INVESTIGATING EFFECTS OF EMPHASIS ON THE MODELED AUDITORY BRAINSTEM RESPONSE

JOHN KYLE COOPER
SAM ZHAO
University of Rochester

As we learn more about how complex speech is processed in the brain, we can integrate these speech encoding mechanisms into technology such as hearing aids in order to provide better hearing outcomes for people who experience hearing-loss. In this study, we examine the well-documented phenomena of focus, which in English produces a contrastive change by emphasizing one word in a sentence. We aim to demonstrate how emphasized speech is processed in the brain through the analysis of modeled auditory brain responses to emphasized and non-emphasized speech. Our results demonstrate that the human auditory perception of emphasis is influenced primarily by the sound level of the speech, which yields larger neural population activity and a faster response time in the human auditory brainstem. Additionally, our results demonstrate that when the vowel stimuli without and with emphasis are presented at the same sound level, the small upward shift in the frequency content (i.e., fundamental and formant frequencies) observed in the emphasized vowel influences the auditory perception of emphasis with increased neural population activity and a faster response time in the human auditory brainstem.

1 Introduction

1.1 Emphasis in the English Language

Focus is a well-documented grammatical contrast marking the information structure of an utterance by making a word or unit of speech phonologically prominent (Halliday 1967, Crystal 1969, Cruttenden 1997, Ladd 2008). Vowels in these prominent words have special properties that make them emphatic: the vowels are longer, the vowel targets are hyper-articulated and the pitch is higher (Fourakis, 1991; Lindblom, 1963; Moon & Lindblom, 1994; Ladd and Morton, 1997). This may result in the F1 and F2 space of emphasized vowels being larger. So, a low vowel may

be realized lower and further back than the same vowel in the non-emphasized word. A low vowel is produced with a higher F1 and a lower F2 as well as a higher pitch. Examples of emphasized words are in caps in (1) and (2)

(1) **SAM** sent the email.
(2) Sam sent the **EMAIL**.

Observe in sentence 1 that the emphasis of the name *Sam* fixes the listener's attention to the importance of Sam and not the actions performed by Sam. In sentence 2 the emphasis of the object *email* shifts the listener's attention to the importance of the email and not that Sam sent the email.

1.2 The Auditory Brainstem Response

In order to investigate the effects of emphasis on the human auditory perception of speech, the neural population activity in the human auditory periphery can be measured. This neural population activity is called the auditory brainstem response (ABR) and can be measured using electroencephalography (EEG). The ABR is a scalp-recorded auditory evoked potential that occurs approximately 10 ms after a transient auditory stimulus (e.g. a click) is presented to the listener (Maddox & Lee, 2018). An example of the ABR is illustrated in Figure 1. This scalp potential consists of specific positive and negative peaks that have been commonly labeled as waves I-VII. Wave I corresponds to neural population activity in the auditory nerve (AN), Wave III to activity in the cochlear nucleus (CN), and Wave V to activity in the lateral lemniscus and the inferior colliculus (IC) (Møller et al, 1995; Verhulst et al, 2018).

Currently, in order to obtain the ABR waveform (waves I-VII), the ABR is measured and averaged over thousands of repetitions (i.e. trials) of the auditory stimulus (Maddox & Lee, 2018). Averaging the EEG trials is necessary due to the low signal-to-noise ratio (SNR) that is inherent in EEG; however, we take advantage of the stochastic property of noise to obtain an estimate of the ABR waveform. Noise is inherently random and contains positive and negative amplitudes at random points along the time course of the noise. Therefore, when a large number of trials of a noisy signal are added together the random positive and negative peaks in each of the trials of the noisy signal cancel each other out and the positive and negative peaks that are not random will remain, which results in an estimate of the underlying signal. The ABR paradigm stated above only uses clicks as the auditory stimulus. In order to measure the ABR to speech, the complex ABR (cABR) paradigm can be used (Skoe & Kraus, 2010). The cABR paradigm uses syllabic speech (e.g. ~40 ms "da") for the auditory stimulus.

While the measurement of the ABR to clicks and cABR to syllabic speech are well understood processes, the acquisition of the ABR to continuous speech is a process that is still in development. The difference between these two processes begins with the stimulus used. A click is a transient stimulus (it ends as soon as it begins) and syllabic speech has a short duration, while continuous speech has a much longer duration. The ABR is relatively quick (~10 ms); therefore, in order to account for the long duration of the speech stimuli, an encoding model is necessary to obtain the measurement of the ABR to continuous speech (Maddox & Lee, 2018). The encoding model is a linear system solved using linear regression with a continuous speech stimulus as the input and the EEG recording as the output. The regression framework of the encoding model is shown in Figure 2. Both the cABR and the encoding model capture the slow fluctuations in the syllabic and

continuous speech stimuli, respectively, which commonly corresponds to the envelope (i.e., amplitude modulation) of the speech stimuli (Dolphin, 1997). This phenomenon is termed as the frequency following response (FFR) or the envelope following response (EFR) (shown in Figure 3). The FFR/EFR has been shown to be driven primarily by the IC (Smith et al, 1975).

