INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Neural responses to speech-specific modulations derived from a spectro-temporal filter bank

Marina Frye[1,3], *Cristiano Micheli*[2], *Inga M. Schepers*[2,3]
Gerwin Schalk[4,5], *Jochem W. Rieger*[2,3], *Bernd T. Meyer*[6]

[1]Medizinische Physik, Carl von Ossietzky Universität, Oldenburg, Germany
[2]Applied Neurocognitive Psychology, Carl von Ossietzky Universität, Oldenburg, Germany
[3]Cluster of Excellence Hearing4all, Oldenburg, Germany
[4]National Center for Adaptive Neurotechnologies, Wadsworth Center, Albany, NY, USA
[5]Albany Medical College, Albany, NY, USA
[6]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, USA
{marina.frye, cristiano.micheli, inga.maren.schepers,
jochem.rieger}@uni-oldenburg.de, bernd.t.meyer@jhu.edu

Abstract

This paper analyzes the application of methods developed in automatic speech recognition (ASR) to better understand neural activity measured with electrocorticography (ECoG) during the presentation of speech. ECoG data is collected from temporal cortex in two subjects listening to a matrix sentence test. We investigate the relation of ECoG signals and acoustic speech that has been processed with spectro-temporal filters, which have been shown to produce robust and reliable representations for speech applications. The organization of spectro-temporal filters into a filter bank allows for a straight-forward separation into spectral or temporal only, as well as true spectro-temporal components. We find electrodes positioned over the superior temporal gyrus that is associated with the auditory cortex to show significant specific high gamma activity to fine temporal and spectro-temporal patterns present in speech. This indicates that representations developed in machine listening are a suitable tool for the analysis of biosignals.

Index Terms: speech perception, ECoG, automatic speech recognition, robust feature extraction

1. Introduction

Findings about the auditory system have influenced research in automatic speech recognition (ASR), which often resulted in more robust machine listening [1, 2]. Although a closer connection between ASR and human speech recognition (HSR) has been promoted earlier to further our understanding of speech processing in humans and machines [3], there is a comparatively small number of studies that bring back ASR technology that profited from auditory insights to better understand HSR; important exceptions are for instance [4] and [5].

In a study presenting physiological data from the primary auditory cortex (A1) in mammals, the use of 2D Gabor functions was proposed to model time-frequency representations that elicit high firing rates in neurons (i.e., spectro-temporal receptive fields, STRFs) [6]. This motivated Kleinschmidt and colleagues to apply 2D Gabor functions in feature extraction in ASR [7]. When organized in a filter bank that evenly covers spectral and temporal modulation frequencies (as proposed in [8, 9]), conventional baselines were outperformed in speech-related tasks such as ASR based on deep learning [10], voice-activity detection [11], and speaker identification [12]. We assume these filters (shown in Fig. 1) to be speech-specific due to their success in the previously mentioned applications, and investigate if the auditory-inspired representations can be useful for explaining neural activity obtained from electrocorticography (ECoG).

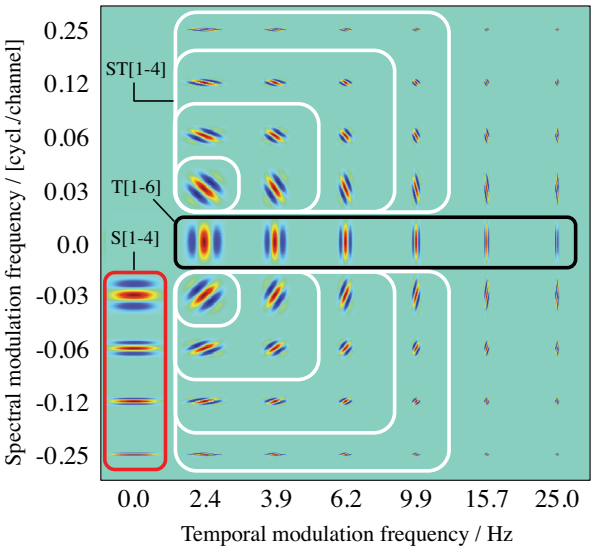

Figure 1: Filter bank of 2D spectro-temporal Gabor filters applied to the speech signals. The filter output is compared to the high gamma band activity of ECoG data and analyzed w.r.t. spectral, temporal, and spectro-temporal filter outputs.

