EFFECTS OF NASALITY AND UTTERANCE LENGTH ON THE RECOGNITION OF FAMILIAR SPEAKERS

Julien Plante-Hébert[1], Victor J. Boucher[2]

Laboratoire de sciences phonétiques de l'Université de Montréal[1, 2]
julien.plante-hebert@umontreal.ca[1], victor.boucher@umontreal.ca[2]

ABSTRACT

The present study examines the effects of nasality and utterance length on memory of familiar speakers using the technique of voice line-ups. With this technique, presented speakers have similar speech F0, dialect, and age range, and they utter the same material. Sets of voice line-ups were elaborated each containing 10 male voices (1 target "familiar" voice and 9 "filler" voices). In each set, speakers produced given utterances of four different lengths, with varying numbers of nasal sounds. Participants (n = 44) were selected on the basis of their familiarity with the target voice. They were asked to identify the familiar voice within line-ups. The results show that both utterance length and nasality positively influence voice recognition but these effects only begin after hearing four or more syllables. This suggests that speaker recognition requires a few syllables and may not operate as quickly as processes of visual recognition.

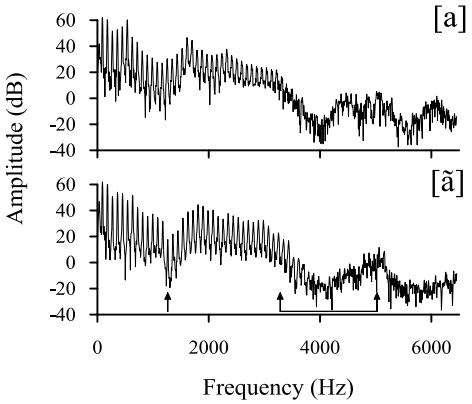
Keywords: voice recognition, memory of speech voice line-ups, forensic phonetics, exemplar theory

1. INTRODUCTION

Research has established that listener's memory of spoken material is not limited to linguistic features but involves "sensory episodes", which include information relating to individual characteristics of voices and speech along with utterance context (see [7] and [8]). Moreover, it is known that there is an incremental learning of these sensory episodes (e.g. [6]), which explains why one can recognize *familiar* voices and speech. In terms of the acoustic information that listeners may store, studies of speaker identification generally refer to two types of attributes of speech signals, which as such involve different processes. The first type relates to the spectral attributes of voices, whereas the second type bear upon the temporal characteristics relating to speech articulation. Though these different acoustic attributes co-occur, in speech they can variably influence speaker recognition.

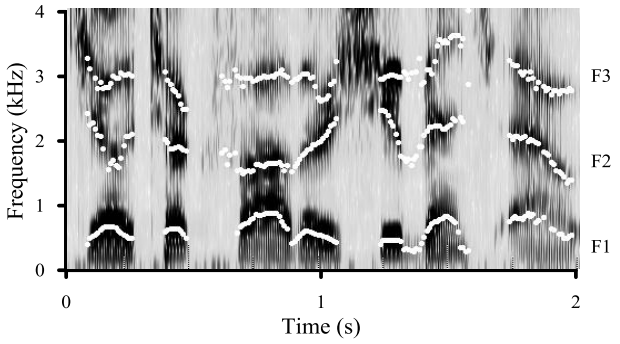
In particular, it has been suggested that spectral information relating to nasality can have a greater effect on the recognition of voices than oral productions (e.g. [1]). Acoustically, the difference between "nasal" and "oral" sounds basically involves characteristic antiresonances, such as those illustrated in Figure 1. These antiresonances occur because of a coupling with the nasal cavities and hence may provide more information on speaker identity than oral sounds - though this has not been confirmed in controlled contexts.

Figure 1. Spectra of the French oral [a] and nasal [ã] showing approximate location of antiresonances (arrows).



As for the temporal attributes that refer to speech motions, these relate directly to changes in spectral aspects of speech in the course of an utterance. The effect is seen in Figure 2 and is distinctly marked by changes in formant frequencies (especially F2). However, research on the effects of these attributes bears some confusion on the time frame that is used by listeners in recognizing a speaker. On this issue, studies in forensic phonetics often refer to work by Bricker [3] and Pollack [14], which concluded that it is the number of articulatory configurations present in the stimuli that can increase identification rates and that utterance length itself is an accessory factor. On the other hand, producing a "number of articulatory configurations" logically entails producing numbers of syllables or segments implying changes in utterance length.

Figure 2. Spectrogram showing that information on articulatory motion is reflected in formant changes in the course of producing an utterance.



