Cortical characterization of the perception of intelligible and unintelligible speech measured via high-density electroencephalography

Rene L. Utianski[1], John N. Caviness[2], and Julie M. Liss[1]

[1]Department of Speech and Hearing Science, Arizona State University

[2]Department of Neurology, Mayo Clinic- Arizona

Running title: Cortical characterization

Keywords: EEG, speech, perception, disordered, intelligibility, listening strategy

Corresponding address:

Rene Utianski

2211 10th Street and Myrtle

Tempe, AZ 85287

Email: rutiansk@asu.edu

Phone: (480) 965-2374

Abstract

High-density electroencephalography was used to evaluate cortical activity during speech comprehension via a sentence verification task. Twenty-four participants assigned true or false to sentences produced with 3 noise-vocoded channel levels (1-unintelligible, 6-decipherable, 16-intelligible), during simultaneous EEG recording. Participant data were sorted into higher- (HP) and lower-performing (LP) groups. The identification of a late-event related potential for LP listeners in the intelligible condition and in all listeners when challenged with a 6-Ch signal supports the notion that this induced potential may be related to either processing degraded speech, or degraded processing of intelligible speech. Different cortical locations are identified as neural generators responsible for this activity; HP listeners are engaging motor aspects of their language system, utilizing an acoustic-phonetic based strategy to help resolve the sentence, while LP listeners do not. This study presents evidence for neurophysiological indices associated with more or less successful speech comprehension performance across listening conditions.

1.0 Introduction

Understanding speech in typical communication is generally effortless, but listening in acoustically adverse situations exposes a wide range of performance variability among healthy listeners [1,2]. It may be hypothesized that this variability is a reflection of the ways in which listeners recover words (or larger units of speech) from the impoverished acoustic signal to extract meaning. A useful model for investigating this variability is the dual stream model of speech perception [3,4,5]. Briefly, the dorsal stream contains the articulatory/ motor networks of the frontal lobe and the ventral stream contains the conceptual/ semantic networks of the temporal lobe. Anatomically, the dorsal stream spans an area of the posterior Sylvian fissure, projecting toward the frontal regions; the ventral stream projects toward the posterior medial temporal gyrus. It is proposed that these two active, bilateral, pathways converge for understanding speech. Previous fMRI work has shown dorsal stream (frontal) activation for processing pseudowords and ventral stream (temporal) activation for complex sentences [6]. While such research demonstrates differential activation for processing isolated units of speech, it is unknown whether or how this extends to processing degraded speech for meaning. If the behavioral performance data map meaningfully to cortical activation patterns within the dual streams, this may lead to an explanation of performance variability. For example, disproportionate activation of the dorsal stream may be indicative of attending to fine structure of the degraded signal, which may have benefit or cost for recovering the spoken message. High-density electroencephalography (EEG) provides both temporal and spatial resolution to identify cortical activation patterns that may distinguish more and less successful listeners.

We used a sentence verification task (true/false) for noise vocoded speech at three levels of intelligibility. The degradation was intended to induce performance variability among healthy

listeners to reveal poorer and better groups. The sentence verification task required both deciphering the speech and rendering a true/false decision, thereby tapping potential differences in recruitment of the dual streams. Simultaneous EEG acquisition allowed for examination of the associated cortical activation patterns, where it was expected that listeners who take advantage of both acoustic-phonetic and semantic information would show activation in both the dorsal and ventral streams, with better performance in the sentence verification task.

2.0 Methods

2.1 Participants

Participants were 24 undergraduate and graduate students (21 female) from Arizona State University. Ages ranged from 20- 48 (mean=25). Listeners self-reported English as native language; right-handedness; and a negative history for speech, language or hearing disorders. Hearing thresholds were within normal limits (detection of 125Hz- 4000Hz < 15dB in each ear), per pure tone screening conducted before the experiment. Listeners received $20 for participation.