Cortical responses captured by ECoG measurements have recurrently been investigated and were reported to increase in power with the energy of the auditory stimulus during speech perception (e.g. [13, 14, 15]). Further, temporal patterns of auditory stimuli seem to be closely tracked by the energy in the high gamma frequency band of electrical cortical data [16, 17, 18]. Additionally, spectral modulations, which define

http://dx.doi.org/10.21437/Interspeech.2016-1327

speech-specific characteristics such as formant frequencies and the combination of both, spectro-temporal modulations, are represented by neural responses in different cortical regions [19]. Former ECoG studies with high-density electrode grids established a link between fundamental acoustic properties (such as short-term spectra) and neural activity [20, 21] while others used ASR-based methods to decode categorical units of speech (such as phonemes or words) from brain data [22]. In this study, we investigate if a relation can be established between ECoG data obtained with more commonly applied low-density grids of 1 cm inter-electrode-spacing and speech-specific representations that have a direct acoustic-physical link (in terms of spectral and temporal modulations frequencies) to the original time signal. In the next section, we describe how ECoG measurements in two patients were performed, how spectro-temporal Gabor features are calculated and grouped, and how the relation of data was analyzed. In Sections 3 and 4 the results for different filters groups as well as individual filters are presented and discussed; Section 5 concludes the paper.

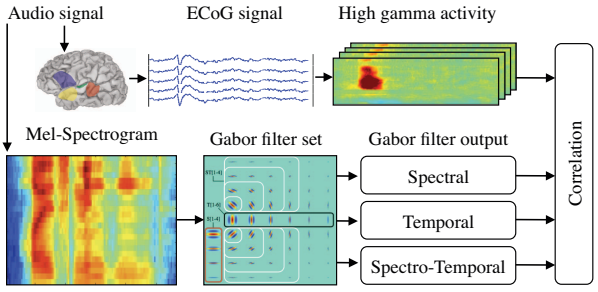

Figure 2: Illustration of the analysis framework. For a larger view on filter selection, see Fig. 1.

2. Methods

2.1. Subjects, speech material and recordings

ECoG data was collected from two subjects who participated in an audiovisual (AV) recognition task in the Schalk Lab at Albany Medical Center (Albany, New York, USA). The experiment was approved by the Institutional Review Boards of the Albany Medical Center.

The AV material consists of English matrix sentences recorded specifically for a set of ECoG experiments. The sentences in the matrix test follow the pattern (subject)(verb)(numeral)(adjective)(object). For each of these five groups, ten word alternatives exist which enables the generation of 10^5 different sentences (e.g., *Peter buys eight wet stones*), which are syntactically correct but semantically unpredictable. The recorded words were the same as published in [25], with an additional catchword from any of the mentioned categories (see above). AV recordings of one female speaker were conducted with a high-quality microphone (Røde M2) and a camcorder with the speaker's face centered in the frame.

Free-field speakers, a desktop PC with a monitor, and a mouse were used for stimulus presentation and response logging in the patients' rooms. The experiment was divided into three blocks of 70 sentences each. Each sentence was followed by a target word from the complete word inventory of the matrix test. Subjects were asked to decide whether the target word had been presented in the preceding sentence by pressing a yes or no button to ensure the subjects' attention. Presentation was either audio-visual or audio-only (A); in the latter case, a still picture of the speaker's face was shown instead of the video.

Both subjects were implanted with subdural multi-electrode grids (Ad-TechMedicalCorp., Racine, WI) due to epileptic treatment and gave informed consent to participate in the study. One patient was implanted on the left (79 electrodes) and the other on the right (91 electrodes) hemisphere with electrode grids that consisted of silicon slips with platinum-iridium electrodes (4 mm in diameter, 2.3 mm exposed) embedded at an inter-electrode distance of 1 cm. Both subjects had normal cognitive capacity and were right-handed. ECoG signals were amplified by a g.HIamp system (g.Tec, Graz, Austria), online band-pass filtered between 0.3 and 500 Hz, digitized at 1000 Hz and stored with the general-purpose software BCI2000 [23].

Noisy channels as well as those with artifacts arising from epilepsy were identified by visual inspection of both the raw potential and its frequency decomposition and excluded from further analysis. Remaining channels were re-referenced with a common average and band-stop filtered to eliminate line noise and its harmonics. Segments during which no experimental material was presented were discarded to minimize computational costs. High gamma band power [70-110 Hz] was calculated from a power-spectrogram estimated with the FieldTrip toolbox (see [24]) using multiple Slepian tapers. The resulting high gamma activity during speech perception was z-scored by mean subtraction and variance removal.

