INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Computational modeling of auditory brainstem responses derived from modified speech

Tzu-Han Zoe Cheng[1], *Paul Calamia*[2]

[1]Department of Cognitive Science, UC San Diego, La Jolla, California, USA
[2]Reality Labs Research at Meta, Redmond, Washington, USA

tzcheng@ucsd.edu, pcalamia@gmail.com

Abstract

The auditory brainstem response (ABR) is a powerful neurophysiological measure to diagnose hearing deficits along the auditory pathway. Wave I of the ABRs is particularly critical for assessing early hearing loss, though hard to observe in humans. The major downside of ABR is that most protocols are very boring since they use thousands of clicks to elicit ABRs. Here, we derived modeled ABRs with continuous speech from an audiobook. Unlike other studies involving computationally intensive modification that made their speech stimuli unnatural-sounding and unlikely to be used in real-life applications, we applied a fast and efficient algorithm that enhances speech transients to better elicit ABRs. Using the auditory periphery model that simulates human brains, we derived ABRs from our transient speech and showed a significantly larger Wave I-V ratio compared to other stimuli. These results demonstrated a potential of assessing hearing conditions in a more objective and naturalistic way.

Index Terms: EEG, ABR, speech perception, auditory periphery model, Transients-enhanced speech

1. Introduction

1.1. Auditory brainstem responses

The auditory brainstem response (ABR) is a powerful measurement to assess hearing condition objectively and passively. Deriving ABRs from real-time conversations, rather than from laboratory tests with specialized stimuli, would have a great impact on augmented reality and virtual reality (i.e. AR/VR) devices on the market which aim to compensate for hearing loss, as well as on hearing aids in non-clinical settings. Despite the importance, few studies have achieved this goal. ABRs consist of Waves I-VII where each individual wave component is associated with a different subcortical structure along the auditory pathway, and can be used to estimate audiograms for individuals [1, 2]. Waves I, III, and V (triggered by activity at the auditory nerve, cochlear nucleus, and inferior colliculus, respectively; see [3]) are most distinguishable and used in human studies. Among them, Wave I is particularly important, though very hard to derive in human subjects, since it is directly generated from the peripheral auditory nerve, and thus can be used to assess hearing loss in the early stage of auditory processing (e.g. inner-ear neural deficits and synaptopathy). For example, reduced Wave I amplitudes have been found to be correlated with aging in animal models [4], human speech perception in noisy backgrounds [5, 6] and synaptopathy [7, 8, 9, 10, 11, 12].

1.2. Modified speech as an ABR stimulus

Traditionally ABRs are derived from brief, non-speech sounds such as clicks or tone bursts that can elicit a broadband frequency response. More recent research has shown the ability to produce an ABR from naturally uttered speech, yet these approaches only reveal later and larger ABR waves such as Wave V [13, 14, 15]. For the first time, [16] derived clear Waves I, III, and V using "peaky" speech, generated by aligning glottal pulses across speech harmonics for pre-recorded English stories. Such phase alignment of the speech signal modifies the time-domain waveform to be peaky and click-like (Figure 1b), but does not significantly alter the spectral properties of speech so that it is still fully intelligible while sounding metallic and unnatural. A more serious issue faced by this modification is its computationally intensive pre-processing, which makes it impractical for real-time applications such as hearing aids and hearing-loss-correcting AR/VR devices during natural conversation. However, other speech-modification approaches might have better real-life applications. For example, one potentially exploitable feature in speech is the transients, the brief transitions such as onsets and offsets of the consonants, which are important cues for identifying and discriminating speech sounds. So-called "transient speech" can be generated by emphasizing the transients in the speech signal, and has been found to be computationally efficient to generate and to have behavioral benefits such as higher intelligibility in a noisy background [17, 18]. The characteristics of speech transients mimicking clicks (i.e., brief and thus broadband) may be a better feature to elicit ABR [19, 20].