2.2. Speech Stimuli

A healthy 46 year-old female recorded stimuli in a sound-attenuating booth via headmount microphone (Plantronics DSP-100). Her acoustic speech characteristics were typical of age and sex peers. She read from a computer screen 240 sentences, half of which were “true” (e.g. zebras have stripes) and half “false” (e.g. donkeys have wings). Speech was acquired via TF32 ([7]; 16-bit, 44kHz) and saved for editing (using [8]). All sentences were three words, with the last word determining the veracity of the sentence. Average sentence duration was 1959ms (SD= 213ms). Sound files were RMS normalized prior to vocoding. For the experiment, a third of the sentences

were vocoded using a 1-channel (unintelligible), 6-channel (moderately intelligible), or 16-channel (intelligible) noise vocoder (PRAAT [9]). Thus, there were 40 sentences at each combination of intelligibility level and veracity. Noise vocoding was chosen to degrade the speech signal because it minimizes the availability of phonetic information, while preserving other properties of the speech signal (e.g. rhythmic structure); but without adding to the speech signal (e.g. speech in noise).

2.3 Task

Data were collected, analyzed, and interpreted at ASU, in accordance with approved IRB protocols. Participants sat in a hard-backed chair at a comfortable distance from the monitor. STIM2 [10] delivered sentences through inter-aural headphones (90dB) and the STIM audiobox system ensured proper synchronization of audio delivery to the participant and EEG recording via Scan [11].

On-screen visual prompts guided the participants through the experiment. Following each sentence presentation, participants pressed true or false keyboard buttons. Pre-training ensured task understanding and sufficient muscle relaxation. Each participant then responded true/false for six blocks of 40 sentences (240), whose order of presentation was randomized within blocks, and order of blocks (1-6) was partially counterbalanced amongst participants. Breaks were permitted.

2.4 Electroencephalographic Recording

Neuroscan Acquire (v4.5, Compumedics Neuroscan, Charlotte, USA) [11], with128-channel QuickCap, was used for EEG recordings. Accurate electrode positioning was assured through measurement and positioning of Cz, Fz, and Pz, according to 10-20 system. Recordings were acquired with a 1000 Hz-sampling rate and low-pass filtered below 200 Hz. After

recording, a 60Hz notch filter minimized effects of electrical artifact. Impedance of all electrodes was well below 5kΩ. Continuous recordings were examined for physiologic and non-physiologic artifact. Artifact reduction through linear derivation minimized the presence of blinks. All recordings were of high quality and used in subsequent epoching.

The continuous file for each individual was epoched for each condition- the 40 sentences of each 1-, 6-, and 16- channel intelligibility levels. Epochs were created for 300ms prior to the onset of the sentence to 1500ms following onset. Each individual epoch was examined for artifact and those with high levels were removed. On average, 80% of recorded epochs were utilized to create an average file for each condition for each participant. The selected files were concatenated to create a grand average for all listeners, for each condition. Ten electrodes were omitted from the average due to high artifact. Only averaged files were subsequently used to facilitate strong signal to noise ratios.

2.5 Data Analysis

2.5.1 Behavioral Accuracy. STIM2 (Compumedics Neuroscan, Charlotte, USA) presented sentences and recorded participant response. Microsoft Excel was used for subsequent analyses. All data were reviewed and incorrect and no response items were considered together. Therefore, only correct responses, provided within the appropriate time interval were considered as correct. As our previous work has demonstrated a wide range of listener performance for 6-channel condition, we anticipated and found this in the current behavioral data. We selected an 80% accuracy level as the threshold for successful performance based on pilot data to create groups of High Performing (HP; n=9) and Low Performing (LP; n=15). Pilot data explored accuracy of responses to a systematic range of noise-vocoded speech, and provided an expectation for average performance on each level of noise-vocoded speech. These data indicate exceptionally

accurate listeners with highly intelligibly noise-vocoded speech (e.g. 16-channel), performed at or above 80% accuracy on the 6-channel speech task. Grouping accomplished sufficient signal-noise ratios for interpretation. The HP and LP group designations were used in subsequent EEG analyses, where analyses described below were completed for all listeners and repeated for both groups. It is of note that participants in each group were of equivalent age (see Table 1), with HP mean of 27 years-old and LP 25 years-old. Thus differences between the groups are unlikely due to age effects but rather listening differences.

