# Neural coding of speaker identity : methodological and ermpirical contributions

# Neural Coding of Speaker Identity

## Methodological and Empirical Contributions

Lars Hausfeld

The work presented in this thesis was funded by the Netherlands Organization for Scientific Research (NWO) and was conducted at Maastricht University.

Production CPI Wöhrmann Print Services, B.V.
ISBN 978-94-6203-505-8

# Neural Coding of Speaker Identity

## Methodological and Empirical Contributions

DISSERTATION

to obtain the degree of Doctor at Maastricht University,
on the authority of the Rector Magnificus, Prof.dr. L.L.G Soete
in accordance with the decision of the Board of Deans,
to be defended in public on Thursday 16th January 2014 at 16:00 hours

by

Lars Hausfeld

**Supervisor**

Prof. dr. Elia Formisano

**Co-supervisor**

Dr. Giancarlo Valente

Dr. Milene Bonte

**Assessment Committee**

Prof. dr. Bernadette M Jansma (Chair)

Prof. dr. Pascal Belin (University of Glasgow, UK)

Prof. dr. Rainer Goebel

Dr. Jonas Obleser (Max Planck Institute, Leipzig, Germany)

# Contents


| | | |
|---|---|---|
| **1** | General Introduction | 7 |
| **2** | Pattern Analysis of EEG Responses to Speech and Voice: Influence of Feature Grouping | 25 |
| **3** | Multiclass fMRI Data Decoding and Visualization Using Supervised Self-Organizing Maps | 49 |
| **4** | Task-Dependent Decoding of Speaker and Vowel Identity in Superior Temporal Cortex | 85 |
| **5** | Auditory Cortical Representations during Speaker Identification in Noise Auditory Scenes | 109 |
| **6** | Summary and Conclusions | 139 |
| | Acknowledgements | 145 |
| | Publications | 153 |
| | Curriculum Vitae | 154 |

# Chapter 1

# General Introduction

## Prologue

Imagine a bag filled with air that is constricted with great pressure to produce an outgoing airflow passing a vibrating membrane. The vibration is transferred across the molecules in the air in the form of a periodic wave. This wave first passes several edges and cave-like structures before it continues to travel a distance of 1.23 meters in open space at a speed of approximately 330 meters per second. After 0.005 seconds a small part of the wave enters another cave-like compound in which it impacts another membrane. The membrane itself sets three small rigid, connected objects in motion. The last of those is connected to a closed and fluid-filled tube where the vibration of the fluid propagates along yet another membrane. Obeying to the laws of mechanics this membrane is bending, which evokes an influx of charged particles into a cell body that - after a cascade of delicate events - initiates an electrical potential within a connected cell. In about 10 milliseconds this electrical potential reaches a cell agglomeration that belongs to a most complex, biological organ. An orchestrated arrival and further propagation of more electrical potentials makes you say after 2 seconds: "Oh, sorry John, what did you say?"

*A mechanistic description of a particular moment in a conversation with one person sharing his view of life and another one daydreaming.*

## Brain processing of Voices

It is actually astonishing that after such an endless chain of events described in the prologue one is saying "Oh, sorry *John*" and not Brian or Peter. It is even more astonishing when one realizes that the briefly mentioned '*orchestrated arrival and further propagation of more electrical potentials*' points towards the large, complex, and intriguing study of the human brain. This thesis combines these two remarks and investigates the brain mechanisms underlying speaker identification.

Recognizing the voice of the person who is talking to us is an ecologically relevant skill. The importance of this skill we mostly learn to appreciate whenever it fails, as it likely puts us in an uncomfortable social situation. Extreme examples of such failures are patients with *phonagnosia*. As a consequence of brain damage, these patients are unable to discriminate and recognize previously familiar voices, despite normal hearing and speech perception (Van Lancker et al., 1989); see also (Garrido et al., 2009).

In normal listeners, identifying a person from his/her voice involves the formation of a *categorical* (abstract) representation of that person's voice, which enables mapping the wide range of acoustical signals that a particular voice can produce to a single entity. Psychophysical and computational modeling studies have suggested a number of properties of the vocal signal that the auditory system may use in order to form such a representation, including the fundamental frequency of the speaker's voice, its timbre and/or its breathiness (e.g. Baumann and Belin, 2010; D. H. Klatt and L. C. Klatt, 1990; Lavner et al., 2000; Murry and Singh, 1980; see Dellwo et al., 2007 for an overview). With the advent of functional neuroimaging and functional magnetic resonance imaging (fMRI) in particular (Bandettini et al., 1992; Kwong et al., 1992; Ogawa et al., 1992), new and promising experimental approaches became available to examine the neural mechanisms underlying the human brain functions, including the processing and recognition of human voices. Following a seminal report by Belin and colleagues (2000), numerous studies have examined the neural processing of voices by comparing brain responses to human vocal sounds and other sound categories using fMRI (Belin et al., 2000; Bonte et al., 2013, chapter 4; Charest et al., 2013; Ethofer et al., 2013; 2009; Latinus et al., 2011; 2013; Moerel et al., 2012). These studies have established that a network of non-primary auditory regions in the superior temporal cortex - often referred to as temporal voice areas (TVAs) – is involved in the processing of human voices (see *Regional tuning to voices*). However, it remains largely unsolved *how* (i.e. based on which acoustic features) and *where* (i.e. in which of the regions) the *identity* of a speaker is derived. It is also unknown whether perceptual representations of speaker identity emerge in specialized

modules within this superior temporal network or, alternatively, from distributed general purpose auditory mechanisms involving also early auditory areas (Formisano et al., 2008).

This thesis describes both methodological and empirical contributions aimed at addressing these open questions. More specifically, the first part of the thesis (chapter 2 and 3) describes novel methods for the multivariate (decoding) analysis of fMRI and EEG data, respectively. While these methods are general and can be applied to the analysis of any dataset, in the work presented here they were developed,, fine-tuned and evaluated using data collected to investigate the neural representation of speaker identity. The two chapters that form the second part of the thesis (chapter 4 and 5), describe empirical studies combining these analysis tools with original experimental paradigms to study the influences of top-down context (chapter 4) and background noise (chapter 5) on the neural representations of speakers. To provide the essential background to the original studies, in the remaining part of this introduction, I will first present a concise overview of the current state-of-the-art in the decoding analysis of fMRI and EEG data and then critically review current neurobiological models of voice processing and speaker identification.

## Decoding Analysis in fMRI and EEG

### Decoding brain states from fMRI Data

The signal obtained with fMRI reflects the metabolic demands that arise during neural activation and cause modulations of the ratio between oxygenated and deoxygenated hemoglobin in the blood (e.g. Logothetis, 2002; Logothetis et al., 2001). As such, fMRI measures neural activity *indirectly*. The combination of high spatial resolution and brain coverage is the major advantage of fMRI compared to other neuroimaging techniques. That is, with fMRI, a measurement of neural activation can be accomplished in about 1-4s in each part of the brain and with a spatial resolution of a few millimeters.

During the first years of fMRI research, the investigations focused on the function and responses of specific brain areas. The general linear model (GLM) statistical framework soon became a standard analysis tool for testing hypotheses on localized response differences and functional specialization (Friston, 1995). Within this framework, a model that contains predictors for the different experimental conditions and for possible confounding factors is fitted to the measured fMRI time series. As the statistical assessment of this model fit and of the response differences is performed independently in each location of the brain

(i.e. in each volumetric measurement element or *voxel*) this approach has been called mass-univariate. Because of the large number of tests involved - most fMRI datasets contain more than 100,000 voxels – this type of analysis needs to be combined with appropriate correction for multiple testing (e.g. family-wise error rate [FWER], cluster-size threshold [Forman et al., 1995], false discovery rate [FDR; Genovese et al., 2002]) to account for false positive outcomes.

In contrast to the mass-univariate approach, a more recent type of analysis aims to detect relations between cognitive or perceptual processes and activation *patterns* covering multiple voxels. In a seminal study by (Haxby et al., 2001), the authors showed that categories of pictures could be distinguished based on voxels covering ventral temporal cortex. In their analysis, average activation patterns for each category were derived from one half of the dataset. To test whether these template patterns were specific to categories of pictures, a cross-correlation analysis was carried out that assigned unlabeled functional images to the category of the template that showed the highest cross-correlation. Comparing the predicted with the true labels revealed a high correspondence, which suggested that spatially distributed activation patterns in the ventral-temporal visual cortex are highly informative of visual categories. Following this study, a large number of methods of pattern recognition and machine learning have been introduced in fMRI research to examine the relation between stimuli and spatial patterns of responses, rather than responses in individual voxels as in the traditional GLM approach. Owing to their higher sensitivity compared to the voxel-wise approach, these methods have quickly gained popularity. Initially, they were variously termed as *pattern-based* or *multivariate decoding* or '*brain reading*'. However, more and more consistently, researchers refer to the approach as *multi-voxel* (or multivariate) *pattern* analysis (MVPA) (see Formisano et al., 2008b; Norman et al., 2006; Pereira et al., 2009 for methodology-focused reviews; for more conceptual descriptions see Haynes and Rees, 2006; Tong and Pratte, 2012).

