Isolating Neural Indices of Continuous Speech Processing at the Phonetic Level

Giovanni M. Di Liberto and Edmund C. Lalor

Abstract The human ability to understand speech across an enormous range of listening conditions is underpinned by a hierarchical auditory processing system whose successive stages process increasingly complex attributes of the acoustic input. In order to produce a categorical perception of words and phonemes, it has been suggested that, while earlier areas of the auditory system undoubtedly respond to acoustic differences in speech tokens, later areas must exhibit consistent neural responses to those tokens. Neural indices of such hierarchical processing in the context of continuous speech have been identified using low-frequency scalp-recorded electroencephalography (EEG) data. The relationship between continuous speech and its associated neural responses has been shown to be best described when that speech is represented using both its low-level spectrotemporal information and also the categorical labelling of its phonetic features (Di Liberto et al., Curr Biol 25(19):2457–2465, 2015). While the phonetic features have been proven to carry extra-information not captured by the speech spectrotemporal representation, the causes of this EEG activity remain unclear. This study aims to demonstrate a framework for examining speech-specific processing and for disentangling high-level neural activity related to intelligibility from low-level activity in response to spectrotemporal fluctuations of speech. Preliminary results suggest that neural measure of processing at the phonetic level can be isolated.

Keywords Hierarchical · Intelligibility · EEG · Noise vocoding · Priming · Natural speech

G. M. Di Liberto (✉) · E. C. Lalor
Trinity College Institute of Neuroscience, Trinity College Dublin, 152–160
Pearse Street, Dublin, Ireland
e-mail: diliberg@tcd.ie

E. C. Lalor
e-mail: edlalor@tcd.ie

G. M. Di Liberto · E. C. Lalor
School of Engineering, Trinity College Dublin, Dublin, Ireland

Trinity Centre for Bioengineering, Trinity College Dublin, Dublin, Ireland

1 Introduction

Speech processing is an active cognitive activity underpinned by a complex hierarchical system (Chang et al. 2010; Okada et al. 2010; Peelle et al. 2010; DeWitt and Rauschecker 2012; Hickok 2012). In particular, evidence for hierarchical speech processing emerged from functional magnetic resonance imaging (fMRI) and electrocorticography (ECoG) studies. Although these methodologies have provided important scientific insights, they have many limitations. For example, fMRI does not allow for the study of the fast temporal dynamics typical of continuous speech, and ECoG studies are limited to patients suffering from severe cases of epilepsy.

Electro- and magnetoencephalography (EEG/MEG), as macroscopic non-invasive technologies, have the potential for further progress on this topic. Traditionally, the low signal-to-noise ratio (SNR) has hampered the experiments conducted with these instruments, limiting the study of speech processing to repeated presentations of simple short sounds, which elicit neural responses known as event related potentials (ERPs). Although studies based on ERPs have provided important insights on this topic, the human auditory processing mechanisms are tuned to process continuous speech and respond differently to discrete sounds (Bonte et al. 2006). Recent studies have provided encouraging evidence on the capability of EEG and MEG to track neural correlates of the low-frequency amplitude envelope of continuous speech (Aiken and Picton 2008; Lalor and Foxe 2010; Ding and Simon 2012), a finding that has proven useful for investigating the mechanisms underlying speech processing (Luo and Poeppel 2007) and its dependence on attention and multisensory integration (Power et al. 2012; Zion Golumbic et al. 2013). However, the causes of this envelope-tracking phenomenon remain unclear.

Our recent research effort has focused on the identification of neural indices of continuous speech processing at different levels of this hierarchical system. In particular, phonetic-features of speech have been shown to capture information that does not emerge when using models based on the envelope, suggesting that EEG is indexing higher-level speech processing (Di Liberto et al. 2015). Here we seek to build on this work by disentangling the higher-level phonemic processing contributions from those related to the low-level acoustic properties of the speech signal. By utilising priming, we can control how much a degraded speech stimulus is recognised without changing the stimulus itself, exploiting the fact that speech recognition is an active process (Hickok 2012). In doing so, we demonstrate a framework for examining speech-specific processing using EEG.

2   Methods

2.1   Subjects and Data Acquisition

Ten healthy subjects (5 male, aged between 21 and 31) participated in this study. Electroencephalographic (EEG) data were recorded from 128 electrode positions, digitised at 512 Hz using a BioSemi Active Two system. Monophonic audio stimuli were presented at a sampling rate of 44,100 Hz using Sennheiser HD650 head-phones and Presentation software from Neurobehavioral Systems (http://www.neu-robs.com). Testing was carried out in a dark room and subjects were instructed to maintain visual fixation on a crosshair centred on the screen, and to minimize motor activities for the duration of each trial.