2.5.2 Event-related potentials. Utilizing CURRY 7 Multi-Modality Imaging Suite (Compumedics Neuroscan, Charlotte, USA) [12], average files were examined for transient activity, indicated by peaks in mean global field power (MGFP), representative of the power of activity, across all electrodes [13]. Independent components analysis (ICA) assessed the transient in the time interval of interest [14]. Current density reconstruction (CDR), via sLORETA, identified component of interest [15]. This localizes the source of the component, or the underlying neural generator responsible for the transient activity. sLORETA computes minimum norm least squares (MNLS) current density amplitudes (dipole moments) and divides them by their error bars (and squares the result), taking into account the amplitude of activity; therefore, the F-values provided by sLORETA can be interpreted as magnitudes of activity. CURRY assigns a Brodmann's area to the solution to the MNLS problem, which was subjectively validated by the authors.

2.5.3 Frequency Analysis. Frequency spectra for average files were examined via SCAN (v4.5, Compumedics Neuroscan, Charlotte, USA). The time-domain averages were spline fit to ensure the average waveforms consisted of a power-of-2 number of points (2048 points). Each average file was analyzed via Fourier transform to obtain the power spectrum at each electrode,

from the offset of the "early ERP" (see Table 2), to 1500ms, capturing the later processing associated with understanding the sentences. After which, average power of each frequency band was calculated for traditional frequency bands: delta (0-3.5Hz), theta (4-7.5Hz), alpha (8-12.5Hz), and beta (13-30Hz), across all electrodes, for each condition. A traditional event-related (de)synchronization approach was not appropriate, as the pre-stimulus interval is likely not a passive baseline from which to conduct the analyses. Further, raw values can be utilized here, allowing for a more straightforward comparison across conditions.

3.0 Results

3.1 Behavioral Accuracy. One-way analysis of variance (ANOVA) on all participants' accuracy scores assessed condition effects. The main effect of channel number (1-, 6- 16) was significant [$F(2,69)=120.32$, $p=.00$]. Pairwise comparisons reveal differences between all levels of intelligibility ($p < .05$). Descriptive statistics are reported in Table 1. One-way ANOVA, with Bonferroni adjustment, evaluated group performance across three conditions. Main effect of listener-type was significant [$F(5,66)=61.53$, $p = .00$], driven by significant differences in performance in the 6- and 16-channel accuracy scores ($p < .05$). Accuracy of decision making of HP for 6- and 16- channel speech differed from 50% chance performance; for LP listeners, only 16-channel performance differed from chance. As expected, test of the 1-channel (unintelligible) revealed no performance difference ($p = .15$).

3.2 EEG Data

3.2.1 Event-related potentials. Transient activity was examined on grand average files for each condition. Two event-related potentials were identified, with varying latencies and amplitudes, hereafter referred to as the "late ERP" (L-ERP) and "early ERP" (E-ERP). The L-

EPR, related to the processing of degraded speech, occurred approximately 600ms following signal onset. This potential occurred earlier for more degraded speech (i.e. 1-channel occurred earlier than 6-channel) and, within that relationship, this potential was induced later and with lower amplitude in LP. This induced potential was also present for LP listeners processing intelligible16-channel speech. However, this was not noticeably present for HP listeners, for whom consistently less effort was required to understand the signal. "Noticeably present" is defined as transient activity with an appropriate signal-to-noise ratio and substantive component associated with such activity. For the L-ERP elicited by1-channel sentences, all listeners and listener sub-groups demonstrated an event-related potential with diffuse distribution, suggesting nonspecific attentional activation, with no clear focal activation. As this speech is unintelligible, diffuse activity is expected. Still, the timing of this transient activity is shorter, occurred later (585-630ms; 45ms in duration), and had lower amplitude for LP than HP listeners (540-600ms). Comparisons made among timing, location, and powers associated with each, at different levels of intelligibility, across HP and LP groups are reported in Table 2.

