PHONETIC CONTENT INFLUENCES VOICE DISCRIMINABILITY

Attila Andics[1, 2], *James M. McQueen*[2], *Miranda van Turennout*[2, 3]

[1]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[2]FC Donders Centre for Cognitive Neuroimaging, Nijmegen, The Netherlands
[3]Behavioural Science Institute, Nijmegen, The Netherlands
Attila.Andics@mpi.nl, James.Mcqueen@mpi.nl, Miranda.vanTurennout@fcdonders.ru.nl

ABSTRACT

We present results from an experiment which shows that voice perception is influenced by the phonetic content of speech. Dutch listeners were presented with thirteen speakers pronouncing CVC words with systematically varying segmental content, and they had to discriminate the speakers' voices. Results show that certain segments help listeners discriminate voices more than other segments do. Voice information can be extracted from every segmental position of a monosyllabic word and is processed rapidly. We also show that although relative discriminability within a closed set of voices appears to be a stable property of a voice, it is also influenced by segmental cues – that is, perceived uniqueness of a voice depends on what that voice says.

Keywords: voice, segment perception, speaker discrimination and identification

1. INTRODUCTION

Behavioural and neuroscientific studies indicate that voice processing and speech processing are partly independent, but interact at an early stage of processing (e.g., [2]). One example of this interaction is the demonstration of early voice-specific effects on fricative perception ([1,3]). But the other direction of the interaction – whether voice-specific segmental information contributes to voice processing – has been studied less extensively. Remez et al. [5], using sinewave replicas of speech, demonstrated that speaker-specific phonetic information can in certain cases be sufficient for talker identification. But does segmental information contribute to the efficiency of discrimination of natural voices?

We investigated possible segmental effects on voice discrimination from the listener's perspective and from the speaker's perspective. First, we explored whether phonetic content influences the voice discrimination performance of listeners. Second, we examined whether segmental cues influence the relative discriminability of different voices. One can find a voice that is more or less distinguishable from other voices, but does this depend on what words the voices say?

These questions were addressed in a voice discrimination experiment. Dutch listeners were presented with a list of Dutch CVC words, spoken by Dutch speakers, and were asked to decide whether each word was spoken in the same or a different voice as the preceding word. Segmental content was controlled using eight words which were made by factorially combining two onset consonants, two vowels, and two coda consonants.

2. METHOD

2.1. Participants

Twelve native Dutch listeners with no known hearing disorders participated.

2.2. Stimuli

Thirteen speakers were chosen. To reduce nonsegmental (e.g., fundamental frequency) variability of the voices, the speakers were selected from a relatively homogenous group: young male non-smoking native speakers of Dutch with no recognizable regional accents and no speech problems (age range: 18-30). Segmental overlap between the words was systematically varied using the words *met* [mɛt], *mes* [mɛs], *mot* [mɔt], *mos* [mɔs], *let* [lɛt], *les* [lɛs], *lot* [lɔt] and *los* [lɔs]. The recordings were sampled at 44100 Hz, 16 bits per sample. Average amplitude was equalized over all stimuli. Average syllable duration was 565 ms.

2.3. Procedure

Stimuli were presented via headphones binaurally, at a standard, comfortable listening level. To make the task harder, stimuli followed each other at a relatively fast pace (2400 ms between syllable onsets), and a pink noise was presented after each syllable (from 600 ms till 2400 ms after every syllable onset).

Subjects were instructed to listen to two-minute long blocks of these CVC words. A same/different forced-choice one-back task was used. Listeners had to decide whether the word they heard was pronounced by the same voice as the preceding word or by a different voice. That is, listeners had to make a decision after every syllable they heard, except for the first one within each block. Assignment of left and right index fingers to same and different buttons was balanced across subjects. The experiment lasted 51 minutes, excluding a short practice session and self-paced breaks between blocks.

2.4. Design

Stimulus presentation was blocked by word, so within one block only one of the eight words appeared. One block consisted of 53 stimuli (that is, 52 comparisons), and there were 24 such blocks. Every listener heard all possible voice pairings for each of the eight words during the experiment. To balance response biases as much as possible, half of the voice comparisons required a “same” response and half of them a “different” response. To achieve that equal distribution, every same-voice pair was presented six times per word, and every different-voice pair was presented exactly once per word. There were at most three same or different pairs in a row. To ensure that responses were based on voice processing rather than auditory change detection, six different utterances of each word from each speaker were used, each of these utterances appeared only twice during the experiment, and these two identical stimuli were always separated by at least one full block. Stimulus ordering was otherwise random and varied across listeners. Altogether 1248 responses were collected per listener.