A preliminary analysis based on the correlation of ECoG high gamma activity and the short-term energy of the acoustic waveform resulted in identical correlation results for Conditions A and AV up to the second decimal place ($r_A = r_{AV} = 0.53$). With this result and with the scope of this study in mind, ECoG data of both conditions were grouped in the further data analysis.

2.2. Speech-specific spectro-temporal filters

Gabor features of the speech material are calculated by processing mel-spectrograms of the input signal by a number of 2D modulation filters. Filtering is performed by calculating the 2D convolution of the filter and the spectrogram. The result of the time-aligned convolution for all filters is used as feature vector. Gabor filters are defined as the product of a complex sinusoidal function $s(n,k)$ (with n and k denoting the time and frequency index, respectively) and an envelope function $h(n,k)$. In this notation, the complex sinusoid is defined as

$$s(n,k) = \exp\left[i\omega_n(n-n_0) + i\omega_k(k-k_0)\right].$$

and the Hann function that we chose as envelope (with the parameters W_n and W_k for the window length) is given by

$$h(n,k) = \left(\frac{1}{2} - \frac{1}{2}\cdot\cos\left(\frac{2\pi(n-n_0)}{W_n}\right)\right) \cdot \left(\frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi(k-k_0)}{W_k}\right)\right).$$

The periodicity of the carrier function is defined by the radian frequencies ω_k and ω_n, which allow the Gabor function to be tuned to particular directions of spectro-temporal modulation, including diagonal modulations. For this study, an arrangement in a filter bank [8, 9] was chosen due to the good results that

were obtained in various speech tasks with this specific implementation [10, 11, 12]. Note that the filter bank covers not only true spectro-temporal (diagonal) filters, but also purely spectral and temporal patterns; these were designed to be orthogonal to each other to comply with the independence requirement of many speech classifiers. To dissect the influence of each filter type, the filters are segmented as shown in Fig. 1. This segmentation resulted in four spectral filters (S1-S4), six temporal filters (T1-T6) and four spectro-temporal filter *groups* (ST1-ST4) that are obtained by starting with the lowest ST modulation frequency and choosing the adjacent filters for the next group.

2.3. Data analysis

To analyze the relation between the filter bank output and the corresponding ECoG data, mel spectrograms of the speech input were processed with all filters as it is standard procedure in speech applications (see, i.e., [9]). The average absolute amplitude was determined for each filter individually, and a filter-specific event was triggered when the filter output value exceeded 80% average activity. A segment of 40 ms duration was then selected from the high gamma ECoG signal, starting from the time index of the trigger plus a time delay between 10 and 200 ms to explore the temporal relation between stimulus and ECoG response. The calculation of a mean of the segment results in one scalar from the ECoG data per activation-triggered event. The average and standard error for these samples of filter-specific activations are presented in the next section. In the plots the mean (red bar), standard error (box) and percentile (whiskers) of the data samples are shown. Significance markers were determined by Bonferroni-corrected two sample Student's t-tests and their level indicated by stars.

3. Results

3.1. Sensitivity to spectral and temporal patterns in the temporal lobe

The segmented neural responses evoked by the filter sets shown in Fig. 1 were examined in electrodes with the strongest auditory responses during a preliminary analysis of across-sentence averaged high gamma band power. These electrodes are located over the temporal lobe. Only one electrode of all chosen electrodes shows few sporadic incidences of low-level significant differences between different *spectral* filters. This indicates a lack of systematic effects for spectral filters in electrodes covered by the grid; this data is hence not shown.

On the other hand, *temporal* modulations represented in the filter bank were found to be encoded in the ECoG data: As Figure 3 depicts, several electrodes show multiple statistically significant differences in neural response. In Patient A, the neighboring Electrodes 40 and 45 ('E40' and 'E45'), which are associated with the belt regions of the auditory cortex, show especially characteristic responses identified by the multiple highly significant Student's t-tests. Other electrodes respond characteristically to one filter group in particular, including one electrode (E76, Patient A) positioned in the anterior STG (see inset in Fig. 3(a)). Interestingly, this electrode responds strongest to higher temporal modulations in the speech signal. In contrast, other electrodes positioned in the middle STG (e.g. E40 in Patient A, E48 in Patient B), are more responsive to lower modulation frequencies. These are generally ascribed to the word and syllable rates of continuous speech and are very important for both human and automatic speech recognition [26].