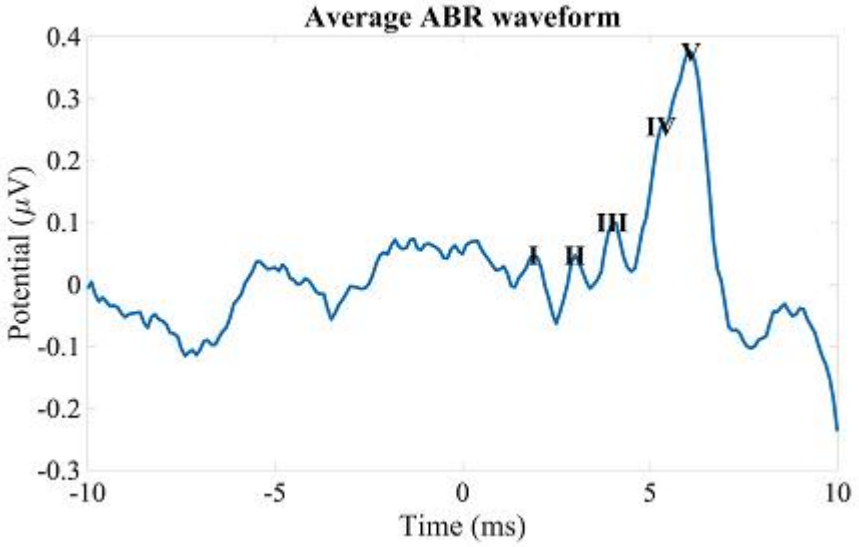

Figure 1. Average ABR to a click stimulus with annotations for Waves I-V.

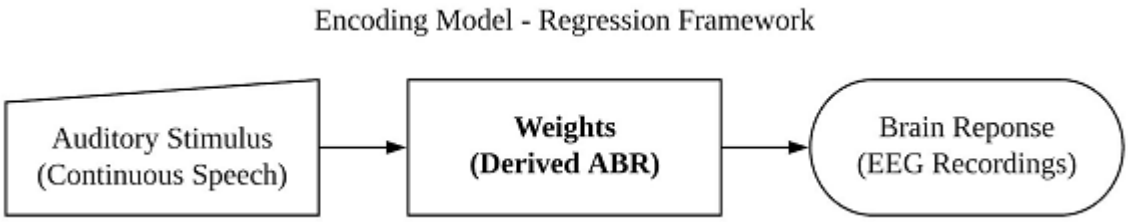

Figure 2. Schematic of the regression framework for the encoding model described in Maddox & Lee (2018).

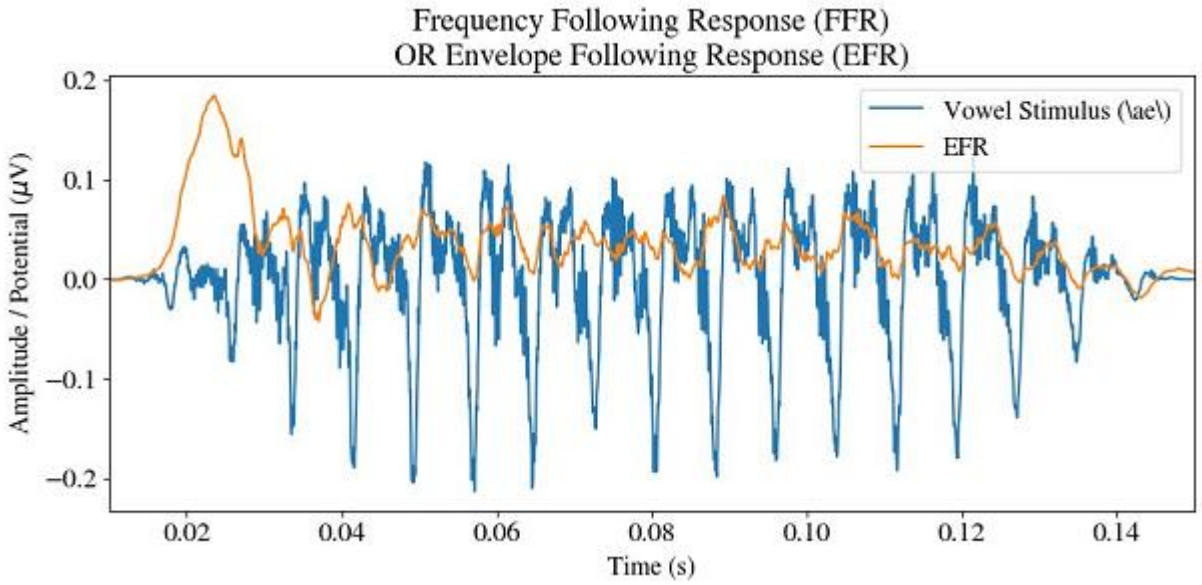

Figure 3. Demonstration of the frequency following response (FFR) or envelope following response (EFR). The estimated EFR (computed using the Verhulst et al (2018) model) to the vowel stimulus (\ae\ in "Sam") follows the envelope of the vowel stimulus.

1.3 Computational Modeling of the Human Auditory Periphery

Encoding models of the human auditory brainstem are useful for understanding how humans process continuous speech. However, a disadvantage of the encoding models is the requirement of EEG recordings from human subjects. Therefore, computational models of the human auditory periphery have been developed using physiologically relevant components in order to better understand the encoding models before recording EEG on human subjects.