3.2.2 Frequency Analysis. Frequency spectra for average files were examined for each condition, separately for all listeners, and both groups. While no statistics were performed, HP v. LP showed differences in overall levels of delta and theta activity. As cortical activity is measured on the small scale of microvolts, a difference of more than 2 microvolts may be substantial. Further, differences in the ratio of theta to alpha activity appear between HP and LP across all three levels of intelligibility. For 6-channel processing intervals, the theta/alpha power ratio is 1 for HP and .73 for LP; for 16- channel processing intervals, this difference is magnified with theta/alpha power ratios of 2.8 for HP and .78 for LP. Higher levels of theta activity seen in

HP listeners drive these differences, which are likely secondary to attention to different aspects of the speech signal.

4.0 Discussion

Overall, there were differences in the characteristics of induced, event-related potentials during the processing of degraded speech, when comparing higher- and lower-performing listeners in a sentence verification task. Characteristics of the E-ERP for 16-channel speech are consistent with the P3b, more commonly known as the P300 [16]. For HP, superior parietal lobe was localized as the neural generator responsible for the activity, as determined by CDR via sLORETA. This is consistent in timing and location with the P300, documented in the literature for processing intelligible speech elicited in an oddball paradigm. However for LP this activity occurred later and was localized to the angular gyrus, an area linked with internal monologue of written words, which may include internal repetition as a strategy for resolving content [17]. This warrants further investigation with an approach yielding higher spatial specificity, particularly as activation occurs bilaterally.

The 6-channel source localization data revealed group differences in listening strategies, likely responsible performance patterns. LP data showed left transverse temporal gyrus, or Heschl's gyrus as the neural generator underlying the potential (Figure 1.a). This area of the brain is strongly associated with speech encoding, namely semantic tasks. This suggests a strategy that may work for non-degraded speech may fall short when processing degraded speech. HP data showed, bilateral activity in the inferior frontal gyrus for 6-channel speech (see Figure 1.b), suggesting involvement of cortical areas traditionally utilized in speech production. This is consistent with the dual stream model of speech processing and the involvement of

articulatory/ motor networks and attention to acoustic-phonetic properties of speech [4], and perhaps a more appropriate and successful listening strategy for degraded speech. Again, this ERP occurs later for LP listeners (680- 705ms after onset of the sentences) as compared to HP listeners (565- 595ms following onset). Finally, an L-ERP in response to 16-channel speech is present for LP but not HP. This is consistent with studies of peripheral hearing disorders that show late or small responses when speech processing is impaired, where the ability to receive or process the speech signal is degraded [see 18].

These data are in line with differential recruitment of dual pathways that are associated with behavioral performance, particularly in the 6-channel challenging condition. Here, a strong temporal activation pattern for LP is associated with attempts at semantic processing, and HP has a strong frontal-dominated activation pattern. This suggests listeners who perform well here are engaging the motor aspects of their language system. This disproportionate activation of the dorsal stream may indicate attention to the fine structure of the degraded signal, which may facilitate recovering the spoken message. Future studies will focus on relating this outcome to phonetic accuracy and error in transcription tasks made by listeners with this profile of activation and explicit analyses related to anatomical regions of interest.

The frequency components of the activity associated with processing degraded speech revealed a relationship between the ratio of theta activity to alpha activity and task performance. We saw changes in theta activity, with higher levels of theta activity present when listeners performed well with 6- and 16-channel sentences. Relationships between theta activity and degraded processing should be explored as a potential biomarker of degraded processing. Previous research shows associations between theta oscillations and processing rhythmic aspects of speech [19]. Perhaps in addition to attending to the acoustic-phonetic details, HP listeners also

focus on the amplitude envelope of the speech signal, which is preserved in noise-vocoded speech.