2.2   Stimuli and Experimental Procedure

Audio-book versions of two classic works of fiction read by the same American English speaker were partitioned into speech snippets, each with a duration of 10s. 120 snippets were randomly selected for the experiment. In order to alter the intelligibility of the speech, a method known as noise-vocoding was implemented (Shannon et al. 1995). This method filters the speech into a number of frequency-bands, and uses the amplitude envelope of each band to modulate band-limited noise. In our experiment, frequency-bands were logarithmically spaced between 70 and 5000 Hz. To determine the number of frequency-bands to use, an offline intelligibility test was performed for each subject using a separate set of 40 speech snippets. The number of frequency-bands was chosen such that subjects could understand an average of two words in each snippet. Each EEG trial consisted of the presentation of 3 speech snippets (Fig. 1). A standard trial consisted of a first presentation (snippet a) in which the speech was degraded using noise-vocoding as previously described. Therefore it was largely unintelligible. The second snippet (snippet b) contained the same speech, but in its original clear form. The noise-vocoded version was then played again (snippet c). As such, this third condition was affected by priming because the original clear speech was played immediately before and because the speech was vocoded at the limit of intelligibility. Importantly, the speech snippets used to determine the number of vocoding frequency-bands served also to reduce the adaptation to the noise-vocoded speech during the experiment (Davis et al. 2005).

In order to measure the effect of priming, we also introduced deviant trials. These trials consisted of a modified version of (a) and/or (c), where a random chunk of ~5s was replaced with words from a different trial. In both cases, the deviant probabilities were set to 10 %. The participants were asked to identify the first and the second speech vocoded streams as a standard or a deviant presentation (after the presentations (b) and (c), respectively), using a level of confidence from 1 to 5 (from 'definitely a deviant' to 'definitely a standard').
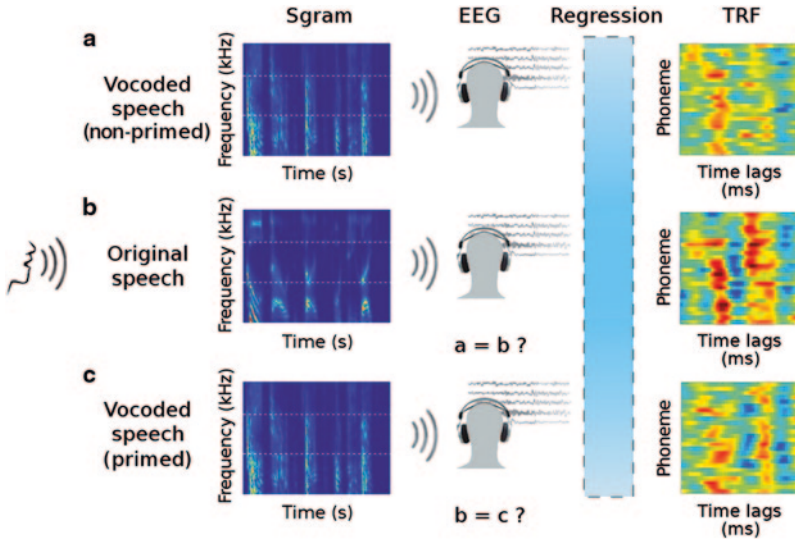
Fig. 1 EEG data were recorded while subjects listened to groups of three 10s long speech snippets. In the standard trial, the first (**a**) and third (**c**) speech streams of each group were the noise-vocoded version of the original one (**b**). The deviant presentations consisted of a modified version of (**a**) and/or (**c**). After both (**b**) and (**c**), the participants were asked to identify the first and the second speech vocoded streams as a standard or a deviant presentation. Linear regression was used to fit multivariate temporal response functions (TRFs) between the low-frequency EEG and each representation of the speech stimulus

2.3 EEG Data Analysis

The EEG data were inspected to identify channels with excessive noise or motor artefacts. Data were then digitally filtered using a Chebyshev type-2 band-pass filter with pass-band between 1 and 15 Hz and down-sampled to 128 Hz. Also, data was referenced to the average of the two mastoid channels.