Verhulst et al (2018) presented a model of the human auditory periphery (cochlea to brain stem) that outputs an estimated human ABR to an input stimulus (e.g. click or speech). The model outputs an estimate of Wave I, Wave III, Wave V of the human ABR and an estimate of the EFR. A flow chart diagram of the Verhulst et al (2018) model is shown in Figure 4. The model estimates these waveforms by first passing the input stimulus through a first-order middle-ear bandpass filter and a transmission line cochlear model. The transmission line model is a common approach to modeling the cochlea that discretizes the area along the length of the basilar membrane and describes this area in terms of coupled mass-spring-damper-elements (Altoè, 2014). The output of the cochlear model is then passed through an IHC-AN synaptic complex model, which yields the AN firing rates (r_{AN}). The AN firing rates are then passed through a same-frequency bushy cell model to yield the firing rates for the CN (r_{CN}) and the IC (r_{IC}). In order to compute the ABR Wave I, the r_{AN} for CFs between 112 Hz and 12 kHz are summed. The ABR Wave III is computed using r_{CN} and ABR Wave V using r_{IC}. The EFR is the weighted sum of the ABR Waves I , III, and V. The weights applied to Waves I, III, and IV are according to human ABR ratios.

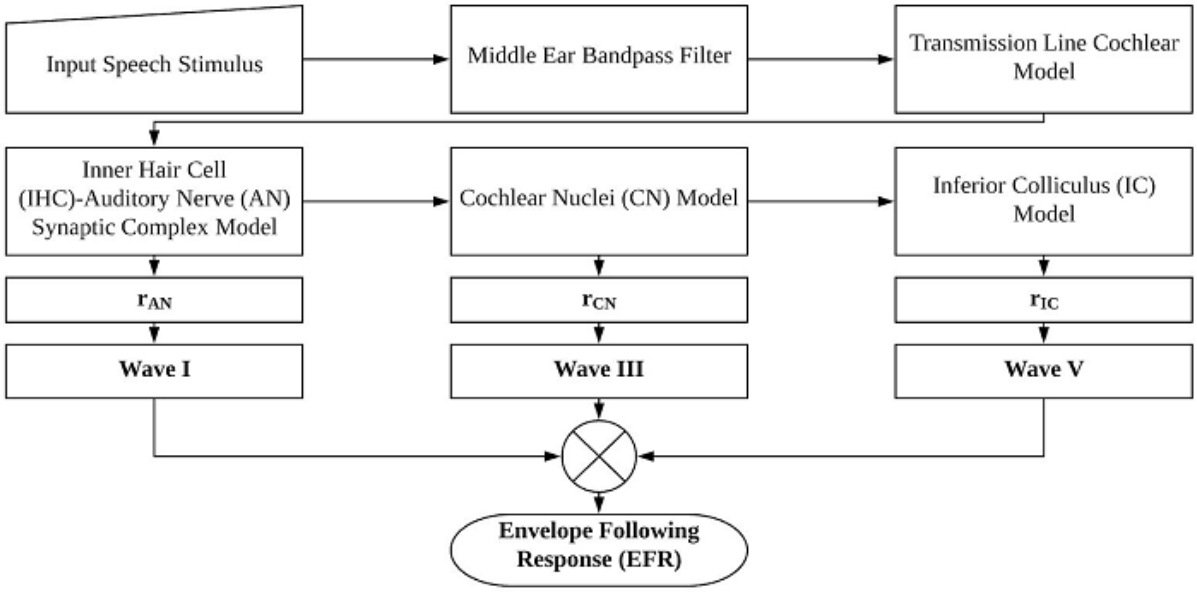

Figure 4. Flow chart diagram of the Verhulst (2018) model.

2 Methods

2.1 Stimuli

For our study, we chose to use isolated vowels for our stimuli. These vowels were extracted from spoken names without and with emphasis (shown in Figure 5). The spoken names and extracted vowels were Sam, Pat, and Todd (respectively, IPA: \æ\, \æ\, and \ɑ\; Hillenbrand: \ae\, \ae\, and \aw\). On average the extracted vowels had a duration of 150 ms. The fundamental and formant frequencies of each vowel stimulus are listed in Table 1. In order to ensure the vowels without and with emphasis were presented at the proper sound level, the vowels were divided by their respective root mean square (RMS) value, in order to achieve an RMS of 1 for each vowel (see the equation in 3). The amplitude of each vowel was then converted into Pascals and amplified to the desired sound pressure level (SPL) in dB using the equation in (4).

(3) $y = \frac{y_0}{RMS_{y0}}$

(4) $y_p = y \times p_0 \times 10^{L/20}$

In equations (3-4), y_0 is the input stimulus, y is the input stimulus with an RMS of 1, y_p is the input stimulus converted to the units of pascals, p_0 is the reference sound pressure and is 2×10^{-5} Pa (a standard reference for the units of dB SPL, and L is the desired sound level in dB SPL. Two scenarios were created to compare the effects of sound level. The first scenario reflected real life speech since the vowels with no emphasis were presented at a sound level of 70 dB SPL and the vowels with emphasis were presented at a sound level of 76 dB SPL (Engineering Toolbox, 2005). The second scenario eliminated the difference in sound level between the vowels without and with

emphasis by presenting both at a sound level of 70 dB SPL. The vowels were windowed using the *tukeywin* function in MATLAB (with a default cosine fraction [r] of 0.5) to avoid transients at the onset and offset of the vowel stimuli (shown in Figure 6). The Verhulst (2018) model is sensitive to transients, which dominate the modeled human auditory periphery response.

