Phoneme-level processing in low-frequency cortical responses to speech explained by acoustic features

Daube, Christoph[a] (c.daube.1@research.gla.ac.uk)
Ince, Robin A. A.[a] (robin.ince@glasgow.ac.uk)
[a]Institute of Neuroscience and Psychology, University of Glasgow
62 Hillhead Street, Glasgow G12 8QB, UK
Gross, Joachim[a,b] (joachim.gross@wwu.de)
[b]Institut für Biomagnetismus und Biosignalanalyse, Westfälische Wilhelms-Universität,
Malmedyweg 15, 48149 Münster, Germany

Linear encoding models constructed to explain human cortical responses to speech have the potential to provide insights into the mechanisms of speech comprehension. It has been shown that combining annotated linguistic features with acoustic features of the speech signal can consistently improve the prediction of brain responses. Here we aim to replicate these effects in source-level magnetoencephalography (MEG) data to ask if the contribution made by linguistic features could be explained by more comprehensive models considering acoustic features only. We thus compare the predictive performance of several acoustic feature spaces of varying dimensionality with that of an annotated linguistic feature space. While we replicate the effect of increased performance when combining annotated features with spectrograms over spectrograms alone, we also obtain similar increases with Gabor-filtered spectrograms and even stronger increases with the combination of spectrograms and their temporal gradients. We then find that the predictions of this best acoustic model are highly redundant with those of the annotated feature space. We conclude that annotated feature spaces are a great as benchmarks. However, we stress that for the understanding of the computations underlying cortical responses to speech, models specifying transformations of the acoustic input are necessary.

Keywords: Encoding models; speech; mid-level; phonemes; MEG

Introduction

Explaining dynamic neuronal responses to speech via linear models considering various nonlinear transformations of the stimuli as features has recently led to fascinating hypotheses about the involved brain regions, temporal dynamics and computations (Di Liberto, O'Sullivan & Lalor, 2015; de Heer, Huth, Griffiths, Gallant & Theunissen, 2017). The underlying rationale is that insights about the locus and dynamics of the biological implementation of the computations involved can be gained from spatiotemporal maps of weights and predictive power, while the architectures of the nonlinear transformations are thought to shed light on the computations themselves. The employed feature spaces are then chosen to cover a hypothesized hierarchy, allowing dissociable descriptions of early, acoustically driven and later, more abstract processing stages.

A compelling finding obtained with this strategy is that cortical responses as measured by electroencephalography (EEG, Di Liberto, O'Sullivan & Lalor, 2015) or certain regions as measured with functional magnetic resonance imaging (fMRI, de Heer et al, 2017) are best predicted with so-called articulatory or phonemic feature spaces. They are linguistically motivated and require a labelling and temporal alignment of a separate transcript of the text and the acoustical stimulus waveform. In this way, they aim to capture an intermediate stage of processing between low-level acoustic stimulus properties and high-level semantic meaning. This opens up the avenue of assessing if there are feature spaces based on acoustic models which can explain this advantage in prediction performance. Doing so could further elucidate to what degree the phenomenon is a signature of an invariant processing stage related to acquired, internal knowledge of a language or more simply a reflection of bottom up encoding of physical properties of the stimulus.

Using source-level MEG data, we here consider a range of pre-existing acoustical feature spaces and evaluate their predictive performances relative to the current state of the art in the literature: a model based on both spectrograms and articulatory features (Di Liberto, O'Sullivan & Lalor, 2015). To then elucidate what acoustical properties could underlie the performance boost of the annotated features, we assess to what degree the predictions of acoustical models share information about observed recordings with predictions of the annotated model.

Methods

21 healthy young participants (native speakers of English, 12 female, mean age 24.05 years, age range 18 – 35 years) listened to a narrative of 55 minutes duration ("The Curious Case of Benjamin Button", public domain recording by Don W. Jenkins, librivox.org) while their brain activity was recorded with a 248 channel magnetometer MEG system (MAGNES 3600 WH, 4D Neuroimaging). The session was split into 6 blocks of equal duration and additionally included the repetition of the last block. Preprocessing was done using the fieldTrip toolbox (Oostenveld, Fries, Maris & Schoffelen, 2011). We manually removed and subsequently interpolated artefactual channels, replaced squid jumps with DC patches, filtered the signal with a fourth-order zero-phase butterworth high-pass filter with a cutoff-frequency of .5 Hz and downsampled the data to 125 Hz. We then performed ICA to identify and remove components reflecting eye and heart activity and further downsampled the data to 40 Hz. Next, we generated corrected-sphere volume conductor models from individual anatomical MRI scans and computed spatial filters for a grid of points in source space of 5 mm resolution using the LCMV beamformer algorithm with 5% regularisation. We correlated the response to the last block with that from its repetition to identify story-responsive regions in source space (de Heer et al, 2017).