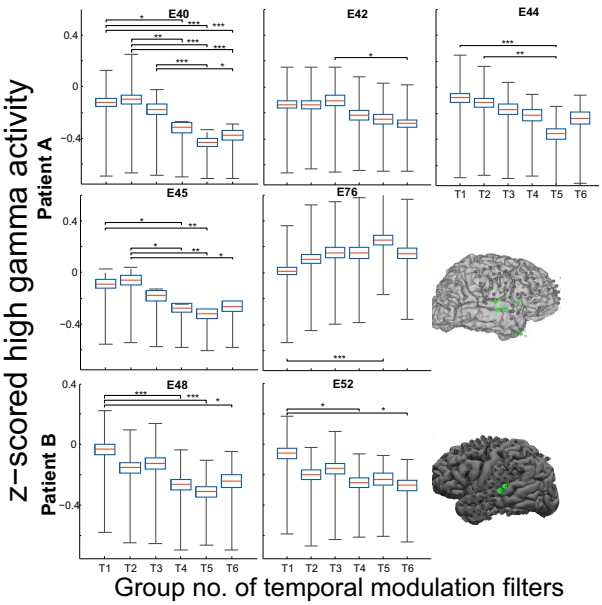

Figure 3: Z-scored activations for electrodes over the temporal lobe of both patients show specificity for temporal modulations of the speech signal. Inset figures show locations of electrodes with significant differences.

3.2. Sensitivity to spectro-temporal patterns in the temporal lobe

The predominant neighboring Electrodes E40 and E45 of Patient A also show characteristic activations to true spectro-temporal filters grouped according to Fig. 1. Figure 4 depicts the sets of responses to the four different groups in E40 and E44 in Patient A. The effects in E40 can be observed for long intervals after the onsets of the filters in this electrode and also in E41 for an interval of 140-180 ms post-stimulus; both electrodes are located at the posterior STG. It is noteworthy that E41 did not show significant specificity towards purely temporal modulations, thereby excluding the possibility of temporal modulations being captured by the ST filter groups. This is in line with earlier findings that showed individual and distinct trends for each filter, which can be explained by the choice of filters to minimize covariance between these. Significant differences of all other groups to the Group ST4, with strong group-specific responses in almost all time frames post-stimulus are also observed in E44. These differences can also be observed in the interval of 170-200 ms after the stimulus in E45. Electrode 44 thus shows an increased specificity towards spectro-temporal modulations in comparison to purely temporal modulations, whereas E45 is found to respond less specifically (cf. Fig. 3).

In Patient B, specific responses are mainly focused in one electrode (E52) located over the medial STG with similar response patterns, allowing for a closer inspection of the temporal evolution of the neural response (see Figure 5). E48 shows similar but attenuated specificity also in later time frames of the observed window. In comparison to the encoding times generally assumed in for auditory processing of around 100 ms for

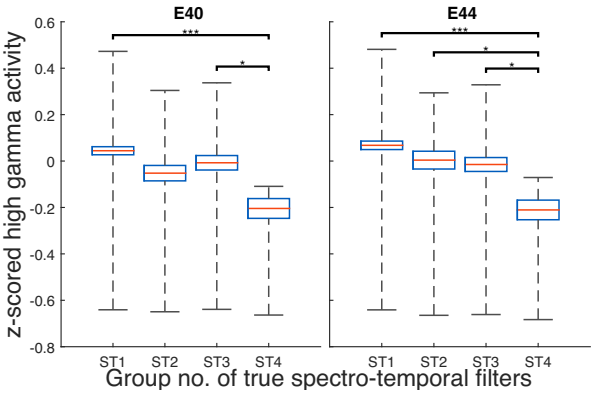

Figure 4: Average high gamma activity response to spectro-temporal filters from the Gabor filter bank in Patient A. Significant differences were found in electrode 40 (E40) and E44. A time delay between stimulus and response of 150 ms was chosen for this figure. However, stable significant selectivity is observed in a larger time window as shown in the next figure.