Independently of the exact type, the procedure of MVPA analysis includes the following three steps:

*1) Extracting and selecting features*. Feature extraction refers to the definition of the set of variables on which the multivariate analyses will be based. An example of commonly-employed features is the amplitude of responses to single presentations of a stimulus (single-trial responses), which can also be estimated as the coefficient (beta) of a model fit. These features are extracted at each voxel. Because the number of variables used may have a large influence on the analysis outcomes, MVPA often includes the important step of feature (voxel) selection (or reduction). The selection of responses may be limited to cortical

or subcortical region-of-interests (ROI) on the basis of anatomical or functional criteria. Alternatively, data-driven selection approaches can help to avoid that noisy voxels (i.e. voxels that are not informative) decrease performance (see Guyon and Elisseeff [2003] for feature selection via *ranking/filtering*).

2) *Training a model using machine-learning algorithms*. In this step, using a training data set, a model (or called classifier) is trained based on the relation between the feature set derived from 1) and the known condition labels. One important distinction of these models is whether they are linear or non-linear. In linear classifiers, the model is expressed as a linear combination of features. Linear classifiers are most abundant in fMRI as they are computationally efficient and work well for ill-posed problems when the number of features is much larger than the number of samples (which most often affects fMRI). Another advantage of linear classifiers relates to the straightforward mapping of the contribution of single features for classification. Non-linear classification approaches are used in particular when the number of voxels/features is small and linear class separation seems unlikely.

Whereas 1) and 2) are presented here – in the interest of simplicity - as separate steps, it should be noted that they may also be combined. So-called *wrapper* methods combine feature selection and model training by choosing recursively features for upcoming classification based on their impact on previous classification models. One example of a wrapper method is recursive feature elimination (RFE, Guyon et al., 2002) that iteratively removes the least contributing features and has been applied successfully for MVPA studies in fMRI (De Martino et al., 2008; Hanson and Halchenko, 2008).

3) *Testing the model's performance on independent data*. As a last analysis step, an independent (test) dataset is used to determine the capability of the classifier – as derived from step (2) – to correctly classify data not used during the training. The model predicts labels of independent feature patterns; the subsequent comparison of these predictions with the actual labels determines the performance of the classifier. In many studies the training and evaluation sets are taken from the same acquired dataset that is split into subsets many times. The model training and testing is done for each *cross-validation* split and their average performance is reported.

**Decoding stimuli and brain states from EEG (and MEG) Data**

Electroencephalography (EEG) and Magnetoencephalography (MEG) are measurement techniques that are commonly used to study the time course of

neural information processing in the human brain with high temporal resolution. In contrast to fMRI, EEG and MEG measure synchronized neural activity of a large number of cells *directly*.

Most studies that use EEG (and/or MEG) to examine cognitive or sensory phenomena rely on the comparison of averaged responses to repeated presentations of experimental conditions. This can be done in the temporal domain (event-related potentials [ERPs] or fields [ERFs], respectively) and/or in the frequency domain (event-related desynchronization and synchronization) (Pfurtscheller and Lopes da Silva, 1999). The averaging of the responses appears necessary because the signal-to-noise ratio (SNR) of EEG signals is generally low (MEG offers a larger SNR) and the expected difference between responses to different conditions are very subtle. Along with the strong tradition of the 'intuitive' analysis of averaged event-related potential/fields (ERP/ERF), this might explain why pattern recognition techniques - which rely on good estimates of single trial responses – have only scarcely been used (although there are a few recent exceptions both with MEG (Howard and Poeppel, 2010; Luo and Poeppel, 2007; Rieger et al., 2008) and EEG (Kerlin et al., 2010).

Pattern recognition techniques have been used in the development of EEG-based brain-computer interfaces (BCI), where EEG signals are used on-line for the control of computer and other devices. EEG-based BCI systems allow paralyzed and locked-in patients to communicate and interact with their environment (Birbaumer, 2006; Wolpaw et al., 2002). One important distinction between the formerly described BCI application and studies in cognitive neuroscience is that the intended contrast of brain activity between conditions in BCI is maximized to assure reliable performance whereas in cognitive neuroscience experimental conditions evoke brain responses with subtle differences in most cases. Importantly, BCI systems rely on similar data processing and analyses tools as MVPA fMRI-studies, i.e. feature extraction and selection, model training and evaluation (for reviews please see Bashashati et al., 2007, Besserve et al. 2007, Blankertz et al. 2010, Lotte et al. 2007, van Gerven et al. 2009).

**The methodological contributions of this thesis**

**Chapter 2** reports original methodological developments aimed at extending a MVPA framework to neuro-cognitive EEG studies (Hausfeld et al., 2012). Compared to MVPA studies in fMRI, the feature extraction step for EEG and MEG datasets offers more alternatives. In fMRI, a measure of response amplitude usually constitutes the input to classification algorithms. In EEG and MEG, however, different *types* of features can be considered that range from signal amplitude in the temporal domain (Rieger et al., 2008) to coherence measures

(Besserve et al., 2007), and power or phase information in the frequency domain (Kerlin et al., 2010; Luo and Poeppel, 2007; Rieger et al., 2008). Moreover, more sophisticated data transformations like wavelet coefficients (Rieger et al., 2008) might also compose the input of the pattern recognition algorithm. To complicate further matters, features can be *grouped* in various ways along, for example, the channel and (spectro) temporal domain. Chapter 2 describes the evaluation of six types of pattern analyses deriving from the combination of three types of feature selection in the temporal domain (predefined windows, shifting window, whole trial) with two approaches to handle the channel dimension (channel wise, multi-channel). These different types of analyses are combined with a Gaussian Naïve Bayes classifier to examine a multi-subject EEG data set from a study aimed at understanding the task dependence of the cortical mechanisms for encoding speaker's identity and speech content (vowels) from short speech utterances (Bonte, Valente, & Formisano, 2009).

**Chapter 3** describes a novel method for fMRI MVPA. So far, most fMRI studies employing multivariate pattern decoding in the context of experimental designs with more than two conditions transformed the multiclass classification problem into a series of binary problems. Furthermore, for decoding analyses, classification accuracy is often the only outcome reported. However, the analysis of the topology of activation patterns in the high-dimensional features space may provide additional insights into the underlying brain representations. The method developed in Chapter 3 is based on a supervised variant of self-organizing maps (SSOM; Kohonen, 2001) and can be used for decoding and visualizing voxel patterns of fMRI datasets consisting of multiple conditions. Using simulations and real fMRI data, this new SSOMs-based approach is evaluated in the context of decoding analyses of single data sets as well as in analyses involving multiple cross-validation splits and/or multiple subjects. In particular, it is applied to a challenging 3-class fMRI classification problem with datasets collected to examine the neural representation of human voices at individual speaker level (3 speakers).

## Voice identity representation in auditory cortex

### Regional tuning to voices

Voices are the most relevant sounds of our daily life, conveying multidimensional information including communication messages, emotional content and cues on person identity (Belin et al., 2004; Campanella and Belin, 2007). The existence of brain regions with a preference for voices over other complex sounds in both humans (Belin et al., 2000) and non-human primates (Petkov et al., 2008) indicates the ecological importance of voices. Voice-sensitive

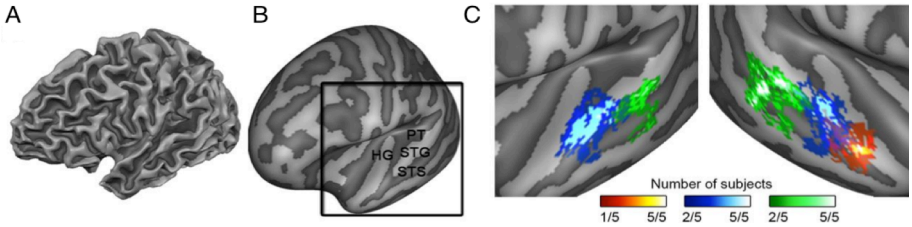



**Figure 1. Voice Areas as Determined by a Localizer.** The localizer contrasted voiced and speech stimuli with sounds of other categories (here: animals, tools and nature sounds). Panel A shows the boundary of cortical grey and white matter. Panel B depicts the inflated hemisphere and anatomical landmarks in auditory cortex. Probabilistic maps of voice areas are presented in panel C. In particular bilateral clusters of higher voice activation are found in posterior and middle STG/STS and in right anterior STG/STS. HG – Heschl's gyrus, PT – planum temporale, STG – superior temporal gyrus, STS – superior temporal sulcus. Adapted with permission from Moerel et al. (2012).

regions or *temporal voice areas* are found in bilateral middle and posterior superior temporal gyri (STG) and sulci (STS), as well as in the right anterior STG/STS (Fig. 1, Belin et al., 2000; Bonte et al., 2013, chapter 4; Charest et al., 2013; Ethofer et al., 2013; 2009; Latinus et al., 2013; 2011; Moerel et al., 2012). While the left anterior STS/STG may also show voice-sensitivity, this region is found less consistently across subjects. Although most details remain unknown, it is hypothesized that speaker identity is derived in this network of areas (or in the auditory cortex in general, see Formisano et al., 2008a) through the processing and binding of a unique combination of acoustic features which are characteristic of vocal signals. The following paragraph describes the acoustic features most commonly associated with the neural processing of voices.