3. RESULTS

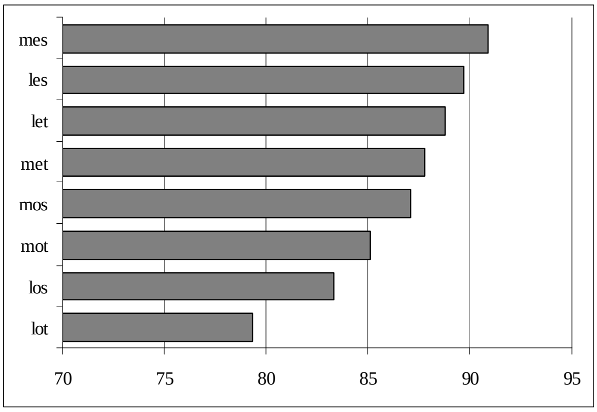
3.1. Overall performance

Overall proportion of correct responses was 87.2%, with a similarly high proportion for same-voice pairs (88%) and different-voice pairs (86.5%). Individual overall hit rates varied between 78.7% and 94.7%, ranging from a responder with a strong “same” bias (98.6% for same-voice pairs and 60.1% for different voice-pairs) to a responder with a clear “different” bias (70.1% for same-voice pairs and 98.6% for different-voice pairs). This listener bias was independent of phonetic content. Average response time was 799 ms for same-voice pairs and 855 ms for different-voice pairs.

3.2. Hit proportion per word and per segment

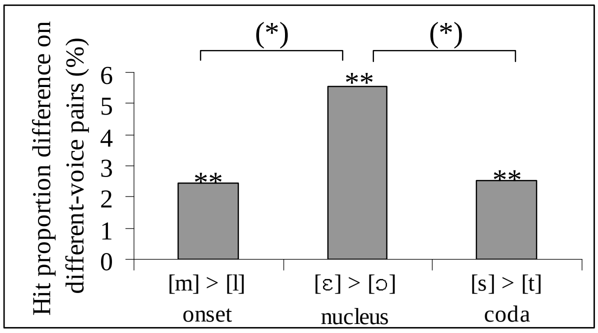
Phonetic contributions to voice discrimination performance were investigated by comparing responses for each word. There were differences in the hit proportion of responses to different-voice pairs between words (see Fig. 1), ranging from 79.3% for [lɔt] to 90.9% for [mɛs].

Figure 1: Same or different voice? Hit proportion of responses to different-voice pairs per word (% correct)



The nature of the CVC stimuli made it possible to examine this effect at the segmental level by varying one of the segmental positions while collapsing across segments in the remaining two positions. Fig. 2 shows the segmental contributions to the voice discriminating benefit of [m] in onset position, [ɛ] in vowel position and [s] in coda position over [l], [ɔ] and [t] respectively. Note that these benefit effects are highly significant for all segmental positions (paired samples t-tests, two-tailed, $p < 0.005$) and that the nucleus change seems

Figure 2: Segmental contribution to voice discrimination performance

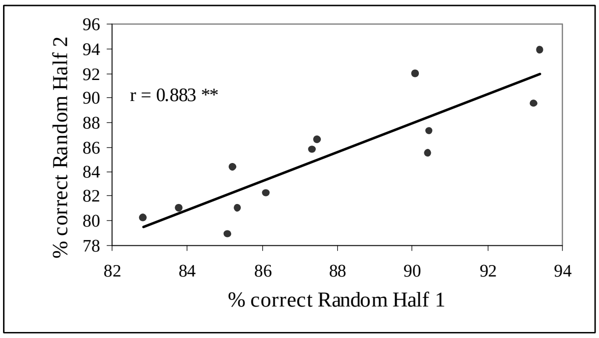


to make a greater difference than either the onset or the coda change (paired samples t-tests, two-tailed, marginally significant: p = 0.065 for nucleus versus onset, p = 0.086 for nucleus versus coda).

3.3. Hit proportion per voice

Discriminability of a voice was investigated by comparing the hit proportion of responses to different-voice pairs for each voice. This measure was calculated by collapsing different-voice trials for each voice across all pairs in which that voice was a member. This way we gained a perceptual rating of the thirteen voices, ranging from the voice which was the most difficult to distinguish from the rest (81.5% correct) to the voice which was the most easily discriminable from the other voices (93.7% correct). To check the reliability of this rating, the same perceptual measure was calculated after randomly splitting the listeners into two groups. Fig. 3 shows the high positive linear correlation of two ratings of voices based on data from these two random halves of the set of listeners (r = +0.883, p < 0.01).

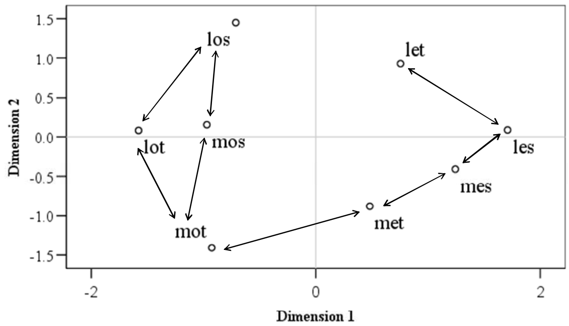
Figure 3: Correlation of hit proportion per voice between two random halves of listeners (% correct)



3.4. Multidimensional scaling of words

To investigate the possible effect of segmental cues on the perceived discriminability of a voice, the discriminability ratings of voices described above were also calculated separately for each word. The correlation coefficient of voice ratings for two given words was considered to be a proximity measure (the higher the correlation, the closer the ratings based on those words are). Inversion of this proximity measure results in a distance measure. Distances were calculated for every word pair (the smaller the distance, the closer the words are with respect to their contribution to voice discriminability). We then performed a multidimensional scaling of the words based on those distances (SPSS ALSCAL using a Euclidean distance model; stress = 0.098, RSQ = 0.919). Fig. 4 shows the resulting two-dimensional map. To illustrate the segmentally determined nature of this map, arrows are added to connect words that corresponded simultaneously to a small perceptual distance (neighbours in the map) and to a minimal physical distance (one segment difference). Note that there are many such arrows. This suggests that voice discriminability is strongly determined by segmental properties. If that were not the case, there would have been fewer arrows, or none at all.