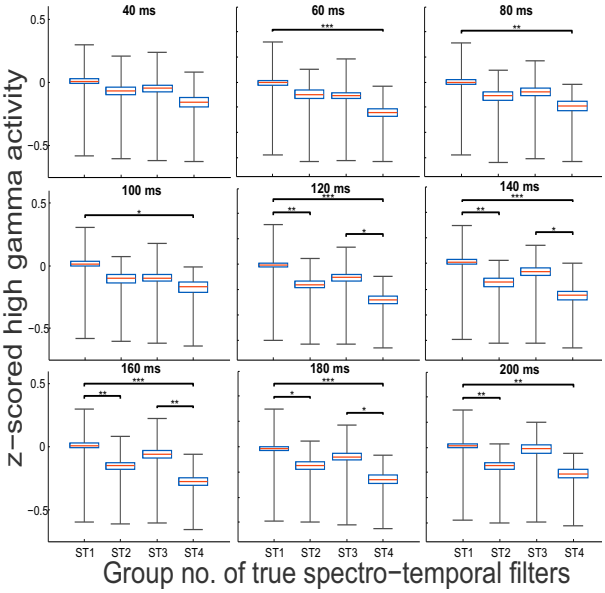

Figure 5: Temporal evolution of neural response in E52 of Patient B to grouped true spectro-temporal filters show characteristic activity especially for the interval of 120 to 180 ms.

cortical areas these evolve far later in the time window. A possible explanation that has to be confirmed in future research is a decelerated encoding of ST features.

4. Discussion

The placement of electrode grids is determined by the medical treatment, hence it is not guaranteed that cortical areas associated with speech perception (see [27, 17, 18, 20, 21]) are covered in ECoG experiments. Since significant differences in encoding of spectral modulations was shown earlier in the literature, we assume that a different coverage might have resulted in differences between spectral filters (which was not found for the current electrode placement). However, significant differences in z-scored high-frequency band power were found for different modulations, indicating that the corresponding electrodes cover cortical tissue that encodes the specific characteristic of the speech signal.
However, temporal modulations which are important for speech perception were especially well-represented and showed highly discriminable responses. Further, several electrodes were selective to groups of spectro-temporal filters, especially for Groups ST1 and ST4. ST1 covers relatively slow changes of the time-frequency representation which relates to the coarse structure of speech and also to vowel transients. Group ST4 covers higher modulation frequencies and hence produces high activations for fine-grained spectro-temporal detail, such as modulations of the fundamental frequency.

In similar research, localized areas tuned to specific modulations in the spectrograms of speech signals as well as onset-sensitive areas in the posterior STG were found using spectro-temporal receptive fields (STRF) [19]. The authors reported an encoding of temporal modulations in the auditory cortex. In this study, electrodes found to be specific to temporal modulations are correspondingly placed over the middle and posterior part of the STG, but for Patient A we also found feature-specific activations the anterior STG. In contrast to [19], electrodes specific to spectro-temporal patterns are also found to be located in the middle and the posterior part of the STG close to the belt areas. In future research, the differences between the STRF-based approach and regularized Gabor filterbank approach will be explored to explain if these differences arise from differences in filter properties, electrode placement, or other factors.

5. Summary and conclusion

ECoG data of two patients implanted with low-density electrode grids was analyzed with respect to the onsets of speech features borrowed from automatic speech recognition (ASR): Speech was converted to spectro-temporal features extracted with a Gabor filter bank, which represent physical-acoustic properties of speech, i.e., spectral, temporal, and spectro-temporal modulations. We analyzed the relation of filter groups to ECoG data by using a high filter activation as a trigger and compared the corresponding high gamma ECoG data of these segments. While our data showed no specific selectivity for spectral modulations, we found significant differences between purely temporal and spectro-temporal filter categories in the middle and posterior regions of the superior temporal gyrus. Conclusively it can be said that ASR-motivated features selected for this analysis allow for an assessment of neural data derived from low-density electrode grids. On this basis, further experiments with other features motivated by machine learning seem feasible and could further the understanding of encoding mechanisms in the human cortex.

6. Acknowledgements

This work was funded by the DFG (SFB/TRR 31 “The Active Auditory System”) and by Google via a Google faculty award to Hynek Hermansky (CLSP at Johns Hopkins University). The authors thank Susann Bräuer for recording the audio-visual stimuli used in this experiment.

7. References

[1] Hermansky, H. (1998). "Should recognizers have ears?," Speech Commun., 25, 3–24.

[2] Stern, R. M., and Morgan, N. (2012). "Hearing is believing: Biologically inspired methods for robust automatic speech recognition," IEEE Signal Process. Mag., 29, 34–43.

[3] Scharenborg, O. (2007). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun., 49, 336–347.

[4] Cooke, M. (2006). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am., 119, 1562–1573.

[5] Scharenborg, O., ten Bosch, L., Boves, L., and Norris, D. (2003). "Bridging automatic speech recognition and psycholinguistics: Extending Shortlist to an end-to-end model of human speech recognition," J. Acoust. Soc. Am., 114, 3032.