### Voice Characteristics

The human voice is a product of several organs including the lungs, the larynx with glottal folds, the pharynx, the mouth and the nasal cavities (Fig. 2). The vocal tract is situated above the larynx (supra-laryngeal) and shaped by jaw, tongue and lips. An influential theory of speech production is the source-filter model that treats speech output as a two-stage process of a sound source and an acoustic filter (Fant, 1960; Stevens, 1999; Titze, 2008). The sound source (also called glottal source) comprises the airflow from the lungs that passes the rapidly and periodically opening and closing vocal folds. The resulting complex tone has a characteristic *fundamental frequency* (F0) that is determined by the glottal pulse rate (GPR). The perceptual correlate of the sound's F0 is *voice pitch*. Subsequently, the sound passes through the vocal tract which acts as a filter by changing its size and shape and thereby reducing or enhancing energy at specific frequencies

(resonance). Frequencies that are selectively enhanced are called *formants*. Formants are important for the identification of voiced speech sounds (i.e. when vocal folds are non-constricted and vibrate), such as vowels. Usually four or five formants can be extracted from voiced speech sounds. During speech, the tension of vocal folds changes as well as the state of vocal tract. This leads to modulations of fundamental frequency and formant frequencies that are referred to as F0 and formant contours, respectively.

The anatomy of the vocal organs thus determines to a large extent the characteristic sound of a speaker's voice. An important perceptual cue to distinguish and/or recognize speakers is the fundamental frequency of their voice (e.g. Baumann and Belin, 2010). Because the length of the vocal folds determines the rate with which they open and close, voice pitch is typically low in adult man, intermediate in adult women and high in children. Besides voice pitch, the length of the vocal tract (VTL) provides information about speaker identity by shifting formant frequencies (e.g. Smith and Patterson, 2005; Turner et al., 2009).

When listening to individual vowel sounds (/a/, /i/, and /u/), listeners seem to rely mostly on a voice's fundamental frequency, and additionally on the first formant (F1) for the recognition of female speakers and on the fourth (F4) and fifth formants (F5) for the recognition of male speakers (Baumann and Belin,

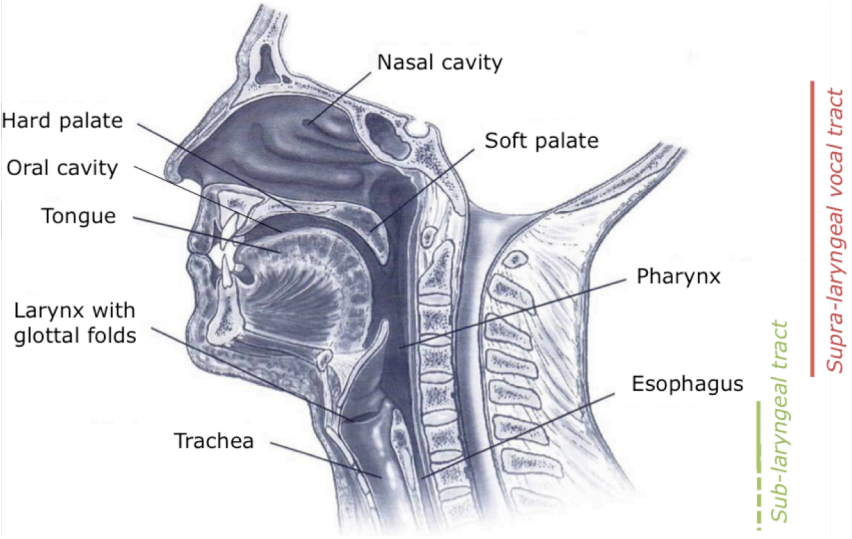



**Figure 2. Human Speech Apparatus.** It is divided into sub-laryngeal tract, larynx and supra-laryngeal tract. The sub-laryngeal tract consists of lungs, diaphragm (both not shown) and trachea (or air tube). The larynx includes the glottal folds. The supra-laryngeal tract begins above the larynx and contains pharynx, oral and nasal cavities, the tongue, palate, jaw and lips.

2010). Instead, when listening to acoustically modified versions of the vowel /a/ pronounced by different familiar speakers, vocal tract parameters (i.e. formants and related measures) may be more important than glottal source features (i.e. fundamental frequency/voice pitch; Lavner et al., 2000). Depending on the type of speech stimuli, we may also use other voice characteristics including spectro-temporal regularity as expressed by harmonics-to-noise ratio (HNR, e.g. Bruckert et al., 2010) or, as it is often the case in everyday situations, dynamic voice cues like prosody or intonation (Murry and Singh, 1980; Schultz, 2007; Van Dommelen, 1990). Furthermore, due to the unique combination of acoustic features characterizing a person's voice, different acoustic cues may be weighted differently depending on the speaker and context (Lavner et al., 2000).

**Neural encoding of voice identity**

When we hear a familiar voice, in most cases, we have no trouble in recognizing the person at hand. This ability is striking as our auditory system is faced with the problem of associating a large range of utterances that differ in various dimensions to a single voice. For example, you could hear "Hello there!" or "Oh no!" and still be certain that it was Sophie shouting. Being able to say *who* uttered something thus entails an adaptive model of voices that uses various and flexible dimensions of speech cues.

A possible model for the representation of voice identities in the human brain suggests an organization that encodes the perceived similarity or dissimilarity of voices. This organization could be conceptualized as a high-dimensional space, whose entries indicate individual voices. Both behavioral (Bruckert et al., 2010; Latinus and Belin, 2011; Papcun, 1989) and imaging evidence (Andics et al., 2013; 2010; Latinus et al., 2013) suggests that the origin (and dimensions) of this hypothetical voice-representative space is not based on absolute, physical or stimulus-inherent, features but is rather understood in terms of relative distances with respect to an average, prototypical voice. Accordingly, this type of representation has been termed norm-based or mean-based coding of voices. Interestingly, faces – the visual counterpart of voices (Belin et al., 2004) – seem to be represented in a norm-based manner as well (Leopold et al., 2006; 2001; Loffler et al., 2005).

Independently of whether the distances between identities in the representational space are expressed in absolute or relative terms, one may conceive each voice identity as an auditory category. Recognizing a person from his/her voice involves mapping the wide range of acoustical signals that a particular voice can produce to an abstract (categorical) representation of that person's voice. The formation of such abstract representations requires a

reduction of within-category distances and maximization of between-category distances (Harnad, 1987). Recently, Ley and colleagues (2012) used fMRI and MVPA to show that learning of newly and artificially established auditory categories promotes the formation of categorical sound representations in the auditory cortex. A similar mechanism can be hypothesized for the neural encoding of familiar voices, which might entail small distances between representations of distinct utterances from the same speaker and large distances across different voices (i.e. a supra-individual space). This provides a rationale for studying the neural encoding of voices using an MVPA approach. The successful decoding of speaker identity from brain responses to stimuli that differ largely in terms of acoustic content and backgrounds, can be taken as evidence for the invariance (or robustness) of the underlying neural representations to the sensory differences, and thus point to a categorical representation of voice identity.

**The empirical contributions of this thesis**

The study described in **chapter 4** employed a decoding approach to investigate how attention to speaker or vowel identity modulates the spatial pattern of auditory cortical responses to the same speech sounds. The stimulus design was similar to the EEG data analysed in **chapter 2**. In this case, however, children voices (a boy and a girl) were employed in addition to an adult male voice. This allowed investigating the processing of children voices that, unlike adult voices, are not readily distinguished based on fundamental frequency and whose identification additionally relies on formant frequencies (Bennett and Weinberg, 1979; Perry et al., 2001). The effects of selectively attending to relevant acoustic features on the response patterns were tested by asking subjects to perform a delayed-match-to-sample task on either speaker or vowel identity. We expected to find more information in activation patterns about the task-relevant compared to the task-irrelevant stimulus dimension.

In **chapter 5** we applied the decoding algorithm developed in **chapter 3** to investigate the robustness to noise of representations of speaker identity in auditory responsive areas of the temporal cortex. Apart from using highly varying vocalizations with large dynamics in pitch and formant contours, we investigated how the representations of individual speakers change when the relevant vocal signals are mixed with stationary white noise or with background noise as encountered in real life situations (e.g. in a cafeteria or train station).

## References

Andics, A., McQueen, J.M., Petersson, K.M., 2013. Mean-based neural coding of voices. NeuroImage 79, 351–360.

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

Bandettini, P.A., Wong, E.C., Hinks, R.S., Tikofsky, R.S., Hyde, J.S., 1992. Time course EPI of human brain function during task activation. Magn Reson Med 25, 390–397.

Bashashati, A., Fatourechi, M., Ward, R.K., Birch, G.E., 2007. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. J Neural Eng 4, R32–R57.

Baumann, O., Belin, P., 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol Res74, 110–120.

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., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Bennett, S., Weinberg, B., 1979. Acoustic correlates of perceived sexual identity in preadolescent children's voices. J Acoust Soc Am 66, 989–1000.

Besserve, M., Jerbi, K., Laurent, F., Baillet, S., Martinerie, J., Garnero, L., 2007. Classification methods for ongoing EEG and MEG signals. Biol Res 40, 415–437.

Birbaumer, N., 2006. Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. Psychophysiology 43, 517–532.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.-R., 2010. Single-trial analysis and classification of ERP components — A tutorial. NeuroImage 1–12.

Bonte, M., Frost, M.A., Rutten, S., Ley, A., Formisano, E., Goebel, R., 2013. Development from childhood to adulthood increases morphological and functional inter-individual variability in the right superior temporal cortex. NeuroImage 83, 739–750.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Curr Biol 20, 116–120.

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

Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2013. Cerebral processing of voice gender studied using a continuous carryover fMRI design. Cereb Cortex 23, 958–966.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.

Dellwo, V., Huckvale, M., Ashby, M., 2007. How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification, in: Speaker Classification I, Lecture Notes in Computer Science. Springer.

