Text for the poster presented at:

XXIX International Evoked Response Audiometry Study Group (IERASG-2025)

Understanding early and late transient responses to speech: experiments using synthetic speech.

Ángel de la Torre[1], Isaac M. Alvarez[1], Nicolás Müller[2], Francisco Chiequero[2], Juan Martín-Lagos[2], José L. Vargas[2]

[1]*Department of Signal Theory, Telematics and Communications, University of Granada, Spain. Research Centre for Information and Communication Technologies (CITIC-UGR), University of Granada, Spain.*
[2]*ENT Service, Hospital Universitario Clinico San Cecilio, Servicio Andaluz de Salud, Granada, Spain.*
E-mail: atv@ugr.es (A. de la Torre), isamaru@ugr.es (Isaac M. Alvarez)

The estimation of auditory evoked responses requires synchronous averaging (if responses are not overlapped) or a deconvolution (if they are) under the assumption of a convolutional system (i.e. each event provides always the same response). Conditioning the hearing system for consistent evoked responses is easy by presenting a simple stimulation pattern (for example, using isolated clicks with constant level). In spite of its clinical value, such repetitive stimuli fall far away from a natural stimulation. On the other hand, the estimation of consistent responses based on natural speech is difficult due to the variability of the speech signal. In this work we investigate the auditory responses evoked by synthetic controlled speech (based on a reduced set of isolated synthetic phonemes) in order to better understand what responses could be expected using natural speech.

Background

Recording auditory responses evoked with natural speech is challenging due to the inherent variability of the speech signal, and the values of the typical parameters associated to normal speech. The fundamental frequency is in the range 90-260 Hz, which means that the interval between adjacent glottal pulses (inter-stimulus-interval, ISI) is between 3.8 and 11 ms (i.e. a high stimulation rate according to conventional stimulation parameters for ABR recording). On the other hand, even though the average articulation rate provides around 10-12 phonemes per second, the duration of the phonemes covers a wide range (between 40 and 150 ms).

This difficulty due to the variability of speech explains the limitations of conventional methods for providing consistent evoked responses to speech. Currently, responses to speech-like stimulation are limited to transient responses using short and isolated vowels or syllables, frequency following response to fundamental frequency in stationary phonemes, speech-evoked envelope following responses, and more recently brainstem responses to glottal pulses in manipulated continuous speech (for example, using "peaky speech").

Methods

The stimulation material includes 4 different patterns: 3 synthetic phonemes (/a/, /i/, /s/) and click bursts. The vowels are synthesized under a stimulation-filter paradigm, where stimuli correspond to synthetic glottal pulses and the filters are designed according to the typical 1st and 2nd formants of these vowels. The /s/ is synthesized as filtered noise. These phonemes are selected in order to include two different voiced phonemes, a fricative unvoiced phoneme. The click burst is included as reference (since the responses to clicks are well known in the literature). The use of synthetic audio signals for stimulation provides a controlled scenario.

Figure with details of the stimulation signal: The figure represent a portion of the stimulation signal (in blue), details of the four stimulation patterns (in red), and a spectrogram of the stimulation signal (figure with colors). We can see that the click bursts are compound of rarefaction clicks and cover all the frequency range. The /s/ phoneme is a noise-like pattern, with dominance of high frequency components (a wide spectral peak at 4 kHz and another one around 6 kHz). Each synthetic vowel consist in a repetitive pattern, where the impulsive response of the filter is repeated at each glottal pulse. The pattern of the impulsive response is connected to the spectral distribution: the /i/ vowel was synthesized with two formants, at 250 Hz and 3.0 kHz, while the /a/ vowel was synthesized with two formants at 850 Hz and 1.3 kHz. The synthetic phonemes have a random duration, are randomly presented and are separated by a short silence.

The evoked responses include deconvolution of fast events (responses to the glottal pulses of the vowels, or to the clicks of the click-bursts) and slow events (responses to the phonemes, or to the click-burts). The deconvolution was estimated using the multi-response deconvolution method (performed in the reduced subspace associated to Latency Dependent Filtering, appropriate for recovering responses of the complete auditory pathway) [De la Torre et al. JASA 155, 3639-3653 (2024); JASA 151, 3745-3757 (2022); JASA 148, 599-613 (2020)].

