An adaptive gain model for predicting auditory brain activity in mice outperforms standard methods for predicting cortical speech tracking in human EEG

Adèle Simon[1], Adrian N. L. Sahani[2], Maria Chait[1], and Jennifer F. Linden[1,3]

[1]Ear Institute, University College London, London, UK

[2]Department of Mathematics, King's College London, London, UK

[3]Department of Neuroscience, Physiology & Pharmacology, University College London, London, UK

Abstract

Human brain activity tracks slow fluctuations in continuous sound, including speech, and modelling this relationship provides insight into how the brain encodes naturalistic auditory input. Traditional approaches to modelling the tracking of speech in the brain with linear regression often use the amplitude envelope of the stimulus as a regressor, implicitly assuming that neural responses scale linearly with sound intensity. However, previous studies of central auditory processing in both animals and humans have shown that adaptive mechanisms adjust auditory responses based on recent sound history. Here, we show that a simple nonlinear transformation of the stimulus envelope, derived from studies of mouse auditory thalamic responses to temporally varying sounds, significantly improves modelling of human cortical speech tracking. This Adaptive Gain transformation essentially normalizes sound level by its recent context. Improvements in modelling of cortical speech tracking using the Adaptive Gain representation are robust across experiments, participant groups, and languages, indicating that short-timescale adaptation to recent sound level is a general feature of auditory processing. We further show that the optimal adaptation time constant for human cortical responses to continuous speech is approximately 50–100 ms, longer than previously observed for auditory thalamic responses to temporally varying sounds in mice. In summary, the Adaptive Gain transformation is a mathematically simple alternative to the standard envelope representation that captures dynamic adaptation in auditory processing and reliably improves prediction of human cortical EEG signals during listening to continuous speech.

1 Introduction

In humans listening to continuous speech, cortical EEG activity tracks slow fluctuations in the auditory signal. This "cortical speech tracking" phenomenon is believed to reflect the aggregation of successive evoked responses to transient acoustic events, such as word and syllable onsets or sudden changes in sound energy, within the ongoing speech stream [3, 45]. The amplitude envelope representation of speech captures power variations that can drive such evoked responses, and several studies have highlighted a link between the envelope of an incoming sound and the neural response measured using EEG [30, 54, 64, 22]. The most common approach to exploring such links between an audio representation (such as the amplitude envelope) and a neural response involves modelling the relationship between the two and assessing how accurately such models can predict neural responses to novel audio inputs [12, 43]. For human auditory cortical responses to continuous speech, this approach typically involves linear modelling to fit a "Temporal Response Function" (TRF) relating an audio representation of the speech signal (typically the amplitude envelope) to EEG activity [39, 16].

The amplitude envelope representation of the audio signal has been the most commonly used TRF regressor in cortical speech tracking studies [30, 54, 64, 22] because there is substantial evidence that the envelope is central to speech perception [52, 3, 57, 44]. However, the envelope alone is acknowledged to be an overly simplistic representation for characterizing auditory temporal processing and the nonlinear transformations occurring along the auditory pathway. By using more complex representations of the audio signal that incorporate known features of auditory processing and comparing the predictive performance of the corresponding TRFs, it is possible to explore which acoustic features of continuous speech might be tracked by the auditory system. Similar strategies have been used to evaluate different models of the acoustic features driving cortical [27, 40, 51, 9, 60] and subcortical responses [36, 51, 42].

One limitation of using the envelope as a stimulus representation in studies of continuous speech is that this approach assumes the cortical response depends solely on the absolute value of speech amplitude at a given moment. This assumption may be incorrect, as it overlooks adaptive mechanisms that adjust neural responses to transient events based on their context [41, 61, 2]. Such adaptive mechanisms have been documented across multiple stages of the auditory pathway [50, 46], including at both subcortical levels [61, 37, 20] and cortical levels [60, 4, 2, 53, 58, 59, 18], and evidence for adaptive mechanisms has also been noted at the perceptual level [56, 33]. Adaptation to auditory context is thought to be critical for sound perception in noisy environments [34, 8], including speech-in-noise situations [26].

Here, we asked whether a simple adaptive nonlinear function of the stimulus envelope previously derived from a model of temporal processing in the auditory thalamus of anesthetized mice [6] could be used to improve cortical tracking of continuous speech in awake humans listening to audiobooks. The

"Adaptive Gain" transformation essentially normalizes the amplitude envelope by recent sound level weighted by a function that decays exponentially in preceding time. Adaptive Gain can interpreted as representing the moment-by-moment responsiveness of the central auditory system to sound level; it is high when sound levels have recently been low, and low when sound levels have recently been high. We find that even when applied only to the broadband stimulus envelope and with parameters optimized originally for modelling neural population responses in the auditory thalamus of anesthetized mice, the Adaptive Gain representation improves human cortical speech tracking across experiments, participant groups, and languages. We also identify optimized Adaptive Gain parameters for human studies that enhance cortical speech tracking even further.

The results demonstrate that cortical speech tracking is strongly influenced by nonlinear adaptation to sound level on a timescale of tens to hundreds of milliseconds, and that a simple, easily applied nonlinear transformation of the stimulus envelope enables improved prediction of human EEG responses to continuous speech.

2 Methods

2.1 Data

2.1.1 Dataset description

Two separate datasets [55, 31] were used for the present study. Both were publicly available datasets. They were chosen based on the following criteria: (i) EEG was recorded during speech listening, with continuous segments lasting at least 30 seconds; (ii) recording was done during listening to clean speech, without additional noise; (iii) participants did not report any hearing or neurological disorders; (iv) the raw EEG data was available; and (v) the original audio used in the experiments was available.

The first dataset [55] analysed here is a subset of the data collected and used in [54]. It consists of EEG recordings from 21 native Danish speakers listening to audiobooks in Danish and in Finnish without background noise for approximately 6 minutes for each language (divided into 70-second trials). Audiobooks were played on loudspeakers at 60 dBA measured at the listening position. Participants' cortical responses while listening were recorded continuously at 2400 Hz with a g.HIamp-Research (g.tec Medical Engineering GmbH, Austria) system and active electrodes positioned according to the standard 10-20 system.

The second dataset [31] is a subset of the collected data used for [29, 30]. It comprises data recorded in 18 native English speakers listening to 10 minutes of natural continuous English speech and 10 minutes of natural continuous Dutch speech without background noise, through Etymotic ER-3C insert tube earphones at 76dBA. Participants' cortical responses were continuously recorded

at 1 kHz through a 64-channel scalp EEG system using active electrodes (actiCAP, BrainProducts, Germany) and a multi-channel EEG amplifier (actiCHamp, BrainProducts, Germany). The electrodes were positioned according to the standard 10-20 system and referenced to the right earlobe.

Both datasets included data recorded while participants listened to speech in a language they understood and also to speech in a foreign language they did not understand. We focus here primarily on analysis of data for understood speech, to avoid effects due to language understanding or variation in attention. However, we also analyzed data for not-understood speech and obtained similar results, as shown at the end of the Results section.

2.1.2 EEG preprocessing

All EEG data were preprocessed using custom-written Matlab code (R2022a; The MathWorks, Natick, MA, USA) combined with the EEGLAB toolbox [23]. The EEG data were re-referenced to an average reference. The data were filtered with first a 3rd-order high-pass Butterworth filter with a cutoff frequency of 0.5 Hz, then a 5th-order low-pass Butterworth filter with a cutoff frequency of 15 Hz [17]. Both filters were implemented using the *filtfilt* function in Matlab, to ensure zero phase distortion [19]. Data were downsampled to 200 Hz to reduce processing time and separated into trials. The first and last 5 seconds of each trial were removed to exclude the onsets and offsets of the speech stimuli [17]. Finally, EEG signals were z-score-normalized before estimation of Temporal Response Functions relating the EEG to different audio representations.

2.2 Audio Representations

Three audio representations were extracted from the original audio stimuli and used as regressors for TRF modelling: Envelope, Logarithmic Envelope (LogEnv), and Adaptive Gain.

After extraction, all audio representations were downsampled to 200 Hz to match the EEG sampling frequency using the decimate function in MATLAB (version R2022a), which includes an 8th-order low-pass Chebyshev Type I infinite impulse response anti-aliasing filter with a cut-off frequency of 79 Hz. As the absolute values differ between the three audio representations, all audio representations were z-score-normalized to standardize them relative to overall temporal variability. Examples of the different audio representations of noise and speech are shown in Figure 1.

2.2.1 Envelope

The envelope corresponds to the slower energy fluctuations that are tracked by the auditory cortex [38, 1] and are known to be crucial for speech intelligibility [52]. It has been shown that cortical speech tracking is dominated by the processing of the amplitude envelope [47]. The envelope is also

the most commonly used audio representation in speech tracking research [30, 54, 64, 22]. Therefore, the envelope served as a reference for assessing whether other audio representations could enhance TRF modelling performance.

The envelope was extracted from the speech signals by applying a full-wave rectification to the original sound pressure waveform $s(t)$ as:

$$Env(t) = abs(s(t)) \tag{1}$$

No additional low-pass filtering was applied following the full-wave rectification, as a low-pass filter (with a cut-off frequency of 79 Hz) was applied as a part of the downsampling.