The speech stimulus was then transformed into various feature spaces. We used the GBFB toolbox (Schädler, Meyer & Kollmeier, 2012) to obtain 31-channel Log-Mel-Spectrograms ("LMS") and summed these across the spectral dimension to also obtain the amplitude envelope ("Env"). Additionally, we filtered the spectrograms with 455 2D Gabor filters of varying centre frequencies as well as spectral and temporal modulation frequencies ("GBFB"). As a last acoustic feature space, we computed half-wave rectified first derivatives of the envelope and spectrogram feature spaces ("ReDe(Env)" and "ReDe(LMS)", Hertich, Dietrich, Trouvain, Moos & Ackermann, 2012). To construct annotated feature spaces, we used the Penn Phonetics Lab Forced Aligner (Yuan & Liberman, 2008) to align the text material to the stimulus waveforms, providing us with onset times of phonemes comprising the text. These were manually corrected using Praat (Boersma, 2001) and subsequently transformed into a 23-dimensional binary articulatory feature space ("Art", de Heer et al, 2017). Finally, we discarded the information about phoneme identity to obtain a one-dimensional binary feature space of phoneme onsets ("Art1D").

We used these seven feature spaces in the following combinations: Env, ReDe(Env), LMS, LMS+ReDe(LMS), LMS+GBFB, LMS+Art1D, LMS+Art.

To perform a linear mapping from our feature spaces to the recorded MEG signals, we used ridge regression (Crosse, Di Liberto, Bednar & Lalor, 2016) in a 5-fold nested cross-validation framework (Varoquaux, Raamana, Engeman, Hoyos-Idrobo, Schwartz, & Thirion, 2017). This allowed us to tune hyperparameters controlling the temporal extent and the amount of L2 regularisation of the ridge models in the inner folds, yielding optimised models for each considered grid point. For the joint feature spaces consisting of multiple subspaces, the temporal extent and L2 regularisation was optimised individually for each subspace to obtain the best possible prediction performance. Bayesian Adaptive Direct Search (Acerbi & Ma, 2017), a recent blackbox optimisation algorithm, was used to efficiently sample the multidimensional hyperparameter space.

To evaluate the model performances for each outer fold of the 21 participants and focus on the differences between the feature spaces, we used a Bayesian hierarchical linear model with random effects for participants and fixed effects for folds, hemispheres and feature spaces (Bürkner, 2017), allowing us to assess posterior distributions of the beta estimates of the categorical variable feature space.

Finally, to assess the amount of redundancy and unique information each acoustical feature space had with the annotated feature space about the observed MEG time series, we used Partial Information Decomposition (PID) based on common change in surprisal (Ince, 2017) within each fold of each participant.

Results

Correlations of repeated chapters peak in bilateral auditory cortices

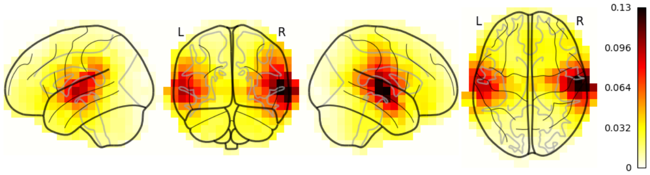

Figure 1: Grand average of correlations of brain activity of the last block and its repetition.

To identify grid points where MEG responses were reproducibly activated by the stimulus, we correlated the responses to one chapter with the responses to its repetition in source space (de Heer et al, 2017). These correlations peaked in regions well in accordance with typical localisations of bilateral auditory cortices (AC, Figure 1). We focussed the following analyses on the grid point of peak correlation within each hemisphere.

