33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Classification of English Vowels using Speech Evoked Potentials

Amir Sadeghian, Hilmi R. Dajani, *Member, IEEE* and Adrian D. C. Chan, *Senior Member, IEEE*

Abstract— **The objective of this study is to investigate whether Speech Evoked Potentials (SpEPs), which are auditory brainstem responses to speech stimuli, contain information that can be used to distinguish different speech stimuli. Previous studies on brainstem SpEPs show that they contain valuable information about auditory neural processing. As such, SpEPs may be useful for the diagnosis of central auditory processing disorders and language disability, particularly in children. In this work, we examine the spectral amplitude information of both the Envelope Following Response, which is dominated by spectral components at the fundamental (F0) and its harmonics, and Frequency Following Response, which is dominated by spectral components in the region of the first formant (F1), of SpEPs in response to the five English language vowels (\a\,\e\,\ae\,\i\,\u\). Using spectral amplitude features, a classification accuracy of 78.3% is obtained with a linear discriminant analysis classifier. Classification of SpEPs demonstrates that brainstem neural responses in the region of F0 and F1 contain valuable information for discriminating vowels. This result provides an insight into human auditory processing of speech, and may help develop improved methods for objectively assessing central hearing impairment.**

I. Introduction

Analyzing the response of scalp-recorded Auditory Brainstem Responses (ABR) to clicks and other transient stimuli has been a key tool for clinicians and researchers to diagnose hearing impairments and to understand the human auditory neural system [1]. The *transient response* refers to the initial component of the ABR (usually up to 20 ms) after the onset of the stimulus. On the other hand, when using periodic sound stimuli, such as amplitude modulated tones and synthetic vowels, an additional response is formed after the transient response, called the *sustained response*. Recent studies have shown that the sustained response provides additional information regarding the state of the central auditory neural system, especially in children with language and learning problems and potentially in other populations with central processing disorders [2,3].

Previous studies on speech-evoked ABRs or Speech Evoked Potentials (SpEPs) have mainly focused on the understanding of underlying auditory neural activity during speech processing, origination of SpEPs, and new techniques for diagnosis of hearing impairment [4,5]. For instance, a possible clinical application would be in hearing assessment of infants. Currently hearing assessment is limited by diagnostic tests, which usually employ artificial signals like tones or clicks that do not allow a clear assessment of auditory function for speech communication. While there are tests of speech perception that rely on subjective responses, these are of no value for assessing the hearing of infants and uncooperative individuals. SpEPs could thus fill the need to objectively assess auditory performance in these cases. Currently, however, there is limited understanding of SpEPs and how they relate to processing of different speech sounds. This study addresses this question by demonstrating a firm relation between different vowels and the corresponding SpEPs. It shows that the SpEPs of the five English vowels can be discerned through a basic classification method, which as far as the authors can determine, is the first attempt at speech recognition using SpEPs. This finding suggests that the SpEPs carry useful information for discriminating speech stimuli.

II. methods

A. Evoked Potential Data Acquisition

This research was approved by the University of Ottawa Research Ethics Board. Four male subjects (age range: 25-45 years) participated in this experiment. Subjects had no known hearing disorder, and normal hearing thresholds of 15 dB or less were confirmed in both ears through an audiometric test using a Clinical Audiometer (model AC40, Interacoustics, Eden Prairie MN, USA) at 500, 1000, 2000, and 4000 Hz.

Five synthetic vowel stimuli (\a\,\e\,\ae\,\i\,\u\) were generated using a formant synthesizer [6], with each stimulus 300 ms in duration. The parameters of the stimuli (first 3 formant frequencies, formant bandwidths, and relative formant amplitudes) followed those determined in previous work for male speakers [6,7]. The fundamental frequency (F0) of all vowels was set to 100 Hz. Only the first three formants (F1, F2, and F3) of each vowel were used, since these formants are the most dominant. The stimuli were generated with a sampling frequency equal to 48 kHz.

