Predictors for Estimating Subcortical EEG Responses to Continuous Speech

Joshua P. Kulasingham[1], *Florine L. Bachmann*[2], *Kasper Eskelund*[3], *Martin Enqvist*[1], *Emina Alickovic*[1,2,*], *Hamish Innes-Brown*[2,4,*]

[1]Automatic Control, Department of Electrical Engineering, Linköping University, Sweden
[2]Eriksholm Research Centre, Snekkersten, Denmark, [3]Oticon A/S, Smørum, Denmark,
[4]Department of Health Technology, Technical University of Denmark, Lyngby, Denmark
[*]Equally contributed as senior authors

name.surname@liu.se, flln,eali,hain@eriksholm.com, ksee@demant.com

Abstract

Perception of sounds and speech involves structures in the auditory brainstem that rapidly process ongoing auditory stimuli. The role of these structures in speech understanding can be investigated by measuring their electrical activity using scalp-mounted electrodes. Typical analysis methods involve averaging responses to many short repetitive stimuli. Recently, responses to more ecologically relevant continuous speech were detected using linear encoding models called temporal response functions (TRFs). Non-linear predictors derived from complex auditory models may improve TRFs. Here, we compare predictors from both simple and complex auditory models for estimating brainstem TRFs on electroencephalography (EEG) data from 24 subjects listening to continuous speech. Predictors from simple models result in comparable TRFs to those from complex models, and are much faster to compute. We also discuss the effect of data length on TRF peaks for efficient estimation of subcortical TRFs.

Index Terms: Auditory Brainstem Response (ABR), deconvolution, speech-ABR, temporal response function (TRF), EEG.

1. Introduction

The human auditory system consists of several subcortical and cortical structures that rapidly process incoming sound signals such as speech. Electroencephalography (EEG) measurements of the aggregate activity of these neural structures have been instrumental in understanding the mechanisms underlying normal hearing and hearing impairments [1, 2, 3]. One important measure is the morphology of the auditory brainstem response (ABR), and the amplitude and latency of ABR peaks have been widely used in many clinical settings such as neonatal hearing screening [4]. However, conventional methods to detect the ABR rely on averaging responses over multiple trials of non-natural, short stimuli such as clicks, chirps or speech syllables [5].

Recently, a method to estimate ABR-like responses to continuous ongoing speech was developed [6, 7], allowing for the exploration of subcortical mechanisms using responses to ecologically relevant speech stimuli. One method to estimate these subcortical responses is the temporal response function (TRF), a linear encoding model of time-locked neural responses to continuous stimuli [8]. TRFs have been widely used for estimating *cortical* responses to speech [9, 10, 11, 12, 13], but few studies have investigated *subcortical* TRFs [6, 14, 15, 16, 17].

Electrical responses that are generated in the brainstem and measured at the scalp are small compared to the amplitude of the on-going EEG (low SNR). They are subsequently difficult to detect, requiring a large amount of data for reliable TRFs [6, 14]. Additionally, subcortical processes rapidly time-lock to fast stimulus fluctuations, and a measurement system with precise synchronization (sub-millisecond) between the stimulus and the measured EEG is essential to detect these responses. Another concern is that TRFs which linearly map the sound stimulus to the EEG ignore several highly non-linear processing stages in the auditory periphery [18]. One study has recently shown that predictors derived from auditory models that incorporate non-linear stages can lead to improved subcortical TRFs [19]. Another recent study showed that auditory-model-derived predictors outperform previously used envelope predictors even for *cortical* TRFs [20].

In this work, we compared predictors derived from auditory models in terms of their suitability for estimating subcortical TRFs. We computed predictors from simpler filterbank models [21] with or without adaptation [22] and compared them to a more complex auditory nerve model [23] that has been previously used to fit subcortical TRFs [19]. TRFs were estimated from EEG data recorded from 24 participants listening to continuous speech. Prior work indicates that the most prominent feature of subcortical TRFs is the wave V peak [6, 16]. This peak was used as the primary measure of performance in our study. Additional measures such as the computational time taken to generate predictors and the amount of data required for fitting TRFs for each predictor type are also reported.

