Neural coding of speaker identity : methodological and ermpirical contributions

Chapter 6

Summary and Conclusions

Summary

Voices convey information about speech content ('*what* is being said'), as well as the affective state ('*how* it is said") and individual characteristics ('*who* is saying it') of speakers (Belin et al., 2004; Campanella and Belin, 2007). This thesis focuses on the last aspect, and specifically on the brain's processing of information that vocal signals convey on the speaker's identity. The brain mechanisms enabling us to tell *who* is speaking are only partially understood. The research presented in this thesis contributes to this field (1) by introducing and evaluating new methods for *decoding analysis* in electroencephalography (EEG) and functional MRI (fMRI), and (2) by applying these methods to examine neural representations of speaker identification in human listeners. In particular, we were interested in the representations that are robust to the large acoustic variability associated with the virtually infinite number of utterances and in the possible interference of background noise.

Methodological Contributions

Chapters 2 and **3** present the development and evaluation of pattern recognition techniques for EEG and fMRI data analysis. More specifically, **chapter 2** illustrates and compares different ways to perform single-trial decoding as an analysis tool for EEG data. Six types of pattern analyses – resulting from the combination of three types of feature selection in the temporal domain (predefined windows, shifting window, whole trial) with two approaches in the channel dimension (channel wise, multi-channel) – are considered. These analyses were applied to EEG data collected to examine the task dependence of the cortical mechanisms for encoding speaker identity and speech content (vowels). Results show that a different grouping of features helps to highlight complementary aspects (i.e. temporal, topographic) of information in the EEG data. The shifting window/multi-channel approach could trace both the early build-up of neural information reflecting speaker or vowel identity and the late and task-dependent maintenance of relevant information reflecting the performance of a working memory task. Since it makes use of the high temporal resolution of EEG (or MEG), the shifting window approach with sequential multi-channel classifications was found to be an appropriate choice for tracing the temporal profile of neural information processing.

Decoding analysis as performed in fMRI studies in most cases investigates whether cortical representations of different cognitive or perceptual states can be separated in a high-dimensional space as defined by multivoxel activation patterns. Visualizing the topology of these patterns may be informative, especially when

classifying more than two conditions. This motivates the development described in **chapter 3,** which introduces a novel method to decode fMRI datasets using a supervised form of self-organizing maps (SOMs). The feasibility of this method for decoding and visualizing high-dimensional fMRI data was evaluated with data simulations and real data from a voice identification experiment. To exploit the visualization possibilities offered by SOMs, one approach is proposed to visualize the classification model and the corresponding classification performance both at single-subject and at group level. In the latter case, single-subject SSOMs are summarized to form a single subject SSOM and subsequently SSOM units of single subjects are mapped into group space. Overall, the analyses show that the SSOMs-based method offered both a good capability to perform multiclass decoding and to convey information about the underlying data topology within one step of analysis.

Empirical Contributions

In the second part of the thesis (i.e. **chapters 4** and **5**) two different aspects of speaker identity processing are investigated. In particular, the effects of (1) context-specific behavioral demands and (2) interfering background sounds on cortical representations of speaker identity are examined. The former was studied in **chapter 4** by decoding fMRI responses to vowel utterances spoken by different speakers while participants were asked to recognize either speakers or vowels. Results showed that information about speaker identity or speech content was only contained in cortical representations while subjects performed the respective task (i.e. the brain-based decoder was able to classify the identity of a speaker during the speaker task and, similarly, the correct vowel during the vowel task). Regions most important for speaker identity classification were early auditory cortex and mid to anterior (right) STG/STS whereas for vowel classification early auditory cortex, bilateral superior temporal plane and mid to posterior STG/STS were most involved. The outcomes showed that context-specific demands led to different processing of the same physical stimuli which was expressed in distributed activation patterns rather than localized activation changes.

To investigate the effect of background noise on representations of speaker identity (**chapter 5)**, short non-linguistic vocalizations were presented in auditory scenes containing artificial and natural background noise while acquiring fMRI responses. We aimed at decoding speaker identity by making use of SSOMs as developed in **chapter 3**. Results showed that speaker identity could be decoded for vocalizations without background noise and with white noise but not within natural noise. In addition, activation patterns evoked by stimuli without noise could be used to decode speaker identity for sounds with white noise and vice

versa. These results suggested that cortical representations were robust to changes in speech content and to added white noise. In contrast, natural noise seemed to interfere with speaker representations more severely, which might be due to its richer spectro-temporal structure as compared to white noise or due to differences in the neural processing required to segregate two (or more) meaningful and ecologically relevant auditory objects. These findings provide evidence for activation patterns in temporal cortex that encode speaker identity in an abstract manner which generalizes across non-linguistic vocalizations and is robust to white noise masking. Further research is needed to gain more insight into the neural representations of speaker identity when vocal sounds are accompanied by natural background.