Linear regression (http://sourceforge.net/projects/aespa) was used to create a mapping between the EEG and five different representations of the speech stimuli:

1. The broadband amplitude *envelope (Env)* of the speech signal, which was calculated as $Env = (x_a(t))$, $x_a(t) = x(t) + jx(t)$, where $x_a(t)$ is the complex analytic signal obtained by the sum of the original speech $x(t)$ and its Hilbert transform $x(t)$.
2. The *spectrogram (Sgram)* was obtained by partitioning the speech signal into three frequency-bands logarithmically spaced between 70 and 5000 Hz according to Greenwood's equation (70–494–1680–5000 Hz, the same used for the vocoder), and computing the amplitude envelope for each band.

3. The *phonemic (Ph)* representation was computed using forced alignment (Yuan and Liberman 2008), given a speech file and the correspondent orthographic transcription broken into 26 phonemes in the International Phonetic Alphabet (IPA). A multivariate time-series composed of 26 indicator variables was then obtained.
4. The *phonetic features (Fea)* encoding is a linear mapping of the phonetic representation into a space of 18 features (Mesgarani et al. 2014), which describe specific articulatory and acoustic properties of the speech phonetic content (Chomsky and Halle 1968). In particular, the chosen features are related to the manner and place of articulation, to the voicing of a consonant, and to the backness of a vowel.
5. Finally, we propose a model that combines *Fea* and *Sgram* (*FS*) by applying a concatenation of the two representations.

For each representation of the speech, the result is a set of regression weights referred to as multivariate temporal response functions (TRFs). *k*-fold cross-validation ($k=10$) was employed to build predictions of the EEG signal from the TRFs of each distinct speech representation model. The prediction accuracies were obtained as an average of the correlation values over a set of 12 best predictive electrodes (6 on the left side of the scalp and the symmetrical counterparts on the right), on a time-lag window from 0 to 250 ms (Lalor et al. 2006; Di Liberto et al. 2015).

The study was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

3   Results

The main idea underlying the experimental paradigm was to take advantage of the priming effect to create two conditions in which the same speech stimuli were unintelligible and intelligible (condition *a* and *c* respectively; the prime is referred to as condition *b*). Such an experiment has the potential to disentangle acoustic and phonetic models and to allow us to address the non-overlapping cortical activity encoded by their TRFs.

3.1   Behavioural Results

The behavioural results (Fig. 2a) confirm that the experiment is creating these two conditions. Specifically, standard trials are detected with a higher confidence level (paired *t*-test, $p=3.1\times10^{-5}$) while no difference emerged for deviant trials ($p=0.41$). The distribution of responses (Fig. 2b, *top*) confirms that the participants
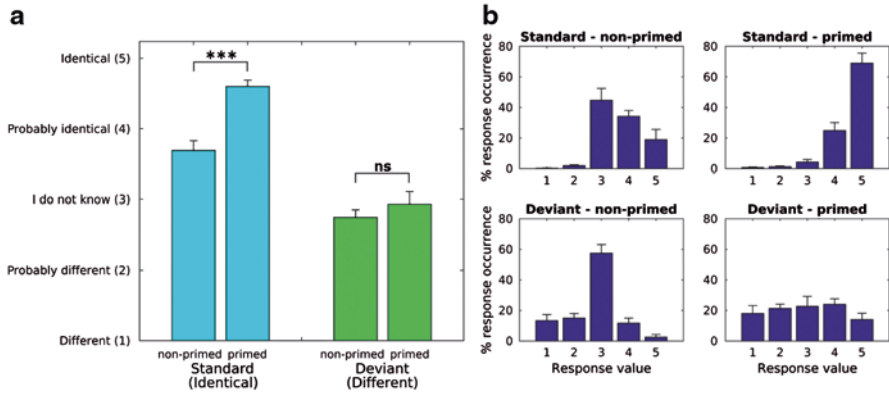
Fig. 2 Subjects were presented with sequences of vocoded-original-vocoded speech snippets and were asked to identify the two noise-vocoded streams (non-primed and primed conditions) as standard or deviant presentations by comparing them with the original speech snippet. Responses consisted of a level of confidence from 1 ('different') to 5 ('identical'). (**a**), the confidence in the identification of the standard trials was higher in the primed case for standard trials ($p<0.0005$), while no significant difference emerged for the deviants ($p>0.05$). (**b**), compares the distribution of response values across subjects in the primed and non-primed conditions, for standard and deviant trials

were more confident in the detection of standard trials; therefore they could better understand the vocoded speech in the primed condition. Interestingly, the deviant trials were not confidently detected (Fig. 2b, *bottom*), in fact they consisted of a vocoded speech snippet with a ~5s portion inconsistent with the original one, for which the prime has no effect. The hypothesis that subjects understand more the noise-vocoded speech presented after the prime is met for standard trials, therefore the analysis is conducted on them.