2.2.2 Logarithmic Envelope (LogEnv)

The Adaptive Gain computation relies on the broadband envelope of the audio signal, logarithmically scaled, extracted as follows:

$$LogEnv(t) = \log_{10}(Env(t) + \epsilon) \tag{2}$$

where Env(t) is the envelope of the audio signal as defined in the previous subsection, and ϵ is a small positive constant included to ensure that the argument of the logarithm remains positive. In both human and animal studies, it has been shown that central auditory encoding of the intensity of an auditory signal is better described using logarithmic than linear representations of sound intensity [28, 2, 48, 3]. However, previous work failed to find an improvement in speech tracking for the logarithmic representation compared to the envelope [27, 9]. As the adaptive gain computation incorporates a logarithmic transformation of the envelope, we included the LogEnv representation among the audio representations compared to the Adaptive Gain representation in order to determine the extent to which logarithmic transformation alone might account for any improvements in speech tracking performance.

2.2.3 Adaptive Gain

The Adaptive Gain transformation can interpreted as capturing dynamic responsiveness of the central auditory system to sound level on a millisecond timescale; it is high when sound levels have recently been low, and low when sound levels have recently been high. Computation of Adaptive Gain is a key step in simple 6-parameter model of auditory temporal processing originally developed to predict neural population activity evoked by gap-in-noise stimuli in the auditory thalamus of anaesthesized mice [6]. The Adaptive Gain calculation is not intended to model specific physiological or cellular processes but to describe the adaptive intensity gain control performed by the auditory pathway through a nonlinear

normalization [14].

The Adaptive Gain function $AG(t)$ can be computed as follows:

$$AG(t) = \frac{1}{1 + \sum_{j=t-M}^{t} w_A(t-j)(\sum_{k=j-N}^{j} w_I(j-k) LogEnv(k))} \tag{3}$$

where $LogEnv(k)$ is the logarithmic envelope transformation of the audio signal as defined in the previous subsection. w_I is the integration window defined by:

$$w_I(a) = \begin{cases} K_I \cdot e^{-a/\tau_I} & if\ 0 \leq a \leq N \\ 0 & otherwise \end{cases} \tag{4}$$

where N is the maximum duration of the integration window and is set to $5 * \tau_I$ to ensure $> 99\%$ decay of the exponential at the earliest time points falling within the window. K_I is chosen such that $\sum_{a=0}^{N} w_I(a) = 1$.

w_A is the adaptation window defined in a similar manner:

$$w_A(b) = \begin{cases} K_A \cdot e^{-b/\tau_A} & if\ 0 \leq b \leq M \\ 0 & otherwise \end{cases} \tag{5}$$

where M is the maximum duration of the integration window and is set to $5 * \tau_A$. K_A is defined to have $\sum_{a=0}^{N} w_I(a) = 1$.

The two successive convolutions $w_I * w_A$ within the Adaptive Gain calculation can be simplified to the following mathematical form (see Supplementary Information for details):

$$v_{IA}(t) = K_{IA}(e^{-t/\tau_A} - e^{-t/\tau_I}) \tag{6}$$

with τ_A and τ_I as defined above, and K_{IA} a single constant defined such that $\sum_{t=0}^{\infty} v_{IA}(t) = 1$. This simplification produces a single smooth weighting function of time preceding the current moment in the sound signal. The $v_{IA}(t)$ function is zero at $t = 0$ and rises rapidly to a peak at a preceding time point, then falls more gradually across time steps further into the past.

The Adaptive Gain function $AG(t)$ can be rewritten more simply in terms of a convolution between this single smoothing function $v_{IA}(t)$ and $LogEnv(t)$ as defined previously:

$$AG(t) = \frac{1}{1 + v_{IA}(t) * LogEnv(t)} \tag{7}$$

As can be seen in Equation 6, the integration and adaptation windows τ_I and τ_A are mathematically equivalent and interchangeable in the computation of the Adaptive Gain. They were, however, designed

to describe distinct operations in the larger model of temporal processing developed in [6] to predict neural population responses to gap-in-noise stimuli in the mouse auditory thalamus. Therefore, we have retained the original labelling of integration and adaptation time constants in our descriptions here, even though the distinction is irrelevant for the Adaptive Gain computation in isolation.

For our initial analysis of human cortical speech tracking using the Adaptive Gain representation of continuous speech, the values of the variables defining the length of the integration and adaptation windows were fixed to values similar to those that had been identified as optimal for predicting neural population responses to gap-in-noise stimuli in the auditory thalamus of anesthetized mice [6] ($\tau_{A\ mice} = 10$ ms and $\tau_{I\ mice} = 6$ ms). As τ_I was originally defined to be an integration window representing more fundamental limits for temporal integration of the stimulus envelope (e.g., on the order of a few milliseconds, close to the limits of gap-in-noise detection for both mice and humans), we set $\tau_{I\ human} = \tau_{I\ mice} = 5$ ms (rounding the original mouse-optimized value of 6 ms in [6] down to 5 ms to align with the 200 Hz downsampled EEG sample rate used here). The adaptation time constant τ_A is theoretically less constrained, and the mouse-optimized values may not be optimal for modelling adaptation at the cortical level in humans. Thus, a second step in the analysis involved an estimation of the optimal $\tau_{A\ human}$ values. Optimization of the adaptation parameter was done with a grid search to find the values of τ_A that yielded the best performance for a forward TRF model. The values tested in this grid search were $\tau = 5, 10, 25, 50, 75, 100, 250, 500$ ms. Increasing τ_A will flatten the shape of the v_{IA} function, which increases the amount of previous data included in the smoothing. Increasing τ_A also introduces a delay in the peak of the v_{IA} function, as illustrated in Supplementary Information.

Adaptive Gain functions as a form of dynamic intensity gain control, and therefore varies inversely with sound intensity. When sound levels are low, the Adaptive Gain values will go high, and conversely, when sound levels are high, the Adaptive Gain will go low. To facilitate comparison with envelope and LogEnv representations, Adaptive Gain values were multiplied by -1 and then all three representations were scaled to the same range for visualization, so that the variations of envelope, LogEnv and Adaptive Gain evolved in the same direction.

Comparison of the envelope, LogEnv and adaptive gain representations in Figure 1 reveals that the adaptive gain function smooths out sharp changes and rapid fluctuations in sound intensity, such as the abrupt noise onsets and offsets in Figure 1A and the rapid fluctuations within the speech signal in Figure 1B. Applied to speech signals, the adaptive gain transformation produces a much smoother representation than the envelope or logarithmic envelope transformation (Figure 1B). The adaptive gain representation also serves to highlight the onsets and offsets of sound segments surrounded by periods of silence, such as the beginnings and endings of words and sentences in continuous speech

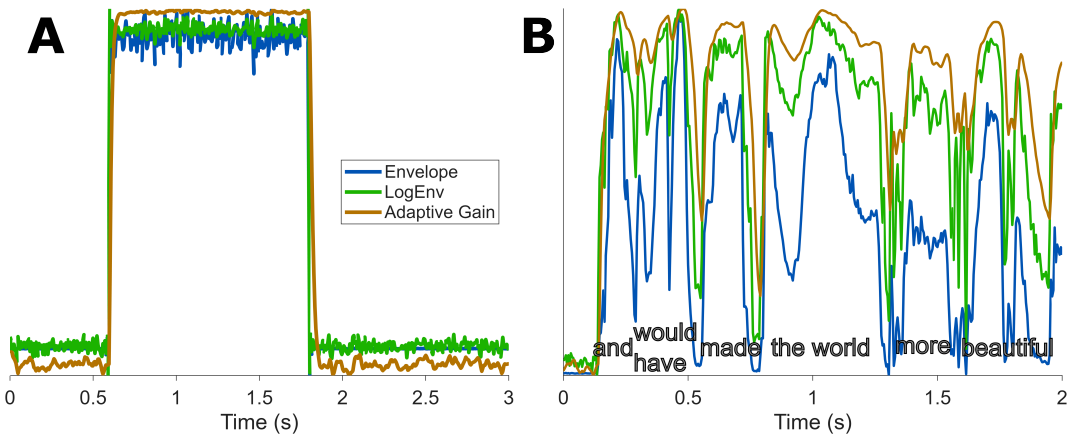


Figure 1: Examples of the audio representations compared in the current study. For visualization purposes, Adaptive Gain values were vertically flipped (multiplied by -1) and then all audio representations were shifted and scaled to the same 0–1 range, so that temporal variations in the different representations could be more easily compared. **A**, Envelope, LogEnv and Adaptive Gain representations of a burst of white noise in silence. **B**, Envelope, LogEnv and Adaptive Gain representations of two seconds of continuous speech.

(Figure 1B).

2.3 Linear cortical modelling

2.3.1 Forward modelling

The temporal response function $TRF(\tau)$ is the estimated linear response function that transforms a stimulus representation $r(t)$ into a neural response $n(t)$ [39], as expressed in the following equation:

$$n(t) = TRF(\tau) * r(t) + \epsilon(t) \tag{8}$$

where $*$ is the convolution operator and the error term $\epsilon(t)$ corresponds to the residual neural response that is not explained by the model. Here, $r(t)$ was either the $Env(t)$, $LogEnv(t)$ or $AG(t)$ representation of the audio signal and TRF modelling was performed separately for each.

Since the relationship between any stimulus feature and the resulting EEG response is not instantaneous, TRFs are always estimated by taking into account multiple time lags. We used a time-lag window for τ spanning the interval from -100 ms to 600 ms for all stimulus representations and TRF models.