Subjects were seated comfortably in an acoustical booth and performed 6 trials for each vowel during which they were asked to stay relaxed, while keeping their eyes open. In one trial, subjects were presented 500 repetitions of the vowel at a repetition rate of 3.1/sec. Responses were coherently averaged over the 500 repetitions prior to further

Manuscript received April 14, 2011.

A. Sadeghian is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: asadeghi@connect.carleton.ca).

H. R. Dajani is with School of Information Technology and Engineering (SITE), University of Ottawa, Ottawa, ON K1N 6N5, Canada (e-mail: hdajani@site.uottawa.ca).

A. D. C. Chan is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada (e-mail: adcchan@sce.carleton.ca).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

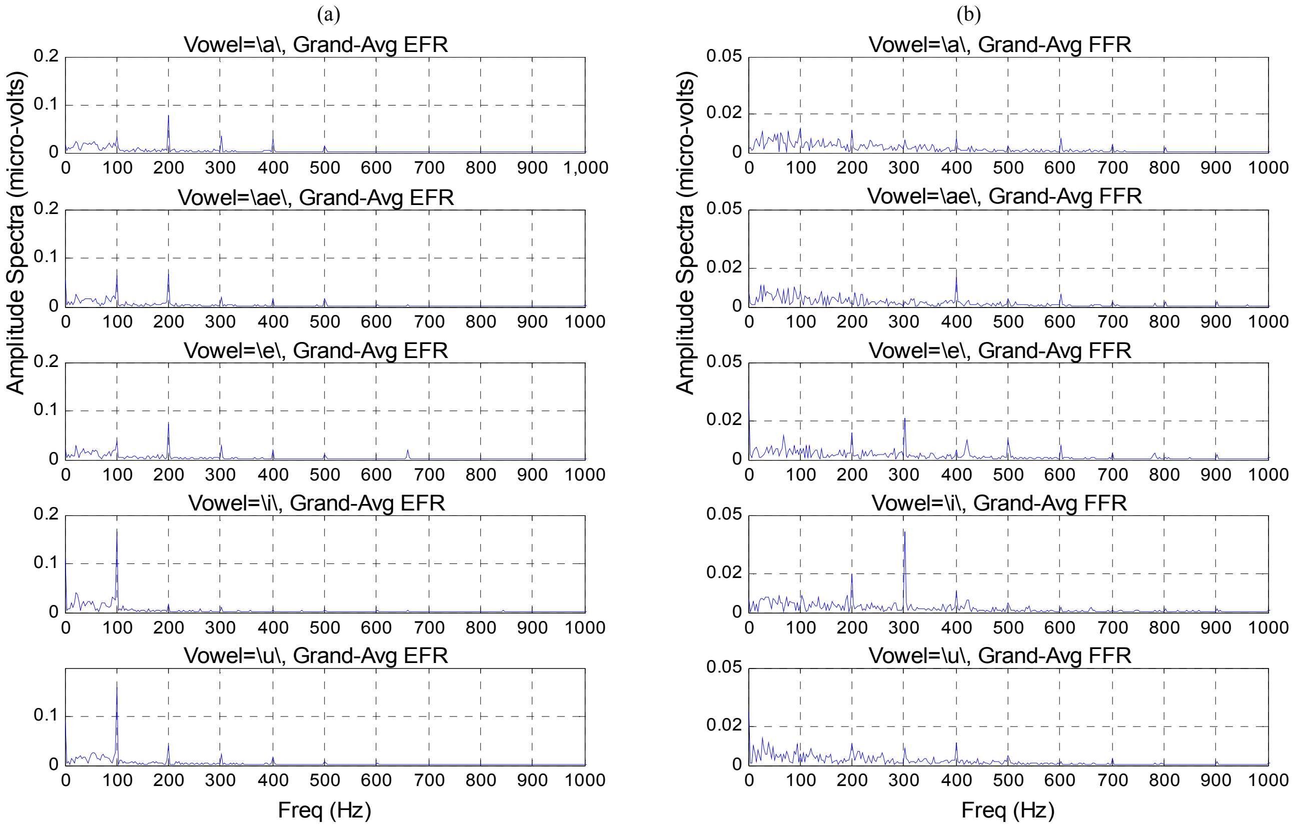

Fig. 1. Single-sided amplitude spectra (up to 1000 Hz) of the SpEPs for all vowels averaged over all trials and all subjects (grand-averages) for a) Envelope Following Response (EFR) and b) Frequency Following Response (FFR).