Ethofer, T., Bretscher, J., Wiethoff, S., Bisch, J., Schlipf, S., Wildgruber, D., Kreifelts, B., 2013. Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus. NeuroImage 76, 45–56.

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P., 2009. Decoding of emotional information in voice-sensitive cortices. Curr Biol 19, 1028–1033.

Fant, G., 1960. The Acoustic Theory of Speech Production. De Gruyter Moulton.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C., 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): Use of a cluster-size threshold. Magn Reson Med 33, 636–647.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008a. "Who" is saying "what?" Brain-based decoding of human voice and speech. Science 322, 970–973.

Formisano, E., De Martino, F., Valente, G., 2008b. Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magn Reson Imaging 26, 921–934.

Friston, K.J., 1995. Statistical parametric maps in functional imaging : A general linear approach. Hum Brain Mapping 2, 189–210.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D., Duchaine, B., 2009. Developmental phonagnosia: a selective deficit of vocal identity recognition. Neuropsychologia 47, 123–131.

Genovese, C.R., Lazar, N.A., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870–878.

Guyon, I., Elisseeff, A.2., 2003. An introduction to variable and feature selection. J Mach Learn Res 3, 1157–1182.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene selection for cancer classification using support vector machines. Mach Learn 46, 389–422.

Hanson, S.J., Halchenko, Y.O., 2008. Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. Neural Comput 20, 486–503.

Harnad, S., 1987. Categorical Perception: The groundwork of cognition. Cambridge University Press.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat Rev Neurosci 7, 523–534.

Howard, M.F., Poeppel, D., 2010. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. J Neurophysiol 104, 2500–2511.

Kerlin, J.R., Shahin, A.J., Miller, L.M., 2010. Attentional gain control of ongoing cortical speech representations in a "Cocktail Party". J Neurosci 30, 620–628.

Klatt, D.H., Klatt, L.C., 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. J Acoust Soc Am 87, 820–857.

Kohonen, T., 2001. Self-organizing maps. Springer.

Kwong, K.K., Belliveau, J.W., Chesler, D.A., Goldberg, I.E., Weisskoff, R.M., Poncelet, B.P., Kennedy, D.N., Hoppel, B.E., Cohen, M.S., Turner, R., 1992. Dynamic magnetic resonance imaging of human brain activity during primary sensory stimulation. Proc Natl Acad Sci U S A 89, 5675–5679.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Front Psychology 2, 175.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex 21, 2820–2828.

Latinus, M., McAleer, P., Bestelmeyer, P.E.G., Belin, P., 2013. Norm-based coding of voice identity in human auditory cortex. Curr Biol 23, 1075–1080.

Lavner, Y., Gath, I., Rosenhouse, J., 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Commun 30, 9–26.

Leopold, D.A., Bondar, I.V., Giese, M.A., 2006. Norm-based face encoding by single neurons in the monkey inferotemporal cortex. Nature 442, 572–575.

Leopold, D.A., O'Toole, A.J., Vetter, T., Blanz, V., 2001. Prototype-referenced shape encoding revealed by high-level aftereffects. Nat Neurosci 4, 89–94.

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., Formisano, E., 2012. Learning of new sound categories shapes neural response patterns in human auditory cortex. J Neurosci 32, 13273–13280.

Loffler, G., Yourganov, G., Wilkinson, F., Wilson, H.R., 2005. fMRI evidence for the neural representation of faces. Nat Neurosci 8, 1386–1391.

Logothetis, N.K., 2002. The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal. Philos Trans R Soc Lond B Biol Sci 357, 1003–1037.

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157.

Lotte, F., Congedo, M., Lécuyer, A., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. J Neural Eng 4, R1–R13.

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

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.

Murry, T., Singh, S., 1980. Multidimensional analysis of male and female voices. J Acoust Soc Am 68, 1294–1300.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci 10, 424–430.

Ogawa, S., Tank, D.W., Menon, R., Ellermann, J.M., Kim, S.G., Merkle, H., Ugurbil, K., 1992. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. Proc Natl Acad Sci U S A 89, 5951–5955.

Papcun, G., 1989. Long-term memory for unfamiliar voices. J Acoust Soc Am 85, 913–925.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: A tutorial overview. NeuroImage 45, S199–S209.

Perry, T.L., Ohde, R.N., Ashmead, D.H., 2001. The acoustic bases for gender identification from children's voices. J Acoust Soc Am 109, 2988–2998.

Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., Logothetis, N.K., 2008. A voice region in the monkey brain. Nat Neurosci 11, 367–374.

Pfurtscheller, G., Lopes da Silva, F.H., 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 110, 1842–1857.

Rieger, J.W., Reichert, C., Gegenfurtner, K.R., Noesselt, T., Braun, C., Heinze, H.-J., Kruse, R., Hinrichs, H., 2008. Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. NeuroImage 42, 1056–1068.

Schultz, T., 2007. Speaker Characteristics, in: Speaker Classification I, Lecture Notes in Computer Science. Springer.

Smith, D.R.R., Patterson, R.D., 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J Acoust Soc Am 118, 3177–3186.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol 19, 498–502.

Stevens, K., 1999. Acoustic Phonetics, Current studies in linguistics. MIT Press.

Titze, I.R., 2008. Nonlinear source–filter coupling in phonation: Theory. J Acoustic Soc Am 123, 2733-2749.

Tong, F., Pratte, M.S., 2012. Decoding patterns of human brain activity. Annu Rev Psychol 63, 483–509.

Turner, R.E., Walters, T.C., Monaghan, J.J.M., Patterson, R.D., 2009. A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. J Acoust Soc Am 125, 2374–2386.

Van Dommelen, W.A., 1990. Acoustic parameters in human speaker recognition. Lang Speech 33, 259–272.

van Gerven, M., Hesse, C., Jensen, O., Heskes, T., 2009. Interpreting single trial data using groupwise regularisation. NeuroImage 46, 665–676.

Van Lancker, D.R., Kreiman, J., Cummings, J., 1989. Voice perception deficits: neuroanatomical correlates of phonagnosia. J Clin Exp Neuropsychol 11, 665–674.

Wolpaw, J., Birbaumer, N., McFarland, D., 2002. Brain-computer interfaces for communication and control. Clin Neurophysiol 113, 767–791.

# Chapter 2

# Pattern Analysis of EEG Responses to Speech and Voice: Influence of Feature Grouping

*Corresponding publication:*

Hausfeld, L., De Martino, F., Bonte, M.,& Formisano, E. (2012). Pattern Analysis of EEG Responses to Speech and Voice: Influence of Feature Grouping. *Neuroimage*, 59(4), 3641-3651.

## Abstract

Pattern recognition algorithms are becoming increasingly used in functional neuroimaging. These algorithms exploit information contained in temporal, spatial, or spatio-temporal patterns of independent variables (features) to detect subtle but reliable differences between brain responses to external stimuli or internal brain states. When applied to the analysis of electroencephalography (EEG) or magnetoencephalography (MEG) data, a choice needs to be made on how the input features to the algorithm are obtained from the signal amplitudes measured at the various channels. In this article, we consider six types of pattern analyses deriving from the combination of three types of feature selection in the temporal domain (predefined windows, shifting window, whole trial) with two approaches to handle the channel dimension (channel wise, multi-channel). We combined these different types of analyses with a Gaussian Naïve Bayes classifier and analyzed a multi-subject EEG data set from a study aimed at understanding the task dependence of the cortical mechanisms for encoding speaker's identity and speech content (vowels) from short speech utterances (Bonte, Valente, & Formisano, 2009). Outcomes of the analyses showed that different grouping of available features helps highlighting complementary (i.e. temporal, topographic) aspects of information content in the data. A shifting window/multi-channel approach proved especially valuable in tracing 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. Because it exploits the high temporal resolution of EEG (and MEG), such a shifting window approach with sequential multi-channel classifications seems the most appropriate choice for tracing the temporal profile of neural information processing.

## Introduction

Electroencephalography (EEG) and magnetoencephalography (MEG) are commonly used to study the time course of neural information processing in the human brain with high temporal resolution. In most cases, EEG/MEG studies rely on the comparison of averaged responses to repeated presentations of experimental conditions either in the temporal domain (event-related potentials [ERPs] or fields [ERFs], respectively) and/or in the frequency domain (event-related desynchronization and synchronization) (Pfurtscheller and Lopes Da Silva, 1999). Often, the statistical analyses (and related inferences on neural processing) are limited to a-priori specified (spectro-) temporal windows of interest – at channel or estimated source level - and therefore only a small subset of the measured signal is actually utilized.

This article illustrates several approaches to EEG data analysis based on pattern recognition (e.g. Bishop, 2007; Duda et al., 2001). In contrast to the conventional approach where a single dependent variable is examined (univariate statistics), these techniques exploit the information content in patterns of dependent variables (features), which are extracted from the measured signals. Pattern recognition allows analyzing EEG data in a more exploratory and data-driven manner and - similar to the recent developments in fMRI (e.g. Haynes and Rees, 2006) - promises to complement conventional approaches for EEG/MEG analysis.