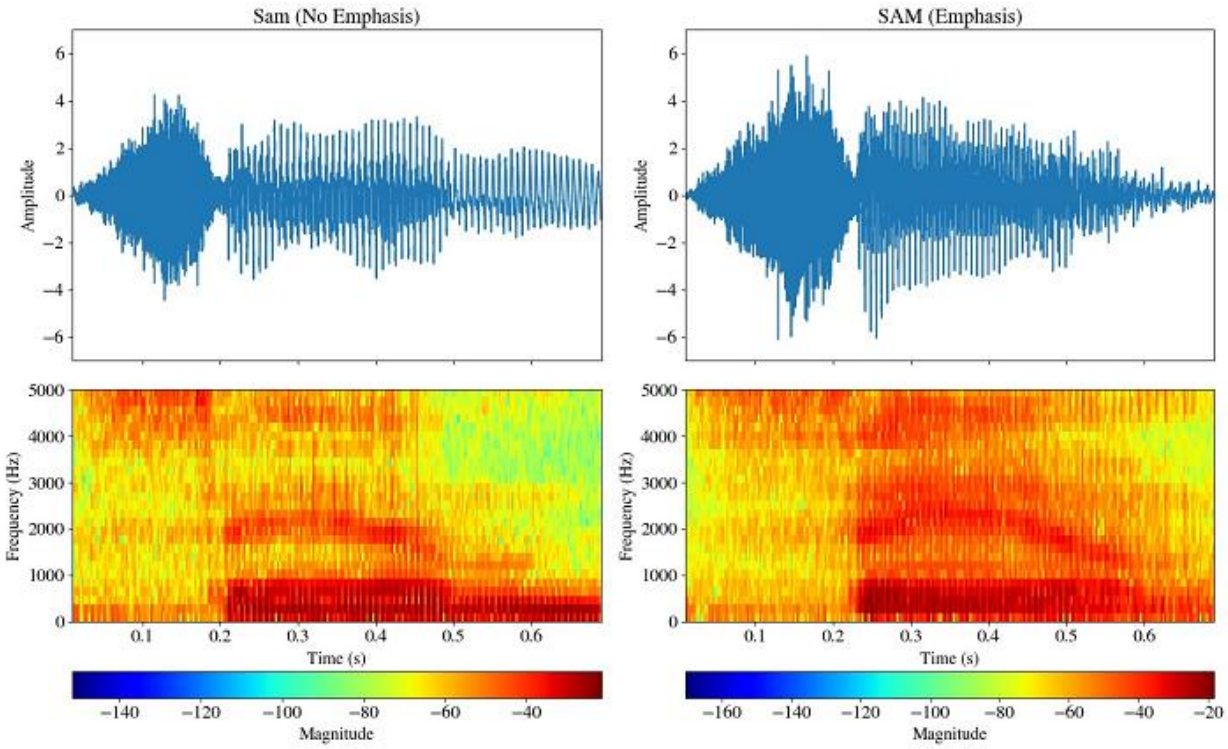

Figure 5. Time series and spectrogram of (*left*) Sam without emphasis and (*right*) SAM (with emphasis). The duration of the \ae\ vowel in Sam without emphasis is 286 ms and with emphasis is 343 ms. The F1 and F2 of the \ae\ vowel in Sam without emphasis is 514 Hz and 1886 Hz and with emphasis 534 Hz and 1901 Hz. The pitch of the \ae\ vowel in Sam without emphasis is 121 Hz and with emphasis is 128 Hz

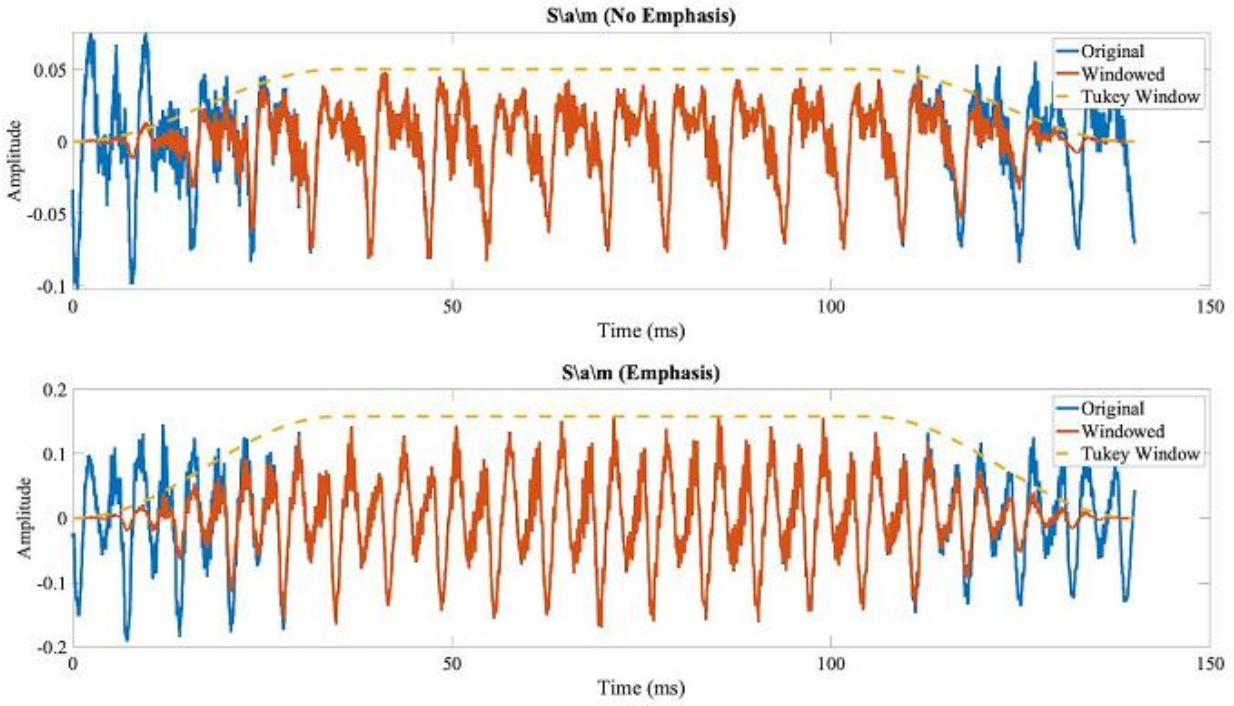

Figure 6. Original and Tukey Windowed vowel, \ae\ in Sam, without (*top*) and with emphasis (*bottom*).