analysis. A BioMARK v.7.0.2 system was used to present the stimuli and record the SpEPs. Each vowel was presented at a calibrated level of 80.5 dB SPL by adjusting an internal calibration factor in the BioMARK system, with the calibration performed using a Brüel & Kjær 2235 sound level meter. Stimuli were presented using Etymotic ER 2 insert earphones. Three gold-plated Grass electrodes were used in this experiment; the recording electrode was placed at the vertex (Cz), the reference electrode was placed on right earlobe, and the ground electrode was placed on the left earlobe. Electrode impedances were kept below 5 kΩ during the recording by monitoring the impedance at the start and end of each trial. Vowels were presented using alternate polarity, at 48 kHz with a 16-bit resolution. For each trial, SpEPs were recorded with a sampling frequency of 3202 Hz for a duration of 319.8ms starting at stimulus onset.

B. *Analysis*

The sustained response in the SpEP can correspond to the EFR or FFR depending on how the response signals are analyzed. SpEPs from the opposite polarity recordings were used to calculate the Envelope Frequency Response (EFR), by taking the average between the responses of the original stimulus and the inverted polarity stimulus, and the Frequency Following Response (FFR), by taking the average between the response of the original stimulus and the negative response of inverted polarity stimulus [9].

The EFR primarily reflects auditory neural phase-locking to the envelopes of the speech stimuli, which are modulated at F0 [9,10]. Fig. 1a shows the amplitude spectra of EFR for each vowel averaged across all trials and subjects (i.e. grand-average EFRs)**.** This figure confirms that there are robust peaks at harmonics of F0. On the other hand, the FFR is formed as a result of auditory neural phase-locking to the formants of a speech stimulus. Spectral analysis of the FFR shows that strong peaks occur at harmonics of F0 near the formant frequencies [11,12]. Fig. 1b illustrates the FFR spectral amplitudes for each vowel averaged across all trials and subjects (the grand-average FFRs) and it is dominated by harmonics near the F1 frequencies listed in Table I.

TABLE I
FIRST FORMANT FREQUENCIES OF ALL VOWEL STIMULI

Vowels	First Formant
\a\	700 Hz
\ae\	660 Hz
\e\	570 Hz
\i\	270 Hz
\u\	300 Hz

In this work, the responses at second and third formants probably played little role in the analysis, because several of them were beyond the upper cut-off frequency of the band-pass filter on the BioMARK system of 1000 Hz, and beyond the phase-locking limit of the probable main generator of SpEPs in the upper brainstem [8].

C. Feature Selection

In order to classify the SpEPs of the different vowels, the amplitudes of the EFR and FFR spectrum at harmonics of F0 were used as signal features. The frequency spectrum was determined using the Discrete Fourier Transform (DFT) of the coherently averaged response in each trial, containing 1024 data points. We only considered spectral values at harmonics of F0 between 100 Hz and 700 Hz, inclusive, for EFR and between 200 Hz and 800 Hz, inclusive, for FFR. Therefore, the feature vectors had 8 amplitude feature elements for the EFR or FFR. The frequency ranges were obtained empirically by looking at the amplitude spectra of the samples of all vowels. As can be seen in Fig. 1, the peaks at harmonics start diminishing after 700 Hz for EFR and after 800 Hz for FFR. Also, we did not consider the FFR amplitudes at 100 Hz because there was no robust peak at this frequency.

D. Classification

Linear Discriminant Analysis (LDA) was employed for classification [13]. We had five classes corresponding to the five vowels and each class had 24 sets of SpEPs samples (6 trials corresponding to 500 stimulus repetitions per subject × 4 subjects). Leave-one-out with no replacement was used to train and test; that is, training was performed on all samples except one, which was used to test. The leave-one-out was repeated such that each of the 120 SpEP samples (5 vowels x 6 trials/vowel x 4 subjects = 120) was tested.