Performance boost over spectrograms with annotated features, but also with acoustic features

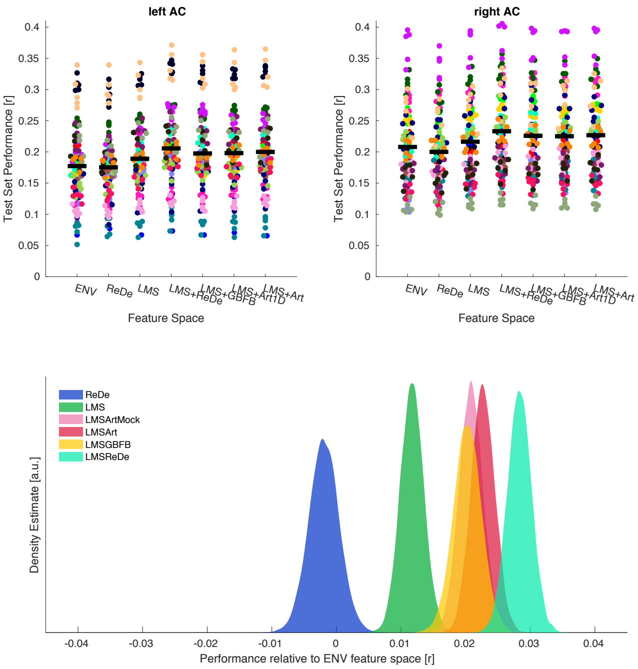

Figure 2: *Top*: Raw test set performances in left and right AC. Each colour is one participant, each dot is one outer fold. *Bottom*: Draws from posterior distribution of beta estimates for fixed effects of feature spaces relative to baseline feature space envelope.

The performances of our models exhibited relatively large inter-participant variability and comparatively small variability across feature spaces (Figure 2, top). To focus on our main interest of systematic differences across feature spaces, we used Bayesian hierarchical linear modelling and extracted the betas of fixed effects for the feature spaces (Figure 2, bottom).

We replicated previous results demonstrating an increase in prediction performance when combining linguistically motivated annotated articulatory features with spectrograms (red). When we discarded the phoneme identity and thereby reduced the annotated features to phoneme onsets (pink), we obtained comparable performances. We also obtained comparable performances with a combination of spectrograms and Gabor-filtered spectrograms (yellow). Finally, we achieved the best prediction performances when we instead combined the spectrograms with their temporal gradients (turquoise). These results raised the question whether the acoustic features explain the same or different aspects of the responses as the annotated features.

Performance boosts of annotated features are highly redundant with those of a combination of spectrograms and their temporal gradient

PID (Ince, 2017) aims to disentangle redundant, unique and synergistic contributions of two source- about a target variable. We used the outer fold predicted MEG signal from two different models as the two source variables, and the corresponding recorded MEG signal as the target variable. This analysis therefore revealed how much the two feature spaces predicted the same parts of the MEG signal (Redundancy), and how much each predicted distinct from the other (unique Information). We fixed LMS+Art as one source, and considered the PID for the different acoustic feature spaces (Figure 3). We observed a gain in Redundancy of the multi-dimensional feature spaces such as LMS or LMS+GBFB over the one-dimensional Env or ReDe(Env). The highest Redundancy was achieved by the LMS+ReDe(LMS), reaching almost 100% (normalised by the information of the prediction from LMS+Art about the observed MEG). Furthermore, we observed a reduction in unique Information of LMS+Art going from the one-dimensional to the multi-dimensional acoustic feature spaces. For LMS+ReDe(LMS), we even observed slightly more unique Information of the acoustical feature space than of the annotated feature space, whose unique information was distributed around zero. On a group level, these patterns were highly similar between left and right ACs.