We corroborate the findings of [19] by confirming that the predictor derived from a complex model of the auditory nerve outperforms the rectified speech predictor. Interestingly, our results indicate that predictors from simpler models can reach almost the same performance for estimating wave V peaks as complex models, with the added advantage of being more than 50 times faster to compute. These simpler models, combined with TRF analysis, could lead to efficient algorithms for future 'neuro-steered' hearing aids [24], and encourage the use of more ecologically relevant continuous speech stimuli in clinical applications.

2. Methods

2.1. Experimental setup

EEG data was collected from 24 participants with clinically normal hearing (14 males, mean age 37.16, standard devation 9.64). All participants provided informed consent and the study was approved by the ethics committee for the capital region of Denmark (journal number 22010204). EEG data was recorded while participants were seated listening to continuous segments from a Danish audiobook of H.C. Andersen adventures read by Jens Okking. The 2-channel audio was averaged to form a mono audio channel which was then highpass filtered at 1kHz using a 1st order Butterworth filter. The participants were instructed to relax and listen to the story. There were 8 trials, each consisting

of 4 to 5 minute segments from the audiobook. The single channel speech segments were scaled to have the same root mean square (r.m.s.) value as a 1 kHz pure tone at 72 dB SPL, and was presented using an RME Fireface UCX soundcard (RME Audio, Haimhausen Germany) and Etymotic ER-2 (Etymotic Research, Illinois, USA) insert earphones, which were shielded using a grounded metal box to avoid direct stimulus artifacts on the EEG. Later analysis confirmed that stimulus artifacts were not present in the estimated TRFs.

2.2. EEG data collection and preprocessing

A Biosemi 32-channel EEG system was used with a sampling frequency of 16,384 Hz and a fifth order cascaded integrator-comb anti-aliasing filter with a -3 dB point at 3276.8 Hz. Additional reference electrodes were placed on the mastoids and earlobes, as well as above and below the right eye. Data analysis was conducted in MATLAB (version R2021a) and the Eelbrain Python toolbox (version 0.38.1) [25] using only the Cz channel placed at the scalp center, which was re-referenced to the average of the two mastoid electrodes. The data was highpass filtered using a first order Butterworth filter with cutoff frequency of 1 Hz. To remove power line noise, the signal was passed through FIR notch filters at all multiples of 50 Hz until 1000 Hz, with transition bandwidths of 5 Hz. Simple artifact removal was performed by zeroing out 1 second segments around parts of the EEG data that had amplitudes larger than 5 standard deviations (s.d.) above the mean, similar to prior work[6]. Finally, only the data from 1 to 241 seconds of each trial was used for further analysis to avoid onset effects and to have the same amount of data in each trial.

Detecting subcortical responses requires precise synchronization between the EEG and the audio stimuli. Hence, to avoid trigger jitters and clock drifts, the output of the audio interface was fed to the BioSemi Erg1 channel via an optical isolator to maintain electrical separation between the mains power and the data collection system (StrimTrak, BrainProducts GmbH, Gilching, Germany). The recorded signal from the StimTrak was used to generate predictors for the TRF analysis.

2.3. Auditory models

Predictors were computed using several auditory models, described below in order of increasing complexity. For all models, the input was the audio stimulus, as recorded by the StrimTrak system. The lags inherent in the output of each model were accounted for by shifting the generated predictors to maximize the correlation with the rectified speech predictor, similar to prior work [19]. Since brainstem responses are largely agnostic to stimulus polarities, a pair of predictors were generated for each model, using an input stimulus pair with the original stimulus and the stimulus with opposite sign. In line with prior work [6, 19], TRFs were fit to each polarity separately and then averaged together.

2.3.1. Rectified speech (RS)

Previous studies have shown that the rectified speech signal can be used to estimate subcortical TRFs to continuous speech [6]. This method was used for the first predictor pair, termed RS, which was formed by rectifying the speech stimulus (and the stimulus with opposite sign).

2.3.2. Gammatone spectrogram predictor (GT)

Incoming sounds undergo several stages of non-linear processing in the human ear and cochlea. The gammatone filterbank is a simple linear approximation of this system [26]. A gammatone filterbank consisting of 31 filters from 80-8000 Hz with 1 equivalent rectangular bandwidth (ERB) spacing was applied to the stimulus pair and the resulting amplitude spectra were averaged over all bands to generate the second predictor pair, which was termed GT. The Auditory Modeling Toolbox (AMT) version 1.1.0 [27] (function auditoryfilterbank with default parameters) was used.