Estimation of the model $TRF(\tau)$ for each of the three stimulus representations was performed by computing a lagged linear regression between the audio representation $r(t)$ and the EEG signal $n(t)$ using the MATLAB mTRF toolbox [16], which incorporates ridge regularization [63]. Regularization parameters were separately optimised for each model through cross-validation across regularization values in the range $[10^{-10}, 10^{10}]$. All models were participant-specific, trained and tested on data from

only one participant.

2.4 Model evaluations

Multiple models were trained for each participant and each audio representation, using segments of 30 seconds of audio along with the corresponding EEG measurements. The training followed a leave-one-out approach, allowing each segment to serve once for testing while the remaining segments were used for training the model. All n values (i.e. $n = 12$ for data from [55] and $n = 40$ for data from [31]) of prediction accuracy obtained for one participant were averaged to yield one average prediction accuracy value per participant and audio representation. These values were compared using Wilcoxon signed-rank tests. Paired comparison tests were used to take into account potential variation between participants, due to differences in the amount of data available for training for each dataset (6 minutes available for data from [54] and 20 minutes for data from [29]) and also due to individual variability in cortical speech tracking.

The performance of the models was assessed using the Pearson correlation between measured EEG $n(t)$ and the predicted EEG response $n'(t)$ obtained through the convolution of the audio representation $r(t)$ and the Temporal Response Function $TRF(\tau)$. EEG prediction accuracies will vary across EEG electrode channels depending on how related the data on those channels are to the stimulus representation. The modelling performance on some irrelevant channels may not differ from chance level and might hinder the assessment of models when looking at the prediction accuracy across all channels [17]. Here, TRF models computed using different audio representations were compared based on their prediction accuracy for the EEG signal at the FCz channel only. This *a priori* choice of channel was motivated by experimental evidence showing that fronto-central channels exhibit strongest cortical tracking of speech envelope [27, 54, 64, 13].

In addition, TRF weights were used to explore whether spatial and temporal patterns of the predicted EEG signals across electrode positions on the head differed depending on the audio representation. As the ridge regularization factors λ, which vary between audio representations and participants, impact the amplitude of the TRF, the TRF weights were z-score normalized for all features and all participants before being averaged across participants for analysis.

2.4.1 Noise floor calculation

To determine the statistical significance of the prediction accuracy and coefficients estimated for the TRF models, we computed noise floors using null models. Null models were created by delaying the audio representation data relative to the EEG data, to ensure that the cortical measurement could not be related to the audio representations when training the null models. The delay was randomly

selected from a range between 7.5 and 15 seconds for each leave-one-out iteration of model fitting and cross-validation. All other aspects of the analysis (i.e., use of participant-specific models, cross-validation approach, ridge-regularization parameters, 30 s data segments) were similar for null models and actual models. Results from participant-specific null models were averaged to obtain a noise floor to which prediction accuracy or coefficient values for actual TRF models could be compared.

2.5 Data analysis

As some of the data were not normally distributed, non-parametric data analysis was preferred. Wilcoxon rank-sum tests were used to assess if the prediction accuracy distributions differed from the noise floor distributions. Wilcoxon signed-rank tests were used for paired within-subject comparisons between audio representations. All tests were two-tailed, and all results were considered significant with $\alpha = 0.05$. Exact p-values are given when higher than 0.0001. Results are shown without Bonferroni correction of p-values but all main conclusions were robust to Bonferroni correction.

3 Results

3.1 Adaptive Gain representation of speech signal improves prediction of cortical speech tracking

We first aimed to determine whether the Adaptive Gain transformation originally derived to improve prediction of mouse auditory thalamic responses to temporally varying sounds could also improve prediction of human EEG responses to continuous speech. For each participant in two independent cortical speech-tracking datasets [54, 29], we estimated temporal response function (TRF) models through linear regression of the EEG signal (at electrode FcZ) with respect to three different transformations of the speech signal: Envelope, LogEnv and Adaptive Gain. TRF models for each of the three audio representations were then evaluated through forward modelling of the EEG signal using segments of continuous speech that had not been used to fit the models (i.e., using cross-validation). Prediction accuracies of the TRF models were quantified with Pearson's coefficient of correlation between actual and predicted EEG signals.

3.1.1 Pooled data

Analysis of pooled data from both datasets (Figure 2A) showed that prediction accuracies for TRFs estimated using Envelope, LogEnv or Adaptive Gain audio representations were all significantly better than chance (Wilcoxon rank-sum tests, $p < 0.001$ for comparison to noise floor; see Methods for details

of noise calculations). However, comparisons between the models revealed that standard Envelope-based TRF models (median prediction accuracy 0.033) were outperformed by both LogEnv models (median= 0.0356; Wilcoxon signed-rank test, $z = -4.1028, p < 0.0001, r = -0.657$) and Adaptive Gain models (median= 0.0368; $z = -4.033, p < 0.0001, r = -0.6458$). Adaptive Gain models also outperformed LogEnv models ($z = -3.1259, p = 0.0018, r = -0.5005$), indicating that logarithmic transformation of the envelope (an initial step in the Adaptive Gain transformation) cannot entirely account for the improvements in prediction accuracy achieved with the Adapative Gain representation.

These results demonstrate that TRFs estimated using the Adaptive Gain representation of the sound signal predict cortical speech tracking more accurately than those estimated using either the Envelope or the LogEnv representation. Moreover, the results suggest that both components of the Adaptive Gain transformation—logarithmic representation of the stimulus envelope and short-timescale temporal adaptation to sound level—help to improve prediction of cortical speech tracking beyond what is possible with the Envelope representation alone. In control analyses (results not shown), we found that these results were robust to changes in the parameters used for defining the Envelope representation (e.g., 15 Hz instead of 79 Hz low-pass filter).

3.1.2 By dataset

Similar patterns of results were found in each dataset when analyzed separately (Figure 2B). An overall difference in modelling performance was observed between the two datasets included in this study. This difference can be mainly explained by the amount of data available per participant in each dataset.

In Dataset 1 [54], TRFs estimated using the Adaptive Gain audio representation (median prediction accuracy = 0.0265) outperformed those estimated using either the Envelope representation (median= 0.0185; Wilcoxon signed rank test, $z = -3.0761, p = 0.0021, r = -0.6712$) or the LogEnv representation (median= 0.0253; $z = -2.9718, p = 0.0030, r = -0.6485$). The LogEnv representation also yielded significantly better results than the Envelope ($z = -2.9023, p = 0.0037, r = -0.6333$).

In Dataset 2 [29], TRFs estimated using the Adaptive Gain representation again outperformed those estimated using the Envelope (median prediction accuracy, Adaptive Gain = 0.0492, Envelope = 0.0418; Wilcoxon signed-rank test, $z = -2.3735, p = 0.0176, r = -0.5594$). The LogEnv representation (median = 0.0469) also again yielded significantly better TRF results than the Envelope ($z = -2.6783, p = 0.0074, r = -0.6313$). For this dataset, the difference in modelling performance between TRFs estimated using Adaptive Gain versus LogEnv audio representations did not reach statistical significance ($z = -1.3718, p = 0.1701$); however, the trend was consistent with higher prediction accuracy for models based on the Adaptive Gain representation.