[6] Qiu, A., Schreiner, C., and Escabi, M. (2003). "Gabor analysis of auditory mid- brain receptive fields: spectro-temporal and binaural composition," Journal of Neurophysiology, 90, pp. 456-476.

[7] Kleinschmidt, M. and Gelbart, D. (2002). "Improving word accuracy with Gabor feature extraction," in Proc. Interspeech, pp. 25-28.

[8] Schädler, M.R., Meyer, B.T., Kollmeier. B. (2012). "Spectro-temporal modulation subspace-spanning filter bank features for robust automatic speech recognition," J. Acoust. Soc. Am. Volume 131, Issue 5, pp. 4134-4151.

[9] Meyer, B. T., Ravuri, S. R., Schädler, M. R., Morgan, N. (2011). "Comparing different flavors of spectro-temporal features for ASR," in Proc. Interspeech, pp. 1269-1272.

[10] Castro Martinez, A.M., Moritz, N., Meyer, B.T. (2014). "Should deep neural nets have ears? The role of auditory features in deep learning approaches," in Proc. Interspeech, pp. 2435-2439.

[11] Tsai, T. J., and Morgan, N. (2012). "Longer Features:They do a speech detector good," Proc. Interspeech 2012, Portland, OR, USA.

[12] Lei, H., Meyer, B., Mirghafori, N. (2012). "Spectro-temporal Gabor features for speaker recognition," in Proc. ICASSP.

[13] Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., Schalk, G. (2013). "The Tracking of Speech Envelope in the Human Cortex," PLoS ONE, 8, pp. 53398

[14] Steinschneider, M., Nourski, K.V., Fischman, Y.I. (2013). "Representation of Speech in Human Auditory Cortex: Is it Special?," Hearing Research, 305, pp. 57-73.

[15] Nourksi, K.V., Reale, R.A., Oya, H., Kawasaki, H., Kovach, C.K., Chen, H., Howard, M.A., Brugge, J.F. (2009) "Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex," Journal of Neuroscience, 49, pp. 15564-15574.

[16] Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., Merzenich, M.M. (2001). "Speech comprehension is correlated with temporal response patterns recorded from auditory cortex," in Proc. National Academy of Sciences, 98, pp.13367-13372.

[17] Crone, N.E., Boatman, D., Gordon, B., Hao, L. (2001). "Induced electrocorticographic gamma activity during auditory perception," Clinical Neurophysiology, 112, pp.565-582.

[18] Canolty, R.T. (2007). "Spatiotemporal dynamics of word processing in the human brain," Frontiers in Neuroscience, 1, pp.185-196.

[19] Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., Chang, E.F. (2016). "Human Superior Temporal Gyrus organization of spectrotemporal modulation tuning derived from speech stimuli," Journal of Neuroscience, 36, pp.2014-2026.

[20] Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., Chang, E.F. (2012). "Reconstructing speech from human auditory cortex," PLoS Biology, 10.

[21] Mesgarani, N., Chang, E.F. (2012). "Selective cortical representation of attended speaker in multi-talker speech perception," Nature, 485, pp.233-236.

[22] Herff, C., Heger, D., de Pesters, A., Telaar, D., Brunner, P., Schalk, G., Schultz, Tanja. (2015). "Brain-to-Text: decoding spoken phrases from phone representations in the brain," Frontiers in Neuroscience, 9.

[23] Schalk, G., McFarland, D.J., Hinterberger, T., Birbaumer, N., Wolpaw, J.R. (2004). "BCI2000: a general-purpose brain-computer interface BCI system," in IEEE Transactions on Biomedical Engineering, 51, pp.1034-1043.

[24] Oostenveld,R., Fries, P., Maris, E., Schoffelen, J.-M. (2011). "FieldTrip: Open Source Software for advanced analysis of MEG,EEG and invasive electrophysiological data," Computational Intelligence and Neuroscience, 2011, pp.1-9

[25] Zokoll M., Hochmuth S., Warzybok A., Wagener K.C., Buschermöhle M., Kollmeier, B. (2013). "Speech-in-noise tests for multilingual hearing screening and diagnostics," American Journal of Audiology, 22, pp. 175 – 178.

[26] Kanedera, N., Arai, T., Hermansky, H., and Pavel, M. (1997). "On the importance of various modulation frequencies for speech recognition," Proc. Eurospeech, pp. 1079-1082.

[27] Hickok, G., Poeppel, D. (2007). "The cortical organization of speech processing," Nature Reviews Neuroscience, 8, pp.393-402