2.3.3. Simple model without adaptation (OSS)

The next predictor pair, termed OSS, was generated using the auditory model provided in [22], which is based on the adaptation model in [21]. The implementation in AMT (function osses2021) was used. This model consists of an initial headphone and outer ear pre-filter (stage 1), a gammatone filterbank (stage 2), and an approximation of inner hair cell transduction using rectification followed by lowpass filtering (stage 3). The next stage of the model consists of adaptation loops (stage 4), which approximate the adaptation properties of the auditory nerve. The initial prefilter was omitted since it is not required for stimuli presented with insert earphones. The adaptation stage was also omitted for this version of the model. Therefore only stages 2 and 3 were used, and the resulting signals with 31 center frequencies (similar to GT) were averaged together to form the predictor pair.

2.3.4. Simple model with adaptation (OSSA)

The adaptation loops (stage 4) of the previous auditory model [22] were now included (i.e., stages 2, 3 and 4 were used). The 31 channel output from the adaptation loops were averaged together to generate a pair of predictors, termed OSSA.

2.3.5. Complex model (ZIL)

Finally, a more complex auditory model [23] was used to generate predictors, and was termed ZIL. This model has been recently used to estimate subcortical TRFs [19] and consists of several stages approximating non-linear cochlear filters, inner and outer hair cell properties, auditory nerve synapses and adaptation. The implementation in the Python cochlea package [28] was used with 43 auditory nerve fibers with high spontaneous firing rates and center frequencies logarithmically spaced between 125 Hz and 16 kHz, in line with previous work [19]. To speed up computation, an approximation of the power-law adaptation was used [19]. The outputs of this model are the mean firing rates of the auditory nerves, which were averaged to form the final predictor pair.

2.4. Temporal response function estimation

The TRF is the impulse response of the neural system. TRFs were fit for each predictor using the frequency domain method outlined in previous studies [6, 14] and shown in eq. (1).

$$TRF = \mathcal{F}^{-1}\left\{\frac{\sum_{i=1}^{N} w_i \mathcal{F}\{x_i\}^* \mathcal{F}\{y_i\}}{\sum_{i=1}^{N} \frac{1}{N} \mathcal{F}\{x_i\}^* \mathcal{F}\{x_i\}}\right\} \quad (1)$$

Here, $\mathcal{F}$ denotes the Fourier transform, N is the number of trials, x_i, y_i and w_i are the predictor, EEG signal and weight for

trial i, and $*$ denotes the complex conjugate. The trial weights w_i were set to be the reciprocal of the variance of the EEG data of trial i normalized to sum to 1 across trials. In line with prior work [14], this was done to down-weight noisy (high variance) EEG trials. This frequency domain method results in TRFs with lags from $-T/2$ to $T/2$ where T is the data length.

Two TRFs were estimated separately for each predictor pair, and then averaged together. These TRFs were then band-pass filtered between 30-1000 Hz using a delay compensated FIR filter and then smoothed using a Hamming window of width 2 ms. The smoothing step was necessary since this unregularized TRF approach resulted in noisy estimates for the OSS and OSSA models (see Discussion). The TRF segment from -10 to 30 ms was extracted for further analysis. Finally, the baseline activity (mean of the TRF segment from -10 to 0 ms) was subtracted from each TRF. To investigate the effect of data length, TRFs were estimated on consecutively increasing number of trials (4, 8, 12, ..., 32 minutes). This resulted in 8 TRFs for each predictor that allowed for quantifying the improvement of TRF estimation with increasing data length.

2.5. Performance metrics and statistical tests

The most prominent feature of ABR TRFs is the wave V peak that occurs around 5-10 ms [6, 15, 16]. The amplitude of this wave V peak was used as the primary metric for comparing TRFs from each predictor type. The SNR of the wave V peak was computed, similar to prior work [6]. First, the TRF peak between 5-10 ms was automatically detected, and the power in a 5 ms window around the peak was computed as a measure of the signal power S. Next, the noise power N was estimated as the average TRF power in 5 ms windows in the range -500 to -20 ms. Finally, the wave V SNR was computed as $SNR = 10\log_{10}(S/N)$.

The amplitudes and latencies of the TRF wave V for each predictor for each participant were also extracted. The consistency of individual wave V was investigated using correlations of wave V amplitudes and latencies across predictors. The overall execution times to generate each predictor and the Pearson correlations between predictors are also reported in Table 1.