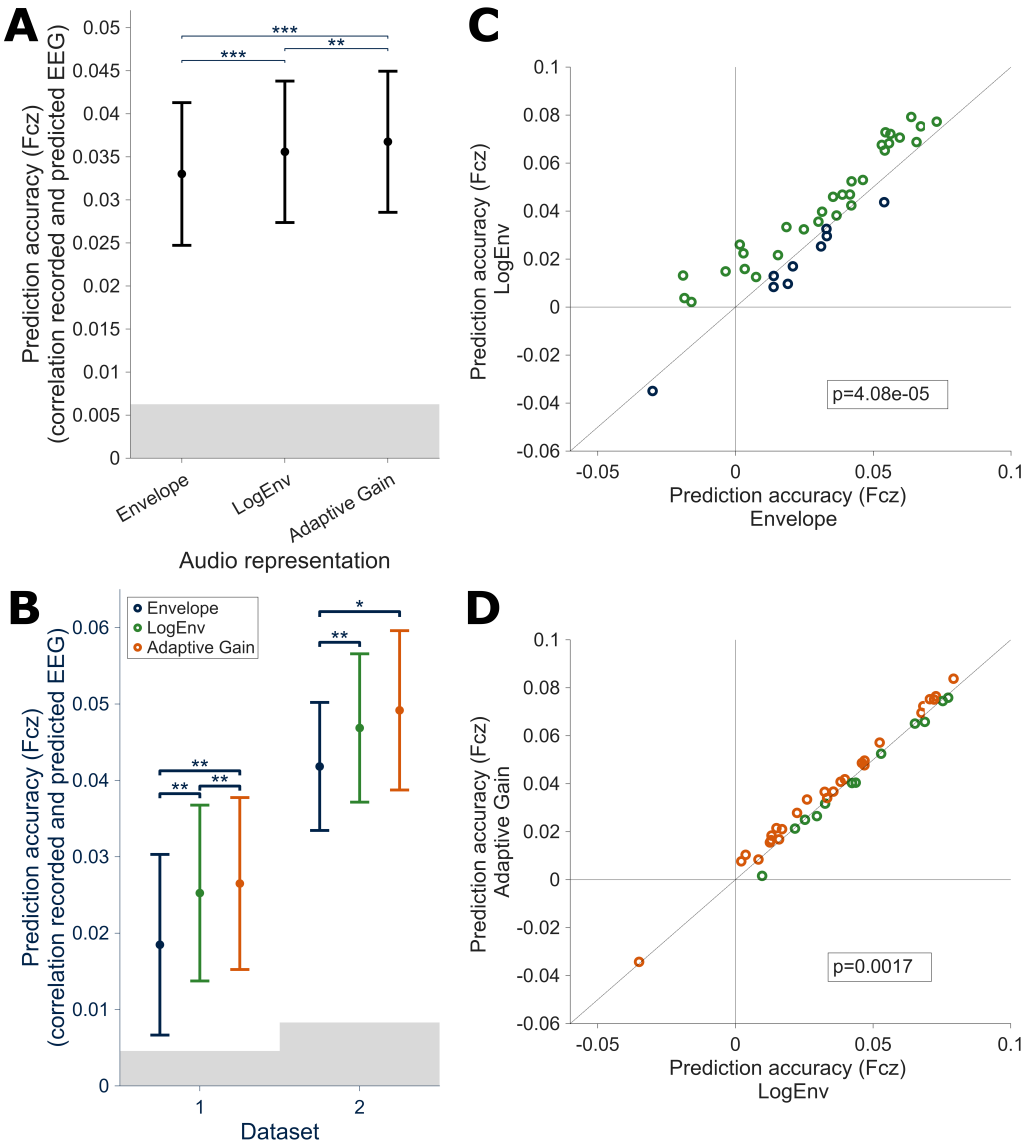


Figure 2: Prediction accuracies for TRF models trained on Envelope, LogEnv, or Adaptive Gain representations of speech. The prediction accuracy corresponds to Pearson's coefficient of correlation between the measured EEG signal on electrode FcZ and the EEG signal predicted for electrode FcZ based on the incoming speech signal and TRF for each audio representation. **A-B**, Median prediction accuracy. Error bars represent 95% confidence intervals (CI). Gray area corresponds to the noise floor obtained with null models. **A**, Median prediction accuracy for data pooled across the two independent datasets. **B**, Median prediction accuracy for each dataset. Dataset 1 corresponds to [54] and Dataset 2 corresponds to [29]. **C-D**, Scatterplots comparing prediction accuracy between Envelope and LogEnv representations (C) or between LogEnv and Adaptive Gain representations (D). Each circle corresponds to data from one participant; p-values show results of Wilcoxon signed-rank tests. Colours correspond to the legend in B and indicate which of the two audio representations compared in the scatterplot produced higher prediction accuracy for each participant.

3.1.3 By participant

Visualization of data from individual participants in the combined datasets (Figure 2C-D) confirmed that both components of the Adaptive Gain transformation—logarithmic transformation of the envelope and short-timescale temporal adaptation to sound level—improved prediction of cortical speech tracking. Scatterplots comparing prediction accuracies of TRFs estimated using different audio representations showed that modelling performance was improved by the LogEnv relative to Envelope representation for 30 out of 39 participants (Figure 2C, green circles), and improved by the Adaptive Gain relative to the LogEnv representation for 27 out of 39 participants (Figure 2D, orange circles). Notably, the improvements in modelling performance provided by these input transformations were relatively robust to high inter-individual variability in overall TRF modelling success, which is a potentially useful feature given that such variability is common within and across EEG recording studies.

3.2 Optimal adaptation time constant for Adaptive Gain representation is longer for humans than mice

The Adaptive Gain transformation was originally derived as part of a temporal processing model optimized for predicting auditory thalamic responses to temporally varying sounds in anesthetized mice [6]. We wondered if the optimal adaptation time constant τ_A for the Adaptive Gain transformation might be different for predicting cortical speech tracking in awake humans. To address this possibility, we compared the prediction accuracies of TRF models estimated using the Adaptive Gain representation with τ_A values of 5 ms (the minimum, equivalent to the more fundamental integration time constant τ_I), 10 ms (the mouse-optimized τ_A value), 25 ms, 50 ms, 75 ms, 100 ms, 250 ms or 500 ms. Prediction accuracies of all models were significantly better than chance ($p < 0.001$), as confirmed by Wilcoxon rank-sum tests comparing model performance to the noise floor. Model performance for the Adaptive Gain representation with the mouse-optimized τ_A value (10 ms) was then compared to model performance for Adaptive Gain representations with all other τ_A values, using Wilcoxon signed-rank tests.

Results showed an increase in model prediction accuracy as the value of τ_A increased until it reached a maximum for $\tau_A = 100$ ms, then decreased for longer values of τ_A (see Figure 3A). Prediction accuracy was significantly higher than that achieved with the mouse-optimized $\tau_A = 10$ ms (median= 0.0368) for $\tau_A = 25$ ms (median=0.0381; Wilcoxon signed-rank test, $z = -4.7726, p < .0001, r = -0.7642$), $\tau_A = 50$ ms (median= 0.0402; $z = -4.6331, p < .0001, r = -0.7419$), $\tau_A = 75$ ms (median= 0.0419; $z = -4.3679, p < .0001, r = 0.6994$), and $\tau_A = 100$ ms (median= 0.0431; $z = -3.7539, p < .000141, r = 0.6011$). However, prediction accuracy for $\tau_A = 5$ ms (median= 0.0365) was lower than for the

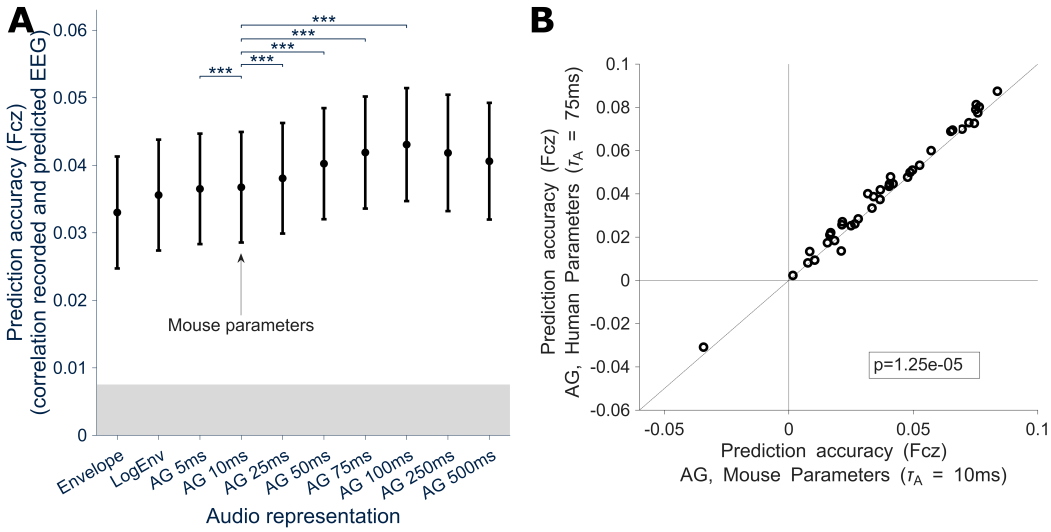


Figure 3: Prediction accuracies for TRF models trained on Envelope, LogEnv and Adaptive Gain representations of speech with varying τ_A values. Conventions as in Figure 2. **A**, Median prediction accuracy ± 95% CI. Brackets and asterisks indicate significant differences in prediction accuracy between TRF models trained on Adaptive Gain representations with the mouse-optimized value for the adaptation time constant τ_A versus other values (Wilcoxon signed-rank tests, $***p < 0.001$). **B**, Scatterplot comparing TRF prediction accuracy for the Adaptive Gain representation with a human-optimized adaptation time constant ($\tau_A = 75$ ms) versus mouse-optimized ($\tau_A = 10$ ms) parameters. Each circle corresponds to data from one participant; p-value shows result of Wilcoxon signed-rank test.

original mouse-optimized parameter ($z = 4.3679, p < .0001, r = -0.6994$). Model prediction accuracy was not significantly different from that obtained with the original mouse-optimized $\tau_A = 10$ ms for $\tau_A = 250$ ms (median= 0.0418; $z = -1.9397, p = 0.0524$) and $\tau_A = 500$ ms (median= 0.0406; $z = -0.8233, p = 0.4103$).

These results indicate that the Adaptive Gain transformation is more useful for predicting cortical speech tracking in humans with an adaptation time constant τ_A of 25-100 ms rather than the $\tau_A = 10$ ms value originally derived in studies of anesthetized mice. We chose $\tau_A = 75$ ms as a human-optimized value for further investigations, since time constants of 50, 75 and 100 ms all produced models with prediction accuracy statistically indistinguishable from the peak. As illustrated by the fact that almost all data points are above the diagonal line in Figure 3B, model prediction accuracy was indeed higher for nearly all participants when using this human-optimized τ_A value rather than the original mouse-optimized value.

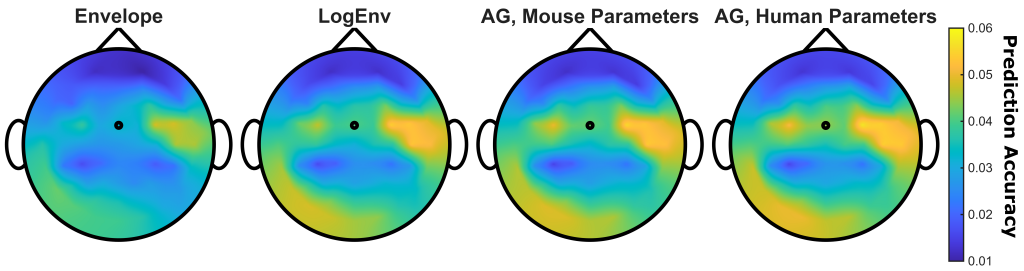


Figure 4: Topographical maps of TRF model prediction accuracy for each audio representation and for each EEG electrode. Black circles correspond to the position of electrode FcZ.