3.2   EEG Predictions and Model Comparison

128-channel EEG data were recorded from 10 subjects as they performed the task (Fig. 1). In this experiment, the original speech (condition *b*) was used as a prime for the second presentation of the noise-vocoded speech (condition *c*). Also, the results obtained in condition *b* provide a baseline with which to compare with previous research (Di Liberto et al. 2015). Indeed, the main interest relies on the comparison between vocoded speech in the primed (*a*) and in the non-primed (*c*) conditions. In particular, we studied the standard trials, in which the difference between the primed and non-primed conditions can originate only in the EEG signal and it is related to an improved intelligibility (Fig. 2a).

   TRF models were fit using linear regression on the standard for each speech representation (*Env*, *Sgram*, *Ph*, *Fea*, and *FS*) in the primed and non-primed conditions. These models were then compared in terms of their ability of predicting the EEG data (see *Methods*). Interestingly, the phonetic models show an improvement
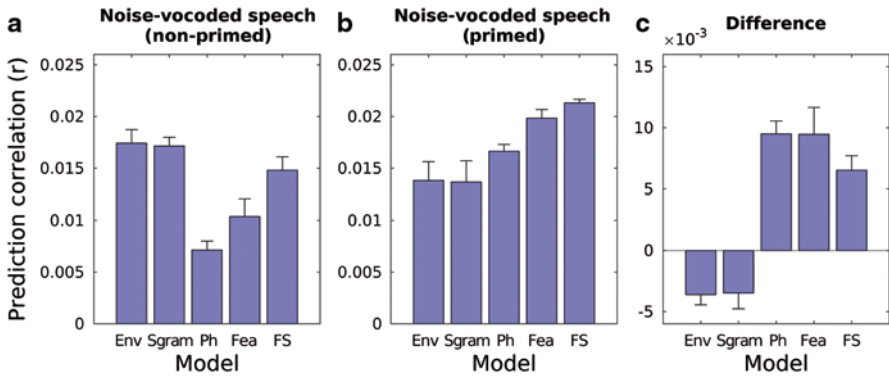
Fig. 3 The models *Env*, *Sgram*, *Ph*, *Fea*, and *FS* were compared using the Pearson correlation index between the EEG signal and its prediction as a quality indicator. The results for the non-primed and primed presentations of the vocoded speech are reported in (**a**) and (**b**) respectively, while (**c**) shows their difference (b−a). The error-bars represent the standard deviation of the Jack-knife distribution

in their EEG predictions unseen for the ones based only on acoustical information (Fig. 3). Importantly, this effect is driven by the whole distribution and not only by single subjects (Jack-knife analysis, two-way repeated measures ANOVA, Greenhouse-Geisser sphericity correction, $df = 1.8$, $F = 66.6$, $p < 0.0005$).

Further analysis can be performed on the model weights returned by the linear regression, which can be interpreted as the activation of macroscopic areas of the cerebral cortex in response to continuous speech at specific delays (time-lags). In the case of the current dataset, the three conditions can be compared in terms of response to *Env* (not shown) and to *Ph* and *Fea* (Fig. 4a and b).

4   Discussion

The ability of processing natural speech correctly is of extreme importance for humans as it represents the main method of direct communication between individuals. Although social interaction can be performed in other ways, deficits in any of its aspects can be the cause of discomfort, isolation, and depression. The awareness of this issue is essential for avoiding such consequences, however the diagnosis can be problematic, especially in the less severe cases. Certainly, methodologies for studying this articulated hierarchical system have the potential to drive the understanding of new insights into its mechanisms and to serve as an instrument of diagnosis for neural disorders related to speech. In particular, we are seeking the consolidation of a dependent measure capable of quantifying the higher-level processing of speech contribution to the scalp recorded neural activity.