Figure: AEP recording system: The responses were recorded with our modular and flexible AEP recording system based on off-the-shelf consumer electronics. An RME UCX audio interface synchronously provides the stimulation and captures the EEG. Stimulation is provided with headphones. The EEG is recorded with electrodes at Fz (active), right mastoid (reference) and middle forehead (ground), and the signal is amplified with a TritonAudio FetHead microphone preamplifier. A computer running a MatLab script prepares the stimulation, sends the stimulation signal and processes the recorded EEG to deconvolve the responses. Because of the use of headphones (instead of insertion earphones with long air tubes), the recorded EEG is contaminated with stimulation artifact (which is useful as reference).

Results

The study includes responses recorded from 12 subjects. The responses were estimated from 12 minutes of recording from each participant. Responses from fast events were estimated within a latency window of 200 ms, while a latency window of 1000 ms was used for slow events.

Figure: Analysis of responses to short-term and long-term events: For the configured recording time, the total number of long-term events (phonemes or click-bursts) was 1000 (around 250 of each pattern). Each vowel model included around 7500 glottal pulse events, and the click-burst model included around 6000 clicks. The plot represents, in the left panel, the responses to the short events: amplitude as a function of the latency (in logarithmic scale). The responses to the glottal pulses of the /a/, the /i/ and to the clicks are represented in the top, center and bottom plot, respectively. The stimulation artifact can easily be identified at the beginning of each response (with different waveforms according to the type of event). Waves III (at 4.5 ms) and V (at 6.5 ms) can also be identified (wave III is not clear for the vowel /a/ due to the stimulation artifact). Some middle latency responses can also be observed (PAM or P0 at 12 ms, PA at 30 ms, and PB/P1 at 55 ms).

The plots in the right panel represent the responses to the long-term events (in blue). The responses to the short-term events are also represented (red lines) for reference (note that the vertical scale changes between both panels). The PB/P1, P2 and P3 cortical responses are clearly observed in this participant. The PAM is also observed in these plots, as well as a PAM artifact. The presence of the PAM (and its artifacts) in the long-term responses is consequence of the instability of this response (it is a not-systematic response: it is observed always with the same latency but it is not always present, since

it is strongly affected by adaptation). As a consequence, the PAM does not match a convolutional model, and therefore, the PAM component represented in the short-term figure does not model the PAM activity, but an average PAM response. Some residual PAM is present in many events (because of the non-consistent convolutional model), which causes the observed PAM component and PAM artifact in the long-term responses. If PAM response was a convolutional process (i.e. always the same waveform, with the same amplitude and latency, and always present for each glottal pulse or each click), the PAM would be correctly modeled in the early responses (responses to short-term events) and no PAM (nor PAM artifact) would be present in the late responses (responses to long-term events).

Figure: Individual results for each participant: The plots in the right side of the poster include the responses for each participant, with the responses to the short-term events (glottal pulses of /a/ and /i/, and clicks) in red and the responses to the long-term events (/a/, /i/, /s/ phonemes, and click-bursts) in blue. In the red plots, the stimulation artifact, the ABR and the MLR components can be identified, while in the blue plots, the cortical responses are observed. Some participant exhibit no PAM, but this component is very relevant in some others.

Figure: Responses to synthetic speech included in DEMO: This test is very interesting to be used in demonstrations for dissemination purposes. On one hand the description is easy to be understood by the audience (it is easy to identify the four different patterns used for stimulation, and it is easy to distinguish between the voiced and unvoiced phonemes and to understand the glottal pulses). On the other hand, this test provides very peripheral responses (waves III and V) and very central responses (P2, P3) easily in the same experiment, and therefore it is interesting to describe the neural activity of the complete auditory pathway associated to the hearing perception. Additionally, the use of synthetic phonemes provides some connection between the stimulation provided in this experiment and the perception of the speech. The figure represent a screenshot of a dissemination demo based on this experimental design, which has been used in several outreach events for primary and secondary education students.

Discussion and conclusions

Regarding the expected responses for continuous speech, it should be remarked that the responses associated to the short-term events are very similar among the different stimulation patterns (/a/, /i/ or clicks): essentially the same waves, with similar amplitudes and latencies), in spite of the differences in the waveform and spectral content. The PAM wave (and the PAM artifact), when present, makes the interpretation of the responses difficult, particularly for the long-term responses. The late responses are also relatively similar among the different stimulation patterns. It is remarkable a systematic delay of the PAM component associated to the /s/ phoneme (with respect to the other long-term events). Finally, the long-term responses are strongly affected by low-frequency noise (mainly associated to miogenic activity). This preliminary study provides valuable information for a better understanding of the speech perception and for the development of protocols based on speech signals to assess the auditory function.