Table 1. Fundamental (F0) and formant (F1, F2, and F3) frequencies of the vowel stimuli (measured using *Praat*).

Word	Vowel (IPA)	Emphasis	F0 (Hz)	F1 (Hz)	F2 (Hz)	F3 (Hz)
Pat	\æ\	No	131	667	1631	2653
		Yes	150	714	1626	2660
Sam	\æ\	No	134	386	2094	2719
		Yes	141	333	2277	2732
Todd	\ɑ\	No	131	592	914	2632
		Yes	148	591	987	2708

2.2 Adjusting the Pitch of the Emphasized Vowel

Pitch adjustment was performed on the emphasized vowel \ae\ in "Sam" to demonstrate if the best modulation frequency in the IC model yielded different effects on the resultant EFR. We adjusted the pitch of the vowel with emphasis to have the same pitch as the vowel without emphasis. The pitch shift was accomplished using the *shiftPitch* function in MATLAB. The SPL value for the vowel with and without emphasis were both 70 dB.

2.3 Using the Verhulst et al (2018) model

We downloaded the Verhulst et al (2018) model code from GitHub (https://github.com/HearingTechnology/Verhulstetal2018Model). We modified the MATLAB files "ExampleSimulation.m" and "ExampleAnalysis.m" to use our vowel stimuli as an input and output the modeled ABR. Also, we created two python scripts "create_input.py" and "analysis.py" in order to run the model and analyze the results solely in python.

3 Results & Discussion

The estimated EFR of paired vowels without emphasis and with emphasis were compared in two scenarios as shown in Figure 7. A peak is referred to as dominant if its amplitude was greater than the rest of the peaks. For example, the first peak in each panel was considered a dominant peak. Dominant peaks of each EFR waveform were compared in amplitude and latency. The amplitude was the voltage difference between peak voltage value and the baseline. The latency was the time interval between start point and the peak. The right plot shows a real scenario where the SPL value of the emphasized vowels is greater (~ +6 dB) than vowels without emphasis. Both SPL values were set as 70 dB in the other scenario as shown at the left column, ensuring the sound level of the vowel without and with emphasis was identical. The identical amplitude scenario enabled us to identify differences that cannot be explained by sound level. The amplitude and latency comparison results were shown in Figure 8.

3.1 Peak Amplitude

If the sound levels were identical, the dominant peak amplitude of no emphasis group was greater than that of emphasis group respectively for each of the three vowels as shown in the left panel of Figure 8. Increasing the SPL value of the emphasized vowels makes the dominant peak amplitude of the corresponding EFR larger than the dominant peak amplitudes of the vowels with no emphasis for \æ\ in "Pat" and "Sam". A larger dominant peak amplitude is attributed to increased neural population activity in the auditory system.

3.2 Peak Latency

At a real-life sound level, the latency of the dominant peak of the EFR to the vowels with emphasis is less than the latency of the dominant peak of the EFR to the vowels without emphasis as shown in the right panel of Figure 8. Lower latency indicates the auditory system responds more quickly to vowels in the emphasis group. This may explain why human attention can be drawn by emphasized speech to some degree.

