Effects of speech modalities in acquiring neural markers of voice recognition: An ERP experiment using voice lineups

Julien Plante-Hébert[1], Victor Boucher[2], Boutheina Jemel[3]

Laboratoire de sciences phonétiques de l'Université de Montréal[1,2], École d'orthophonie et d'audiologie de l'Université de Montréal[3]

julien.plante-hebert@umontreal.ca[1], victor.boucher@umontreal.ca[2], Boutheina.jemel@umontreal.ca[3]

ABSTRACT

Event-related potentials (ERPs) were used to ascertain the influence of contextual information in acquiring neural components of voice recognition. During a training phase, 18 participants had to learn the voices of three speakers selected according to standards of voice lineups. Each voice was trained in a particular modality: audio (A), audiovisual (AV), audiovisual with speaker-listener interaction (AVI). These voices were presented during EEG recordings along with an untrained voice which served as a baseline. Analyses of ERPs revealed that all training conditions led to significant changes on a P2 and a late positive component (LPC). The LPC showed a non-enhancing "face overshadowing effect" in the AV condition which was cancelled in the AVI condition. Combined with previous observations, the results indicate that multisensory information accompanying voice learning does not affect early components of voice *recognition* but does affect a LPC of voice *identification* when multisensory information extends to speech interaction.

Keywords: speaker familiarity, event-related potentials (ERP), voice learning, voice lineups

1. INTRODUCTION

Conflicting reports on the role of facial information in voice recognition have emerged from different sectors of research. Diverging results owe principally to the variability of stimuli that are used and the amount of training given to listeners in experiments of voice familiarization. In terms of the stimuli, investigations bearing on the reliability of ear- and eyewitness testimony refer to protocols such as voice lineups that greatly reduce the variability of face and voice stimuli in identification tasks (e.g., [1-4]). Such protocols are not the general practice in neurophysiological investigations but are nonetheless quite relevant in determining the effects of audio and audiovisual training on neural markers of voice recognition.

For one thing, salient attributes of faces or speech can bias listener's attention on a particular modality and their associative memory of a voice, which can impact neural responses of voice recognition. As for the amount of training required to obtain responses that reflect a recognition of "familiar" voices, it is unclear whether responses to famous voices or voices learned in laboratory settings resemble those evoked by intimately familiar voices [5]. The learning of the latter voices is obviously different not only in terms of the amount of training but also in terms of multimodal experiences that accompany intimately familiar voices and which generally include speech interaction with individuals. Indeed, much of the problem in defining neural markers of voice recognition has to do with the inherent variability of personal experiences with voices.

To address this problem, a previous study compared ERPs on trained and intimately familiar voices that were all quite similar as in standard voice lineups [6]. That study revealed differences in early components P2-N250 and a late positive component (LPC) for trained and intimately familiar voices but where LPC stood out as a marker of voice recognition relating to the *identity of a speaker.*

The present study builds upon these results and aims to determine the effects of modalities of voice learning on ERPs of voice recognition where voice variability is again minimized in accordance to standard voice lineups. To investigate modality effects, three learning or "training" conditions are used involving different modalities of stimuli presentation. These include an audio modality (A) where listeners are trained on heard voices; an audiovisual modality (AV) in which voices are trained by viewing and hearing speakers; and an audiovisual modality with interaction (AVI) where voices are trained by hearing, seeing, and repeating what speakers say.

2. METHODOLOGY

2.1 General design

The design included a voice-training phase followed by an experimental phase involving EEG recordings. In the training phase, participants learned to identify, by way of symbols on a keypad, three target voices, V1, V2, V3, which were presented in a particular modality (A, AV or AVI). Thus, there were three trained voices (TVs), individually learned in three

different modality conditions. The modalities in which each voice was learned varied across participants as well as their order of first appearance. Once the voices were acquired, participants proceeded to the experimental phase and EEG was recorded during audio presentations of the TVs and of one untrained voice (UV), V4, which served as a baseline in the analysis of ERPs.

