INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Electrophysiological correlates of familiar voice recognition

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

[1] Laboratoire de sciences phonétiques, Université de Montréal, Montréal, Canada
[2] Laboratoire de recherche en électrophysiologie cognitive, Hôpital Rivière-des-Prairies, Montréal, Canada

julien.plante-hebert@umontreal.ca, victor.boucher@umontreal.ca, boutheina.jemel@umontreal.ca

Abstract

Our previous work using voice lineups has established that listeners can recognize with near-perfect accuracy the voice of familiar individuals. In a forensic perspective, however, there are limitations to the application of voice lineups in that some witnesses may not wish to recognize the familiar voice of a parent or close friend or else provide unreliable responses. Considering this problem, the present study aimed to isolate the electrophysiological markers of voice familiarity. We recorded the evoked response potentials (ERPs) of 11 participants as they listened to a set of similar voices in varying utterances (standards of voice line ups were used in selecting voices). Within the presented set, only one voice was familiar to the listener (the voice of a parent, close friend, etc.). The ERPs showed a marked difference for heard familiar voices compared to an unfamiliar set. These are the first findings of a neural marker of voice recognition based on voices that are actually familiar to a listener and which take into account utterances rather than isolated vowels. The present results thus indicate that protocols of near-perfect voice recognition can be devised without using behavioral responses.

1. Introduction

Recognizing someone familiar strictly by hearing their voice is a common human ability. This ability can be crucial when applied in legal investigations where voice data is sometimes the only means by which to identify suspects. Speaker recognition is, however, not well understood as a human process. The once popular idea that voices contain a set of static features analogous to finger prints [1] has been refuted by experts [2], and is discredited in legal circles [3]. One of the main reasons for this is that acoustic parameters of the voice, such as fundamental frequency (F_0) or other markers, can vary within speakers just as much as across speakers [4]. Moreover, variations in acoustic markers can arise from the very conditions by which speech samples are gathered. For example, the length of speech samples or the time or day at which samples are taken can alter acoustic measures of speakers' voices (for other such confounds, see [5-7]).

It is also important to note that, despite decades of research, automatic speaker recognition still presents error rates that prevent its application in legal contexts without additional validation from other methods [8]. For instance, a recent report comparing accuracy rates of various methods of automatic speaker recognition showed that acceptable positive recognition rates could not be obtained without raising the false alarm rates [9-11]. Moreover, available technologies require lengthy samples along with highly controlled experimental and recording conditions, which are hardly practical in legal settings [9].

On the other hand, human-based voice recognition can be highly accurate, even in relatively noisy conditions and with short samples of conversational speech [12]. In fact, we showed in a previous study that, on four-syllable utterances, listeners are able to recognize a familiar voice with accuracy rates of over 99 % even with recordings reflecting signals of cell-phone quality [13].

Humans thus have a capacity to precisely recognize voices and this appears early in development, perhaps even in prenatal phases as some research suggests [14]. However, the use of human-based voice recognition presents inherent limitations when applied in the field of forensic phonetics, especially in the identification of suspects. In these situations, listeners who recognize a familiar voice may not collaborate in verbally identifying an individual or else may contribute unreliable information. In these situations, methods that rest on neural responses to sensory signals, such as Evoked response potentials (ERPs) may be used.

On these applications, some studies have demonstrated ERPs in speaker recognition using very short speech samples, such as single long vowels [15]. However, research has shown that long speech samples (of up to 2000 ms) can lead to improved recognition of familiar voices [13, 16]. Our own studies show that, although voice recognition is possible using single syllables, accuracy rates are low, compared to the rates obtained on longer samples (see Figure 1 where 100 % recognition was obtained on 4-syllable utterances).

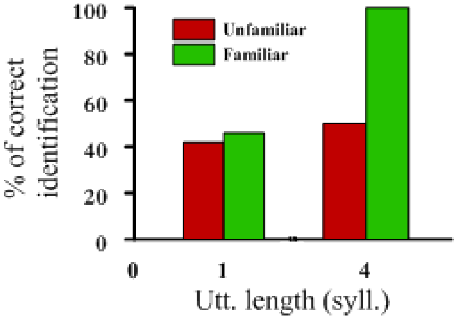

Figure 1: *Effect of utterance length on speaker recognition for unfamiliar voices (red) and familiar voices (green). Adapted from data in [13].*

It is worth noting that, in studies comparing the effects of sample duration, "length" is measured not so much in terms of milliseconds but in terms of the number of presented

http://dx.doi.org/10.21437/Interspeech.2017-1392

articulatory configurations, generally associated with the number of "syllables" [17]. The implication is that, in speaking over some length of time, varying articulatory configurations can provide spectral information relating to the speaker's idiosyncrasies and physiology, and this, in effect, can contribute to accurate recognition. The contribution of this dynamic spectral information need not entail long samples, however, as Figure 1 suggests. A recent review of the literature [18] concerning voice familiarity also concludes that, in general, the stimuli used in research on voice recognition are too short and that there is a need to investigate effects over phrase-length contexts.