1.3. Computational Modeling of the ABR

As an alternative to assessments with human subjects, which have faced severe restrictions due to the COVID-19 pandemic, modified speech could be evaluated through computational simulations of ABRs rather than in-lab electroencephalography (EEG) recordings with physical contact. A well established computational model of the human auditory periphery by [21], which includes comprehensive auditory models mimicking real nerve responses, can simulate ABRs and has been verified using recorded human data. The model includes high, middle and low spontaneous rate neurons to simulate Wave I, and also has cochlear nucleus and inferior colliculus components to simulate Waves III and V, respectively. Without requiring human subjects to come into the lab for data collection, the auditory periphery model offers a viable alternative to compare ABR waveforms elicited from different speech stimuli.

10.21437/Interspeech.2023-593

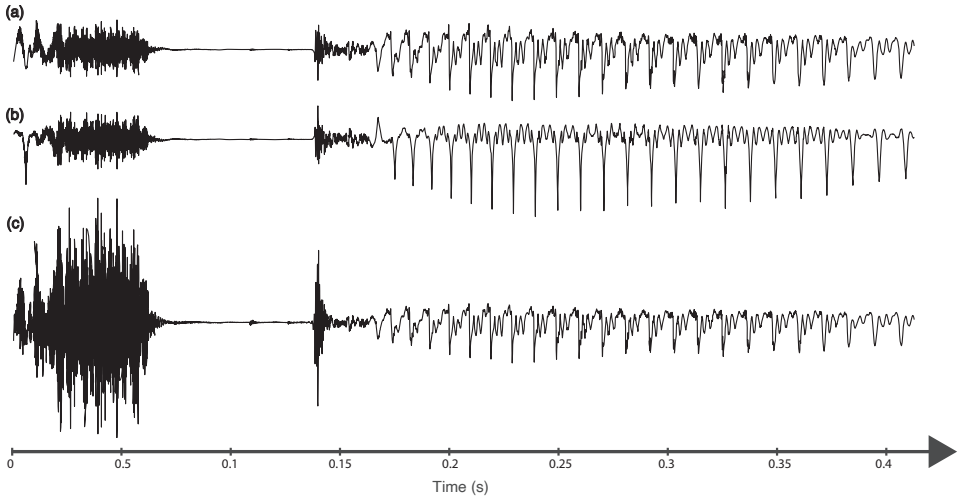

Figure 1: *A short segment of speech ("This kind") fed into the auditory periphery model. (a) Unaltered speech: the original speech signal. (b) Peaky speech: emphasizes the glottal pulses to be "peaky" and "click-like". (c) Transient speech: focuses on the onset and offset of the consonants, emphasizing the "noise-like" section in the speech.*

1.4. Current study

In our study, we implemented a computationally efficient speech modification algorithm, transient speech, to elicit speech-derived ABRs. Such an approach may allow for ABR-based hearing assessments in more natural conditions with little to no burden placed on the user. We compared the simulated ABRs from the auditory periphery model [21] with unaltered speech, peaky speech, transient speech, and non-speech clicks as stimuli. Results showed clear ABRs from our computationally efficient transient speech. Furthermore, a more prominent Wave I was elicited by the transients than unaltered speech and peaky speech. To the best of our knowledge, this is the first study using fast and efficient transients-enhancement algorithm to successfully derive ABRs, especially Wave I. These results demonstrated the great potential of efficiently implemented and natural-sounding transients-enhanced speech to derive ABRs, providing a new way to assess hearing conditions more objectively and naturally.

