Data-driven methods reveal the generalizing mechanisms of speech processing in naturally varying soundscapes

Moritz Boos (moritz.boos@uol.de)
Applied Neurocognitive Psychology Lab, University of Oldenburg, Ammerländer Heerstrasse 114
Oldenburg, 26129 Niedersachsen Germany

Jörg Lücke (joerg.luecke@uol.de)
Machine Learning Group, University of Oldenburg, Ammerländer Heerstrasse 114
Oldenburg, 26129 Niedersachsen Germany

Jochem Rieger (jochem.rieger@uol.de)
Applied Neurocognitive Psychology Lab, University of Oldenburg, Ammerländer Heerstrasse 114
Oldenburg, 26129 Niedersachsen Germany

Abstract

Humans rarely encounter speech without background noise. However, research on the cortical mechanisms of speech processing mostly focusses on individual speech features in isolation, which might not generalize to a more naturalistic environment. To examine the mechanisms of speech processing in natural soundscapes, we use unsupervised learning to infer spectro-temporal patterns that are adapted to the statistics of speech in noise. Using these patterns, we predict fMRI activity (n=20) evoked by a long auditory stimulus with voxel-wise encoding models and find a latent space of predicted brain activity that is shared between participants and represents the perceived noise level of the stimulus. Activity in the latent space forms two clusters, one representing stimuli with varying noise level related to the presence of simple, time-frequency separable patterns, and another consisting of stimuli with uniformly low perceived noise level. The cluster representing noisy stimuli explains variance in secondary auditory areas, through reduced activation for very noisy stimuli, while the cluster consisting of clear speech explains variance in both primary and secondary auditory areas. This shows how features adapted to speech in a natural soundscape relate to differences in the subjective percept of noise and the resulting dichotomy in brain activity.

Keywords: fMRI; auditory; encoding models; unsupervised learning; speech processing

Introduction

Every day humans encounter a multitude of auditory environments in which speech is embedded in an auditory background. However, most research on the cortical mechanisms of speech processing focusses on clear or computer-altered speech (Mattys et al., 2012; Holdgraf, De Heer, et al., 2016). This research has generated insights on the neuronal processing of speech (Friederici, 2012; DeWitt and Rauschecker, 2012) by isolating the effect of a-priori chosen features (like phonemes) on brain activity, but it is not clear which speech representations remain relevant for the neuronal processing of speech in natural soundscapes.

Previous studies show that visual features adapted to the statistics of natural images outperform a-priori chosen features in the prediction of functional magnetic resonance imaging (fMRI) data (Güçlü and Gerven, 2014) and sparse coding approaches can find complex auditory representations of clear speech or environmental sounds that resemble neuronal receptive fields (Młynarski and McDermott, 2017), but so far no study combined statistically inferred auditory features with predictive models of fMRI data to show how features adapted to the natural statistics of speech in an auditory background relate to cortical activity.

We fill this gap by training voxel-wise encoding models (Naselaris et al., 2011; Holdgraf, Rieger, et al., 2017) to predict fMRI activity elicited by a long, naturalistic stimulus of speech in a continuously varying soundscape from auditory features learned by a sparse coding approach. To account for inter-individual differences in neuroanatomy, we functionally link individuals across shared voxel responses (Lashkari et al., 2010).

This allows us to find functional components – voxel response patterns –, that are highly similar across participants and whose predicted activation form two clusters. One cluster consists of predicted brain activity due to speech stimuli that are homogenously rated as clear and subsequently represented by learned auditory features with higher complexity. Stimuli that are perceived as noisy, but with varying levels of noisiness, lead to predicted brain responses that form a second cluster shared across individuals. In this cluster, stimuli represented by simpler, time-frequency separable, spectro-temporal patterns are rated as more noisy.

These clusters predict brain activity in different regions: one cluster predicts variation in mostly secondary auditory areas for higher perceived noise ratings, while the cluster of clear speech stimuli predicts variation in primary and secondary auditory cortex.

Methods

Data

All data were acquired from the OpenfMRI portal, at `http://openfmri.org/dataset/ds0001113`. Hanke et al. (2014) presented the two hour long German audio-description of the movie ”Forrest Gump” to twenty participants while fMRI activity was recorded in a 7-Tesla sccanner.

Data-driven stimulus representation

We use unsupervised learning (Murphy, 2012) to statistically infer spectro-temporal patterns – basis functions – that represent the stimulus. We choose a sparse coding (Olshausen and Field, 1997) approach, because sparse coding tends to produce patterns that are similar to the tuning properties of visual (Olshausen and Field, 1997) and auditory (Młynarski and McDermott, 2017) sensory neurons. To learn basis functions from auditory stimuli, we use binary sparse coding (BSC) which constrains their activations to be binary – either a pattern is present or not – similar to the spike of a sensory neuron (Henniges et al., 2010). Using this method, we find 200 basis functions that decompose the audio data into a set of sparsely activated spectro-temporal patterns with differing complexity. Each pattern is 100ms long, with ten time-steps consisting of 48 Mel frequencies.