3.3 Topographical maps and TRF shapes suggest that Envelope, LogEnv and Adaptive Gain representations capture similar speech-tracking mechanisms

The Adaptive Gain transformation was developed to capture biologically relevant adaptive normalization in time. Our hypothesis is that the Adaptive Gain transformation suppresses and enhances features in the speech signal that would also be suppressed or enhanced by central auditory temporal processing mechanisms. This hypothesis assumes that the underlying speech-tracking mechanisms modelled with TRFs using the Envelope, LogEnv, and Adaptive Gain representations as regressors should be qualitatively similar, even if model performance is quantitatively different. To test this hypothesis, we compared topographic maps of prediction accuracy and the TRF shapes (i.e., the weights of the TRF models) obtained using the different audio representations as regressors.

Figure 4 compares the maps of the prediction accuracy obtained with models trained on Envelope and LogEnv representations and Adaptive Gain representations either with the original adaptation time constant from mouse studies ($\tau_A = 10ms$) or the human-optimized adaptation time constant identified here ($\tau_A = 75$ ms). All four maps present a similar pattern of increased prediction accuracy on the frontocentral electrodes, with lateralization on the right hemisphere and improved prediction accuracy also on the left occipital electrodes. This pattern is consistent with previous work on auditory TRFs modelled on the stimulus envelope or representations of slow temporal amplitude variations in sound ([54, 32, 47, 24, 40, 11, 27]). The similarity of patterns here suggests that the mechanisms modelled with the four audio representations are comparable. On the electrodes that are best predicted with all models, an increase of prediction accuracy can be seen and follows the trend observed on FCz, emphasising that using the Adaptive Gain representation, especially with a longer length of the adaptive window, leads to a better-performing model than using the Envelope or the LogEnv representations.

The weights of the models, which represent the shape of the Temporal Response Functions (TRFs),

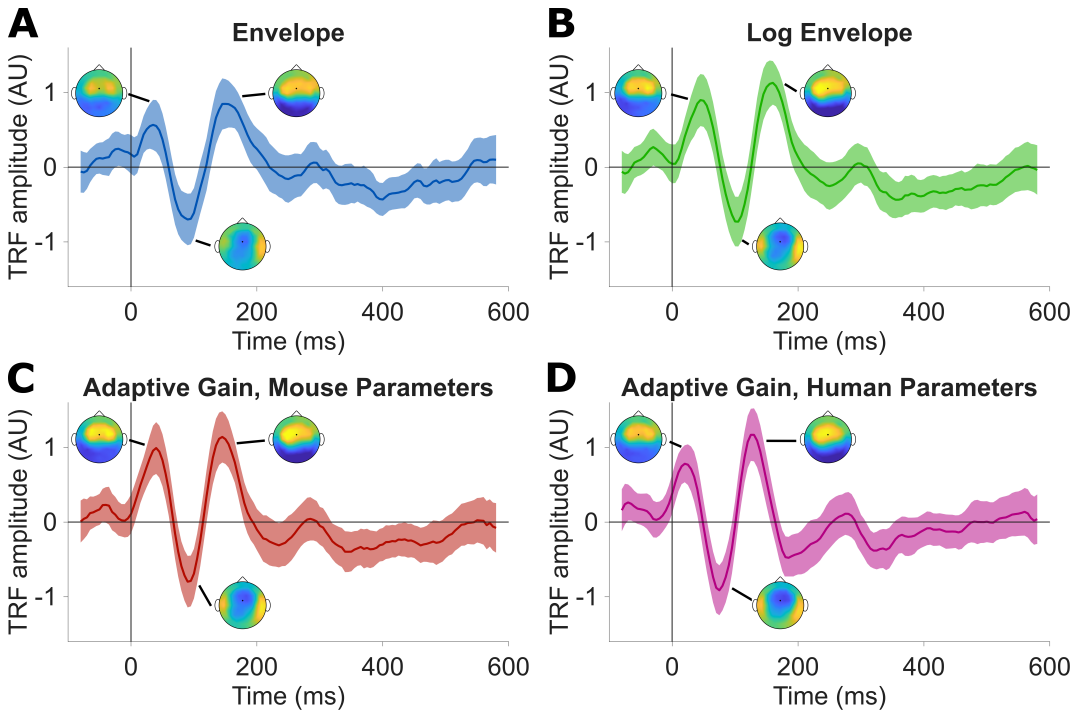


Figure 5: Temporal response functions (TRFs) derived from linear regression between EEG data and the different representations of the speech stimuli: Envelope (A), LogEnv (B), Adaptive Gain with mouse-optimized parameters (C) and Adaptive Gain with human-optimised parameters (D). Shaded areas correspond to 95% CI. Topographic maps show the spatial distribution of the activities for the P50, N100 and P200 components.

also show notable similarities (Figure 5). For all four audio representations analyzed, the TRFs recorded at the FCz electrode exhibit a typical P50-N100-P200 complex, a pattern associated with auditory event-related potentials. The topographic mapping of the TRF models indicates that this activity is primarily concentrated around the fronto-central electrodes, which has also been observed in other studies [54, 32, 24, 40, 25, 15, 11]. However, there are differences in both amplitude and latency. In terms of amplitude, the TRFs are comparable for models based on the Envelope, logarithmic envelope (LogEnv), and Adaptive Gain using mouse parameters. Notably, the model that employs Adaptive Gain with parameters optimized for humans shows higher amplitudes for all three main peaks. Analysis of the latency of the peaks reveals a consistent time shift for the two Adaptive Gain conditions. This time shift is likely attributable to the Adaptive Gain transformation itself; the delay due to smoothing with the function v_{IA} is shorter for $\tau_{A\ mouse} = 10$ ms than for $\tau_{A\ human} = 75$ ms.

3.4 Adaptive Gain representation also improves prediction of cortical speech tracking for not-understood speech

Assuming the Adaptive Gain transformation captures fundamental mechanisms of temporal processing in the central auditory system shared by mice and humans (with possibly different time constants),

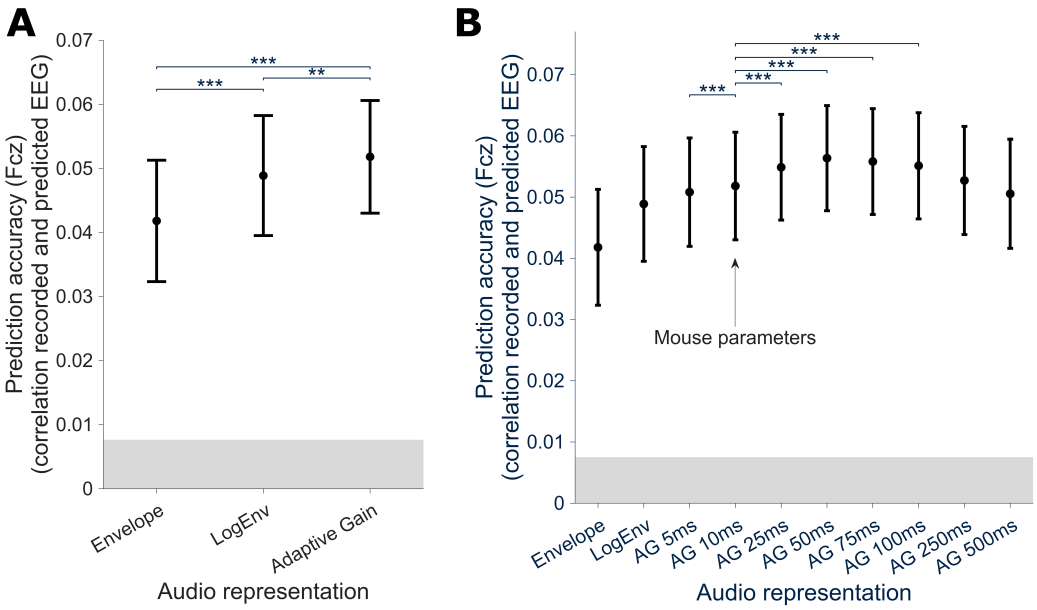


Figure 6: Prediction accuracies obtained on the remaining data of both datasets, in which participants listened to speech in not-understood languages (Finnish for data from [55] and Dutch for data from [31]). Conventions as in Figure 2 and Figure 3. **A**, Median prediction accuracy ± 95% CI for TRF models trained on Envelope, LogEnv and Adaptive Gain representations with the original mouse-optimized parameters (compare with Figure 2A). **B**, Median prediction accuracy ± 95% CI for Envelope and LogEnv representations and for Adaptive Gain representations with varying τ_A values (compare with Figure 3A).

then it should to improve cortical speech tracking for speech that was not understood as well as for understood speech. We tested this hypothesis using remaining data from both datasets, in which participants listened to speech in languages they did not understand (Figure 6). Results were very similar to those obtained for cortical speech tracking of understood languages (compare to Figure 2A and Figure 3A). Again, time constants of 50, 75 and 100 ms all produced models with prediction accuracy statistically indistinguishable from the peak (although here the peak median prediction accuracy occurred at $\tau_A = 50$ ms instead of $\tau_A = 100$ ms in Figure 3A). We did not pursue a more detailed comparison of results for understood and not-understood languages because of the difficulty of controlling for levels of attentional engagement. However, we note that TRF prediction accuracies were very similar for understood and not-understood languages.