2. Methods

2.1. Stimuli

The stimuli fed into the auditory periphery model was a single click and three modifications of speech with the same content. The click ABR was simulated using 1 click (0.1 ms single monophasic square wave) from the example code included in the model. The speech stimuli included 40 trials of peaky speech, transient speech, and unaltered speech as the control (see Figure 1). The amplitudes of all speech stimuli were normalized to have the same rms value. The unaltered and peaky speech were downloaded from [16], which were originally extracted from the audiobook The Alchemyst [22], read by a male narrator. The transient speech was modified from the unaltered speech. The unaltered speech was first high-pass filtered with a cutoff of 700 Hz, then decomposed using wavelet packets with a Daubechies-18 wavelet to a depth of 4, resulting in 16 sub-band wavelet packets. For each packet, transients were defined as an abrupt change of the energy calculated by the Euclidean norm of the first derivatives of MFCCs across frequency sub-bands for each time frame [17, 18, 23]. As in [17], we only emphasized the transients of the unvoiced consonant interval but not the spectral transients in the vowels, the latter of which will create unwanted, noticeable fluctuations in loudness. An unvoiced consonant interval was defined as having small short-time energy (i.e., lower in unvoiced consonants than voiced vowels) and a higher zero crossing rate (i.e. higher in unvoiced consonants than voiced vowels) compared to the average across the entire speech signal. This "consonant detection" algorithm was run before the transient detection. The detected transient portions of the speech were enhanced by doubling the amplitude, which kept the stimuli within reasonable volume. Note that the peaky-speech algorithm and the transient-speech algorithm emphasized very different portions of the speech (see Figure 1): the former emphasized the vowel glottal pulses while the latter emphasized the sudden transitions of the consonants.

2.2. ABRs Modeling

We modeled ABR waveforms with the auditory periphery model from [21] with default settings for normal hearing, including 13 high spontaneous rate (70 spikes/s), 3 middle spontaneous rate (10 spikes/s) and 3 low spontaneous rate (1 spike/s) fibers. Due to the computation time, we only used the last 16s of the speech stimuli as input. We focused on Waves I, III, and V as the model's output. The speech ABRs were derived by deconvolution [15, 16, 24], which is based on the assumption of a linear relationship between the speech input and the EEG output with the ABRs considered to be the auditory system's impulse response. Deconvolution was conducted between the model outputs and the regressors, which were a click , glottal pulse train, rectified transients and rectified speech signal, respectively for the click, peaky speech, transient speech, and unaltered speech, normalized to a range of [0 1]. All computation was done in Matlab.

2.3. Statistical analysis

To first verify that continuous speech could derive ABR waveforms similar to traditional click-evoked ABRs, we calculated

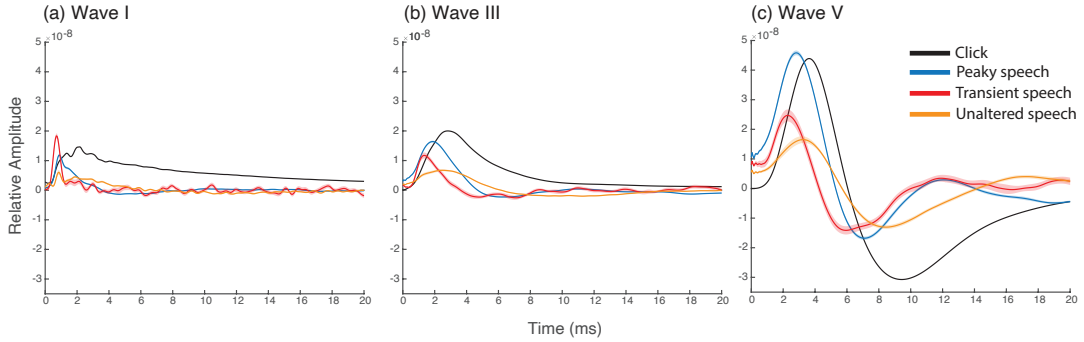

Figure 2: *Model output of one click (black) and (averaged over 40 trials of) peaky speech (blue), transient speech (red), and unaltered speech (yellow). (a) Wave I. (b) Wave III. (c) Wave V. Note that the relative amplitude of the click-evoked ABRs was rescaled by 0.2 to be comparable with the speech-derived ABRs just for the visualization purpose. Shaded errors are SEs.*