EEG has been shown to be sensitive to phonetic-features of speech (Di Liberto et al. 2015). These categorical features capture information to which the envelope of speech is not sensitive, however the actual contribution of higher-level active
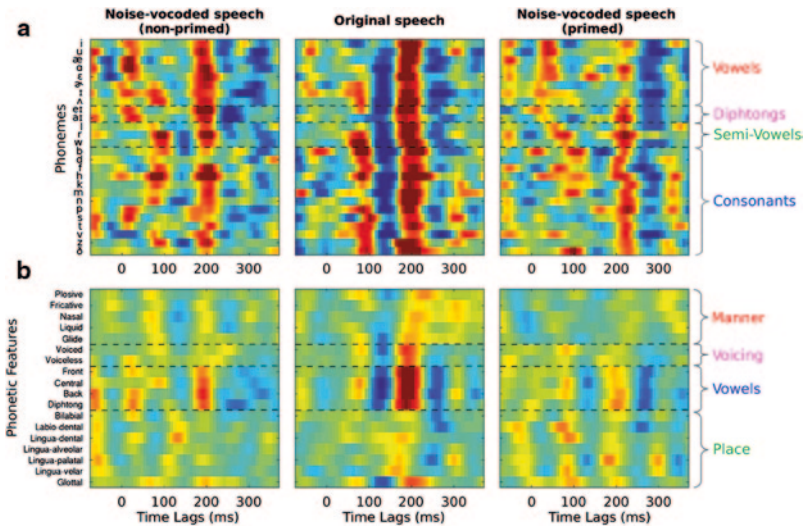
Fig. 4 Temporal response functions (TRFs) in response to *Ph* (**a**) and to *Fea* (**b**) are compared across conditions. The colour-scale relates to the neural activation in response to continuous speech for a specific phoneme or phonetic-feature. The *y-axis* labels are grouped in macroscopic categories which exhibit similar patterns of activations

processes related to intelligibility remains unclear. The framework we have defined here has the potential to disentangle these contributions from low-level EEG responses to spectrogram fluctuations. The results presented here support this hypothesis and motivate further studies.

The definition of a neural dependent measure of speech processing related to intelligibility could find application in clinical areas. For example, previous research (Pisoni 2000) suggests that the inability of deaf children with cochlear implants to discriminate fine phonetic differences in place of articulation and voicing is reflected in their perception of spoken words in terms of broad phonetic categories or functional equivalence classes. Indeed, such models are potentially useful in monitoring the development of patients in those situations and may provide new insights on the mechanisms underlying speech processing.

References

Aiken SJ, Picton TW (2008) Envelope and spectral frequency-following responses to vowel sounds. Hear Res 245(1):35–47

Bonte M, Parviainen T, Hytönen K, Salmelin R (2006) Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. Cereb Cortex 16(1):115–123

Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13(11):1428–1432. [10.1038/nn.2641]. http://www.nature.com/neuro/journal/v13/n11/abs/nn.2641.html#supplementary-information

Chomsky N, Halle M (1968) The sound pattern of English. Harper & Row, New York

Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C (2005) Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. J Exp Psychol Gen 134(2):222

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109(8):E505–514. doi:10.1073/pnas.1113427109

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low frequency cortical entrainment to speech reflects phonemic level processing. Curr Biol 25(19):2457–2465

Ding N, Simon JZ (2012) Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109(29):11854–11859. doi:10.1073/pnas.1205381109

Hickok G (2012) The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. J Commun Disord 45(6):393–402. doi:10.1016/j.jcomdis.2012.06.004

Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci 31(1):189–193. doi:10.1111/j.1460-9568.2009.07055.x

Lalor EC, Pearlmutter BA, Reilly RB, McDarby G, Foxe JJ (2006) The VESPA: a method for the rapid estimation of a visual evoked potential. Neuroimage 32(4):1549–1561

Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54(6):1001–1010

Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic feature encoding in human superior temporal gyrus. Science 343(6174):1006–1010

Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cereb Cortex 20(10):2486–2495. doi:10.1093/cercor/bhp318

Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical processing for speech in human auditory cortex and beyond. Front Hum Neurosci 4:51. doi:10.3389/fnhum.2010.00051

Pisoni DB (2000) Cognitive factors and cochlear implants: some thoughts on perception, learning, and memory in speech perception. Ear Hear 21(1):70–78

Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. Eur J Neurosci 35(9):1497–1503. doi:10.1111/j.1460-9568.2012.08060.x

Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270(5234):303–304

Yuan J, Liberman M (2008) Speaker identification on the SCOTUS corpus. J Acoustical Soc Am 123(5):3878

Zion Golumbic EM, Cogan GB, Schroeder CE, Poeppel D (2013) Visual input enhances selective speech envelope tracking in auditory cortex at a "cocktail party". J Neurosci 33(4):1417–1426