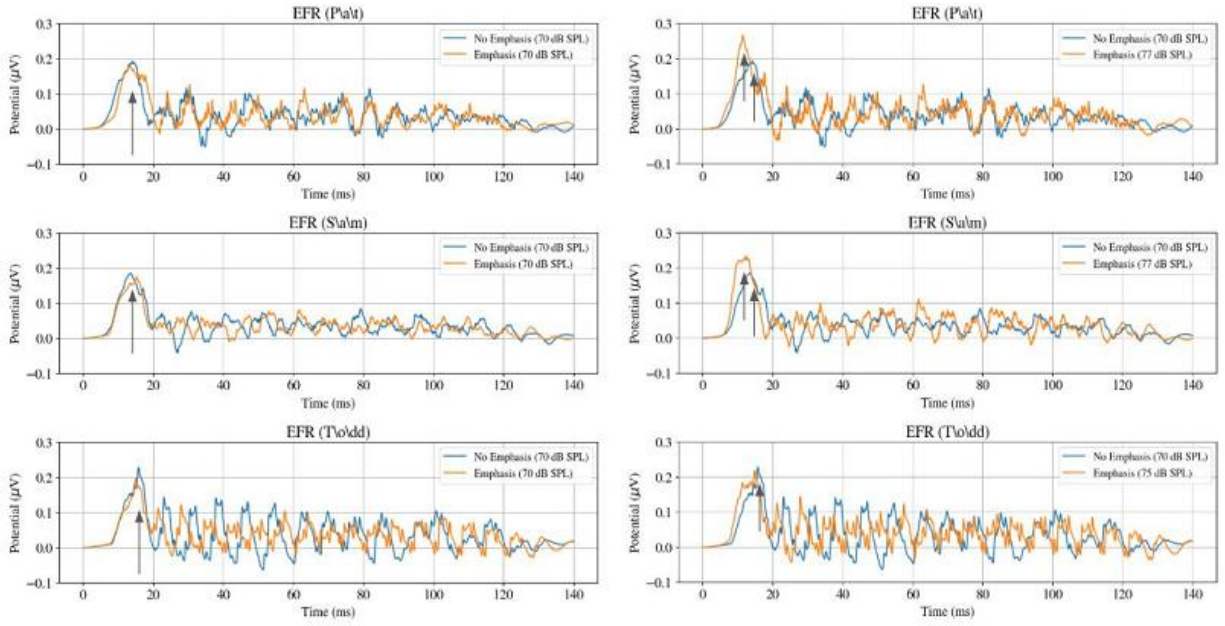

Figure 7. Comparison of the EFR to emphasis and no emphasis stimulus. Left column: EFR of emphasis (70dB SPL) and no emphasis (70dB SPL); right column: EFR of emphasis (76dB SPL) and no emphasis (70dB SPL). Arrows indicate dominant peaks; some arrows merged because two peaks are too close to each other.

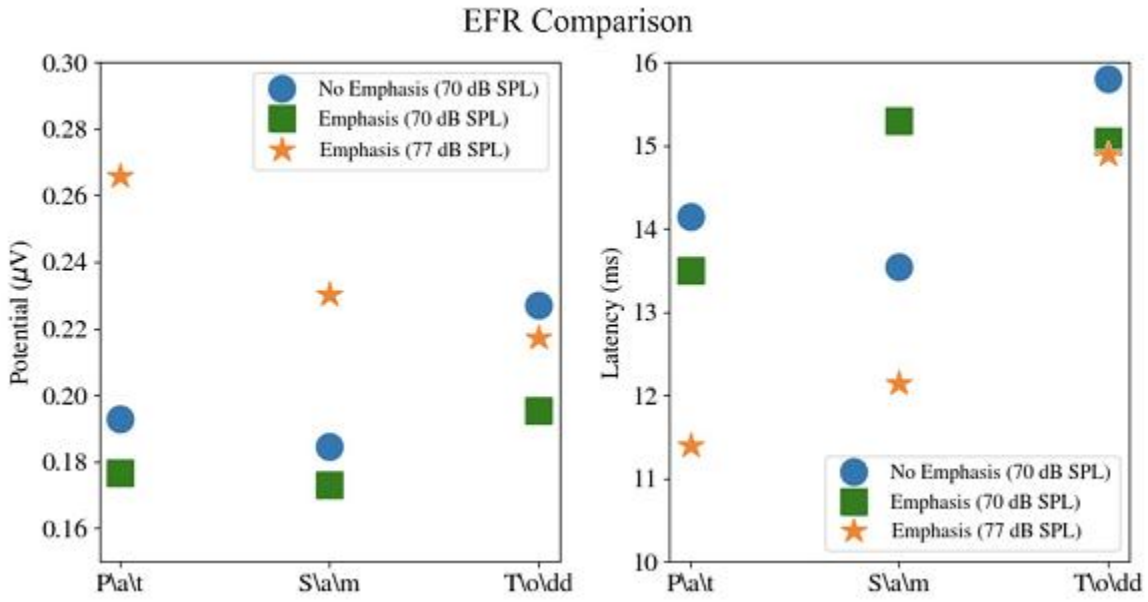

Figure 8. Comparison of the potential (μV) and latency (ms) of the maximum value of the EFR to (*left*) no emphasis (70 dB SPL) and emphasis (76 dB SPL) (*right*) no emphasis (70 dB SPL) emphasis 70 dB SPL) of the vowel in each spoken word.

3.3 EFR Spectrum

We also compared spectra of input stimulus vowels and the steady-state part of EFRs. Respectively, the fast Fourier transformation (FFT) was applied to each group's EFR trimmed to 40-100 ms for \æ\ in "Sam", 40-100 ms for \æ\ in "Pat", and 50-110 ms for \ɑ\ in "Todd" to obtain the frequency spectra of the steady state segments in the EFRs. Also, the FFT was applied to the input vowel stimulus to obtain the frequency spectra of input (shown in Figure 9). The limited amount of data points available to compute the frequency spectrum of each trimmed EFR resulted in poor frequency resolution. Therefore zero-padding 10x the length of the trimmed EFRs was applied to each of the trimmed EFRs to achieve the same frequency resolution in the frequency spectrum of the input stimulus. It was observed that the fundamental frequency (or pitch) differed between the vowels without and with emphasis. The vowels with emphasis had a higher fundamental frequency in both input sound and the EFR. Also, the energy distribution of the frequency spectrum did not change with SPL level.

3.4 Pitch Adjustment

We also investigated if pitch would affect our observations by shifting the pitch of the emphasized vowel. The ultimate goal of this pitch shift was to ensure that any differences observed between the resulting EFRs to the vowels without and with emphasis were not solely driven by a difference in fundamental frequency. A difference between EFRs driven simply by the difference in fundamental frequency would be uninteresting since this difference would be from the response of the periodicity-tuned IC model stage.