the cross-correlation between the model output from clicks and modified speech. Then, amplitude and the latency of each peak were extracted. Please note that the regressors used to compute the deconvolution were in their own arbitrary units, and thus the amplitude of model responses were not comparable among conditions. Therefore, we analyzed the Wave I-V ratio, an index used in previous literature to quantify the sensitivity of the ABRs measurement [25, 26, 27]. The Wave I-V ratio was calculated for each trial and for each speech condition. Repeated-measures analysis of variance (ANOVA) was carried out to compare the Wave I-V ratios among speech conditions. Multiple comparisons were corrected using the Bonferroni correction method (marked as p_b). Some trials did not show clear wave components at the expected latencies; see the Results section and Discussion section for further details.

Table 1: *The maximum normalized cross-correlations and lags (ms) between click and speech.*

Stimuli	*Wave I*	*Wave III*	*Wave V*
Click/Peaky	0.463 (1.050)	0.817 (1.050)	0.802 (0.950)
Click/Transient	0.385 (1.350)	0.625 (1.400)	0.781 (1.700)
Click/Unaltered	0.540 (0.500)	0.776 (0.750)	0.878 (0.600)

3. Results

The model generated clear ABR waveforms from the click stimulus and all speech inputs (Figure 2). We visualized individual ABR components (Waves I, III, and V) separately because the model does not account for all synaptic delays in the auditory pathway and thus does not reflect the accurate temporal structure of the composite ABRs as found in recorded data (see [21] for more details). Moderate and high cross-correlations were found between all three speech ABRs and click ABRs and across all wave components (see Table 1).

The amplitudes and the latencies of individual wave components, Wave I, III and V, are reported in Figure 3, Table 2 and Table 3. To compare the model responses across speech conditions, we analyzed the Wave I-V ratio.

Table 2: *Mean Amplitude and SEs of each peak of Wave I, III and V.*

Stimuli	*Wave I*	*Wave III*	*Wave V*
Peaky speech	1.204E-08(8.744E-11)	1.645E-08(5.106E-11)	4.606E-08(1.309E-10)
Transient speech	1.935E-08(1.375E-10)	1.363E-08(1.252E-10)	3.028E-08(3.658E-10)
Unaltered speech	7.731E-09(1.047E-10)	7.358E-09(5.853E-11)	1.787E-08(2.087E-10)

Table 3: *Mean Latency and SEs of each peak of Wave I, III and V.*

Stimuli	*Wave I*	*Wave III*	*Wave V*
Peaky speech	0.907 (0.000)	1.859 (0.002)	2.816 (0.003)
Transient speech	0.704 (0.002)	1.380 (0.005)	2.317 (0.012)
Unaltered speech	1.615 (0.024)	2.575 (0.022)	3.371 (0.017)

The Wave I-V ratio of click-evoked ABRs was 0.333. For the speech conditions, the mean (standard errors appear in parentheses) of the Wave I-V ratio was largest in the ABRs derived from the transient speech, 0.778 (0.048) compared with the unaltered speech, 0.459 (0.024), and the peaky speech, 0.259 (0.010) (Figure 4). There was a main effect of different speech conditions, $F(2, 114) = 20.480$, $p < .000$, with significantly higher Wave I-V ratio for the transient speech compared with the unaltered speech, $t(39) = 5.946$, $p_b < .000$, $d = 1.712$, and the peaky speech, $t(39) = 10.930$, $p_b < .000$, $d = 2.392$, and higher Wave I-V ratio for the unaltered speech compared with the peaky speech, $t(39) = 5.946$, $p_b < .000$, $d = 1.334$.