III. Results

Table II shows the classification accuracies of three different amplitude feature sets that correspond to EFR, FFR, and EFR+FFR (concatenating the EFR and FFR feature vectors to form a single 16 element feature vector). The combined EFR and FFR amplitude features gave the highest accuracy of 78.33 %. The EFR features appear to discern the vowels better than the FFR features.

TABLE II
Classification Accuracies Of EFR+FFR, EFR, And FFR, Amplitude Features

Amplitude Features	Classification Accuracy
EFR + FFR	78.33 %
EFR	70.83 %
FFR	53.33%

Table III shows confusion matrices for the three different amplitude feature sets. Table III-A shows that vowel \i\ was correctly classified with the highest accuracy of 95.8% and vowel \e\ was correctly classified with the lowest accuracy of 62.5% among all vowels.

IV. Discussion

The ability to discern speech from the SpEPs provides a potentially powerful tool for researchers to better understand the human auditory system. Classification results obtained from this experiment demonstrates that we were able to successfully discern SpEPs of five different vowels with an accuracy of 78.3%, which is noticeably higher than the chance accuracy of 20% (100% / 5 vowels = 20%). Moreover, this study was performed using a listener independent approach (not trained per subject), which makes the results more generally applicable.

The ability to discern the five English vowels using the SpEPs demonstrates that brainstem neural responses in the region of F0 and F1 contain valuable information for discriminating vowels. The high classification accuracy with the EFR in particular is unexpected since the EFR mainly reflects neural activity that corresponds to the source of speech (as opposed to the filter) [5], whereas vowels are usually thought to be perceptually discriminated based mainly on the formant frequencies, and in particular the relative frequencies of F1 and F2 [7].

TABLE III
Confusion Matrices For A) EFR+FFR, B) EFR, And C) FFR Amplitude Features

(A)

EFR+FFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	17	3	4	0	0
	\ae\	0	19	4	0	1
	\e\	5	2	15	0	2
	\i\	0	0	0	23	1
	\u\	1	0	2	1	20

(B)

EFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	15	1	8	0	0
	\ae\	2	18	3	1	0
	\e\	6	2	13	0	3
	\i\	0	0	0	22	2
	\u\	1	0	3	3	17

(C)

FFR		Predicted Vowels				
		\a\	\ae\	\e\	\i\	\u\
Actual Vowels	\a\	9	7	2	0	6
	\ae\	7	15	2	0	0
	\e\	4	1	10	3	6
	\i\	1	0	3	14	6
	\u\	4	1	2	1	16

The EFR amplitude features provide higher classification accuracy than the FFR amplitude features, and thus are more distinctive than the FFR amplitude features. Visually, it appears vowels may be differentiated using the EFR spectrum better (Fig. 1a) than using the FFR spectrum (Fig. 1b). The weakness of the FFR amplitude features could be due to three reasons. The first reason is that only the FFR at F1 was analyzed; any responses at F2 and F3 were omitted, as explained in analysis section. Adding the responses of higher formants (especially F2), if they are available in the SpEP, could improve the FFR features by providing additional distinct information specific to each vowel [7].

The second reason is that the F1 frequencies for vowels \a\, \ae\, and \e\ are similar and also similar for vowels \i\ and \u\ (Table I). This could generate overlapped response peaks at harmonics of F0 around F1 frequencies. For instance, vowels \a\ and \ae\ have the closest F1 frequencies among all vowels. As shown in Table III-C, they were mainly misclassified to each other. However, there are instances in Table III-C that show this may not be applicable to all vowels. For example, vowels \a\, and \e\ were highly misclassified to vowel \u\ even though their F1 frequencies were not similar. The third reason is that the biological variability of the subjects could impact the FFRs such that the auditory phase-locking to F1 was not consistent across subjects. As a result, the FFR spectrum of different vowels may not have been strongly distinguishable when the responses were combined from all subjects. Both the EFR and FFR features provide a classification accuracy that is considerably higher than chance, demonstrating speech information in both signals. Together the EFR and FFR provide a higher classification accuracy than either individually. As indicated in section II-B the EFR and FFR reflect different aspects of the response.