This result demonstrates that the Adaptive Gain transformation improves prediction of cortical speech tracking for not-understood as well as understood speech.

4 Discussion

This study investigated whether a simple Adaptive Gain mechanism previously identified in studies of central auditory temporal processing in mice could be used to improve prediction of human cortical

responses to continuous speech. The results demonstrate that temporal response function (TRF) models estimated using the Adaptive Gain representation of speech outperform models estimated using either the standard Envelope or the logarithmic envelope (LogEnv) representations, indicating that adaptive mechanisms play a significant role in shaping cortical tracking of speech. Moreover, parameter optimization for human EEG revealed that longer adaptation time constants (around 75–100 ms, compared to 10 ms in [6]) yielded the best prediction performance, suggesting that the temporal dynamics of adaptation in the auditory cortex of awake humans differ from those previously observed in the auditory thalamus of anesthetized mice.

These findings extend previous work showing that the use of biologically grounded auditory transformations improves the performance of cortical speech-tracking models [40, 9, 27]. Here, the biologically grounded transformations were nonlinearities in intensity and temporal processing: logarithmic representation of sound intensity (which compressed and smoothed amplitude variations in the sound envelope) and short-timescale adaptive normalization in time (which enhanced responsiveness to sound features occurring when sound levels had recently been low, and suppressed responsiveness to sound features occurring when sound levels had recently been high).

The improvements in TRF modelling performance observed with the Adaptive Gain over the Envelope representation arose from both nonlinearities, as was illustrated by the intermediate improvements achieved with TRF models based on the LogEnv representation alone. These findings are biologically plausible and consistent with previous results from animal and human studies. There is extensive psychophysical and neurophysiological evidence for logarithmic coding of sound intensity in the auditory cortex (e.g. [3, 2]), although previous work on linear modelling of cortical responses has not found that adding logarithmic scaling to the envelope improves performance [9, 10, 27]. The adaptive component of the Adaptive Gain transformation is a novelty for cortical speech tracking studies but also biologically plausible, as it approximates short-timescale neural adaptation processes that are ubiquitous in the auditory pathway [6, 35]. Previous studies in animal models have found that similar adaptive mechanisms improve the prediction of single-unit responses to temporally varying sounds in the auditory cortex [62]. Here, we observed improvements in prediction of cortical speech tracking with the Adaptive Gain transformation compared to both Envelope and LogEnv representations across two independent datasets [31, 55], highlighting the robustness of the findings.

Analysis of topographical maps of prediction accuracy and TRF shapes also provided evidence for the biological plausibility of the Adaptive Gain transformation and its usefulness for modelling cortical speech tracking. The topographic distribution of TRF prediction accuracy for the Envelope representation, with maximal prediction accuracy observed on fronto-central electrodes, was consistent with results reported in previous envelope-tracking studies [54, 47, 32, 11, 27]. Similar topographical

distributions of TRF prediction accuracy were observed for models based on the LogEnv and Adaptive Gain representations, suggesting similarity in the underlying speech-tracking mechanism detected by the models. The similar shape of the TRFs obtained using the Envelope, LogEnv and Adaptive Gain representations also supports this interpretation. For all audio representations, a three-peaked complex, resembling the typical P50-N100-P200 complex found in auditory event-related potentials, can be clearly observed in the TRF morphology, and was also concentrated on the fronto-central electrodes. The slight latency shift observed in the TRF estimated using the Adaptive Gain representation, relative to the TRFs estimated using the Envelope and LogEnv representations, is consistent with the temporal smoothing and delay introduced by the double convolution with integration and adaptation time windows in the Adaptive Gain computation (see Supplementary Information). Taken together, these results support the idea that the Adaptive Gain transformation of speech enhances, rather than replaces, Envelope-based models of slow cortical tracking.

Adaptive mechanisms have been found throughout the auditory pathway, operating over a range of different time scales [49, 46, 34, 50, 61]. Animal studies have revealed that while some adaptive mechanisms are driven by subcortical processing, including at the level of the thalamus [5], additional adaptive mechanisms arise in the auditory cortex on a longer time scale [7, 58, 21]. These observations suggest that the longer adaptation time constant for the human-optimized than mouse-optimized Adaptive Gain transformation might reflect differences between auditory cortex and thalamus (and between speech processing and processing of simpler temporally varying sounds) as well as differences between awake humans and anesthetized mice. Further studies could explore the contributions of subcortical versus cortical adaptation to speech tracking in humans by integrating the Adaptive Gain audio representation with linear modelling of subcortical responses during speech listening [36, 29], to provide a deeper understanding of hierarchical adaptation in the auditory system.

In conclusion: the Adaptive Gain representation is a simple and easly implemented alternative to the Envelope representation that reliably improves TRF modelling performance for continuous speech, demonstrating a role for adaptive temporal normalization in cortical speech tracking. Notably, the Adaptive Gain representation improved prediction of cortical speech tracking even with mouse-optimized parameters, and for not-understood as well as understood speech, suggesting that it captures fundamental properties of central auditory temporal processing. Therefore, the Adaptive Gain representation might be useful as an alternative to the Envelope representation to improve predictions of human EEG responses not only for speech but for any temporally varying sounds. This improvement could translate into practical benefits, for example in applications such as auditory attention decoding in multi-speaker environments or development of novel objective measures for auditory processing disorders.

References

[1] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, 98(23):13367–13372, 2001.

[2] M. B. Ahrens, J. F. Linden, and M. Sahani. Nonlinearities and contextual influences in auditory cortical responses modeled with multilinear spectrotemporal methods. *Journal of Neuroscience*, 28(8):1929–1942, 2008.

[3] S. J. Aiken and T. W. Picton. Human cortical responses to the speech envelope. *Ear and hearing*, 29(2):139–157, 2008.

[4] M. Akritas, A. G. Armstrong, J. M. Lebert, A. F. Meyer, M. Sahani, and J. F. Linden. Nonlinear sensitivity to acoustic context is a stable feature of neuronal responses to complex sounds in auditory cortex of awake mice. *ELife*, 13, 2024.

[5] L. A. Anderson, G. B. Christianson, and J. F. Linden. Stimulus-specific adaptation occurs in the auditory thalamus. *Journal of Neuroscience*, 29(22):7359–7363, 2009.

[6] L. A. Anderson and J. F. Linden. Mind the gap: two dissociable mechanisms of temporal processing in the auditory system. *Journal of Neuroscience*, 36(6):1977–1995, 2016.

[7] H. Asari and A. M. Zador. Long-lasting context dependence constrains neural encoding models in rodent auditory cortex. *Journal of neurophysiology*, 102(5):2638–2656, 2009.

[8] B. D. Auerbach and H. J. Gritton. Hearing in complex environments: auditory gain control, attention, and hearing loss. *Frontiers in neuroscience*, 16:799787, 2022.

[9] W. Biesmans, N. Das, T. Francart, and A. Bertrand. Auditory-inspired speech envelope extraction methods for improved eeg-based auditory attention detection in a cocktail party scenario. *IEEE transactions on neural systems and rehabilitation engineering*, 25(5):402–412, 2016.

[10] W. Biesmans, J. Vanthornhout, J. Wouters, M. Moonen, T. Francart, and A. Bertrand. Comparison of speech envelope extraction methods for eeg-based auditory attention detection in a cocktail party scenario. In *2015 37th annual international conference of the ieee engineering in medicine and biology society (embc)*, pages 5155–5158. IEEE, 2015.

[11] H. B. Borges, J. Zaar, E. Alickovic, C. B. Christensen, and P. Kidmose. Speech reception threshold estimation via eeg-based continuous speech envelope reconstruction. *European Journal of Neuroscience*, 61(6):e70083, 2025.

[12] C. Brodbeck and J. Z. Simon. Continuous speech processing. *Current Opinion in Physiology*, 18:25–31, 2020.

[13] G. Cantisani, A. Chalehchaleh, G. Di Liberto, and S. Shamma. Investigating the cortical tracking of speech and music with sung speech. In *Proc. INTERSPEECH 2023*, pages 5157–5161. ISCA, 2023.

[14] M. Carandini and D. J. Heeger. Normalization as a canonical neural computation. *Nature reviews neuroscience*, 13(1):51–62, 2012.

[15] S. Carta, A. M. Mangiacotti, A. L. Valdes, R. B. Reilly, F. Franco, and G. M. Di Liberto. The impact of temporal synchronisation imprecision on trf analyses. *Journal of Neuroscience Methods*, 385:109765, 2023.

[16] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor. The multivariate temporal response function (mtrf) toolbox: a matlab toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, 10:604, 2016.

[17] 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. *Frontiers in neuroscience*, 15:705621, 2021.

[18] S. V. David and S. A. Shamma. Integration over multiple timescales in primary auditory cortex. *Journal of Neuroscience*, 33(49):19154–19166, 2013.

[19] A. de Cheveigné and I. Nelken. Filters: when, why, and how (not) to use them. *Neuron*, 102(2):280–293, 2019.

[20] I. Dean, N. S. Harper, and D. McAlpine. Neural population coding of sound level adapts to stimulus statistics. *Nature neuroscience*, 8(12):1684–1689, 2005.