Some trials did not yield clear wave components at the expected latencies (see Figure 3); for the purpose of comparison, we excluded the trials which had any wave component more

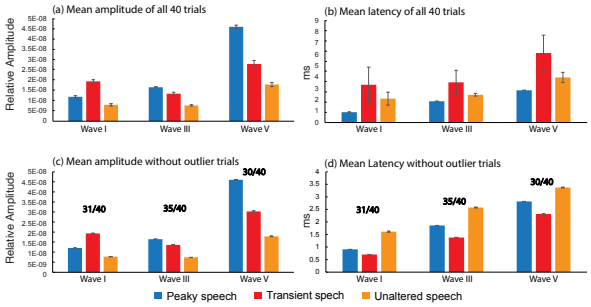

Figure 3: *The average of peak amplitude (a) and latency (b) of Wave I, III and V before and after (c, d) removing the outlier trials across peaky speech (blue), transient speech (red) and unaltered speech (yellow). The number (#/40) on top of the bar chart of (c) and (d) showed the number of trials within 1.5 interquartile ranges of the upper and lower quartile of the averaged latency. Error bars are SEs.*

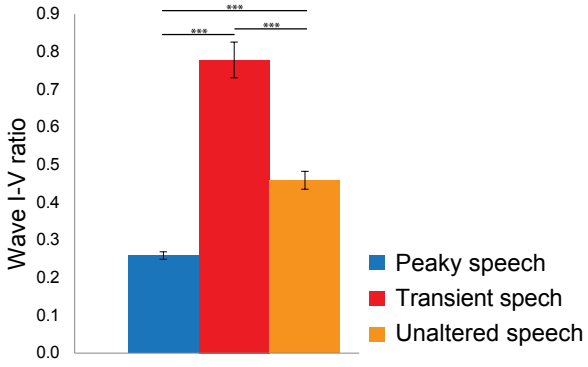

Figure 4: *The average of Wave I-V ratio for peaky speech (blue), transient speech (red) and unaltered speech (yellow). Error bars are SEs. ***p_b <.0001.*

than 1.5 interquartile ranges above the upper quartile or below the lower quartile of the averaged latency and computed the Wave I-V ratio. The statistical results of the Wave I-V ratio among speech conditions remained significant after excluding those trials.

4. Discussion

The current study tested if a computationally efficient speech modification algorithm, the transient speech, can effectively derive ABRs. Accordingly, we compared the derived ABRs from our transient speech to the previously used peaky and unaltered speech [16]. All three speech stimuli generated clear modeled Waves I, III and V as simulated by the auditory periphery model. The transient speech and the peaky speech emphasized different portions of the speech and thus have different model responses. The peaky speech consistently showed Wave I, III and V across trials, while the transient speech showed the most prominent Wave I, which has been particularly challenging to derive from previous studies [25, 26, 27].

Our finding of distinct Wave I in the transient speech may be critical in cochlear synaptopathy or "hidden hearing loss" from putative auditory nerve deficits [7, 8, 9, 10, 11, 12], which do not affect threshold audiometry but rather supra-threshold hearing ability. The transient speech may have great benefit on not only enhanced intelligibility in general but also improved supra-threshold hearing ability for cochlear synaptopathy listeners. Future research will need to establish the relationship between intelligibility and ABRs, and also further investigate the ABRs of hidden hearing loss listeners. Importantly, transients enhanced in the transient speech are not occurring as frequently as the glottal pulses enhanced in the peaky speech, but have higher Wave I-V ratio compared with the peaky speech. The high Wave I-V ratio of the transient speech can be arise from relatively large Wave I or relatively small Wave V, or both, which cannot be concluded based on the data of this study.

Note that before removing the outliers, the latency of the derived wave components were generally more variable in the transient speech than the unaltered and peaky speech (error bars in Figure 3a,b), reflecting the noise across trials. Specifically, some trials showed clear wave components, while some did not. This is consistent with our pilot EEG data (not included here), which also showed more variable transient-derived ABRs across subjects. This may be because transients in the transient speech were more diverse than the glottal pulses in the peaky speech such that transient speech emphasized different consonants with varied build-up and decay rates. More studies may investigate what type of speech transients (e.g. stop, fricative or affricate consonants) could be the best to elicit ABR waveforms. This variability of transient speech also suggests that a longer recording may be needed to more reliably elicit ABRs for hearing assessment.