Statistical analysis was performed using a linear mixed effects (LME) model with wave V SNR as the dependent factor, predictor type and data length as fixed effects, and participant number as a random effect. Two participants were excluded from these tests since they did not have data from all trials. Post-hoc two-tailed paired t-tests with Holm-Sidak correction were used to test for pairwise differences in wave V SNR across predictors for the TRFs fit on all 8 trials.

3. Results

3.1. Subcortical TRFs for predictors derived from auditory models

A comparison of the computational time required to generate each predictor and their correlations with the simplest (RS) and the most complex (ZIL) models are provided in Table 1. The computations were performed on an AMD Ryzen 7 PRO 5850U 1.9 GHz CPU with 32 GB RAM. Note that even the approximate ZIL model is more than 50 times slower than the others.

The TRFs for the five predictors over all 24 subjects are shown in Fig. 1. The TRF shows a clear wave V peak for all predictors. The wave V peak latency slightly varies across the predictor types, even after removing lags arising from the models themselves by shifting each predictor to have the maximum correlation with RS (see Discussion).

Table 1: *Predictor comparison*

Predictor	**Computation Time** (1 s input)	**Correlation with RS**	**Correlation with ZIL**
RS	-	-	0.316
GT	0.0521 s	0.461	0.550
OSS	0.0563 s	0.438	0.496
OSSA	0.0680 s	0.262	0.577
ZIL	4.1208 s	0.316	-

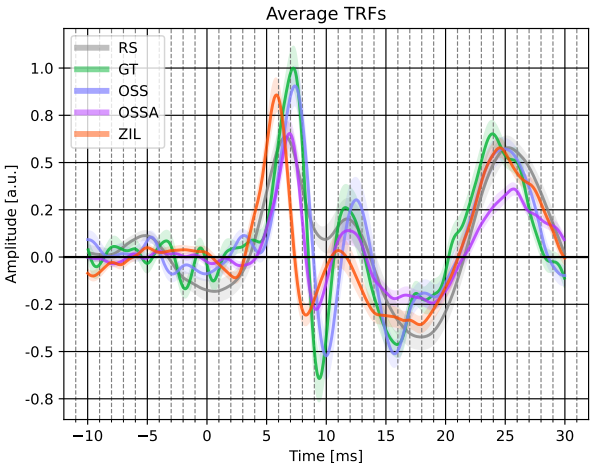

Figure 1: *TRFs for each predictor. The mean and standard error of the mean (s.e.m.) across 24 subjects is shown. Clear wave V peaks are seen for all TRFs.*

3.2. Interaction of data length and predictor type on subcortical TRFs

The amount of data required for clear wave V peaks was investigated by fitting TRFs on an increasing number of 4 minute trials. The wave V SNR was used as the metric to compare TRFs across predictors and data lengths as shown in Fig. 2. Almost all subjects reached above zero SNR with 12 minutes of data for all models except RS, and the LME found a significant interaction between data length and predictor type ($F_{28,588} = 4.66, p < 0.001$). Two further trends were observed; 1) models with filterbanks (GT, OSS) had on average higher SNR compared to RS, 2) models with adaptation and level dependency (OSSA, ZIL) had on average higher SNR compared to filterbanks. Interestingly, wave V SNR of the simpler OSSA model was comparable to the more complex ZIL model. For the TRFs fit on 32 minutes of data, pairwise t-tests with Holm-Sidak correction revealed that wave V SNR was significantly higher for all model conditions compared to RS (all $p < 0.015$), and for OSSA and ZIL vs. OSS and GT (all $p < 0.001$).

3.3. Individual amplitudes and latencies of wave V

Finally, the individual wave V amplitudes and latencies for the OSSA and ZIL predictors were compared as shown in Fig. 3. The OSSA model showed a high degree of correlation with the ZIL model (0.973 for the peak amplitudes and 0.926 for the