Conclusions

The work presented in this dissertation dealt with the multivariate analysis of both EEG and fMRI datasets concerned with speaker identity processing. Different ways to perform decoding of EEG data have been evaluated and an approach to apply self-organizing maps to classify and visualize multiclass fMRI data has been developed. Two original fMRI investigations demonstrated that information on speaker identity is reflected by distributed activation patterns that cover early as well as higher-order auditory cortex. Furthermore, these studies show that the amount of information of speaker or vowel identity is modulated by specific behavioral demands and is robust to distortions by noise with a flat spectral response. Taken together, these findings suggest that speaker identity is jointly encoded by neuronal populations in multiple auditory areas. This is in contrast with results suggesting that speaker identity is exclusively represented in specialized regions on a higher level in the processing hierarchy. The finding that distributed patterns represent speaker identity rather suggests a temporal coding model. A binding of features representing one auditory object by temporal coherence could also help to explain task-specific modulations of cortical representations (chapter 4; see also Bonte et al. [2009] and Elhilali et al. [2009a, 2009b], Shamma et al. [2011]).

One limitation of decoding studies including the ones presented here is that while results reveal *whether* activation patterns are informative for distinguishing experimental conditions, in most cases limited insights are provided on the processing or transformation of stimulus features that underlie this information. It would be interesting to follow a complementary *encoding* approach that predicts brain activity based on hypothesized processing of the sensory stimulus (see Naselaris et al., 2011 for a review and Çukur et al., 2013; Kay et al., 2008; Mitchell

et al., 2008; Moerel et al., 2012; Pasley et al., 2012 for exemplary studies). In combination with high spatial resolution (fMRI) and high temporal resolution (EEG or MEG) data, such an approach would allow formulating testable predictions on which computational model best describes the extraction and processing of features underlying speaker identification. Furthermore, it may help elucidating the specific role of the different cortical auditory areas (early and higher order) and to understand the sources of current decoding outcomes.

References

Andics, A., McQueen, J.M., Petersson, K.M., Gál, V., Rudas, G., Vidnyánszky, Z., 2010. Neural mechanisms for voice recognition. NeuroImage 52, 1528–1540.

Belin, P., Fecteau, S., Bédard, C., 2004. Thinking the voice: neural correlates of voice perception. Trends Cogn Sci 8, 129–135.

Belin, P., Zatorre, R.J., 2003. Adaptation to speaker's voice in right anterior temporal lobe. Neuroreport 14, 2105.

Bonte, M., Valente, G., Formisano, E., 2009. Dynamic and Task-Dependent Encoding of Speech and Voice by Phase Reorganization of Cortical Oscillations. J Neurosci 29, 1699–1706.

Campanella, S., Belin, P., 2007. Integrating face and voice in person perception. Trends Cogn Sci 11, 535–543.

Çukur, T., Nishimoto, S., Huth, A.G., Gallant, J.L., 2013. Attention during natural vision warps semantic representation across the human brain. Nat Neurosci 16, 763–770.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., Shamma, S.A., 2009a. Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron 61, 317–329.

Elhilali, M., Xiang, J., Shamma, S.A., Simon, J.Z., 2009b. Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. PLoS Biol 7, e1000129.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., Itoh, K., Kato, T., Nakamura, A., Hatano, K., Kojima, S., Nakamura, K., 1997. Vocal identification of speaker and emotion activates different brain regions. Neuroreport 8, 2809–2812.

Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355.

Kriegstein, von, K., Eger, E., Kleinschmidt, A., Giraud, A.-L., 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17, 48–55.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A., Just, M.A., 2008. Predicting human brain activity associated with the meanings of nouns. Science 320, 1191–1195.

Moerel, M., De Martino, F., Formisano, E., 2012. Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. J Neurosci 32, 14205–14216.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., Nagumo, S., Kubota, K., Fukuda, H., Ito, K., Kojima, S., 2001. Neural substrates for recognition of familiar voices: a PET study. Neuropsychologia 39, 1047–1054.

Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56, 400–410.

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 Biol 10, e1001251.

Shamma, S.A., Elhilali, M., Micheyl, C., 2011. Temporal coherence and attention in auditory scene analysis. Trends Neurosci 34, 114–123.