2.2 Participants

These were 18 native speakers of Quebec French (9 females) aged between 21 and 30 years (mean=25, s.d.=3). All were dominant right handers [13], had normal hearing as established by an audiometric screening test, and normal memory spans (WMS-III).

2.3 Speech and voice stimuli

The training and experimental stimuli were drawn from a list 260 two-syllable common nouns. Audio and audiovisual recordings were made from the productions of these words by four native speakers of Quebec French that were selected in accordance with standards of voice lineups. Specifically, the four male speakers had no discernible regional accent, no idiosyncratic articulations, and they had similar speaker fundamental frequencies (SF0s) to within 1 semitone (mean: 120.83 Hz).

2.3.1 Experimental stimuli

These stimuli were the audio recordings of the TVs and the UV. Audio presentations for the four voices comprised 60 different words each. Recordings of these produced words were made in a sound-attenuating booth using a Shure (X2u) sound card, at 44.1 kHz and 16-bit, a Lavalier (Audio-Technica, AT831b) and software (Golwave 6.31). To obtain productions of stimuli with similar prosody, a rhythm guide was used, and amplitudes were normalized to /a/ sounds.

2.3.2 Training stimuli

These were audio and audio-video recordings of the same 20 words produced by three of the speakers described above (V1, V2 and V3). The recordings in mp4 format used a webcam (Web HD Pro, Logitech), a 64-bit NVIDIA graphic card (Quadro K5200) and were edited via DaVinci 14 (Blackmagic). In the videos, the head and face of the speakers were similarly positioned on the screen and displayed against a neutral background. The facial features and the clothing of the speakers were also similar. The 20 words serving to train voices V1, V2, V3 were randomly assigned to the three training conditions across participants. Finally, an important aspect of the stimuli is that, during the recording of produced words in the AV condition, the speakers looked down at a mark on the table in front of them whereas, for the AVI condition, speakers' eye gaze was directed at the camera during the production of words before returning to a downward gaze. This was meant to capture the gaze effects that occur during speaker-listener interaction, and such effects prevail even if the speaker is not physically present [7]. Eye gaze served to cue participants' repetition of heard words in the AVI condition (see "Procedure").

2.4 Procedure

2.4.1 Training phase

The training involved an introductory presentation of the stimuli followed by repeated training-test cycles, similar to [8]. Participants sat in front of a laptop and attended to presented audio and audiovisual recordings using insert earphones (EARtone 3A).
In the introductory presentation, 10 identical speech contexts representing voices V1, V2, V3 were played back in the same order, for a total of 30 trials. The three voices were, however, presented in different training conditions (A, AV, AVI) that were randomly assigned and counterbalanced across participants such that all voices and conditions were played back an equal number of times in a training block. The participants were instructed to focus on speaker's characteristics so as to remember their voice. Only for the AVI condition, they were also told to look at a speaker in the eyes and repeat the heard word aloud.
In the training-test cycles, the same 10 speech stimuli used in the introduction were presented in random order. Each cycle consisted of a training block of 30 trials followed by a test block of 30 trials presented using E-prime 3. During the training blocks, participants had to learn to associate a voice with a symbol on the keypad and this symbol was also displayed on slides that accompanied the presented recordings. In the following test blocks, only audio versions of the training stimuli were presented and participants identified the voices using the keypad. The training blocks were repeated until participants could correctly identify speakers on at least 23 of the 30 trials of the test blocks in three consecutive cycles.

2.4.2 Experimental phase

EEG recordings were performed at this phase. Participants sat at approximately 130 cm from a computer monitor displaying a fixation cross and wore ear inserts as in the training phase. The same symbol-coded keypad was also used to record voice-identification responses. The experimental stimuli

consisting of audio recordings only were played back in two continuous blocks separated with a pause. Stimuli were presented using MatLab/Psychtoolbox software routines. The participants were required to rapidly identify, after each heard word, whether the voice was V1, V2, V3, or "other" on the keypad.