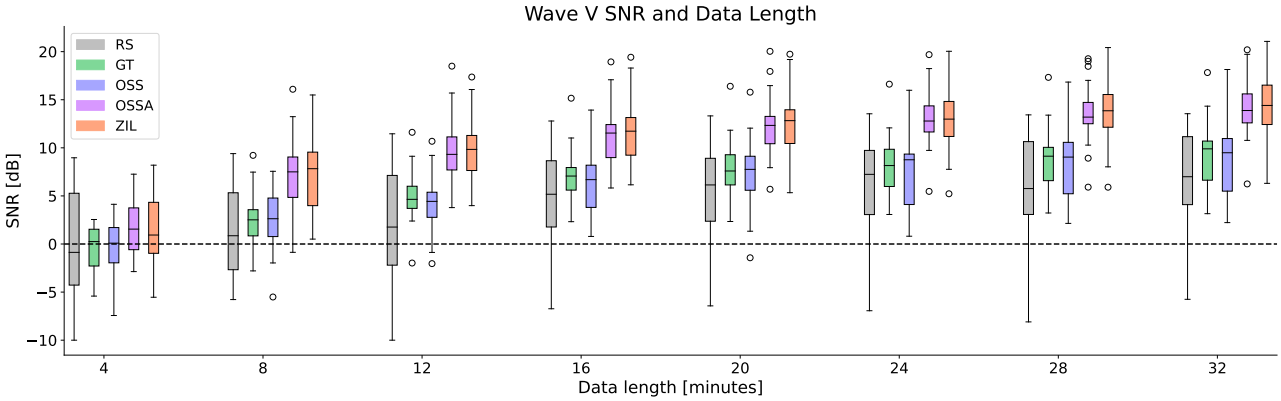

Figure 2: *Effect of predictor type and data length on wave V SNR. Boxplots are shown across 24 subjects.*

peak latencies), confirming that both models provide TRF wave V estimates that are consistent across subjects. However, the ZIL model has a shorter mean latency, also seen in Fig. 1 (see Discussion). Nevertheless, this correlation analysis indicates that the simpler OSSA model may provide a good trade-off between computational efficiency and reliable wave V peaks.

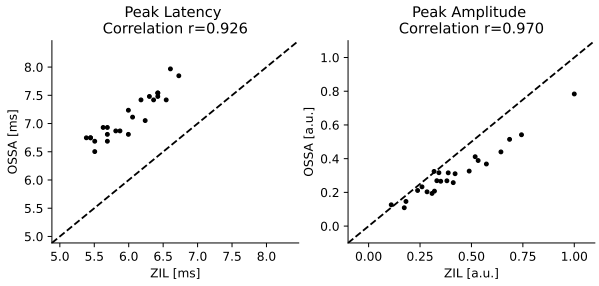

Figure 3: *Individual wave V peak amplitudes and latencies. Scatterplots across subjects and the correlation between OSSA and ZIL are shown.*

4. Discussion

In this work, we compared the suitability of several predictors for estimating subcortical TRFs to continuous speech. Results indicate that the addition of filterbanks and adaptation stages to the predictor models greatly improves estimation of wave V in the TRFs over rectified speech predictors. We show that even simpler models may allow for robust wave V peaks with around 12 minutes of data. These simple models give wave V estimates that are comparable to a more complex model, even though the complex model is 50 times slower to compute. However, it must be noted that OSSA wave V SNRs were comparable to ZIL only after smoothing the TRFs using a 2 ms Hamming window (see Methods), perhaps because the OSSA TRFs were noisier. Other methods such as regularized regression, which is widely used for cortical TRFs [29, 30, 31], or direct estimation of TRF peaks [32] may be able to overcome this issue. Nevertheless, our correlation analysis revealed that these smoothed TRFs resulted in wave V peak amplitudes and latencies for OSSA and ZIL that were consistent across subjects.

This work does not provide an exhaustive list of auditory models or predictors for estimating subcortical TRFs. We also do not directly compare the performance of the auditory models themselves (see [33]), but only evaluate their suitability to generate predictors for subcortical TRFs. Several other models [34, 35] could be utilized to generate predictors, although our work suggests that simple models are reliable enough to fit TRFs with clear wave V peaks.

It must be noted that although the wave V peak was used as the primary metric of performance, the conventional click ABR consists of several other morphological features [36]. The wave V peak was selected here to both be consistent with prior work [6, 19, 16], and because it was the only consistent feature that was detected in all subjects. TRFs using ZIL had shorter wave V peak latencies (see Figures 1 and 3). It is possible that the wave V from the ZIL model is earlier since the ZIL model better incorporates peripheral non-linearities. This may provide a predictor that is similar to intermediate signal representations in the auditory pathway near the wave V generators, which could in turn result in an earlier estimated wave V. Further investigation is needed to disentangle the effects of model lags in order to ascertain whether these latency differences are meaningful properties of the ABR. Future work could also explore if other features of the conventional ABR can be reliably detected using subcortical TRFs.