In following these recommendations, we investigate, in the present study, the recognition of familiar voices by examining listeners' neural responses to different four-syllable phrases constituting common expressions and greetings. Our main objective was to determine whether heard familiar voices elicit specific markers in ERPs compared to a set of unfamiliar voices presenting similar acoustic attributes. It should be noted that isolating such markers is of interest in applications bearing on ear-witness identification involving memory of spontaneous speech. The contexts used as stimuli reflect segments of usual speech, which is unprecedented in the literature.

2. Methodology

2.1. Participants

Eleven participants (9 females), aged between 21 and 43 years (mean = 30.73, s.d. = 5.31) completed the study. They were all native speakers of Quebec French except one speaker who learned Quebec French at four years of age. All were dominant right handers according to a standard questionnaire [19] and had normal hearing as established by an audiometric screening test. A forward and backward digit-span test (WAIS-III) [20] confirmed normal memory performance for all participants. It is useful to note that the recruitment of participants followed a strategy based on their pre-existing familiarity with a target voice. "Familiarity" was assessed via a questionnaire that was validated in a previous behavioral study [13]. All target voices had to present particular characteristics as in a standard voice lineup (see section 2.2), and this further restricted participant admissibility. In short, the above 11 participants all knew a target voice used as stimuli and all target voices had similar acoustic attributes.

2.2. Stimuli

The voice stimuli were eight four-syllable utterances (listed in Table 1) produced by 14 native Quebec-French speakers without any discernible regional accents. Average *Speaking Fundamental* frequency (SF$_0$) for all speakers was highly similar and varied by no more than one semitone. As shown in Table 1, each utterance contained a number of nasal sounds. These sounds have been shown to have some effect on voice recognition likely because they provide information on speaker physiology relating to resonance cavities [17]. The speakers were asked to produce the stimuli in a conversational fashion and these contexts were recorded in a sound-treated booth using an omnidirectional headset microphone (*AKG*, C477 WRL) and a 16-bit external sound card set at a sampling rate of 44,1 kHz (*Fast-track Ultra*, M-Audio).

While recording the stimuli, the speakers followed an audio guide to ensure similar prosodic patterns.

The amplitude of recorded signals was normalized. The audio files were also aligned so that the perceptual-center (P-center) of the first syllable of all utterances was at 200 ms from the beginning of the audio file. This ensured that all the audio signals are aligned in terms of their perceptual onsets rather than their acoustic onset. The P-center is described in [21, 22]. The overall length of the signals was between 725 ms and 911 ms with an average of 793 ms.

Audio files containing the stimuli were arranged in eight blocks, each reflecting an utterance of Table 1. The blocks were ordered such that the first, third, fifth, and seventh blocks contained 240 trials each (the EEG-recording blocks) and the four other alternating blocks contained 60 trials each (behavioral-test blocks).

Table 1: *Stimuli used in regular orthographic Quebec French, their IPA transcription and the number of nasal sounds in each utterance.*

Utterance	**IPA transcription**	**Nasal segments**
Bonjour madame.	[bɔ̃ʒuʁmadam]	3
Combien t'en prends?	[kɔ̃bjɛ̃tɑ̃pʁɑ̃]	4
Comment qu'elle va?	[kɔmɑ̃kavɑ]	2
De temps en temps.	[dətɑ̃zɑ̃tɑ̃]	3
Donne-moi en deux.	[dɔmwazɑ̃dø]	2
J'en connais quatre.	[ʒɑ̃kɔnɛkat]	2
Quand est-ce qu'il vient?	[kɑ̃tɛskivjɛ̃]	2
Quelqu'un t'attend.	[kɛkœ̃tatɑ̃]	2

Moreover, each block contained presentations of one familiar (target) voice (33.33 % of trials), one unknown but frequent voice (33.33 % of trials) and twelve unknown rare voices (each 2.77 % of trials). Within each block, the voices were randomized with the restriction that no consecutive presentation contained the same voice. Note that 11 participants were recruited on the basis that each participant was familiar with one target voice in the stimuli -- but the target voices differed across participants. Thus there were 14 different voices but a particular voice was familiar only to some participants.