Figure 4: Multidimensional scaling of words based on the similarity of their effect on voice discriminability



4. DISCUSSION

4.1. The naturalness of voice discrimination

Listeners were presented with blocks of voices uttering one of eight CVC words and they had to compare each words' vocal identity to that of the previously heard word. All listeners performed far above chance level. This indicates that voice discrimination is an extremely robust ability of human listeners that is readily applicable even in an attentionally demanding and unnatural task.

Interestingly, many listeners had a considerable response bias either for the "same" or for the "different" response, but this effect disappeared after collapsing data over all listeners. Therefore, this variability does not seem to be caused by an inherent biasing factor in the experimental design, but rather by individual variation in how conservative a given listener is when setting up categories for new voices.

4.2. Phonetic content influences voice discrimination performance

Phonetic contribution to listeners' performance was investigated by comparing the hit proportion of responses to different-voice pairs for each word. We found a higher proportion of correct voice discriminations for words containing an onset [m] versus [l], a nucleus [ɛ] versus [ɔ] and finally a coda segment [s] versus [t]. These differences suggest that the phonetic content of speech affects the listener's voice discrimination performance, and this effect is not restricted to certain segmental positions within a CVC word.

Two important observations have to be made here. First, vowel change seems to make the greatest difference, since its effect is marginally higher than the effect of any of the consonant changes. This suggests that vowels may vary more than consonants in the amount of paralinguistic information that they can carry. Further research is required, however, to test whether the present results generalize to other vowels and consonants.

Second, segmental variation in the coda position makes a significant difference to voice discrimination performance. This indicates that listeners do not always make their decisions based on the vowel or based on the first two segments only, but rather they use all segments of a word before making a "same voice" or "different voice" decision. If we now put this result together with the listeners' average response times, we can see that vocal identity information extracted from the coda position is applied quite rapidly: the most acoustic energy of the coda segment is situated around 300-500 ms after syllable onset, and average response time for different-voice pairs is 855 ms, meaning that listeners are able to apply phonetic information to distinguish between voices in less than half a second.

4.3. Discriminability is a stable property of a voice

By comparing proportion of responses to different-voice pairs across voices, we obtained discriminability ratings for every voice. The high correlation of these voice ratings suggest that discriminability, at least relative to other voices within a closed set, is a stable property of a voice. That is, a voice's discriminability rating is independent of individual listener's biases.

Additional analyses (not reported in detail here) examined the correlation between hit proportions on same-voice and different-voice pairs. They showed that utterances of voices that are less discriminable are also less identifiable, that is, they were perceived as the same voice less consistently than the utterances of more discriminable voices.

The discriminability ratings reported here may thus reveal the prototypical organization of voices. In keeping with the nature of prototypically organized categories in for example phonetic categories [4], voices close to the hypothesized prototype-voice are perceived as less discriminable than voices further from the prototype, independently of the individual listener.

4.4. Segmental cues affect the discriminability of voices

Although discriminability of a voice is relatively independent of individual listener biases, it need not be independent from the segmental information that the voice carries. Our results indicate that segmental cues do have an effect on the perceived discriminability of a voice. We presented a multidimensional scaling of the eight words that were used in the experiment, based on the similarity of their effects on the voice discriminability ratings (see Fig. 4). The large number of arrows connecting perceptual neighbours that have a one-segment step distance suggests that this map of words is at least in part structured by segmental cues. That is, certain phonetic contents make some voices more and some other voices less discriminable than what one would expect on the basis of their overall discriminability. In short, perceived typicality or uniqueness of a voice depends on what that voice says.

5. REFERENCES

[1] Eisner, F., McQueen, J.M. 2005. The specificity of perceptual learning in speech processing. *Perception & Psychophysics* 67, 224-238.

[2] Knösche, T.R., Lattner, S., Maess, B., Schauer, M., Friederici, A.D. 2002. Early Parallel Processing of Auditory Word and Voice Information. *NeuroImage* 17, 1493-1503.

[3] Kraljic, T., Samuel, A.G. 2007. Perceptual adjustments to multiple speakers. *Journal of Memory and Language* 56, 1: 1-15.

[4] Kuhl, P.K. 1991. Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics* 50, 93-107.

[5] Remez, R.E., Fellowes, J.M., Rubin, P.E. 1997. Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance* 23, 651-666.