Therefore, we prepared two groups for comparisons between the dominant peaks and frequency spectra of each EFR. The input signals of the first group were the \æ\s in "Sam" of identical SPL (70 dB) without and with emphasis. The input signals of the second group were all the same as the former, except that we adjusted the pitch of the vowel with emphasis so that the pitch was identical. As shown in Figure 10, a comparison between the original and pitch shifted vowel with emphasis reveals the subtle upward shift in frequency content of the original vowel with emphasis yields a higher amplitude and lower latency in the dominant peak of the EFR. In contrast, adjusting the pitch of the vowel with emphasis to the pitch of the vowel without emphasis results in a lower amplitude and a higher latency in the dominant peak of the EFR to the pitch shifted vowel with emphasis. This comparison demonstrates that the subtle upward shift in frequency content results in increased neural population activity and a faster response time in the simulated auditory brainstem.

Differences can be observed in the EFRs to the original vowel with emphasis and the vowel without emphasis (shown in Figure 10). The amplitude of the dominant peak of the EFR to the original vowel with emphasis is lower than that of the vowel without emphasis. This difference in amplitude can be understood using the Verhulst et al (2018) model, since the IC stage of the model has a best modulation frequency of 100 Hz. This means that the model will have a stronger response (i.e. higher amplitude) to vowel stimuli with frequency content near 100 Hz, which is observed in our results since the vowel without emphasis has a fundamental frequency closer to 100 Hz compared to the original vowel with emphasis. However, differences can also be observed

in the EFRs to the pitch adjusted vowel with emphasis and the vowel without emphasis (shown in Figure 10). The amplitude of the dominant peak of the EFR to the pitch adjusted vowel with emphasis is still lower than that of the vowel without emphasis. Therefore, these remaining differences mean that the subtle upward shift in frequency found in the original vowel with emphasis is not the key reason for the differences between the EFRs to the vowels without and with emphasis. These remaining differences can be described as a potential physiological response to this pitch-matching feature.

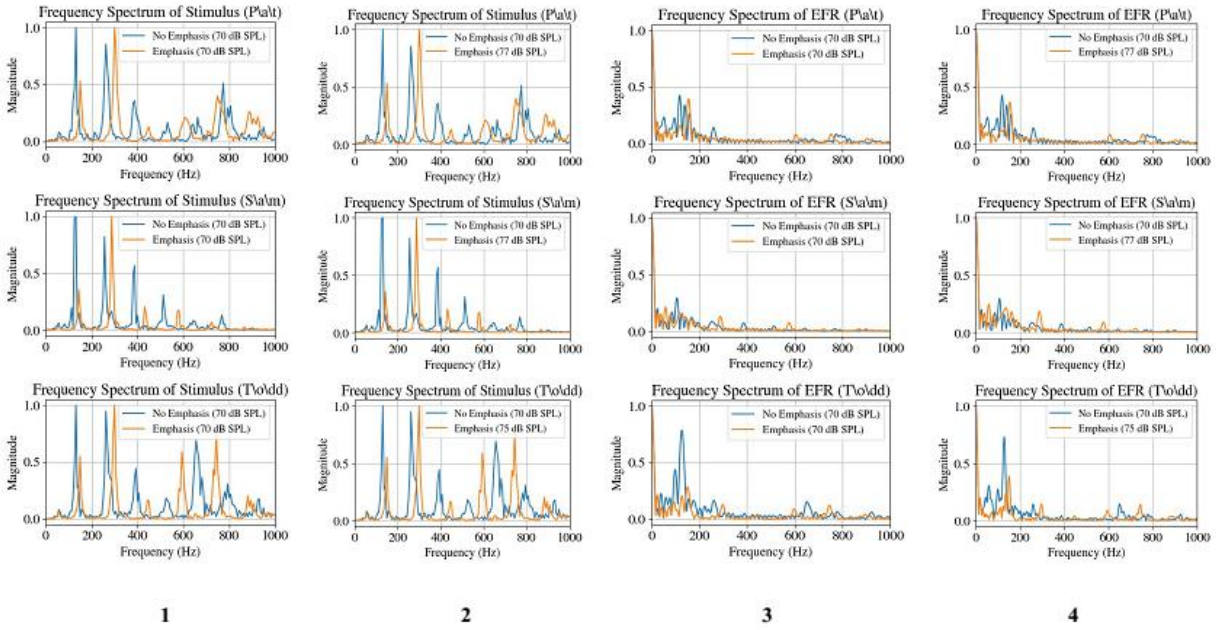

Figure 9. Comparison of the frequency component distribution for both stimulus vowel (left two columns) and corresponding EFR (right two columns). Columns 1 and 3 present the spectrum difference of the vowels without and with emphasis at the same SPL level (70 dB); columns 2 and 4 present the spectrum difference between the vowel with emphasis at ~76 dB SPL and the vowel without emphasis at 70 dB SPL. Disregarding SPL variation, EFR spectrum figures (shown in columns 3 and 4) show that the peak of the fundamental frequency for the vowels without emphasis has a greater amplitude than that of the vowels with emphasis.