A typical application of pattern recognition methods includes three steps, (1) extracting and selecting features (i.e. dependent variables), (2) learning a model with a machine-learning algorithm, and (3) determining the generalization ability of the learnt model using an independent evaluation dataset. In EEG/MEG, various *types* of features can be considered, ranging from signal amplitude in the temporal domain (e.g. Rieger *et al.*, 2008) to power or phase information in the frequency domain (Kerlin *et al.*, 2010; Luo and Poeppel, 2007; Rieger *et al.*, 2008). Specific transformations, such as wavelet coefficients (Åberg and Wessberg, 2007; Rieger *et al.*, 2008), and coherence measures (Besserve *et al.*, 2007) can also be used. Furthermore, features can be differently *grouped* in the (spectral-) temporal and spatial domain. For example, limiting the information to pre-defined temporal windows of interest is essential to many realizations of EEG-based *brain-computer interface* (BCI) systems (e.g. Birbaumer, 2006; Blankertz *et al.*, 2011; Wolpaw *et al.*, 2002). Alternatively, the information contained in a sliding time interval of EEG data can be used, e.g. to detect the occurrence of seizures in epileptic subjects (Schad *et al.*, 2008). Concerning the spatial (channel) domain, many BCI systems employed spatial filters (i.e. linear combinations of channels; see Blankertz *et al.*,

2011) to enhance performances. For the same reason sophisticated feature selection or reduction methods were applied in BCI systems (see Bashashati *et al.*, 2007).

Several machine-learning algorithms have been used to learn the relation between selected features of the EEG/MEG data and experimental labels. These algorithms include simple correlation (e.g. Luo and Poeppel, 2007), support vector machines (SVMs) (Vapnik, 1995), linear discriminants analysis (LDA) (e.g. Duda *et al.*, 2001), and neural networks or Bayesian approaches (Bishop, 2007). Most frequently, learning algorithms are based upon linear models (e.g. Lotte *et al.*, 2007; Rieger *et al.*, 2008; van Gerven *et al.*, 2009) due to their fast computation, robustness and simplicity of results interpretation.

To determine the generalization ability of the computed model, an independent set of test data is required. This can be done at single-subject level, splitting the measured data into training and testing sets (e.g. Luo and Poeppel, 2007) or across subjects, using a subjects' subset for training and the other for evaluating the generalization performance (e.g. Kerlin *et al.*, 2010).

In this study, we consider and evaluate the effects of differently combining and grouping the features in the temporal (*predefined windows*, *shifting window*, *whole trial*) and channel domain (*single channel*, *multichannel*) in the context of a neuro-cognitive EEG paradigm. Using *Gaussian Naïve Bayes* (GNB; Mitchell, 1997) classification, we analyze data from an auditory EEG study aimed at understanding the task dependence of the cortical mechanisms underlying the processing of voice and speech identification (Bonte *et al.*, 2009) and illustrate the results of each possible feature combination in the temporal and channel domain.

## Materials and Methods

Machine-learning approaches for the analysis of neuroimaging data require single trials to be described by an *n*-dimensional vector of features. In our approach, basic features are defined as EEG voltages and include time (samples) and measurement channels (electrodes). In particular, we consider six types of classification analyses derived from combining three types of features grouping in the temporal domain (*predefined windows*, *shifting windows*, *whole trial)* with two approaches to handle the channel dimensions (*single channel*, *multichannel*, see Fig. 1). These different types of analyses can be combined with any classification algorithm (e.g. LDA classifier or SVMs). Here, we use a modified *Gaussian Naïve Bayes classification,* because of its simplicity which implies lower computational costs (e.g. compared to SVM classification) and interpretability of model parameters. We

examine the case of pair-wise classifications of EEG responses to simple vowels (/a/, /i/, /u/) spoken by three speakers (sp1, sp2, sp3) (see 2.3).

**Types of Feature Grouping**

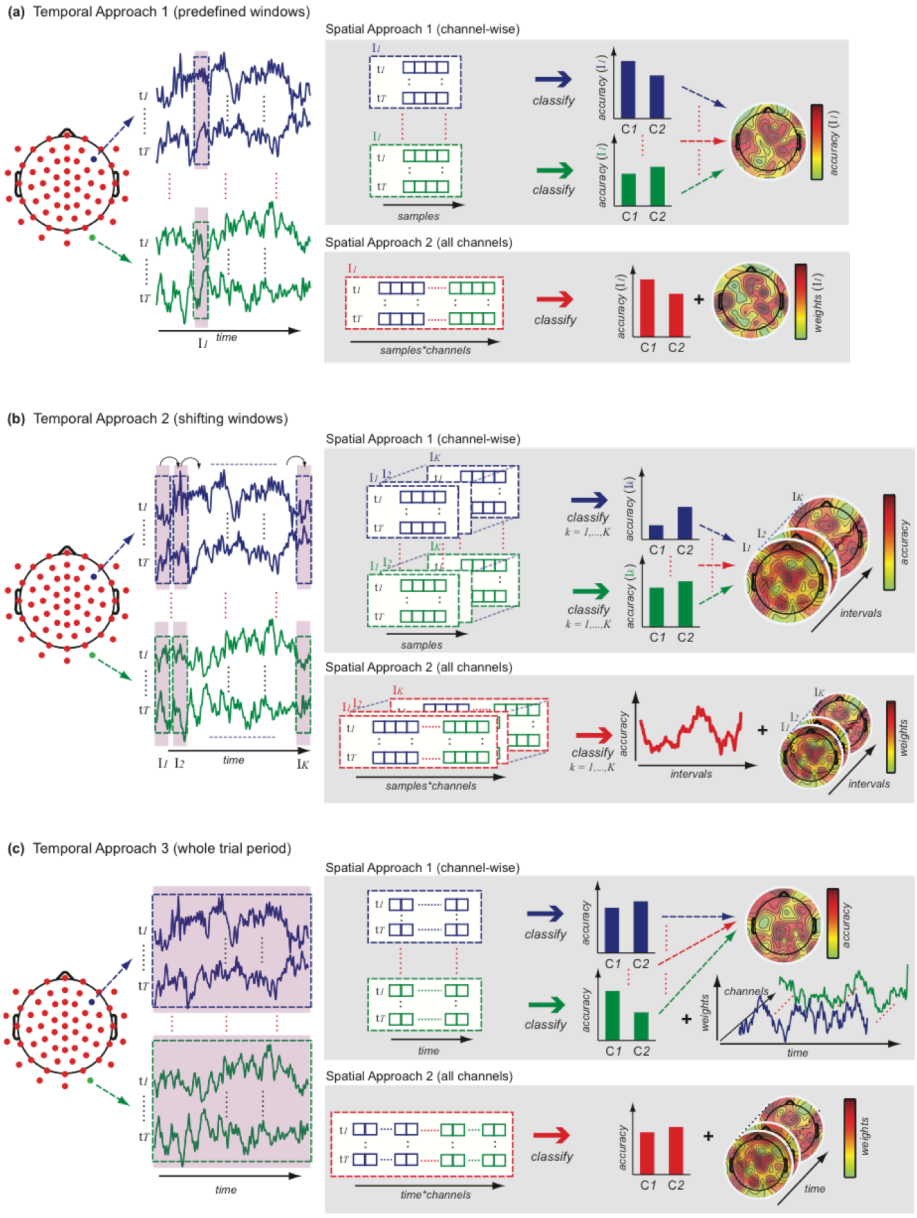
*Predefined Windows.* In the first approach, we use prior hypotheses (e.g. typical ERP windows) to select the temporal windows entering the analysis. As depicted in Fig. 1.a, the temporal samples within a specific interval are used as features to classify single trials either for each of the *K* channels (right upper panel) or for all channels (right lower panel). In the latter case, the feature set is defined by concatenating sampling points of multiple channels. In the case of a channel-by-channel analysis accuracy values are obtained for each electrode. This allows creating a topographic map of classification performance for the predefined intervals. Classifying based on features from multiple channels results in one classification accuracy value. In this case, a topographic map is created from the weights estimated during model training (see eq.4) that indicate the relevance of each electrode contribution to the classification.

*Shifting Windows.* In the second approach (Fig. 1b), the analyses are not restricted to specific latencies and are based upon features from *shifting windows* either on a channel-by-channel basis (right upper panel) or by concatenating features from multiple channels (right lower panel). Results of the single-channel approach can be depicted as a time series of topographic plots indicating classification performance. The multi-channel classification allows retrieving the information content over time (information time-course). A weight vector - indicating the relevance of individual channels - is obtained for each time window.

*Whole Trial Period.* In the third temporal approach (Fig. 1.c) all temporal samples within a trial period are used. Classifications are performed either using the channel-wise (right upper panel) or multi-channel (right lower panel) approach. Results for the channel-wise approach may be used to create a topographic map of the information content within the entire trial period. For the multichannel approach, the analysis returns an overall accuracy value. Weights are defined for each sampling point and channel and thus indicate the temporal and topographical variation of the information content.

---

**Figure 1. Overview of the Six Different Types of Classification Considered in this Study.** Different selections and groupings of data in the spatial and temporal dimension result in different types of classification. (a) *Temporal Approach 1:* Classifications are performed using signal amplitudes within predefined windows of interest, e.g. $I_1$, which need to be specified based on prior hypotheses. (b) *Temporal Approach 2:* Using *K* shifting windows ($I_1$,…, $I_K$), separate classifications are performed, which results in a time-course of the information content (accuracies) (c) *Temporal* …




… *Approach 3:* The overall information content within a trial is estimated using a single classification that employs all samples within the trial period. In addition to the different types of temporal grouping, the spatial dimension can be accounted for either by performing separate classifications at each channel (*Spatial Approach 1, left panels*) or concatenating all channels (*Spatial Approach 2, right panels*), which results in a single classification. $t_1$, …, $t_T$ denote trials of EEG signals, measured at each recording channel. See text for detailed information.