Finally, this work only analyses subcortical responses to *speech* stimuli. Recent work indicates that complex auditory model predictors (ZIL) provide significant advantages over rectified speech when estimating subcortical TRFs for *music* [22]. Future work could investigate the suitability of *simpler* auditory model predictors for estimating TRFs for non-speech stimuli.

5. Conclusions

This work provides a systematic comparison of predictors derived from auditory peripheral models for estimating subcortical TRFs to continuous speech. Our results indicate that simple models with filterbanks and adaptation loops may suffice to estimate reliable subcortical TRFs. Such efficient algorithms may pave the way toward the use of more ecologically relevant natural speech for investigating hearing impairment and for future neuro-steered hearing aids.

6. Acknowledgements

This work was supported by the William Demant Foundation. The authors are also grateful to all participants for their participation in this study.

7. References

[1] E. Alickovic, T. Lunner, D. Wendt, L. Fiedler, R. Hietkamp, E. H. N. Ng, and C. Graversen, "Neural representation enhanced for speech and reduced for background noise with a hearing aid noise reduction scheme during a selective attention task," *Front. Neurosci.*, vol. 14, p. 846, 2020.

[2] T. Lunner, E. Alickovic, C. Graversen, E. H. N. Ng, D. Wendt, and G. Keidser, "Three new outcome measures that tap into cognitive processes required for real-life communication," *Ear Hear.*, vol. 41, no. Suppl 1, p. 39S, 2020.

[3] E. Alickovic, E. H. N. Ng, L. Fiedler, S. Santurette, H. Innes-Brown, and C. Graversen, "Effects of hearing aid noise reduction on early and late cortical representations of competing talkers in noise," *Front. Neurosci.*, vol. 15, p. 636060, 2021.

[4] M. P. Warren, "The Auditory Brainstem Response in Pediatrics," *Otolaryngologic Clinics of N. America*, vol. 22, no. 3, pp. 473–500, 1989.

[5] E. Skoe and N. Kraus, "Auditory brainstem response to complex sounds: a tutorial," *Ear Hear.*, vol. 31, no. 3, pp. 302–324, 2010.

[6] R. K. Maddox and A. K. C. Lee, "Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners," *eNeuro*, vol. 5, no. 1, 2018.

[7] 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.

[8] 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," *J. Neurophys.*, vol. 102, no. 1, pp. 349–359, 2009.

[9] G. Di Liberto, J. O'Sullivan, and E. Lalor, "Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing," *Curr. Biol.*, vol. 25, no. 19, pp. 2457–2465, 2015.

[10] C. Brodbeck and J. Z. Simon, "Continuous speech processing," *Curr. Opin. Physiol.*, vol. 18, pp. 25–31, 2020.

[11] C. Brodbeck, A. Presacco, and J. Z. Simon, "Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension," *NeuroImage*, vol. 172, pp. 162–174, 2018.

[12] J. P. Kulasingham, N. H. Joshi, M. Rezaeizadeh, and J. Z. Simon, "Cortical Processing of Arithmetic and Simple Sentences in an Auditory Attention Task," *J. Neurosci.*, vol. 41, no. 38, pp. 8023–8039, 2021.

[13] J. P. Kulasingham, C. Brodbeck, A. Presacco, S. E. Kuchinsky, S. Anderson, and J. Z. Simon, "High gamma cortical processing of continuous speech in younger and older listeners," *NeuroImage*, vol. 222, p. 117291, 2020.

[14] 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.

[15] F. L. Bachmann, E. MacDonald, and J. Hjortkjær, "A comparison of two measures of subcortical responses to ongoing speech: Preliminary results," *Proc. Int. Symp. on Auditory and Audiological Research*, vol. 7, pp. 461–468, 2019.

[16] F. L. Bachmann, E. N. MacDonald, and J. Hjortkjær, "Neural Measures of Pitch Processing in EEG Responses to Running Speech," *Front. Neurosci.*, vol. 15, p. 738408, 2021.

[17] M. Kegler, H. Weissbart, and T. Reichenbach, "The neural response at the fundamental frequency of speech is modulated by word-level acoustic and linguistic information," *Front. Neurosci.*, vol. 16, 2022.