The overall classification accuracy can probably be improved by increasing the number of stimulus repetitions per trial (i.e. averaging more individual responses to produce a single SpEP sample). This will help to reduce noise in the EFR and FFR spectra and so provide better amplitude features.

V. Conclusions

We have demonstrated that SpEPs of five English vowels are discernable with a fairly high accuracy of 78.33% using LDA for classification and the amplitudes of EFR and FFR as features. Results show that the EFR amplitude features represents each vowel better compared to the FFR amplitude features. The disadvantage of the FFR amplitude features was suggested to be due to the limitation of examining only the responses of F1, having similar F1 frequencies for some vowels, and the biological variability of the subjects. Analysis in this study used the entire SpEP, which contained both the transient and sustained responses. Future work will examine these responses, separately.

Results obtained from this study demonstrate that SpEPs contain useful information about the stimuli. Therefore, this work is a solid baseline for further study of SpEP classification using more complex stimuli, such as words. In addition, these results may be improved by collecting additional data from a larger number of subjects and employing more advanced feature selection algorithms and classification methods. Using more complex classifiers may help to obtain a higher accuracy by providing better decision boundaries for misclassified cases such as SpEPs of vowels \a\ and \ae\. However, given the limited data set in this study, a simple classifier like LDA probably helped to prevent over-fitting.

References

[1] Burkard, R.F., Eggermont, J.J., Don, M. 2007. Section 1: What are auditory evoked potentials?, Auditory evoked potentials: basic principles and clinical application, 1 ed. Lippincott Williams & Wilkins, Philadelphia pp. 7-20.

[2] Russo, N., Nicol, T., Musacchia, G., Kraus, N., Brainstem responses to speech syllables, Clinical Neurophysiology, 2004, 115, 2021-2030.

[3] Johnson, K.L., Nicol, G.T., Zecker, S.G., Bradlow, A.R., Skoe, E., Kraus, N., Brainstem encoding of voiced consonant–vowel stop syllables. Clinical Neurophysiology, 2008, 119, 2623-2635.

[4] Martin, B.A., Korczak, P. and Tremblay, K., Speech-evoked potentials: From the Laboratory to the Clinic. Ear and Hearing, 2008, vol.29, 285-313

[5] Kraus N, Nicol T. Brainstem origins for cortical what and where pathways in the auditory system. Trends neurosci. 2005, vol. 28, No. 4, 176-181.

[6] Klatt, H.D, Software for a cascade/parallel formant synthesizer. J. Acoust. Soc. Am., 1980, Vol. 67, 971-995.

[7] Peterson, E.G., Barney, L.H. 1952. Control Methods Used in a Study of the Vowels. The Journal of the Acoustical Society of America 24, 175-184.

[8] Johnson, K.L., Nicol, G.T., Kraus, N. 2005. Brain Stem Response to Speech: A Biological Marker of Auditory Processing. Ear & Hearing 26, 424-434.

[9] Aiken, S.J., Picton, T.W,. Envelope and spectral frequency-following responses to vowel sounds. Hearing Research, 2008, 245, 35-47.

[10] Dajani, H. R.., Purcell, D., Wong, W., Kunov, H., Picton, T.W., Recording Human Evoked Potentials That Follow the Pitch Contour of a Natural Vowel, IEEE Transactions on Biomedical Engineering, 2005, vol.52, 1614-1618

[11] Krishnan, A. 2002. Human frequency-following responses: representation of steady-state synthetic vowels. Hearing Research 166, 192-201.

[12] Skoe, E., Kraus, N. 2010. Auditory Brain Stem Response to Complex Sounds: A Tutorial. Ear & Hearing 31, 302-324.

[13] Duda, R.O., Hart, O.E., Stok, D.G. 2001. Pattern Classification, 2 ed. Willey Interscience, Toronto(Canada). 114–121.