Using the same speech stimuli, regressors and analysis methods as [16], we compared our modeled ABRs with their recorded ABRs. Consistent with the human ABRs recorded by [16], the modeled ABRs of the peaky speech more reliably showed Wave I, III and V than the unaltered speech, which had significantly smaller amplitude for all wave components. Interestingly, even though the effect was small, our method still showed clear Wave I and III in the unaltered speech condition, which was not visible in the empirical results [16]. This may be due to the noise in real-life EEG collection, which is fundamentally greater than that in computational modeling. The lack of Wave I in their results may be a power issue due to low SNR. More research and data are needed to investigate the recorded ABR derived from the unaltered speech.

The current study applied the auditory periphery model to compare simulated ABRs from unaltered speech, peaky speech, and our transient speech. This computational model, though apparently far from empirical data, is a powerful approach as the first step to test out the modified speech stimuli before running on human subjects. Future studies need to replicate our findings with collected EEG from human subjects.

5. Conclusion

In conclusion, our study demonstrated that continuous speech with varied modifications can be used to derive ABRs. This is the first study to test if the computationally efficient transient speech modification can best elicit ABRs early wave components without potential EEG artifacts that may be picked up in the lab. Importantly, compared to peaky speech and unaltered speech, our transients-enhanced speech derived a clear and significantly higher Wave I-V ratio, which has been difficult to measure reliably in previous studies. These findings shed new light on speech-derived ABRs research, showing the potential of real-time hearing estimations that may have great impact on hearing aids and AR/VR applications.

6. References

[1] M. P. Gorga, T. A. Johnson, J. K. Kaminski, K. L. Beauchaine, C. A. Garner, and S. T. Neely, "Using a combination of click- and toneburst-evoked auditory brainstem response measurements to estimate pure-tone thresholds," *Ear and hearing*, vol. 27, no. 1, p. 60, 2006.

[2] D. R. Stapells and P. Oates, "Estimation of the pure-tone audiogram by the auditory brainstem response: a review," *Audiology and Neurotology*, vol. 2, no. 5, pp. 257–280, 1997.

[3] R. H. Britt and G. T. Rossi, "Neural generators of brainstem auditory-evoked responses," *The Journal of the Acoustical Society of America*, vol. 67, no. S1, pp. S89–S90, 1980.

[4] R. Cai, S. C. Montgomery, K. A. Graves, D. M. Caspary, and B. C. Cox, "The fbn rat model of aging: investigation of abr waveforms and ribbon synapse changes," *Neurobiology of aging*, vol. 62, pp. 53–63, 2018.

[5] N. Bramhall, B. Ong, J. Ko, and M. Parker, "Speech perception ability in noise is correlated with auditory brainstem response wave i amplitude," *Journal of the American Academy of Audiology*, vol. 26, no. 05, pp. 509–517, 2015.

[6] G. C. Stamper and T. A. Johnson, "Auditory function in normal-hearing, noise-exposed human ears," *Ear and hearing*, vol. 36, no. 2, p. 172, 2015.

[7] C. J. Plack, D. Barker, and G. Prendergast, "Perceptual consequences of "hidden" hearing loss," *Trends in hearing*, vol. 18, p. 2331216514550621, 2014.

[8] H. M. Bharadwaj, S. Verhulst, L. Shaheen, M. C. Liberman, and B. G. Shinn-Cunningham, "Cochlear neuropathy and the coding of supra-threshold sound," *Frontiers in systems neuroscience*, vol. 8, p. 26, 2014.

[9] H. M. Bharadwaj, S. Masud, G. Mehraei, S. Verhulst, and B. G. Shinn-Cunningham, "Individual differences reveal correlates of hidden hearing deficits," *Journal of Neuroscience*, vol. 35, no. 5, pp. 2161–2172, 2015.