2.2.1. Pre-test stimuli validation

As an added precaution, we conducted a pre-test involving four volunteers that did not know any of the voices used in the stimuli. We wanted to establish whether equal numbers of presentations for the above voices and contexts created non-specific ERPs. The test conditions were the same as during the present experiment (see section 2.3) and each volunteer was exposed to a total of 10 trials per voice per utterance. The pretest confirmed that, in presenting different utterances and voices a similar number of times, average ERPs did not visually differ across conditions. However, one of the voices had to be removed due to an unexplained difference in the ERPs observed when compared to the other voices. Overall, variations in ERPs under the present test conditions can be related specifically to familiarity and frequency of presentation.

2.3. Procedure

2.3.1. Experimental tasks

Participants were asked to listen to the stimuli in a dimly lit room using insert earphones (*E-A-Rtone 3A*, EAR Auditory Systems). The stimuli were calibrated so as to obtain peak levels of 74 dBa at the insert. The stimuli were played back using *E-prime 1.0* (Psychology Software Tools). The trials were separated by an inter-stimulus interval (ISI) that varied randomly from 500 ms to 650 ms in steps of 50 ms so as to minimize anticipation effects. In listening to these stimuli, the participants were sitting at a distance of 70 inches from a blank computer screen with a fixation cross. They were asked to listen to the stimuli and keep their eyes on the fixation cross. For the behavioral blocks, participants were also required to keep their indexes on the buttons of a mouse and to indicate as quickly as possible if the voice heard during each trial was familiar or unknown by pressing either the left or the right button, respectively (this was reversed for half of the participants).

2.3.2. EEG recordings

The EEG were recorded in eight continuous blocks for each participant using the international 10-20 system with ASA-lab EEG/ERP 64 channels amplifier (ANT neuro) with an online average reference at a 1000 Hz sampling rate. The eye movements and blinks were recorded using four electrodes placed above and below the dominant eye (VEOG) and at the outer canthus of each eye (HEOG). AFz was used as ground and all other 64 channels were kept below a10 kΩ impedance during the recordings.

Offline, the recordings were band pass filtered (0.3-30 Hz) and blinks were removed using ASA software (ANT neuro) offline. All other artefacts exceeding a standard deviation of 20 µV within a sliding window of 220 ms were automatically removed. EEG recordings were then averaged across all blocks according to the conditions (target voice, frequent voice, and rare voice) after being time-locked to the stimuli onset. The average time window was set between 200 ms and 800 ms. A baseline correction was also applied using a 200 ms pre-stimulus window.

3. Results

3.1. Behavioral recognition

The overall recognition rate was 99.27 % across 1913 trials. The false alarm rate was 0.32 % and the misses represented 1.49 % of the trials. All responses exceeding 1300 ms were excluded (16.09 %) and 6 blocks were removed from the analyses owing to an absence of response.

3.2. ERP results

The early ERP components, between stimulus onset and 300 ms, did not show any significant difference between the frequent voice and the rare voices in early components (between stimulus onset and 300 ms). The present analysis therefore focused on the differences observed between the frequent unknown voice (frequent) and the familiar voice (target) for this time window since both voices were presented with equivalent probabilities of occurrence (33.33 % of the time). The comparison between the rare voices and the frequent voice is not addressed in the present report.

Figure 2 shows the remaining ERP variations after subtracting the data of the frequent voice from the data of the target voice. A visually prominent ERP difference peaking between 210 ms and 245 ms is seen on frontal sites, the strongest difference observed being at the electrode FCz (fronto-central site) as indicated by the asterisk on the topography of Figure 2.

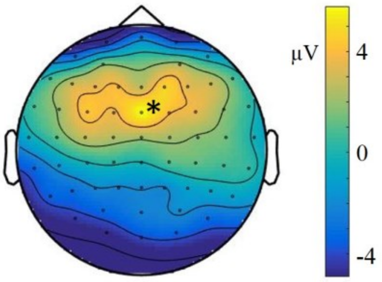

Figure 2: *Top topography showing the average ERP differences (µV) between the target voice and the frequent unknown voice between 210 ms and 245 ms.*

A t-test was applied on the same specific time window (210 ms to 245 ms) using *Fieldtrip* [23], and this showed a significant difference across familiar and unfamiliar (frequent) voices at FCz [t (11) = 2.50, $p < 0.034$], F3 [t (11) = 2.88, $p < 0.030$] and FC3 [t (11) = 3.06, $p < 0.006$]. As illustrated in Figure 3, the ERP waveforms elicited by the target and frequent voices show a large amplitude difference at the level of a positive component peaking over FCz after 200 ms.