[21] I. Dean, B. L. Robinson, N. S. Harper, and D. McAlpine. Rapid neural adaptation to sound level statistics. *Journal of Neuroscience*, 28(25):6430–6438, 2008.

[22] L. Decruy, J. Vanthornhout, and T. Francart. Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of neurophysiology*, 122(2):601–615, 2019.

[23] A. Delorme and S. Makeig. Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. *Journal of neuroscience methods*, 134(1):9–21, 2004.

[24] G. M. Di Liberto, C. Pelofi, S. Shamma, and A. de Cheveigné. Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology*, 41(1):361–364, 2020.

[25] B. Dieudonné, L. Decruy, and J. Vanthornhout. Neural tracking of the speech envelope predicts binaural unmasking. *European Journal of Neuroscience*, 61(1):e16638, 2025.

[26] N. Ding and J. Z. Simon. Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *Journal of Neuroscience*, 33(13):5728–5735, 2013.

[27] D. P. Drennan and E. C. Lalor. Cortical tracking of complex sound envelopes: modeling the changes in response with intensity. *eneuro*, 6(3), 2019.

[28] M. A. Escabí, L. M. Miller, H. L. Read, and C. E. Schreiner. Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *Journal of Neuroscience*, 23(37):11489–11504, 2003.

[29] 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*, 200:1–11, 2019.

[30] O. Etard and T. Reichenbach. Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. *Journal of Neuroscience*, 39(29):5750–5759, 2019.

[31] O. Etard and T. Reichenbach. Eeg dataset for 'decoding of selective attention to continuous speech from the human auditory brainstem response' and 'neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise', 2022.

[32] J. Hjortkjær, J. Märcher-Rørsted, S. A. Fuglsang, and T. Dau. Cortical oscillations and entrainment in speech processing during working memory load. *European Journal of Neuroscience*, 51(5):1279–1289, 2020.

[33] B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani. Adaptation of the human auditory cortex to changing background noise. *Nature communications*, 10(1):2509, 2019.

[34] A. J. King and K. M. Walker. Listening in complex acoustic scenes. *Current opinion in physiology*, 18:63–72, 2020.

[35] C. Kopp-Scheinpflug and J. F. Linden. Coding of temporal information. In F. B. (Ed.) and G. B. V. Ed.), editors, *The senses: a comprehensive reference*, pages 691–712. Academic Press, 2021.

[36] J. P. Kulasingham, F. L. Bachmann, K. Eskelund, M. Enqvist, H. Innes-Brown, and E. Alickovic. Predictors for estimating subcortical eeg responses to continuous speech. *Plos one*, 19(2):e0297826, 2024.

[37] M. N. Kvale and C. E. Schreiner. Short-term adaptation of auditory receptive fields to dynamic stimuli. *Journal of Neurophysiology*, 91(2):604–612, 2004.

[38] E. C. Lalor and J. J. Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European journal of neuroscience*, 31(1):189–193, 2010.

[39] 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*, 102(1):349–359, 2009.

[40] E. Lindboom, A. Nidiffer, L. H. Carney, and E. C. Lalor. Incorporating models of subcortical processing improves the ability to predict eeg responses to natural speech. *Hearing research*, 433:108767, 2023.

[41] M. Lohse, V. M. Bajo, A. J. King, and B. D. Willmore. Neural circuits underlying auditory contrast gain control and their perceptual implications. *Nature Communications*, 11(1):324, 2020.

[42] R. K. Maddox and A. K. Lee. Auditory brainstem responses to continuous natural speech in human listeners. *eneuro*, 5(1), 2018.

[43] A. F. Meyer, R. S. Williamson, J. F. Linden, and M. Sahani. Models of neuronal stimulus-response functions: elaboration, estimation, and evaluation. *Frontiers in systems neuroscience*, 10:109, 2017.

[44] I. J. Moon, J. H. Won, M.-H. Park, D. T. Ives, K. Nie, M. G. Heinz, C. Lorenzi, and J. T. Rubinstein. Optimal combination of neural temporal envelope and fine structure cues to explain speech identification in background noise. *Journal of Neuroscience*, 34(36):12145–12154, 2014.

[45] Y. Oganian, K. Kojima, A. Breska, C. Cai, A. Findlay, E. F. Chang, and S. S. Nagarajan. Phase alignment of low-frequency neural activity to the amplitude envelope of speech reflects evoked responses to acoustic edges, not oscillatory entrainment. *Journal of Neuroscience*, 43(21):3909–3921, 2023.

[46] D. Pérez-González and M. S. Malmierca. Adaptation in the auditory system: an overview. *Frontiers in integrative neuroscience*, 8:19, 2014.

[47] K. D. Prinsloo and E. C. Lalor. General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras. *Journal of Neuroscience*, 42(41):7782–7798, 2022.

[48] M. Rahman, B. D. Willmore, A. J. King, and N. S. Harper. Simple transformations capture auditory input to cortex. *Proceedings of the National Academy of Sciences*, 117(45):28442–28451, 2020.

[49] B. L. Robinson and D. McAlpine. Gain control mechanisms in the auditory pathway. *Current opinion in neurobiology*, 19(4):402–407, 2009.

[50] S. Shamma and J. Fritz. Adaptive auditory computations. *Current opinion in neurobiology*, 25:164–168, 2014.

[51] T. Shan, M. S. Cappelloni, and R. K. Maddox. Subcortical responses to music and speech are alike while cortical responses diverge. *Scientific Reports*, 14(1):789, 2024.

[52] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270(5234):303–304, 1995.

[53] B. Shechter and D. A. Depireux. Response adaptation to broadband sounds in primary auditory cortex of the awake ferret. *Hearing research*, 221(1-2):91–103, 2006.

[54] A. Simon, S. Bech, G. Loquet, and J. Østergaard. Cortical linear encoding and decoding of sounds: Similarities and differences between naturalistic speech and music listening. *European Journal of Neuroscience*, 59(8):2059–2074, 2024.

[55] A. Simon, S. Bech, G. Loquet, and J. Østergaard. Eeg data of continuous listening of music and speech, 2022.

[56] A. J. Simpson, N. S. Harper, J. D. Reiss, and D. McAlpine. Selective adaptation to "oddball" sounds by the human auditory system. *Journal of Neuroscience*, 34(5):1963–1969, 2014.

[57] Z. M. Smith, B. Delgutte, and A. J. Oxenham. Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416(6876):87–90, 2002.

[58] N. Ulanovsky, L. Las, D. Farkas, and I. Nelken. Multiple time scales of adaptation in auditory cortex neurons. *Journal of Neuroscience*, 24(46):10440–10453, 2004.

[59] U. Werner-Reiss, K. K. Porter, A. M. Underhill, and J. M. Groh. Long lasting attenuation by prior sounds in auditory cortex of awake primates. *Experimental brain research*, 168:272–276, 2006.

[60] R. S. Williamson, M. B. Ahrens, J. F. Linden, and M. Sahani. Input-specific gain modulation by local sensory context shapes cortical and thalamic responses to complex sounds. *Neuron*, 91(2):467–481, 2016.

[61] B. D. Willmore and A. J. King. Adaptation in auditory processing. *Physiological Reviews*, 103(2):1025–1058, 2023.

[62] B. D. Willmore, O. Schoppe, A. J. King, J. W. Schnupp, and N. S. Harper. Incorporating midbrain adaptation to mean sound level improves models of auditory cortical processing. *Journal of Neuroscience*, 36(2):280–289, 2016.

[63] 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. *Frontiers in neuroscience*, 12:531, 2018.

[64] N. J. Zuk, J. W. Murphy, R. B. Reilly, and E. C. Lalor. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLoS computational biology*, 17(9):e1009358, 2021.

Supplementary Information:

Mathematics of Adaptive Gain

1 Form of the adaptive gain term

The Adaptive Gain model applies two successive convolutions to the logarithm of the stimulus envelope (LogEnv) and then applies a nonlinear transformation—a $\frac{1}{1+x}$ inversion intended to constrain values to between 0 and 1. Here we attempt to clarify and simplify the Adaptive Gain function by reducing the two successive convolutions to a single smoothing function.

The Adaptive Gain function as defined in [1] is as follows.

$$AG(t) = \frac{1}{1 + (w_A * (w_I * LogEnv))(t)} \tag{1}$$

where $*$ signifies convolution, w_A and w_I are the exponential weight functions detailed in the main text of the current paper, and LogEnv is the logarithm of the stimulus envelope as also defined in the main text.

Convolution (discrete and continuous) is both commutative:

$$(f * g)(t) = (g * f)(t) \tag{2}$$

and associative:

$$((f * g) * h)(t) = (f * (g * h))(t) \tag{3}$$

So using (3) and (2) in (1):

$$\begin{aligned}(w_A * (w_I * LogEnv))(t) &= ((w_A * w_I) * LogEnv)(t) \\ &= ((w_I * w_A) * LogEnv)(t) \\ &= (v_{IA} * LogEnv)(t)\end{aligned}$$

To give a closed form for v_{IA} we perform a continuous convolution, before resampling back to discrete functions. This gives the same result as discrete convolution provided sampling is fast enough (double the highest relevant frequency in the sound pressure waveform $s(t)$, which here would be the upper frequency limit for human hearing). The limits here are given under the assumption that the weight functions continue to ∞. Error terms from truncation and discretization will be calculated subsequently.