5.0 Conclusions

This study presented evidence for different cortical activation patterns that correspond to speech comprehension performance. The identification of a late-event related potential for low-performing listeners in the intelligible condition and all listeners when challenged with moderately intelligible speech supports the notion that this induced potential may be related to either processing degraded speech, or a degraded processing of intelligible speech. In line with known function of cortical areas, different cortical locations are noted as the neural generators involved in the presence of this activity. This suggests listeners who performed well may have used an acoustic-phonetic based strategy to help resolve the sentence through engaging articulatory/ motor networks of their language system. This study supports the notion that HP listeners present with a more complex and effective listening profile than LP listeners. Importantly, it suggests that behavioral performance is an important component to consider in neurophysiological studies of speech perception and comprehension.

5.0 Acknowledgments

This work was supported by National Institute of Health, National Institute on Deafness and Other Communicative Disorders grant 2R01 DC006859, awarded to J. M. Liss and CR5 from the Mayo Clinic foundation for education and research, awarded to Dr. John N. Caviness. A portion of this research was presented before the 163rd Meeting of the Acoustical Society in Hong Kong in May 2012.

References

[1] Choe, Y., Liss, M., Azuma, T., and Mathy, P. (2012). Evidence of cue use and performance differences in deciphering dysarthric speech. *Journal of the Acoustical Society of America, 131* (2), EL112-EL118.

[2] Borrie, S.A., McAuliffe, M.J., Liss, J.M., O'Beirne, G.A. and Anderson, T. (2012) A follow-up investigation into the mechanisms that underlie improved recognition of dysarthric speech. *Journal of the Acoustical Society of America, 132*(2), EL102-EL108. DOI: 10.1121/1.4736952.

[3] Hickok, G. and Poeppel, D. (2000). *Trends in cognitive sciences, 4,* 131-138.

[4] Hickok, G. and Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition, 92,* 67-99.

[5] Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Neuroscience, 8,* 393- 402.

[6] Saura, D., Kreher, B., Schnell, S., Kummerera, D., Kellmeyera, P., Vry, M-S., Umarova, R., Mussoa, M., Glauchea, C., Abeld, S., and Huberd, W. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Science, 105*(46), 18035-18040.

[7] Milenkovic, P.H. (2004). TF32 [Computer software]. Madison: University of Wisconsin, Department of Electrical and Computer Engineering.

[8] Audacity: Free Audio Editor and Recorder [software], version 2.0.5. See http://audacity.sourceforge.net.

[9] Boersma, P. and Weenink, D. (2008). Praat: doing phonetics by computer [software package], version 5.0.18. See http://www.praat.org/

[10] STIM2. Neuroscan, El Paso, TX.

[11] Scan v. 4.5. Compumedics Neuroscan. Charlotte, NC, USA.

[12] CURRY 7: Multimodal Imaging Suite. Compumedics Neuroscan. Charlotte, NC, USA.

[13] Murray, M.M., Brunet, D., and Michel, C.M. (2008). Topographic ERP analyses: a step-by-step tutorial review. *Brain Topography, 20*(4), 249-64.

[14] Lee, T.W., Girolami, M., and Sejnowski, T.J. (1999) Independent component analysis using an extended infomax algorithm for mixed subgaussian and supergaussian sources. *Neural computation, 11*, 417–441.

[15] Pascual-Marqui, R.D. (2002). Standardized low-resolution brain electromagnetic tomography (sLORETA): technical details. Methods and Findings in Experimental and Clinical Pharmacology, *24D*, 5–12.

[16] Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128-2148.

[17] Geschwind, N. (1972). Language and the Brain. *Scientific American, 226*(4), 76-83.

[18] Martin, B.A., Tremblay, K., and Korczak, P. (2008). Speech evoked potentials: From the laboratory to the clinic. *Ear and Hearing, 29*(3), 285- 313.

[19] Ghitza, O. (2011). Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in psychology, 2*(June), 130. doi:10.3389/fpsyg.2011.00130