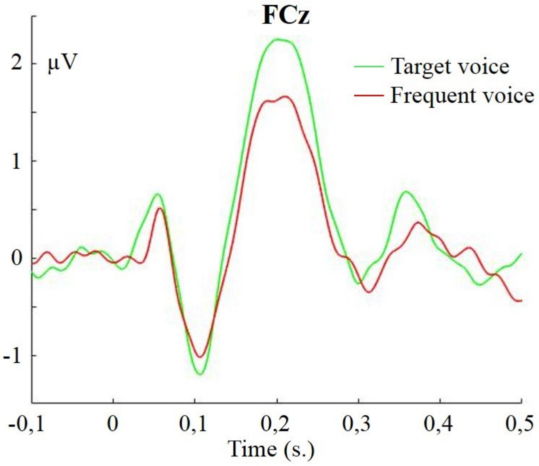

Figure 3: *ERPs for the target voice (green) and the frequent unknown voice (red) on FCz.*

Voice discrimination, as such, may be an earlier process than associating a voice to particular individual (as when verbally identifying a speaker), which may involve later occurring EEG components (see [24] and [25]). These were not examined in the present report.

4. Discussion and conclusion

The aim of the present study was to find electrophysiological markers of familiar voice recognition. The participants ($n = 11$)

took part in an experiment where they had to listen to a set of similar voices including one from familiar speaker.

Behavioral results show that all participants successfully discriminated the familiar voice from the unfamiliar ones with an overall recognition rate of 99.24 %. These results accord with those of our previous study where speaker recognition was above 99 % in the case of familiar voices, where voice familiarity was established using a quantitative scale.

The ERP results presented in this paper indicate the presence of at least one specific component in the processing of familiar voices. A visually marked difference peaking at 210 ms post-stimulus onset was found on frontal electrodes when comparing the time-locked averages for the target voice to averages for the unknown frequently heard voice.

A significant difference on frontal sites during passive listening of speech samples partly corresponds to previously reported results [15] and supports the interpretation that a discrimination process is engaged quite early in hearing the voice with a possible recognition process requiring more time or signal information (as suggested by [24] and [25]). Our findings do not, however, show a difference in mismatch negativity (MMN): note that the frequent and target voices were both presented in similar numbers (each 33.33 % of the trials). It is also important to mention that the audio recordings used in the present study were four syllables long compared to single vowels used in [15]. Thus, the present results suggest that spectro-temporal information extending across syllables is required to achieve high speaker recognition rates (which conforms to our earlier observations – see Figure 1). Interestingly, the positivity seen in Figure 3 corresponds to the average onset of the second syllable of utterances used as stimuli. This suggests that familiar speaker recognition based on sufficient spectro-temporal information generates an increased positivity around 210 ms. Such findings are relevant in establishing the processes underlying the human ability to discriminate voices. The results presented in this paper also stand out from previous works in that the stimuli used involved voices from speakers that were very familiar to the individual participants in comparison to experiments that often use “famous voices” (e.g. [16] and [26]). In a forensic perspective, our findings demonstrate that highly accurate voice recognition in the case of familiar speakers is possible and that human-based methods that do not require overt behavioral responses can be applied. Such applications, of course, carry ethical implications, though no more than if one used a polygraph with obtained consent from an individual.

5. Acknowledgements

We thank anonymous Interspeech reviewers for their insightful comments of this paper. Errors that remain are strictly our own. This research was funded by a FRQNT Team Grant and by SSHRC and FQRSC awards to the first author.

6. References

[1] L. Kersta, "Voiceprint Identification; Bell Telephone Laboratories, Inc., Murray Hill, NJ," *Nature,* p. 1253, 1962.

[2] P. French, "An overview of forensic phonetics with particular reference to speaker identification," *Forensic linguistics,* vol. 1, pp. 144-153, 1994.

[3] W. R. Jones, "Danger-Voiceprints Ahead," *Am. Crim. L. Rev.,* vol. 11, p. 549, 1972.