[10] E. Lobarinas, C. Spankovich, and C. G. Le Prell, "Evidence of "hidden hearing loss" following noise exposures that produce robust tts and abr wave-i amplitude reductions," *Hearing research*, vol. 349, pp. 155–163, 2017.

[11] S. G. Kujawa and M. C. Liberman, "Synaptopathy in the noise-exposed and aging cochlea: Primary neural degeneration in acquired sensorineural hearing loss," *Hearing research*, vol. 330, pp. 191–199, 2015.

[12] M. C. Liberman, M. J. Epstein, S. S. Cleveland, H. Wang, and S. F. Maison, "Toward a differential diagnosis of hidden hearing loss in humans," *PloS one*, vol. 11, no. 9, p. e0162726, 2016.

[13] O. Etard, M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach, "Decoding of selective attention to continuous speech from the human auditory brainstem response," *Neuroimage*, vol. 200, pp. 1–11, 2019.

[14] A. E. Forte, O. Etard, and T. Reichenbach, "The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention," *elife*, vol. 6, p. e27203, 2017.

[15] R. K. Maddox and A. K. Lee, "Auditory brainstem responses to continuous natural speech in human listeners," *Eneuro*, vol. 5, no. 1, 2018.

[16] M. J. Polonenko and R. K. Maddox, "Exposing distinct subcortical components of the auditory brainstem response evoked by continuous naturalistic speech," *Elife*, vol. 10, p. e62329, 2021.

[17] D. M. Rasetshwane, J. R. Boston, and C.-C. Li, "Identification of speech transients using variable frame rate analysis and wavelet packets," in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2006, pp. 1727–1730.

[18] D. M. Rasetshwane, J. R. Boston, C.-C. Li, J. D. Durrant, and G. Genna, "Enhancement of speech intelligibility using transients extracted by wavelet packets," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2009, pp. 173–176.

[19] E. Skoe and N. Kraus, "Auditory brainstem response to complex sounds: a tutorial," *Ear and hearing*, vol. 31, no. 3, p. 302, 2010.

[20] O. Fobel and T. Dau, "Searching for the optimal stimulus eliciting auditory brainstem responses in humans," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2213–2222, 2004.

[21] S. Verhulst, A. Altoe, and V. Vasilkov, "Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss," *Hearing research*, vol. 360, pp. 55–75, 2018.

[22] M. Scott, *The Alchemyst: the secrets of the immortal Nicholas Flamel*, ser. Book 1. New York: New York: Listening Library, 2007.

[23] G. Szwoch, M. Kulesza, and A. Czyżewski, "Transient detection for speech coding applications," *IJCSNS*, vol. 6, no. 12, p. 320, 2006.

[24] E. C. Lalor, A. J. Power, R. B. Reilly, and J. J. Foxe, "Resolving precise temporal processing properties of the auditory system using continuous stimuli," *Journal of neurophysiology*, vol. 102, no. 1, pp. 349–359, 2009.

[25] J. W. Gu, B. S. Herrmann, R. A. Levine, and J. R. Melcher, "Brainstem auditory evoked potentials suggest a role for the ventral cochlear nucleus in tinnitus," *Journal of the Association for Research in Otolaryngology*, vol. 13, no. 6, pp. 819–833, 2012.

[26] A. E. Hickox and M. C. Liberman, "Is noise-induced cochlear neuropathy key to the generation of hyperacusis or tinnitus?" *Journal of neurophysiology*, vol. 111, no. 3, pp. 552–564, 2014.

[27] G. Mehraei, A. E. Hickox, H. M. Bharadwaj, H. Goldberg, S. Verhulst, M. C. Liberman, and B. G. Shinn-Cunningham, "Auditory brainstem response latency in noise as a marker of cochlear synaptopathy," *Journal of Neuroscience*, vol. 36, no. 13, pp. 3755–3764, 2016.