Voxel-wise encoding models

Due to the focus on speech and auditory processing of voxels, we limit our analysis to voxels in the temporal lobe (115421 voxels each of size $1mmx1mmx1mm$ per subject). To account for the delay and temporal integration of the BOLD response we split the movie into six second segments and concatenate the corresponding BSC basis activations to use as time-lagged stimulus representation. For each voxel (and subject) we train a regularized linear regression to predict the fMRI sample located four seconds **after** the time window from the concatenated BSC features. To validate the voxel-wise encoding models and study patterns in predicted brain activity, we use cross-validation to predict the voxel activity for each of the eight runs of the movie. The predicted brain activity – a matrix of fMRI samples N times voxels V – can be compared to the observed brain activity to derive a correlation score between predicted and observed brain activity in each voxel, but we can also use dimensionality reduction to find sets of voxels for which our encoding models make similar predictions. To find these patterns, we use principal component analysis (PCA) to decompose the NxV matrix of predicted brain activity across voxels: each resulting, orthogonal component represents a time-course of predicted activity in response to the stimulus that is shared by multiple voxels. Using dimensionality reduction allows us to go from a matrix of activity in voxels to a lower dimensional matrix of activity in functional components, ordered by their explained variation. This matrix forms the core of our subsequent analysis.

Functional patterns in predicted brain activity are shared between individuals

We now quantify the similarity of these functional patterns between individuals. Therefore we compute the correlation between the time-series of each principal component between participants, i.e. correlation between the time-series of the first component for each set of participants, and show the resulting correlation matrices – one matrix for each component – in Figure 1, as well as the boxplots of the upper diagonal elements of the correlation matrices. The first three components are similar across participants, with a sharp drop in similarity for the fourth and fifth component. Because predicted brain activity is similar between individuals only in the first three components, all further analyses use this three dimensional latent space.

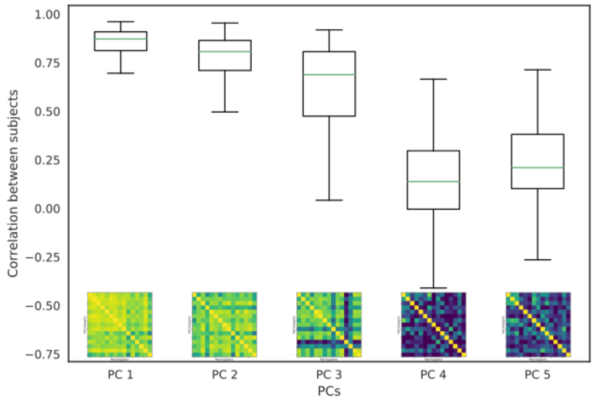

Figure 1: Boxplots of all between-participants correlations between their component time-series and the correlation matrices of the component time-series, where each row and column denote one specific participant.

Functional patterns reveal a dichotomy between clear and noisy speech

By relating components of predicted brain activity across individuals, we find a three dimensional space that can represent predicted brain activity in all participants, but it does not provide an interpretation what causes the distribution of predicted brain activity in this space. Since we are interested in the effect of an auditory background on the neuronal representation of speech, we obtain ratings on the perceived noisiness (on a scale of zero to six) of a subset of the auditory stimulus from 18 participants, that were not involved in the collection of the fMRI data. To relate this saliency of speech versus noise to activation in the space of predicted brain activity, we average the activation of the first three components across participants and show the noise ratings averaged across raters in the first two functional components (Figure 2 A). We use a Gaussian mixture model with two components to find clusters in the first three components and color all data points by their assigned cluster. Due to the different noise ratings between the two clusters (mean noise rating of 1.43 and 3.80, $t = 19.27$, $p < 0.0001$), we term one cluster ”noisy speech” and the other ”clear speech”.

In the first two components the majority of stimuli form a center of mass that includes stimuli rated as clear or slightly noisy (2 A, lowest circle), from which one cluster extends along similar activation of the first and second component (right circle), consisting mainly of clear speech and one cluster mainly along the second component that consists of stimulus parts rated as noisy (upper circle). Activation in the first dimension in this space negatively correlates with stimulus loudness in decibel ($r = -0.78$, $p < 0.0001$). To relate activity in these clusters to predicted brain activity, we reconstruct voxel-wise predictions from fMRI samples belonging to three areas in this space (dashed black circles). Figure 2 B shows the predicted fMRI activity for each area, averaged across samples and participants. Predicted brain activity in the area of clear and quiet stimuli is negative in primary auditory areas and close to zero in secondary auditory areas. Louder and slightly more noisy stimuli lead to predicted brain activity that is positive in both primary and secondary areas, and stimuli that are loud as well as noisy lead to positive predicted brain activity in primary auditory areas, and negative predicted brain activity in secondary auditory areas. Comparing predicted fMRI activity to observed fMRI activity belonging to either cluster (Figure 2 C) shows that both clusters explain variance in primary and secondary auditory areas, however the cluster of noisy stimuli explain more variance in secondary than in primary auditory areas.