Hence, one would expect that identification rates based on articulatory information are dependent on utterance length to some extent. Yet, in the context of usual speech, the length of utterance that listeners need to recognize a familiar voice has not been determined.

With the aim of clarifying the effects of spectral and temporal information on listener's memory of speakers, the present study focuses on the claimed effects of nasality and utterance length. To examine these factors, we refer to the voice line-ups. In this technique, an individual is presented with limited series of similar voices and is asked whether he/she recognizes a particular voice. The use of this technique in paralegal contexts has led to established guidelines [4, 9, 11, 13]. Thus, in creating a voice line-up, one selects speakers with similar speaking F0 and age, who have the same regional dialect, and who produce the same sequences of sounds. While applying these guidelines, the following experiment included in each line-up a voice that was familiar to the listener and speech stimuli reflecting common phrases used in answering or greeting someone in a phone conversation. The guiding hypothesis was that, since nasal sounds offer comparatively more information on the individual attributes of speakers, their presence in utterances would add to dynamic articulatory information as reflected in utterance length effects.

2. METHODOLOGY

2.1. Participants

The listeners that participated in the present experiment were 15 males and 29 females (n = 44) aged between 18 and 65 years. All participants were native speakers of Quebec French and none had diagnosed or obvious hearing problems. The participants were selected on the basis of their familiarity with a target voice as established by a short questionnaire. Hence, a "familiar voice" in the present experiment was defined relatively to the *frequency*, the *recency*, the *duration* and the *period* of spoken contact between the listener and the target voice. Following this selection, speech samples were collected from one individual that was familiar and 9 individuals that were not familiar to the listener.

2.2. Stimuli

2.2.2. Recordings

The stimuli were sets of voice line-ups recorded in a noise-attenuated booth with an Electro Voice A365 microphone and digitized at a sampling rate 44,1 kHz using a 32 bits sound card. Each line-up contained 10 voices (1 target voice and 9 filler voices or *foils*) recorded from men similar in age (20-35 years old) who spoke a similar dialect (Quebec French). All the speaking F0 of voices in a given set of line-ups were within one semitone of the speaking F0 of the target voice.

As for the linguistic content of the stimuli, all speakers in the voice line-ups produced utterances that varied in four given lengths (1, 4, 10 and 18 syllables), and with varying numbers of nasal sounds (oral vs nasal). Table 2 summarizes the different conditions. One notes that, the utterances were mainly composed of oral segments (counted in terms of IPA symbols), but one set contains comparatively more nasal sounds. The utterances in question were familiar greetings in Quebec French such as " *oui* ", " *non* ", " *merci beaucoup* " or " *comment vas-tu?* ".

Table 1: Presented utterances: length (in syllables) and no. of nasal / oral segments (IPA symbols).

utt. length (syll.)	oral segments	nasal segments
1	0/1	1/1
4	1/9	2/8
10	0/22	3/24
18	2/52	12/51

2.2.2. Modifications to reproduce cell-phone conditions

With a view on potential applications in the field of forensic phonetics, the present experiment used recordings that were band-passed filtered to emulate cell-phone band widths. A number of studies suggest that the effect of telephone and portable communication devices needs to be considered when applied to tasks of voice recognition and

identification [2, 5, 12]. In the present case, all the recordings were filtered with a Blackman bandpass filter that reproduced a cellular phone bandwidth between 300 Hz and 3500 Hz. Finally, a just audible background noise (a white noise at a maximum amplitude of 24 dB) was added that did not affect speech perception as such.

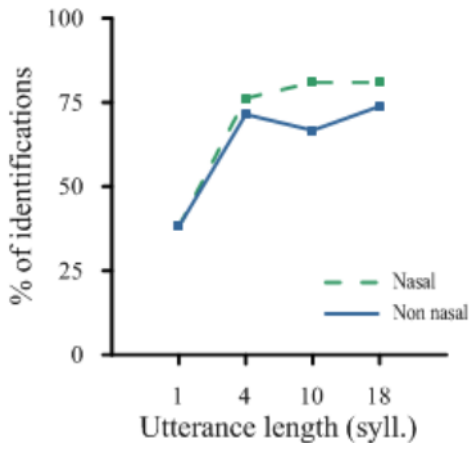
2.2. Procedure

In the task, participants used headphones and listened to the 8 voice line-ups that were played back at a maximum amplitude of 69 dB using a portable computer (32 bits sound card). In the task, the listener clicked the screen to hear the playbacks. They were asked to first listen to all ten voices in a line-up and identify the voice of a familiar individual. After listening to all 10 voices and giving a first identification, the participants were asked to listen again to any voice as many times as they wanted before making their final decision. It was these second answers that were considered in the following results.