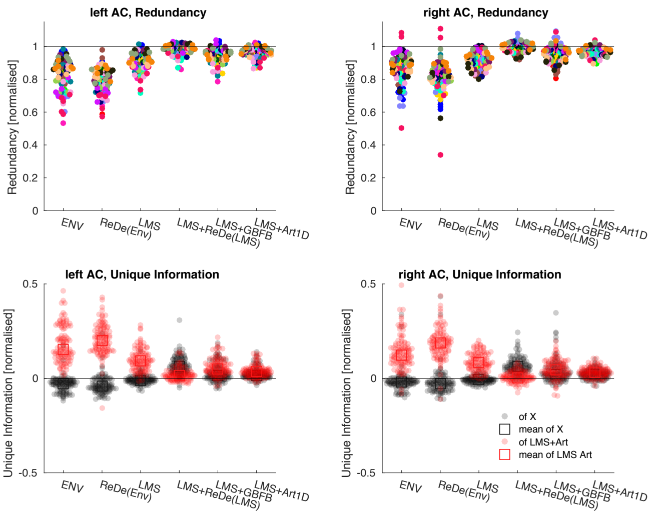

Figure 3: PID with predictions of annotated vs other feature spaces as sources and observed MEG as target for left and right AC; values normalised by MI of prediction of annotated features about observed MEG; each dot is one outer fold of one participant. *Top*: Redundancy; *Bottom*: Unique Information

Conclusion

We found that a model based on acoustic features resulted in better prediction performance of MEG responses than the current state-of-the-art model based on annotated linguistic features. Furthermore, we found that predictions based on the acoustic feature space were highly redundant with those of the linguistic feature space. The amount of unique information was practically zero for the annotated feature space, but weakly positive for the acoustic feature space. This means that both models are providing the same prediction in the same time points. In combination, these results show that what has been interpreted as a signature of prelexical abstraction in low-frequency EEG data is replicable in source level MEG data but can be explained with relatively simple spectrotemporal dynamics.

To let our models adapt optimally to different signal characteristics such as signal-to-noise-ratio at different grid points, we allowed the hyperparameters to be optimised individually at each grid point. Even for our relatively simple linear models, this was a computationally intensive procedure which prevented us from modelling broader regions in source space. While we assume that the nested cross-validation framework chosen here safeguarded us against overfitting, the results we present here are currently restricted to the grid points of peak retest-correlations. These are possibly biased towards low-level processing since our participants might not have paid the same degree of attention to the repetition. However, an inspection of cross-talk- and point spread functions of our spatial filters suggested that the grid points chosen here capture a large part of the activity stemming from bilateral superior temporal gyri.

Our results underscore that annotated linguistic feature spaces are useful tools to explore neuronal responses to speech and serve as excellent benchmarks. Their performance for explaining neuronal responses of high temporal resolution was exceeded by an acoustic feature space that has been shown to increase the robustness of automatic speech recognition systems to noise and reverberation (Kumar, Kanwoo & Stern, 2011). This suggests that during listening to speech, activity in ACs is mainly structured by non-stationarities in the speech signal, reflecting the importance of rapid power changes which characterise this part of our sensory environment.

Acknowledgments

CD is funded by the College of Science and Engineering at the University of Glasgow; JG received support from the Wellcome Trust (UK; 098433). We thank Moritz Boos, Jan-Mathijs Schoffelen, Dale Barr and Christoph Scheepers for helpful discussions.

References

Acerbi, L. & Ma, W. J (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. *Advances in Neural Information Processing Systems, 30,* 1834—1844

Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International, 5, 341—345

Bürkner, P. C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*

Crosse, M. J., DiLiberto, G. M., Bednar, A. & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience, 10:604.*

de Heer, W. A., Huth, A. G., Griffiths T. L., Gallant J. G. & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience, 37,* 6539–6557.

Di Liberto, G. M., O'Sullivan J. A., Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology, 25,* 2457–2465.

Hertrich, I., Dietrich, S., Trouvain, J., Moos, A. & Ackermann, H. (2012). Magnetic brain activity phase-locked to the envelope, the syllable onsets, and the fundamental frequency of a perceived speech signal. *Psychophysiology, 49,* 322–334.

Ince, R. A. A. (2017). Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy, 19.*

Kumar, K., Kanwoo, C., J., & Stern, R. M (2011). Delta-spectral-cepstral coefficients for robust speech recognition. In: *International Conference on Acoustics, Speech and Signal Processing (ICASSP).* Prague, Czech Republic: IEEE

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, 156869*

Schädler, M. R., Meyer, B. T. & Kollmeier, B. (2012). Spectro-temporal modulation subspace-spanning filterbank features for robust automatic speech recognition. *Journal of the Acoustical Society of America, 131,* 4134—4151

Varoquaux, G., Raamana, P. R., Engeman, D. A., Hoyos-Idrobo, A., Schwartz, Y. & Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage, 145,* 166—179

Yuan, J., Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America, 123,* 3878