Figure 3 A shows the average time-frequency separability – a measure of the complexity of receptive fields (Mazer et al., 2002) – of the spectro-temporal basis functions that are used to represent movie segments in the two clusters. In the cluster of noisy speech, an increase in perceived noise level is related to an increased presence of easily time-frequency separable basis functions ($r = .34$, $p < 0.0001$). Figure 3 B shows several examples of spectro-temporal patterns with high and low time-frequency separability, that are used to represent the auditory stimulus.

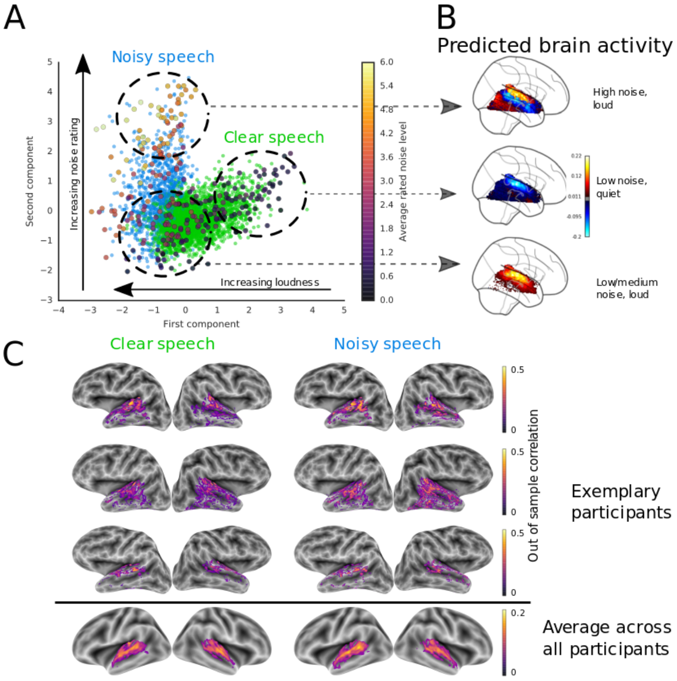

Figure 2: **A** Scatter plot of each movie segment’s activation of the first two components and the average perceived noise level of a subset of the movie. Cluster analysis reveals differing patterns for speech in quiet and speech in noise. **B** Average predicted activity for fMRI samples belonging to three areas in the latent space. **C** Out-of-sample correlation between predicted and observed fmri activity reconstructed from shared clusters in the latent space for stimuli containing speech in noise and speech in quiet, shown for exemplary individuals and averaged across individuals.

Discussion

We use binary sparse coding to infer spectro-temporal patterns that are adapted to the natural statistics of a long auditory stimulus consisting of speech in a natural soundscape. We predict fMRI activity with voxel-wise encoding models from these spectro-temporal patterns and combine the resulting models with dimensionality reduction of (predicted) voxel response time-courses (Lashkari et al., 2010). This allows us to relate the low dimensional distribution of predicted brain activity to perceptual (rated noise levels) and physical (loudness) properties of the stimulus to gain insight into the resulting predicted activation of primary and secondary auditory cortex. We then show which stimulus features predict a higher noise rating (a larger proportion of simple, time-frequency separable spectro-temporal patterns) and show how the presence of this feature explains the observed dichotomy