2.5 EEG recording and analyses

EEG signals were recorded in two 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 and a 1 kHz sampling rate. Eye movements were recorded using four electrodes placed above and below the dominant eye and at the outer canthus of each eye. Electrode AFz was used as ground. Offline, the recordings were band-pass filtered (0.1-30 Hz) and blinks were removed using ASA software (ANT neuro). All other artefacts in the EEG exceeding a standard deviation of 20 µV within a sliding window of 200 ms were automatically removed with EEprobe GUI (version 1.2.0.2, ANT Software). All subsequent analyses including ERP averaging across individual trials and participants as well as statistical analyses were performed using Fieldtrip [9]. EEG recordings were then averaged across blocks according to the training conditions. Only trials associated with correct responses were included in the ERP averages. The average time window of separate epochs was set between 200 ms before and 1000 ms after each stimulus file onset. The 200 ms pre-stimulus interval was used for baseline correction.

Before analyzing the behavioral data, trials with response times (RT) exceeding 2 standard deviations from the average value were excluded (2.08%).

A global fields power analysis presented in Figure 1 (left) served to circumscribe time windows of peak brain activity to peaks representing the P2 component (139 to 239 ms post onset) and the LPC (550 to 900 ms post onset) [10]. For each of the components, smaller short-time-windows were used so as to perform t-test statistical comparisons using a Monte-Carlo method which compares ERP responses on each training condition to ERP responses for the UV baseline condition. The method involves non-parametric cluster-based random permutation tests as implemented in Fieldtrip [11, 12]. Clusters were considered significant at $p < 0.05$. Topographies in Figure 1 (right) therefore reflect the *t*-values obtained in the analyses rather than more common ERP magnitudes and polarities.

3. RESULTS

3.1 Behavioral results

A repeated-measures ANOVA revealed a main effect of training condition on the RTs [$F(3, 45)=11.98$, $MSE=47678.6$, $p<0.000$, $\eta^2=0.444$]. Post hoc tests using Bonferroni correction for multiple comparisons indicated that the strong difference owed to the contrast between TVs in the three training conditions and UV. RTs for all three TV conditions were significantly shorter than those for UV (at $p<0.001$). However, the same post hoc tests did not reveal significant differences amongst the three TV conditions on either correct responses or RTs.

3.2 ERP results

The analyses of the differential responses to TVs were carried within four 25 ms windows surrounding the P2 peak. These yielded significant clusters for all three training conditions when compared individually to UV between 164 and 189 ms. Some significant clusters also appeared for the AV condition between 139 and 164 ms post stimuli onset. All significant clusters within these time windows were located on central and parieto-central sites along the middle line. However, such uniformity of P2 across conditions was not present for the LPC.

The LPC was analyzed using 50 ms time windows for these protracted responses. As illustrated in Figure 1, cluster analyses of the LPC responses to TVs in the