**Gaussian Naïve Bayes Classification**

We report below a short description of GNB classification with reference to EEG data; see Mitchell (1997), for a more complete and general formulation of this algorithm.

Let us consider a supervised learning problem in which we wish to approximate the function $f: X \rightarrow C$ or equivalently $P(C|X)$, where *C* is a Boolean random variable representing the categories in our classification problem and $X = \langle x_1, \ldots, x_n \rangle$ is a n-dimensional feature vector obtained from the EEG data. Using Bayes rule we can write:

$$P(C = c_m | x_1, \ldots, x_n) = \frac{P(x_1,\ldots,x_n|C=c_m)P(C=c_m)}{\sum_j P(x_1,\ldots,x_n|C=c_j)P(C=j)} \quad (1)$$

where $c_m$ represents the $m^{th}$ category. One way to learn $P(C|X)$ is to use the training data to estimate $P(X|C)$ and $P(C)$ and then use eq. 1 to classify any new instance of *X*. The Naïve term is introduced when in the estimation of $P(C|X)$ the *n* features are assumed to be conditionally independent and eq. 1 can be written as:

$$P(C = c_m | x_1, \ldots, x_n) = \frac{\prod_i^n P(x_i|C=c_m)P(C=c_m)}{\sum_j \prod_i^n P(x_i|C=c_j)P(C=c_j)}. \quad (2)$$

Following eq.2 and having estimated $P(x_i|C)$ and $P(C)$ from the training data, any new EEG trial $Y_{new} = \langle y_1, \ldots, y_n \rangle$ can be classified following:

$$C \leftarrow argmax\, P(C = c_m) \prod_{i=1}^{n} P(y_i|C = c_m), \quad (3)$$

where argmax returns the class ($c_m$) with highest probability given $Y_{new}$. In spite of the naïve assumption, the GNB was shown to perform well for many examples of neuroimaging datasets and to be fast and robust (e.g. Pereira *et al.*, 2009). To solve the multi-class classification problem we used a one-versus-one approach, which reduced the problem to a series of binary (2-class) classifications (see 2.4). We furthermore assumed equal covariance matrices of the two classes in the estimation of $P(x_i|C)$. This allowed us to pool the training data set of the two classes in the estimation of the variance of the classes. In order to derive the relative importance of features in the classification problems we estimated weights for each of the n features as:

$$w_i = \frac{1}{\sigma_i^2}(\mu_i^+ - \mu_i^-) \quad (4)$$

where $\sigma_i^2$ represents the estimated variance and $\mu_i^+$ ($\mu_i^-$) represent the mean of the two classes (+, -) for the $i^{th}$ feature (Pereira *et al.*, 2009). For visualization purposes, weights were further transformed by a ranking procedure (values ranged from 1-100 with 1 representing the lowest and 100 the highest weight).

**EEG Experiment and Data**

We illustrate the different types of classification analyses in the context of a recent auditory EEG study aimed at understanding the timing and mechanisms of cortical processing of voice and speech (Bonte *et al.*, 2009). For reader's convenience, essential information on experimental design, EEG measurements and data pre-processing is reported below. A more detailed description can be found in Bonte *et al.* (2009).

*Participants.* Fourteen Dutch undergraduate students (8 female; 1 left-handed) took part in this study. No history of hearing losses or neurological abnormalities was reported. Participants gave their informed consent and received course credits or payment for participation. The study was approved by the Ethical Committee of the Faculty of Psychology at the University of Maastricht.

*Stimuli.* Stimuli were speech sounds of three Dutch vowels (/a/, /i/, and /u/) uttered by three native Dutch speakers (sp1: female, sp2: male, sp3: female). To introduce acoustic variability, for each vowel and each speaker three different tokens were recorded. Stimulus length was equalized to 230ms. Sound intensity levels were equalized by matching RMS values. For analysis, stimuli were either grouped according to speaker identity (*speaker grouping*) ignoring the vowel dimension or according to vowels (*vowel grouping*) ignoring the speaker dimension.

*Experimental Design and Procedure.* Task dependent processing was induced by introducing one-back tasks on either speaker or vowel identity (*speaker and vowel task*). A passive task denoting passive listening of the stimuli was also included but not used in our analyses. For the active tasks, subjects were instructed to respond with a button press every time that the same vowel (vowel task) or the same speaker (speaker task) was presented in two subsequent trials (target trials), which occurred in 6.25% of all trials. Trials including targets, and/or button responses (correct responses, omissions, false positives) were not included in the analysis. Each task involved two blocks amounting to a total of 450 non-target trials. Stimulus onset asynchrony varied between 3.0 and 3.5s. All subjects participated for two EEG sessions and performed either two passive blocks followed by two speaker task blocks or two passive blocks and two vowel task blocks. The order of sessions was counterbalanced across subjects. Before the speaker and vowel tasks a short practice session assured that participants understood the task.

*EEG Recording and Preprocessing.* Data were recorded (sampling rate: 250Hz) in an electrically shielded and sound attenuating room from 61 equidistant electrode positions (Easycap, Montage No.10) relative to left mastoid reference. Impedance levels were kept below 5k**Ω**. Artifacts were removed in two steps. First, artifacts

like high-amplitude, high frequency muscle noise, swallowing, or electrode cable movements were rejected. Second, eye-blinks, eye movements, heartbeat effects were corrected by using ICA as implemented in EEGlab (Delorme and Makeig, 2004; Makeig *et al.*, 2002). For each task, ICA components were decomposed into brain-related activity and non-brain artifacts by visual inspection. Electrode signals were recreated by using all brain-related components (speaker task: 24±4 components; vowel task: 23±4; passive task: 20±4) and baseline corrected (1s before stimulus onset) (see Bonte *et al.*, 2009). Finally, signals were recomputed using the average reference.

**EEG Data Classification**

For all different types of classification we followed a 20-fold cross-validation procedure by assigning randomly selected trials to non-overlapping training ($Train_l$; $l = 1,\ldots,20$) and testing sets ($Test_l$). In order to prevent model learning to be affected by the number of training examples, we made use of a leave-in procedure (i.e. resulting in a constant number of training trials). For each iteration $l$ the training set $Train_l$ consisted of 30 trials per condition whereas the amount of trials in $Test_l$ varied (~15) due to trial rejection. Three binary comparisons were performed for each grouping (i.e. Speaker Grouping: sp1 vs. sp2, sp1 vs. sp3, sp2 vs. sp3; Vowel Grouping: /a/ vs. /i/, /a/ vs. /u/, /i/ vs. /u/).

To evaluate classification performance, we computed the accuracy of predicting class labels for the independent test set for each binary comparison. Accuracy was defined as the percentage of correctly classified trials.

Classification performances and feature weights were averaged over the 20 folds. Accuracies and weights for the speaker and vowel grouping were obtained by averaging results of the respective three binary comparisons. Significance of classification accuracies on individual subject level was obtained with permutation testing (Golland and Fischl, 2003). The empirical null distribution was derived for each classification strategy and subject by repeating the whole classification one thousand times with permuted labels of trials in $Train_l$. In the case of *shifting window* analysis, we made use of the permutation distribution obtained for the *predefined windows* approach to avoid massive computations by computing permutations for each window. We assessed the significance for each channel and window using the channel's most conservative chance level estimation of the five windows examined in the *predefined windows* approach for the respective task and grouping.

At group level, we calculated the significance of classification using a binomial test (e.g. Darlington and Hayes, 2000) with $n = 14$ (number of subjects), $p = 0.05$, and $k$ expressing the number of subjects with a significant ($p < 0.05$) classification

performance according to individual permutation tests. Differences between stimulus groupings [*speaker grouping* - *vowel grouping*] were examined by paired t-tests on the respective classification accuracies for each task (*grouping effect*). For visualization of scalp topographies, only significant channels with at least one significant neighboring channel were considered (i.e. significant but isolated channels were not displayed).

**Parameters for Feature Extraction and Grouping**

For the analysis in pre-defined time intervals (see 2.1.1), we selected five intervals of 60 ms (i.e. N1: 80-140 ms; P2: 170-230 ms; N270: 240-300 ms; P340: 310-370 ms; LateP: 500-560 ms) based on results from Bonte and colleagues (2009) that consisted of 15-16 time samples. Classification was performed following two different strategies: 1) separately for each participant, channel and window (the feature set reduced to either 15 or 16 values per trial); 2) considering all channels together for each subject and window (leading to 915 or 976 features for each trial). In both cases we obtained classification accuracies for each subject and interval. The relevance of a single channel was either accessed by its performance (single channel classification) or averaging feature weights (multichannel approach).

For the classification analysis with *shifting windows* (see 2.1.2), we selected a window length = 60 ms, a sliding step = 10 ms and a trial period from -250 to 810 ms). Finally, the same temporal interval (-250 ms – 810 ms) was used for the *whole trial*-based classification (see 2.1.3). Both classifications were performed considering either single channels or multiple channels. Note that pre-stimulus data was included to compare results (classification performance and feature weights) of time windows containing no information to informative ones.

## Results

**Predefined Windows**

We first considered the classifications of speakers and vowels in five predefined temporal intervals (N1, P2, N270, P340, LateP). Fig. 2.a shows – for the single channel case - group classification results for *speaker* and *vowel grouping* during the *speaker* (top panels) and *vowel task* (lower panels), respectively. To estimate reproducibility across subjects, we created topographic maps depicting, at each channel, the number of subjects with a significant classification performance. For each subject, significance was assessed channel-by-channel by permutation testing and corrected to account for multiple testing [no. of channels] using false discovery rate (FDR [Benjamini and Hochberg, 1995], $q < 0.05$).