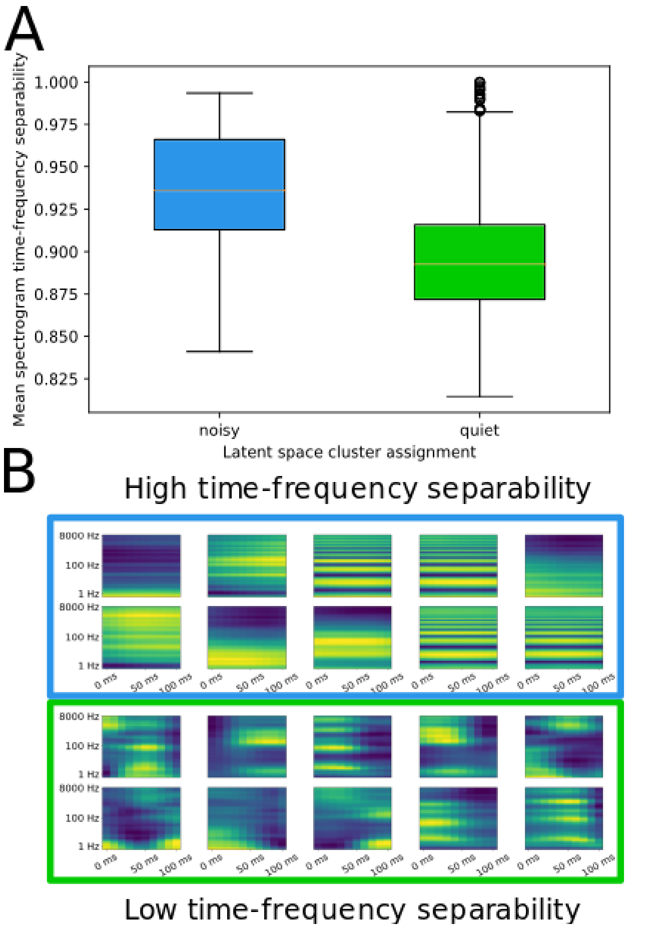

Figure 3: **A** Boxplots of the average time-frequency separability of basis functions that are active in the movie segments belonging either to the noisy or clear speech cluster. **B** Examples of time-frequency separable and nonseparable learned spectro-temporal basis functions.

between clear speech and noisy speech in the stimulus, in predicted brain activity, and in the resulting percept. Finally, the presence of these highly separable basis functions are associated with a reduction in predicted activity in secondary auditory areas. By incorporating subjective ratings of individual stimuli, we provide a reconciliation of results from laboratory settings that show reduced activation in secondary auditory areas for high noise levels (Scott and McGettigan, 2013) with results that show increasing robustness to noise along the auditory pathway (Kell et al., 2018): predicted activation of secondary auditory areas remains positive for low to medium perceived noise levels (Figure 2 A, lowest circle) and only drops in the high noise condition (Figure 2 A, rightmost circle). This provides a perspective on how features adapted to speech in a natural soundscape relate to differences in the subjective percept of noise and the resulting dichotomy in brain activity.

References

DeWitt, Iain and Josef P Rauschecker (2012). “Phoneme and word recognition in the auditory ventral stream”. In: *Proceedings of the National Academy of Sciences* 109.8, E505–E514.

Friederici, Angela D (2012). “The cortical language circuit: from auditory perception to sentence comprehension”. In: *Trends in cognitive sciences* 16.5, pp. 262–268.

Güçlü, Umut and Marcel AJ van Gerven (2014). “Unsupervised feature learning improves prediction of human brain activity in response to natural images”. In: *PLoS Comput Biol* 10.8, e1003724.

Hanke, Michael et al. (2014). “A high-resolution 7-Tesla fMRI dataset from complex natural stimulation with an audio movie”. In: *Scientific data* 1.

Henniges, Marc et al. (2010). “Binary sparse coding”. In: *International Conference on Latent Variable Analysis and Signal Separation*. Springer, pp. 450–457.

Holdgraf, Christopher R, Wendy De Heer, et al. (2016). “Rapid tuning shifts in human auditory cortex enhance speech intelligibility”. In: *Nature communications* 7, p. 13654.

Holdgraf, Christopher R, Jochem W Rieger, et al. (2017). “Encoding and Decoding Models in Cognitive Electrophysiology”. In: *Frontiers in Systems Neuroscience* 11, p. 61.

Kell, Alexander JE et al. (2018). “A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy”. In: *Neuron* 98.3, pp. 630–644.

Lashkari, Danial et al. (2010). “Discovering structure in the space of fMRI selectivity profiles”. In: *Neuroimage* 50.3, pp. 1085–1098.

Mattys, Sven L et al. (2012). “Speech recognition in adverse conditions: A review”. In: *Language and Cognitive Processes* 27.7-8, pp. 953–978.

Mazer, James A et al. (2002). “Spatial frequency and orientation tuning dynamics in area V1”. In: *Proceedings of the National Academy of Sciences* 99.3, pp. 1645–1650.

Młynarski, Wiktor and Josh H McDermott (2017). “Learning Mid-Level Auditory Codes from Natural Sound Statistics”. In: *arXiv preprint arXiv:1701.07138*.

Murphy, Kevin P (2012). *Machine learning: a probabilistic perspective*. MIT press.

Naselaris, Thomas et al. (2011). “Encoding and decoding in fMRI”. In: *Neuroimage* 56.2, pp. 400–410.

Olshausen, Bruno A and David J Field (1997). “Sparse coding with an overcomplete basis set: A strategy employed by V1?” In: *Vision research* 37.23, pp. 3311–3325.

Scott, Sophie K and Carolyn McGettigan (2013). “The neural processing of masked speech”. In: *Hearing research* 303, pp. 58–66.