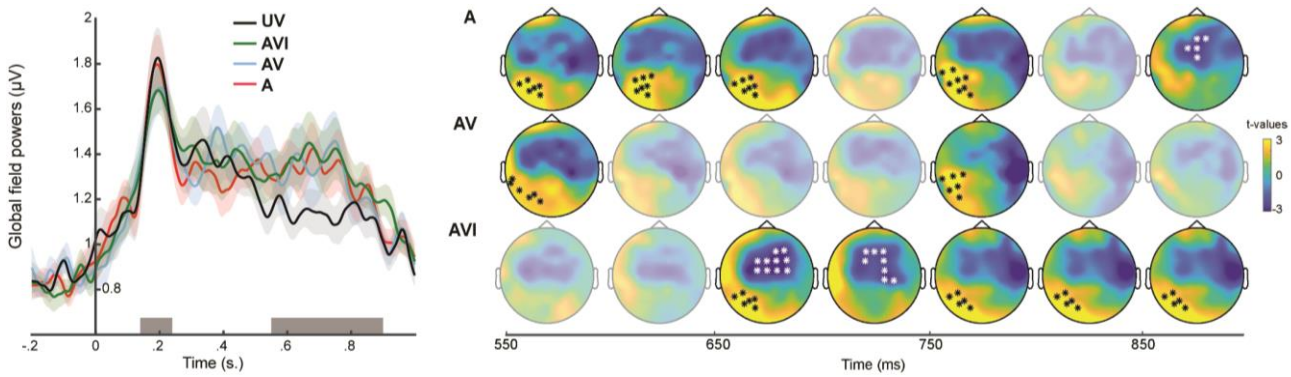

Figure 1: Left: Global fields power analysis averaged from all sites. Right: topographic representations of t-values obtained from the cluster analyses of differential response to TVs and UV across the learning conditions for time windows around the LPC peak. Highlighted electrodes are statistically significant ($p < 0.05$).

three training conditions were significantly different from responses to UV across a wide time interval, and sites of differential activity varied across conditions. Specifically, the cluster analyses showed that the training condition A associated with significant clusters in five of the seven time-windows starting from 550 ms and extending to 900 ms. Significant clusters where mostly located on left-parietal, centro-parietal, and parieto-occipital sites. The AV condition also yielded significant clusters at these sites but only between 550 and 600 ms and 750 ms 800 ms. In sum, for these LPCs, presentations of similar faces in learning voices in the AV training condition *did not enhance* differential ERP responses to TVs and UV in comparison to presentations of voices alone. However, a different pattern of activity arose in the AVI training condition. In this case, significant clusters reflecting differential centro-parietal responses appeared later (starting at 650 ms post-stimuli onset) and shifted to middle and right fronto-central sites before returning to left parietal sites (at 750 ms post onset). This suggests that active speech participation while learning voices of visually presented speakers has a sustained long-lasting effect on voice processing in contrast to passive conditions A and AV.

4. DISCUSSION

The above experiment used narrowly controlled voice and face stimuli to investigate effects of multimodal information on voice identification via three training modalities. After a training phase, participants were asked to identify target speakers on multiple trials. The behavioral results showed that TVs were successfully identified but there were no significant differences between training conditions.

However, in the analyses of the ERPs, responses to TVs were differentiated from those to UV and the Monte-Carlo analyses of these differential responses bore out a particular effect of training conditions A and AV. It will be recalled that A and AV only differed in terms of the presence of visual facial information at training. Both conditions resulted in modulations of P2 and the LPC. This suggests that adding visual face information did not facilitate voice identity encoding and, in fact, *may have impeded such encoding.* Such an effect concurs with research attesting to a "face overshawdowing effect" (FOE), although the effect in the present case may owe to the highly similar facial information of the speakers that was provided and which did not assist in identity discrimination [13, 14]. It should also be noted that the preceding FOE occurred with dynamic facial stimuli (cf. [15]).

As for the effects of AV and AVI, it will be recalled that these two training conditions differed in that, for AV, the listeners did not see the speaker's gaze whereas, in the AVI condition, the listeners were asked to repeat the words produced by the speaker while gazing at the speaker on a monitor. This mode of presentation was designed to partly simulate the effects of speech interaction. As in other training conditions, the AVI condition elicited a P2 and an LPC. Compared to the AV training, however, the AVI training resulted in a wider-ranging response in the LPC window, and also in a greater number of significant clusters that shifted momentarily toward mostly middle and right centro-frontal sites. These results therefore provide evidence that speech interaction in the AVI condition, involving effects of gaze and simulated speech interaction, enhances ERP components at voice recall. Of course a laboratory simulation has its limits and, if anything, might underestimate the effects of actual person-to-person contact on voice identification as reported by [16].