[18] M. Saiz-Alía and T. Reichenbach, "Computational modeling of the auditory brainstem response to continuous speech," *J. Neural Eng.*, vol. 17, no. 3, p. 036035, 2020.

[19] T. Shan, M. S. Cappelloni, and R. K. Maddox, "Music and Speech Elicit Similar Subcortical Responses in Human Listeners," *bioRxiv*, 2022.

[20] E. Lindboom, A. Nidiffer, L. H. Carney, and E. Lalor, "Incorporating models of subcortical processing improves the ability to predict EEG responses to natural speech," *bioRxiv*, 2023.

[21] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *J. Acoust. Soc. Am.*, vol. 102, no. 5, pp. 2892–2905, 1997.

[22] A. Osses Vecchi and A. Kohlrausch, "Perceptual similarity between piano notes: Simulations with a template-based perception model," *J. Acoust. Soc. Am.*, vol. 149, no. 5, p. 3534, 2021.

[23] M. S. A. Zilany, I. C. Bruce, and L. H. Carney, "Updated parameters and expanded simulation options for a model of the auditory periphery," *J. Acoust. Soc. Am.*, vol. 135, no. 1, pp. 283–286, 2014.

[24] S. Geirnaert, S. Vandecappelle, E. Alickovic, A. de Cheveigne, E. Lalor, B. T. Meyer, S. Miran, T. Francart, and A. Bertrand, "Electroencephalography-Based Auditory Attention Decoding: Toward Neurosteered Hearing Devices," *IEEE Signal Process. Mag.*, vol. 38, no. 4, pp. 89–102, 2021.

[25] C. Brodbeck, P. Das, J. P. Kulasingham, S. Bhattasali, P. Gaston, P. Resnik, and J. Z. Simon, "Eelbrain: A Python toolkit for time-continuous analysis with temporal response functions," *bioRxiv*, p. 2021.08.01.454687, 2022.

[26] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *a meeting of the IOC Speech Group on Auditory Modelling at RSRE*, vol. 2, no. 7, 1987.

[27] P. Majdak, C. Hollomey, and R. Baumgartner, "AMT 1.x: A toolbox for reproducible research in auditory modeling," *Acta Acust*, vol. 6, p. 19, 2022.

[28] M. Rudnicki, O. Schoppe, M. Isik, F. Völk, and W. Hemmert, "Modeling auditory coding: from sound to spikes," *Cell Tissue Res.*, vol. 361, no. 1, pp. 159–175, 2015.

[29] D. D. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. De Cheveigne, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Front. Neurosci.*, vol. 12, p. 531, 2018.

[30] E. Alickovic, T. Lunner, F. Gustafsson, and L. Ljung, "A tutorial on auditory attention identification methods," *Front. Neurosci.*, p. 153, 2019.

[31] M. J. Crosse, N. J. Zuk, G. M. Di Liberto, A. R. Nidiffer, S. Molholm, and E. C. Lalor, "Linear modeling of neurophysiological responses to speech and other continuous stimuli: methodological considerations for applied research," *Front. Neurosci.*, p. 1350, 2021.

[32] J. P. Kulasingham and J. Z. Simon, "Algorithms for Estimating Time-Locked Neural Response Components in Cortical Processing of Continuous Speech," *IEEE Trans. Biomedical Eng.*, vol. 70, no. 1, pp. 88–96, 2023.

[33] A. O. Vecchi, L. Varnet, L. H. Carney, T. Dau, I. C. Bruce, S. Verhulst, and P. Majdak, "A comparative study of eight human auditory models of monaural processing," *Acta Acust.*, vol. 6, p. 17, 2022.

[34] H. Relaño-Iborra, J. Zaar, and T. Dau, "A speech-based computational auditory signal processing and perception model," *J. Acoust. Soc. Am.*, vol. 146, no. 5, pp. 3306–3317, 2019.

[35] S. Verhulst, H. M. Bharadwaj, G. Mehraei, C. A. Shera, and B. G. Shinn-Cunningham, "Functional modeling of the human auditory brainstem response to broadband stimulation," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1637–1659, 2015.

[36] T. Picton, S. Hillyard, H. Krausz, and R. Galambos, "Human auditory evoked potentials. I: Evaluation of components," *Electroencephalography and Clinical Neurophys.*, vol. 36, pp. 179–190, 1974.