[4] J. E. Atkinson, "Inter-and intraspeaker variability in fundamental voice frequency," *The Journal of the Acoustical Society of America,* vol. 60, pp. 440-445, 1976.

[5] B. Lindblom, "Explaining phonetic variation: A sketch of the H&H theory," in *Speech production and speech modelling*, ed: Springer, 1990, pp. 403-439.

[6] G. Klasmeyer, T. Johnstone, T. Bänziger, C. Sappok, and K. R. Scherer, "Emotional voice variability in speaker verification," in *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, 2000.

[7] S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, Z. Deng, S. Lee*, et al.*, "An acoustic study of emotions expressed in speech," in *INTERSPEECH*, 2004.

[8] G. S. Morrison, F. H. Sahito, G. Jardine, D. Djokic, S. Clavet, S. Berghs*, et al.*, "INTERPOL survey of the use of speaker identification by law enforcement agencies," *Forensic science international,* vol. 263, pp. 92-100, 2016.

[9] U. S. D. o. Commerce. (2012). *Speaker Recognition Evaluation*. Available: https://www.nist.gov/itl/iad/mig/sre12-results

[10] A. K. Jain, P. Flynn, and A. A. Ross, *Biometric authetification*. USA: Springer, 2007.

[11] J. A. Unar, W. C. Seng, and A. Abbasi, "A review of biometric technology along with trends and prospects," *Pattern Recognition,* vol. 47, pp. 2673-2688, 2014.

[12] S. J. Wenndt, "Human recognition of familiar voices," *The Journal of the Acoustical Society of America,* vol. 140, pp. 1172-1183, 2016.

[13] J. Plante-Hébert and V. J. Boucher, "L’identification vocale: pour une quantification des effets de la familiarité," in *Journée d'Études sur la Parole*, Le Mans, 2014.

[14] A. J. DeCasper and W. P. Fifer, "Of human bonding: Newborns prefer their mothers' voices," *Science,* pp. 1174-1176, 1980.

[15] M. Beauchemin, L. De Beaumont, P. Vannasing, A. Turcotte, C. Arcand, P. Belin*, et al.*, "Electrophysiological markers of voice familiarity," *European Journal of Neuroscience,* vol. 23, pp. 3081-3086, 2006.

[16] S. R. Schweinberger, A. Herholz, and W. Sommer, "Recognizing Famous Voices Influence of Stimulus Duration and Different Types of Retrieval Cues," *Journal of Speech, Language, and Hearing Research,* vol. 40, pp. 453-463, 1997.

[17] J. Plante-Hébert and V. J. Boucher, "Effects of nasality and utterance length on the recognition of familiar speaker," in *18th Internation Congress of Phonetic Sciences*, Glasgow, 2015.

[18] S. R. Schweinberger, H. Kawahara, A. P. Simpson, V. G. Skuk, and R. Zäske, "Speaker perception," *Wiley Interdisciplinary Reviews: Cognitive Science,* vol. 5, pp. 15-25, 2014.

[19] R. C. Oldfield, "The assessment and analysis of handedness: the Edinburgh inventory," *Neuropsychologia,* vol. 9, pp. 97-113, 1971.

[20] D. Wechsler, "WAIS-III: Wechsler adult intelligence scale, administration and scoring manual. San Antonio, Texas: Psychological Corporation," ed: Harcourt Brace, 1997.

[21] S. M. Marcus, "Acoustic determinants of perceptual center (P-center) location," *Perception & psychophysics,* vol. 30, pp. 247-256, 1981.

[22] J. Morton, S. Marcus, and C. Frankish, "Perceptual centers (P-centers)," *Psychological Review,* vol. 83, p. 405, 1976.

[23] R. Oostenveld, P. Fries, E. Maris, and J.-M. Schoffelen, "FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational intelligence and neuroscience,* vol. 2011, p. 1, 2011.

[24] D. R. Van Lancker and J. Kreiman, "Voice discimination and recognition are separate abilities," *Neuropsychologia,* vol. 25, pp. 829-834, 1987.

[25] D. Van Lancker and J. Kreiman, "Unfamiliar voice discrimination and familiar voice recognition are independent and unordered abilities," *UCLA Working Papers in Phonetics,* pp. 50-60, 1985.

[26] D. Van Lancker, J. Kreiman, and K. Emmorey, "Familiar voice recognition: patterns and parameters. Part I: Recognition of backward voices," *Journal of phonetics,* vol. 13, pp. 19-38, 1985.