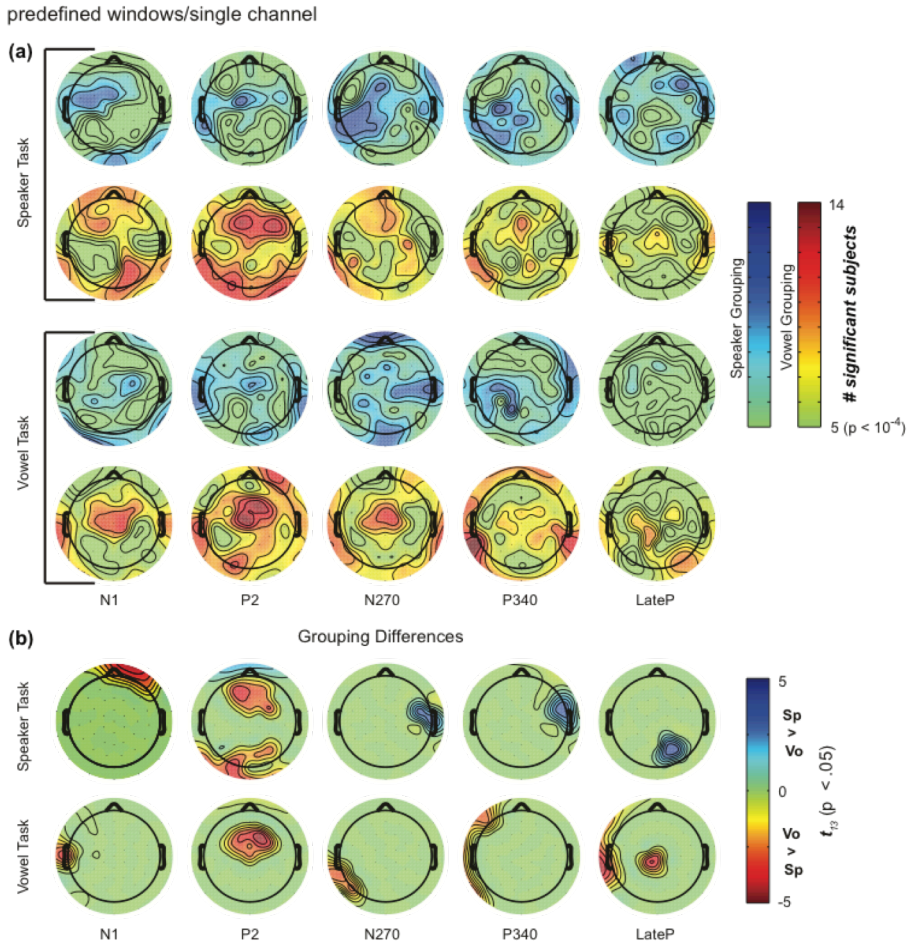



**Figure 2. Results of the Predefined Windows/Single Channel Analysis.** (a) Single channel classification performances are presented - for each interval, task and grouping - by scalp topographies depicting the number of subjects with significant classification accuracy (see text). Values for speaker and vowel grouping are depicted in blue and red colors, respectively. (b) Higher classification performance for speaker vs vowel grouping are indicated by blue and red colors for each channel. Tests were restricted to channels with a significant accuracy for one of the groupings. For visualization of scalp topographies, only channels with at least one significant neighboring channel were considered.

Differences between *speaker* and *vowel groupings* are depicted in Figure 2.b for the two tasks separately. The N1 and P2 topographic maps included several channels showing higher classification performance for the *vowel grouping* during both tasks, with early left lateralization n in the case of the *vowel task,* The later intervals N270, P340 and LateP were characterized by better classification performances for the

dimension relevant for the task. During the *speaker task* better speaker discrimination was observed for right temporo-parietal (N270, P340) and right occipito-parietal channels (LateP). Higher accuracy values for the vowel grouping during the *vowel task* were found at left lateral (N270, P340, LateP) and parietal channels (LateP).

Fig. 3 shows results obtained when features extracted from all channels were employed. Group averaged accuracy values for *speaker* and *vowel groupings* and both tasks are presented in Fig. 3.a together with the average 95% confidence intervals, resulting from permutation tests at single-subject level. Corresponding weight differences between groupings are presented as topographic maps in Fig. 3.b for each of the two tasks. Classification performances for the two groupings and tasks for most of the windows were small but above chance. Largest average accuracies were found in the P2 interval for the classification of vowels both during the *speaker* and *vowel task*. Within this window, a significantly higher accuracy was observed during the *vowel task* in the classification of vowels compared to the classification of speakers (paired t-test, $p = 0.026$). For both tasks the topographies of weight differences were comparable to single channel accuracy differences (Fig. 2.b) but possessed additional channels being more relevant during one of the groupings especially for the task irrelevant dimension (i.e. during the speaker task for vowel grouping and vice-versa).

**Shifting Windows**

To obtain a detailed temporal profile of speaker and vowel discrimination we conducted classification analyses using *shifting windows*. Fig. 4.a shows the results for the *shifting window/single channel* analysis. For display, different channels are arranged along the y-axis of the plot; blue and red color-coding denotes significant differences between speaker and vowel grouping ($p < 0.05$, uncorrected). In addition, topographic plots of accuracy differences (*speaker grouping - vowel grouping*) are shown below for relevant latencies. Statistical tests and color-coding were limited to channels and intervals with speaker and/or vowel classification performance above chance level (i.e. exceeding the most conservative 95% confidence interval in the prvious analysis). During the *speaker task* enhanced classification accuracies for vowels were observed between 150 and 240 ms (frontal, central, posterior channels at [150-200 ms]; frontal, parietal channels at [200-240 ms]). At [230-400 ms] and [500-730 ms] higher classification accuracies for speakers were found. Right temporo-parietal channels discriminated better between speakers during the medium latencies interval and posterior and left lateral channels showed this effect during the later intervals. Accuracy differences for the *vowel task* were characterized by enhanced vowel discrimination at two

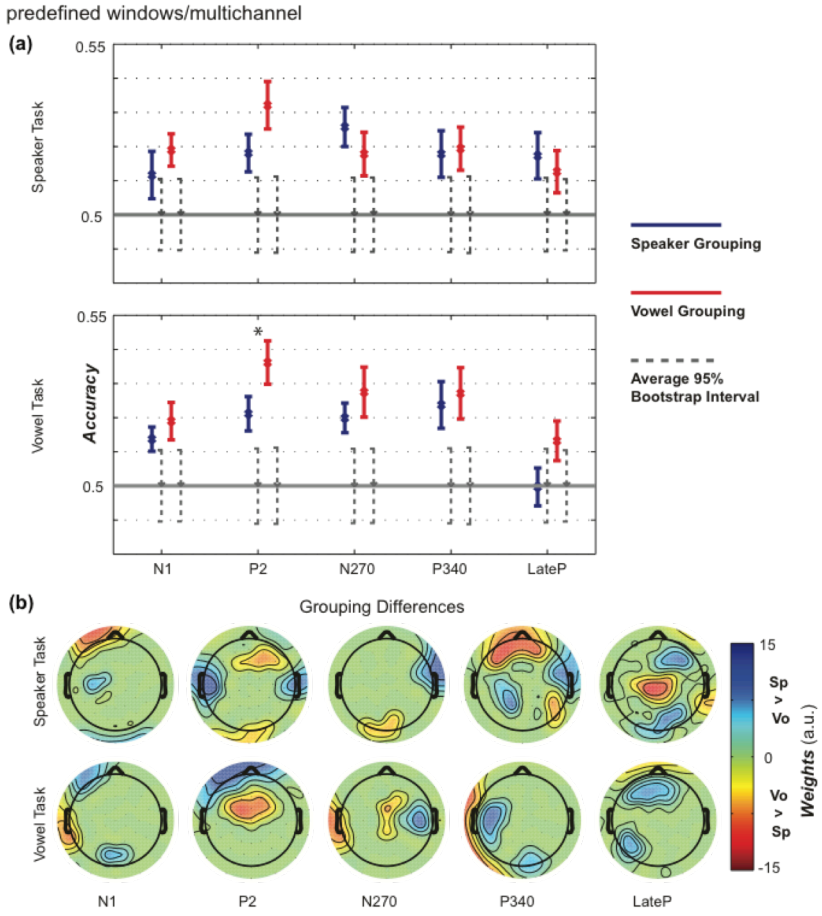



**Figure 3. Results of the Predefined Windows/Multichannel Analysis.** (a) Average classification accuracies are shown for tasks and groupings (red: speaker grouping; blue: vowel grouping) for each window separately. Bars denote average accuracies and SEM. Grey bars show average 95% CIs for individual permutation tests. (b) Weight differences of speaker and vowel grouping for each time window and task are presented in the lower panel. Note the similarities to the results for the corresponding single channel analysis (see Fig. 2.b). Only channels with at least one significant neighboring channel were considered for visualization of scalp topographies.

intervals ([120-230 ms], [450-550 ms]). During the early interval vowels were better classified at central and frontal channels. Central, left temporal and lateral channels classified vowels better than speakers during the late interval.

