features that remain constant even without the benefit of having heard these voices many times under different conditions, and thus without a semantic representation of the specific voice? There is evidence that being able to discriminate between unfamiliar voices, which relies on the initial analyses of acoustic features, can be dissociated from being able to recognize familiar voices, which relies more on the identification of patterns of acoustic features.[9] It is also important to note that there is no single salient acoustic feature that aids discrimination even in just speaking voices, and the features most useful for identifying a specific speaker on any one occasion differ from voice to voice.[10] Thus, the task of discriminating whether a singing voice and a speaking voice belong to the same unfamiliar person or to different unfamiliar people might be quite daunting and might need to rely on a more holistic approach. For instance, Susan Boyle of *Britain's Got Talent* fame wowed audiences in 2009 in part because her singing voice was so different from her speaking voice. In her case, the acoustic feature changes from speaking to singing were salient enough to completely disrupt voice identification.

Accurate identification of unfamiliar voices is particularly important in the domain of earwitness testimony. Recent studies have shown such factors as tone of voice, laughter type, and accent to be crucial to such identification.[11–13] Thus, it is important to explore how singing voices are identified in this realm as well. Further, the ability to match singing and speaking voices may shed more light on the different auditory components of a voice that contribute to our ability to form a unique perceptual identification.

Accepted for publication June 13, 2016.
From the Department of Psychology, American University, Washington, DC.
Address correspondence and reprint requests to Zehra F. Peynircioğlu, Department of Psychology, American University, Washington, DC 20016. E-mail: peynir@american.edu
Journal of Voice, Vol. 31, No. 2, pp. 256.e13–256.e17
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2016.06.004

The main question addressed in this study was whether we can in fact match an unfamiliar singing voice to his or her speaking voice. In addition to matching just speaking or just singing voices, participants indicated whether a pair of voices one singing and one speaking belonged to the same person or not. Further, to date, voice recognition or identification studies have often involved comparisons between only isolated vowels or only full speech phrases.[14] The second question addressed in this study was the contribution of content changes to voice identification and how these content changes might interact with the modality changes from singing to speaking. Thus, pairs of to-be-matched voices that were speaking or singing were sometimes both on a single vowel, sometimes both on a speech phrase, and sometimes one on a single vowel and one on a speech phrase.

METHOD

Participants

In the stimulus collection phase, we first recorded voices speaking and singing a set of phrases. We needed 26 different voices uttering each set of items, two to be used in the practice session and 24 to be used in the experiment. To get these 26 voices, we recorded 56 American University undergraduates, regardless of any singing or music training, and then eliminated 30 of the recordings of individuals who were not able to stay on pitch or tempo while singing or individuals whose voice acoustic properties were outliers (as discussed below). To avoid any possible interactions as a function of gender information, only female voices were recorded. Also because listeners have been shown to be able to extract age information from voices,[15] limits were set on the age range of the recorded voices (18–30) to prevent a participant in the main experiment phase from differentiating any given voice based simply on age information (eg, based on an observation such as “this is an older person’s voice or this is a child’s voice and thus different from what I just heard.”) In the main experiment phase, 36 American University undergraduates (also aged 18–30 to match the age range of the voices and prevent any possible other-age bias from emerging) did the voice matching task. All participants received extra credit in psychology courses, $5, or a chance to win a $50 gift certificate.

Materials

Two phrases, “Doe a deer, a female deer” (*The Sound of Music*) and “Somewhere over the rainbow” (*The Wizard of Oz*) were recorded using an Olympus LS-10 digital audio recorder (Olympus). They were spoken as well as sung and also spoken and sung on the vowel /a/. Thus, there were eight utterances recorded by each voice. Two comprised speaking the phrases using the original words of these two songs, two comprised singing these word phrases, two comprised speaking the same phrases but with the vowel /a/ substituted for each syllable instead of the original words, and two comprised singing the same syllable phrases. Because the main factors in recognizing speaking and singing voices have been found to be the rate of speaking, intensity, and the fundamental frequency (F0),[16,17] all recordings were made with an in-ear metronome set at 80 bpm and the loudness was kept constant; thus, although it could be adjusted by the participants during the experiment to fit their comfort level, the level would be the same for all recorded voices and regardless of speaking or singing. Additionally, C4 was set as the starting pitch for all singing. All recordings were also analyzed for their mean fundamental frequency (F0), the first four formant positions (F1 through F4), and skewness of pitch (used to describe how the fundamental frequency of the voices changes over time) using the freeware *Praat* (www.praat.org). The purpose was to catch any obvious outliers (three standard deviations (SD) away) that made any one voice more distinctive than the others. The mean fundamental frequency of all speaking voices was 195.12 Hz (SD = 20.66), approximately a G3, which is 196 Hz. The mean frequencies of the formants were as follows: F1 was 814 Hz (SD = 49.17), F2 was 1302.56 Hz (SD = 103.35), F3 was 2901.80 Hz (SD = 146.16), and F4 was 3860.94 Hz (SD = 289.09).

In the main experiment, stimuli were presented via *SuperLab 4.5*. Participants used Sony headphones (Sony) with adjustable volume controls and the keyboard of a MacBook Pro computer (Apple) to enter their responses.

Design and procedure

There were two variables of interest: mode and content. Mode referred to the voicing of the stimuli, either singing or speaking. Content referred to what was being voiced, either actual words (of the entirety of both of the recorded phrases) or repetitions of /a/ (again corresponding to the entirety of both of the recorded phrases). There were three listening blocks: mode, content, and control (summary of the design; Table 1). In the mode block the comparison that led to the judgment of whether the two voices were the same or different was always between a singing and a speaking voice. The content was held constant such that if the first voice in the pair spoke words, then the second voice sang words as well, and vice versa. If the first voice in the pair spoke the phrase on /a/, then the second voice sang the phrase on /a/ as well, and vice versa. In the content block the comparison that led to the judgment of whether the two voices were the same or different was between phrases on words and on /a/. The mode was held constant such that if the first voice sang, the second voice also sang, and if the first voice spoke, the second voice also spoke. In the control block, both mode and content were kept the same in that the utterance by the second voice was in the same mode (speaking or singing) and had the same content (words or syllables) as the utterance by the first voice. It is important to note that, in all blocks, the comparisons within each pair were between *different* phrases (ie, if the first phrase in a pair was from *The Sound of Music*, then the second phrase was from *The Wizard of Oz*, and vice versa) so that the comparison would not be reduced to simple template matching. The order within pairs in the mode and content blocks (whether the word or the vowel came first or whether the singing or the speaking came first) was randomized in each block within the constraint that there were equal numbers of each and counterbalanced across participants.

Each block comprised 16 comparisons where half of the pairs were voiced (spoken or sung) by the “same” person and half were voiced by “different” people in each comparison. To achieve this,