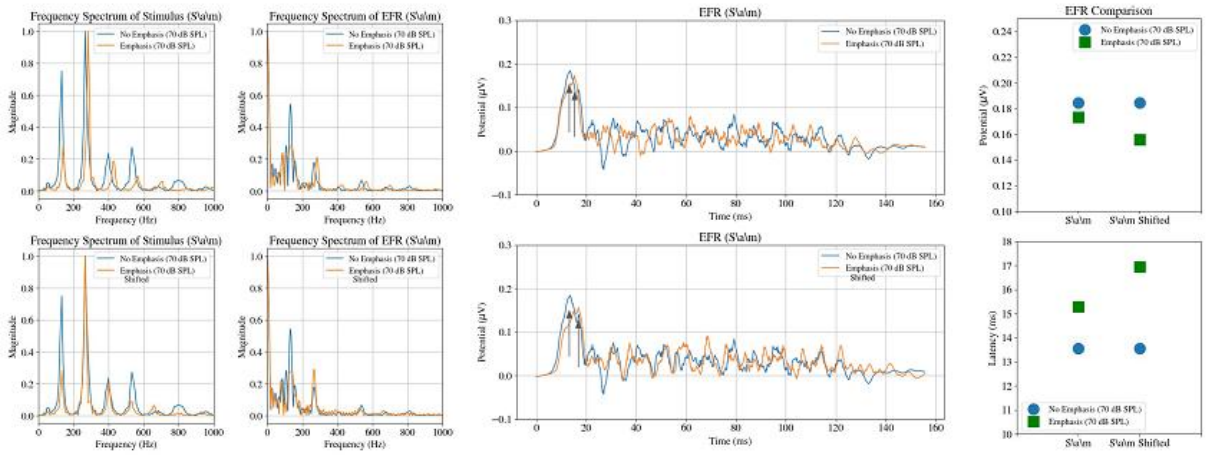

Figure 10. Shifting the pitch of the vowel stimulus with emphasis to match the pitch of the vowel stimulus without emphasis (*left*) results in differences in the corresponding EFR (*middle*). These differences can be observed in the potential (μV) and latency (ms) of the dominant peak in the EFR (*right*).

3.5 Limitations of the Study

We would like to point out a limitation of our study design. We took only the vowel from a speech signal for comparison, ignoring the context. This scenario is different from real life with context. In EFR population response figures, those responses start from zero as the initial status, but our perception of vowels is embedded in a word. This means the initial status for vowel simulation can be other non-zero values. We did not take content dependency into consideration, which calls into question our dominant peak amplitude and latency measurements. Therefore, we must consider that these measurements may only be salient features of the response because there was no preceding sound.

4 Conclusion

In conclusion, our results demonstrate that the simulated neural response to emphasis is influenced by amplitude and frequency content (fundamental frequency and formants). Our results demonstrate that vowel stimuli with a higher sound level elicit larger neural population activity and a faster response time in the auditory brainstem. Furthermore, our results of shifting the pitch of the vowel with emphasis to the pitch of the vowel without emphasis demonstrate that the subtle shift in the frequency content of the emphasized vowel elicits increased neural population activity and a faster response time in the simulated auditory brainstem.

Future work will be focused on running continuous speech stimuli through the entire Verhulst et al (2018) model. Currently, Verhulst and her team are working on using machine learning to simplify the model into a matrix consisting of learned weights and biases. This development will increase the computational efficiency of the model when processing large amounts of data (e.g. continuous speech).

References

Altoè, A., V. Pulkki & S. Verhulst. 2014. Transmission line cochlear models: Improved accuracy and efficiency. *JASA* 136.4: 302-308. doi:10.1121/1.4896416

Beaver, D.I. & B.Z. Clark. 2008. *Sense and Sensitivity: How Focus Determines Meaning.* Malden, MA: Blackwell.

Cruttenden, A. 1997. *Intonation* (2nd ed.). Cambridge.

Crystal, D. 1969. *Prosodic Systems and Intonation in English.* Cambridge University Press.

Dolphin, W.F. 1997. The envelope following response to multiple tone pair stimuli. *Hearing Research* 110.1-2: 1-14. doi:10.1016/s0378-5955(97)00056-7

Engineering ToolBox. 2005. Voice Level at Distance. [online] Available at: https://www.engineeringtoolbox.com/voice-level-d_938.html.

Fourakis, M. 1991. Tempo, stress, and vowel reduction in American English. *JASA* 90: 1816. doi:10.1121/1.401662

Fry, D.B. 1958. Experiments in the perception of stress. *Language and Speech* 1.2: 126–152. doi:10.1177/002383095800100207.

Halliday, M.A.K.. 1967. Notes on Transitivity and Theme in English: Part 2. *Journal of Linguistics* 3.2: 199-244. Retrieved April 19, 2020, from www.jstor.org/stable/4174965

Halliday, M.A.K. 1967. *Intonation and grammar in British English*. The Hague: Mouton.

Ladd, D.R. and R. Morton. 1997. *Phonetics* 25: 313-342.

Ladd, D.R. 2008. *Intonational Phonology* (2nd ed.). Cambridge University Press.

Maddox, R.K. & A.K.C. Lee. 2018. Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners. *eNeuro* 5(1): 441-417. doi:10.1523/eneuro.0441-17.2018.

Møller, A.R., H.D. Jho, M. Yokota & P.J. Jannetta. 1995. Contribution from crossed and uncrossed brainstem structures to the brainstem auditory evoked potentials: A study in humans. *The Laryngoscope* 105.6:596-605. doi:10.1288/00005537-199506000-00007

Moon, S.-J. 1994. Interaction between duration, context, and speaking style in English stressed vowels. 96.1: 40. doi:10.1121/1.410492

Skoe, E. & N. Kraus, N. 2010. Auditory Brain Stem Response to Complex Sounds: A Tutorial. *Ear and Hearing* 31.3: 302-324. doi:10.1097/AUD.0b013e3181cdb272

Verhulst, S., A. Altoè & V. Vasilkov. 2018. Computational modeling of the human auditory periphery: Auditory-nerve responses, evoked potentials and hearing loss. *Hearing Research* 360: 55-75. doi:10.1016/j.heares.2017.12.018