More importantly, a central finding of the present study is that the observed ERPs to trained voices narrowly conform to ERPs found in a previous study involving intimately familiar and trained-to-familiar voices [6]. Both investigations involving similar stimuli have revealed responses of voice recognition and identification in the same time ranges as the P2 and LPC. It is interesting to note that in the above study, all three successfully trained voices, as validated by behavioral results, elicited a P2 that differed significantly from responses to UV. This brings new evidence confirming that the P2 is a valid marker of voice recognition. But the present results also indicate that speaker identification, or access to available semantic information on a speakers' identity when hearing a voice, as measured by the LPC, was not enhanced by the additional presentation of visual stimuli during training. This entails that adding dynamic visual information of speakers' faces may not necessarily enhance voice encoding and recall. In fact, such contextual information can, in some cases, divide the learner's attention as suggested in reports of a FOE in voice recognition tasks. On the other hand, the greater LPC response of AVI as compared to AV shows that the FOE may vary or diminish with the addition of other contextual information such as speech interaction and gaze toward a speaker. Even if some participants reported being distracted by the repetition of verbal forms, a differential LPC in response to TVs was observed. In other words, in the same way [15] showed that the FOE decreased after a certain exposure threshold or amount of training, the present results suggest that FOE could also be reduced by the type and amount of contextual information that is provided.

6. REFERENCES

[1] Hollien, H., Huntley, R., Kunzel, H., and Hollien, P.A. 2013. Criteria for earwitness lineups. *International Journal of Speech Language and the Law*. 2. 143-153.

[2] Nolan, F. and Grabe, E. 1996. Preparing a voice lineup. *International Journal of Speech, Language and the Law*. 3. 74-94.

[3] Wells, G.L., Kovera, M.B., Douglass, A.B., Brewer, N., Meissner, C.A., and Wixted, J.T. 2020. *Policy and procedure recommendations for the collection and preservation of eyewitness identification evidence*. Law and Human Behavior. 44. 3.

[4] Yarmey, D.A. 2014. The psychology of speaker identification and earwitness memory, in: Lindsay, R.C.L., Ross, D.F., Read, J.D., and Toglia, M.P., (eds), *Handbook Of Eyewitness Psychology*. Routledge, 115-150.

[5] Kreiman, J. and Sidtis, D. 2011. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

[6] Plante-Hébert, J., Boucher, V.J., and Jemel, B. 2021. The processing of intimately familiar and unfamiliar voices: Specific neural responses of speaker recognition and identification. *Plos one*. 16. e0250214.

[7] Conty, L., Russo, M., Loehr, V., Hugueville, L., Barbu, S., Huguet, P., Tijus, C., and George, N. 2010. The mere perception of eye contact increases arousal during a word-spelling task. *Social Neuroscience*. 5. 171-186.

[8] Zäske, R., Volberg, G., Kovács, G., and Schweinberger, S.R. 2014. Electrophysiological correlates of voice learning and recognition. Journal of *Neuroscience*. 34. 10821-10831.

[9] Oostenveld, R., Fries, P., Maris, E., and Schoffelen, J.-M. 2011. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*. 2011. 1.

[10] Lehmann, D. and Skrandies, W. 1980. Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and clinical neurophysiology*. 48. 609-621.

[11] Kroese, D.P., Taimre, T., and Botev, Z.I. 2013. *Handbook of monte carlo methods*. John Wiley & Sons.

[12] Maris, E. and Oostenveld, R. 2007. Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*. 164. 177-190.

[13] Stevenage, S.V., Howland, A., and Tippelt, A. 2011. Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*. 25. 112-118.

[14] Tomlin, R.J., Stevenage, S.V., and Hammond, S. 2016. Putting the pieces together: Revealing face–voice integration through the facial overshadowing effect. *Visual Cognition*. 25. 629-643.

[15] Zäske, R., Mühl, C., and Schweinberger, S.R. 2015. Benefits for voice learning caused by concurrent faces develop over time. *PloS one*. 10. e0143151.

[16] Hammersley, R. and Read, J.D. 1985. The effect of participation in a conversation on recognition and identification of the speakers' voices. *Law and Human Behavior*. 9. 71.