Substituting in continuous formulations of the definitions of w_I and w_A given in [1] with appropriate normalizing constants K_I^c, K_A^c for the continuous case:

$$\begin{aligned} v_{IA}(T) &= (w_I * w_A)(T) \\ &= \int_0^T K_I^c \cdot e^{-(T-t)/\tau_I} \cdot K_A^c \cdot e^{-t/\tau_A} dt \\ &= K_I^c K_A^c e^{-T/\tau_I} \int_0^T e^{-t(\tau_I - \tau_A)/\tau_I \tau_A} dt \\ &= \frac{K_I^c K_A^c \tau_I \tau_A}{\tau_A - \tau_I} (e^{-T/\tau_A} - e^{-T/\tau_I}) \end{aligned}$$

We absorb the constant term in the above formulation into a normalization constant K_{IA}^c. It may seem like we've introduced an asymmetry with the negative sign, but in fact as $e^{-T/\tau_A} \geq e^{-T/\tau_I} \Leftrightarrow \tau_A \geq \tau_I$, we have the sign in the numerator always cancelling the sign in the denominator of K_{IA}^c, maintaining the symmetry we expect from (2). Thus without loss of generality we assume $\tau_I \leq \tau_A$ and $K_{IA}^c \geq 0$ to give the closed form:

$$v_{IA}(t) = K_{IA}^c (e^{-t/\tau_A} - e^{-t/\tau_I}) \tag{4}$$

2 Error terms from discretization and truncation of weight functions

To calculate both error types, we look at the difference in integral of v_{IA} between the discrete and continuous case, as v_{IA} is integrated in its use case as a component of a convolution.

2.1 Discretization

As stated above in §1, provided sampling is fast enough for our use case (human hearing) we have no relevant discretization error between performing the convolution discretely, or continuously and then resampling. The sampling performed is at 48 kHz with a period of $\delta \approx 0.02$ ms, which is fast enough to

achieve this (the accepted sample rate for human audio being 44.1 kHz). Thus the only discretization error present is the difference in integral between the continuous case and the discrete case, for which we will calculate a conservative upper bound by summing the absolute value of the differences in each segment. To evaluate this, we use the Mean Value Theorem applied to the continuous function v_{IA}, where $x \in [t, t+\delta]$:

$$(x-t)v'_{IA}(s_x) = v_{IA}(x) - v_{IA}(t)$$

Here $s_t \in [t, x]$ is a variable depending on x. Now we integrate both sides with respect to x from t to $t+\delta$:

$$\int_t^{t+\delta} (x-t)v'_{IA}(s_x)dx = \int_t^{t+\delta} v_{IA}(x)dx - \delta v_{IA}(t)$$

The absolute value of the right-hand side of the equation above is our goal, so assign it to a function DE(t):

$$\begin{aligned} DE(t) &= \Big| \int_t^{t+\delta} (x-t)v'_{IA}(s_x)dx \Big| \\ &\leq \int_t^{t+\delta} (x-t)|v'_{IA}(s_x)|dx \\ &\leq \sup_{t \leq x \leq t+\delta} |v'_{IA}(x)| \cdot \frac{\delta^2}{2} \end{aligned}$$

Summing this over all intervals in the domain of v_{IA}, we get the following error percentages for different values of τ_A with τ_I held at 5 ms.

τ_A (ms)	Area error percentage (%)
5	3.11
10	1.04
25	0.35
50	0.17
75	0.11
100	0.08
150	0.05

N.B.: In lieu of calculating the supremum of the absolute value of the derivative on each interval, which would require continuous data or lots of closed form calculation, we simply take the maximum of the absolute value of the derivative from each pair of endpoints of each interval. This only fails to

achieve the supremum when an inflection point is contained in the interval, which occurs for exactly one interval over the domain of the function. The contribution of a single interval to the total percentage error is on the order of δ^2, so miniscule, and thus we can ignore this slight calculation error.

2.2 Truncation of w_I and w_A

In the main text, we use w_A and w_I functions that are cut off to 0 at t-values bigger than $5\tau_A$ and $5\tau_I$, which produces an error relative to ideal functions that continued to ∞.

To get a sense for the size of this error, we look at the area of v_{IA} cut off by truncation of the weight functions compared to the idealized v_{IA}. Denote the ideal functions calculated to ∞ by v'_{IA}, w'_I and w'_A, so

$$\sum_{t=0}^{\infty} w'_I[t] = \sum_{t=0}^{\infty} w_I[t] + \sum_{t=5\tau_I}^{\infty} w'_I[t]$$

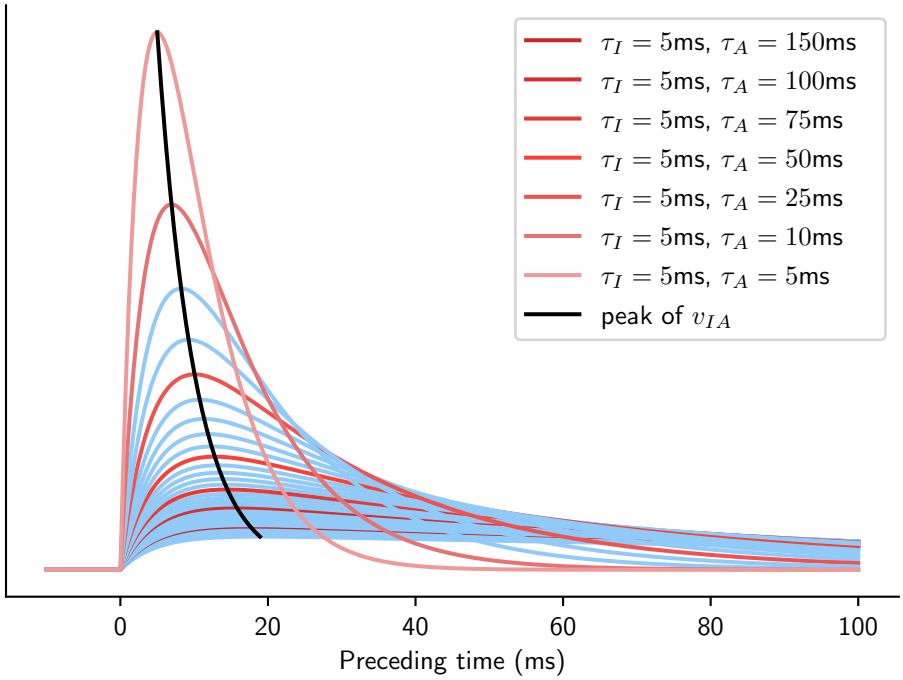
The total sum of a convolution is the product of the total sums of the two components, and convolution is linear in both functions, so we find the total area of v_{IA} removed by the truncated weight functions. In the following, we assume the non-truncated functions are normalized, and we let $\epsilon_I = \sum_{t=5\tau_I}^{\infty} w'_I[t]$ and $\epsilon_A = \sum_{t=5\tau_A}^{\infty} w'_A[t]$.

$$\begin{aligned}
\sum_{t=-\infty}^{\infty} v_{IA}[t] &= \Big(\sum_{t=-\infty}^{\infty} w'_I[t] - \sum_{t=5\tau_I}^{\infty} w'_I[t]\Big)\Big(\sum_{t=-\infty}^{\infty} w'_A[t] - \sum_{t=5\tau_A}^{\infty} w'_A[t]\Big) \\
&= \Big(\sum_{t=-\infty}^{\infty} w'_I[t]\Big)\Big(\sum_{t=-\infty}^{\infty} w'_A[t]\Big) - \Big(\sum_{t=5\tau_I}^{\infty} w'_I[t]\Big)\Big(\sum_{t=-\infty}^{\infty} w'_A[t]\Big) \\
&\quad - \Big(\sum_{t=-\infty}^{\infty} w'_I[t]\Big)\Big(\sum_{t=5\tau_A}^{\infty} w'_A[t]\Big) + \Big(\sum_{t=5\tau_I}^{\infty} w'_I[t]\Big)\Big(\sum_{t=5\tau_A}^{\infty} w'_A[t]\Big) \\
&= 1 - \epsilon_I - \epsilon_A + \epsilon_I\epsilon_A
\end{aligned}$$

$\epsilon_I, \epsilon_A < 0.01$ by choice of truncation at 5τ. Thus our truncation error compared to an ideal case is $< \epsilon_I + \epsilon_A - \epsilon_I\epsilon_A < 0.02 = 2\%$.

3 Character of v_{IA}

We compare the shape of v_{IA} for different values of τ_A given a fixed $\tau_I = 5$ ms:



The major difference between a standard exponential smoother and the v_{IA} smoothing function used here is the delay in the peak. In a standard smoother, the peak is at $t = 0$, whereas here we can see the peak occurs at $t = 10 - 20$ ms. When this function is convolved with the logarithm of the stimulus envelope $LogEnv(t)$ in the adaptive gain calculation, this peak delay means that the maximal contribution to the final convolution at a point T comes from the point in $LogEnv(t)$ with $t \approx T =$ 10–20 ms for $\tau_I = 5$ ms and τ_A values as indicated in the legend above. This is in contrast to the standard smoother which gives maximal contribution from the input at the same point as where the convolution is evaluated.

References

[1] L. A. Anderson and J. F. Linden. Mind the gap: two dissociable mechanisms of temporal processing in the auditory system. *Journal of Neuroscience*, 36(6):1977–1995, 2016.