3. RESULTS

Figure 3 summarizes the results of the various conditions of utterance length and nasality. In considering the length condition, one can see that voice recognition rates are low -- though above chance for one-syllable utterances. However there is a marked rise (of 35.7 %) for utterances of 4 syllables and this tends to level-off beyond 4 syllables (overall there is a 0 % change between 4 and 10 syll., and a 3.5 % change between 10 and 18 syllables).

Figure 3: Identification of "familiar voice" as a function of utterance length and the inclusion of nasal sounds (n = 88 per length)



As for the effects of nasality, one can see in Figure 3 that no effects appear for common one-syllable utterances such as "oui/non" (*yes/no*). A *t*-test applied to all the data showed that the difference for the nasality condition was non-significant [t (44) = 1.755, $p < 0.086$]. However, in removing the short one-syllable utterances, the data show a significant difference [t (44) = 2.507, $p < 0.016$]. This result is surprising given that the stimuli included small numbers of nasal sounds.

4. DISCUSSION AND CONCLUSION

The above results bear two implications with respect to research on voice recognition and memory of speakers. The first is that high rates of voice recognition require a stretch of speech and may not be obtained using monosyllabic utterances. This suggests that spectral information of voices is not the sole factor, and that listeners have a memory of dynamic motion-related attributes. On the other hand, this recognition of attributes does not require of benefit from long sequences. These aspects of the recognition process stand in sharp contrast with other a visual recognition of individuals – such as facial recognition -- which occur within fractions of a second (e.g. [10]).

The second implication bears on the added effect of spectral information linked to nasal cavities. The present experiment included utterances that differed slightly in terms of the number of nasal segments they contained. Yet, significant differences suggest that this information was being processed and linked to a memory of familiar speakers. The overall effect suggests that listeners are picking up on a "sensory episode" [6, 7] that includes information relating to the individual shapes of speaker's articulatory apparatus and not only on dynamic information.

7. REFERENCES

[1] Amino, K., Takashi, O. 2013. Speaker Identification Using Japanese Monosyllables and Contributions of Nasal Consonants and Vowels to Identification Accuracy. *Japan Science and Technology Law Journal*. 18, 13-21.

[2] Betantcourt, K. S., Huntley Bahr, R. 2010. The Influence of signal complexity on speaker identification. *The International Journal of Speech, Language and the Law*. 17, 179–200.

[3] Bricker, P. D., Pruzansky, S. 1966. Effects of Stimulus Content and Duration on Talker Identification. *J. Acoust. Soc. Am*. 40, 1441–1449.

[4] Broeders, A. P. A., van Amelsvoort, A. G. 1999. Lineup construction for forensic earwitness identification: A practical approach. *Proc. 14th International Congress of Phonetic Sciences*. San Fransisco, 1373-1376.

[5] Foulkes, P., Barron, A. 2000. Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*. 7, 180-198.

[6] Gluck, M., Meeter, M., & Myers, C. 2003. Computational models of the hippocampal region: Linking incremental learning and episodic memory. *Trends in Cognitive Science*. 7, 269–276.

[7] Goldinger, S. D. 1996. Words and voices: Episodic traces in spoken word identification and recognitionmemory. *Journal of Experimental Psychology: Learning, Memory and Cognition*. 22, 1166–1183.

[8] Goldinger, S. D. 1998. Echoes of echoes? An episodic theory of lexical access. Psychological Review. 105, 251–279. [6] Hollien, H., Huntley Bahr, R., Harnsberger, J. D. 2014. Issues in forensic voice. *Journal of voice*. 28, 170-184.

[9] Interpol. 2001. Forensic speech and audio analysis forensic linguistics. *Proc. 13th INTERPOL Forensic Science Symposium.* Lyon.

[10] Jemel, B., Schulter, A-M., Goffaux, V. 2009. Characterizing the Spatio-temporal Dynamics of the Neural Events Occurring prior to and up to Overt Recognition of Famous Faces. *Journal of Cognitive Neuroscience*. 22, 2289-3305.

[11] Jessen, M. 2008. Forensic Phonetics. Language and linguistics compass. 2, 671–711.

[12] Nolan, F. 2002. The 'telephone effect' on formants: a response. *Forensic linguistics*. 9, 74-82.

[13] Nolan, F. 2003. A recent voice parade. *Forensic linguistics*. 10, 277–291.

[14] Pollack, I., Pickett, J. M., Sumby, W. H. 1954. On Identification of Speakers by Voice. *J. Acoust. Soc. Am*. 26, 403–406.