Using a *multichannel* approach we extracted the overall information content over time. Fig. 4.b visualizes classification performance of the *speaker* and *vowel grouping* as a function of time for the speaker (right) and vowel task (left). Accuracies were

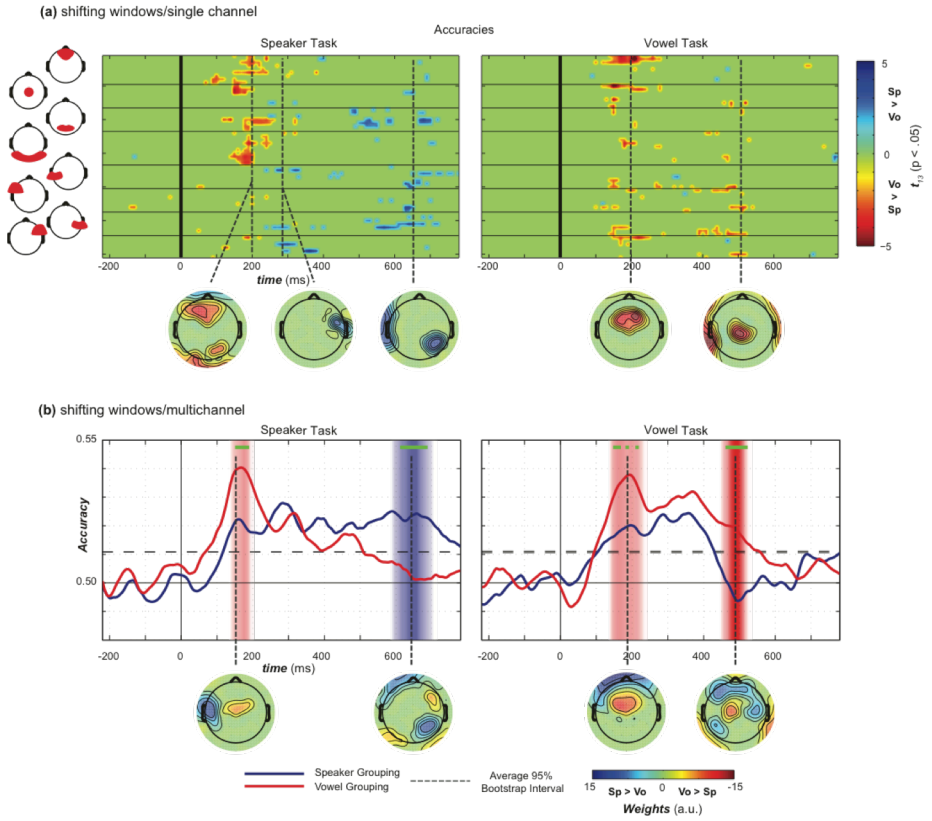



**Figure 4. Results of Shifting Window Analyses.** (a) Single channel analysis: accuracy differences between speaker and vowel grouping are presented for the speaker (left) and vowel task (right panel). For better visualization scalp topographies of accuracy differences at specific latencies indicated by dashed lines are shown below. (b) Multichannel analysis: Classification performances for shifting windows are shown for the speaker (left) and vowel task (right panel). Average accuracy for both speaker (blue) and vowel (red) grouping are depicted (blue and red shadings denote significance of higher speaker and vowel grouping, respectively). Green lines indicate grouping differences with FDR-corrected ($q < 0.1$) significance levels. Horizontal dashed lines depict the most conservative estimation of the chance level in the *predefined windows/multichannel* analysis. Topographic plots show weight differences between the speaker and vowel task at latencies with significant performance differences. For visualization of scalp topographies, only channels with at least one significant neighboring channel were considered.

defined to be above chance when the respective most conservative 95% confidence interval in the *predefined windows/multichannel* analysis was exceeded (see above). Analyses of task differences were limited to intervals with classification performance above chance for at least one grouping. Shadings denote latencies that showed significant differences between *speaker* and *vowel grouping* ($p < 0.05$, uncorrected). Grouping differences that remain significant after correcting for

multiple comparisons [FDR, $q < 0.10$] are indicated by green lines. Enhanced classification of vowels compared to speakers can be noted for an early interval for both tasks (*speaker task* [190-240 ms], *vowel task* [150-240 ms]) with similar topographies of weight differences. Task dependent effects as shown by higher speaker classification during the *speaker task* and higher vowel classification during the *vowel task* are observed at late intervals (*speaker task* [580-700 ms], *vowel task* [480-550 ms]). For the *speaker task* the enhanced speaker discrimination was accompanied by higher weights in right parietal channels whereas the enhanced vowel discrimination during the *vowel task* was characterized by higher weights at central and lateral channels.

**Whole Trial Period**

Next, we considered all samples within the trial period and performed *single channel* and *multichannel* classifications. Results for the single channel analysis are depicted in Fig. 5.a by means of topographic maps of accuracy differences between *speaker* and *vowel grouping* ($p < 0.05$, uncorrected) for both tasks. Maps were restricted to channels that were significant for at least one grouping (FDR-corrected, $q < 0.05$). Fig 5.b shows - for selected channel clusters - the temporal profile of the weights resulting for the classification of speakers and vowel. Time intervals with high values for the weights are those mostly contributing to the classification. Accuracy differences during the *speaker task* showed a left parieto-temporal and a right temporal cluster with enhanced classification performance for speakers. Weight differences for these two clusters were found to be larger for speakers at an early interval [220-260 ms] for both clusters (Fig. 5.b). A later interval [440-480 ms] showed larger weights for the *speaker grouping* for the left parieto-temporal cluster. After ~570 ms larger weights for speakers were observed for both clusters but differences were more pronounced within the right lateral cluster. For the *vowel task* three clusters (a central cluster and both a left and right posterior lateral cluster) showed enhanced accuracies for vowels compared to speakers. Early intervals showing higher weights for vowels were found for two clusters (central at [160-210 ms]; left lateral at [100-190 ms]). For later intervals at [260-300 ms] and [450-500 ms] higher weights for the vowel grouping were observed for the central and left lateral clusters (for the right lateral cluster higher weights for vowels were found at [380-500 ms]; results not shown).

Finally, classifications were performed by employing the full spatio-temporal set of features. Accuracy values for the two tasks and effects (top panel) and corresponding weight differences of selected channels between the two groupings (lower panel) are shown in Fig. 6. For each task both types of groupings were above chance level ($p < 10^{-11}$, for all task by grouping combination). When

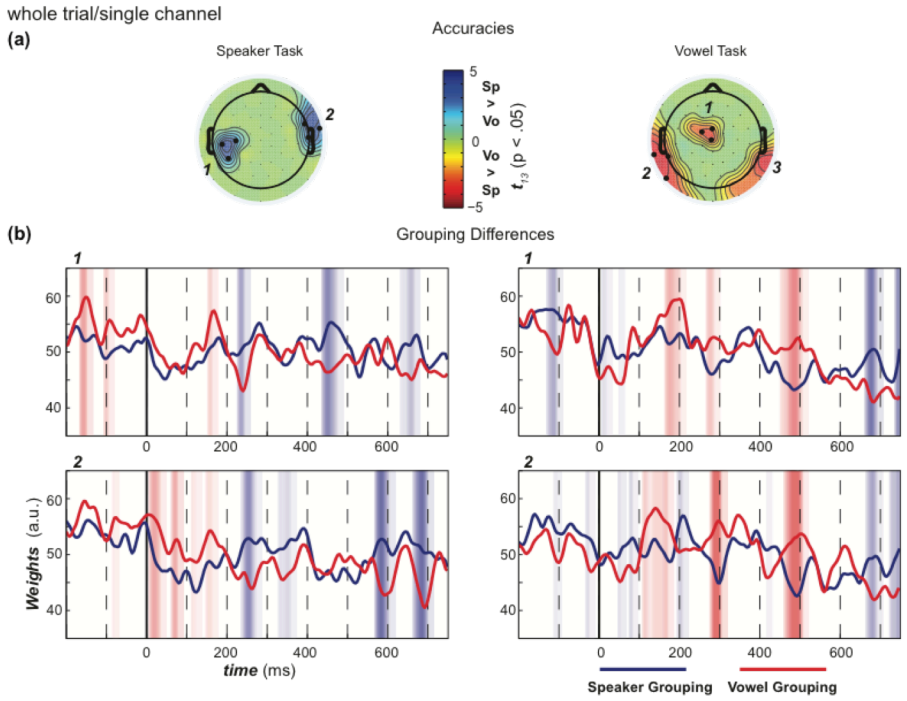



**Figure 5. Results of the Whole Trial/Single Channel Analysis.** (a) Channel-wise differences of classification performances for each task are shown in the upper panel. Tests were restricted to channels that had a significant accuracy value for one of the groupings. (b) For clusters showing significant performance differences in (a), average weight differences of speaker and vowel grouping are shown. Weights for speaker and vowel groupings are depicted in blue and red, respectively. Shadings denote larger speaker (red) and vowel (blue) weights. Only channels with at least one significant neighboring channel were considered for visualization of scalp topographies.

comparing speaker and vowel groupings, larger classification performances could be found for the vowel grouping during the vowel task ($p < 0.001$) but not during the speaker task, during which there was a trend for enhanced speaker classification (Fig. 6.a). Weight differences between speaker and vowel groupings (Fig. 6.b) revealed similar results compared to the accuracies obtained in the *shifting windows/single channel* classification with some differences.

In sum, outcomes of the *single channel* analysis for the *whole trial period* produced maps revealing the spatial distribution of classification differences between speaker and vowel groupings. These were characterized by higher classification performances for the task-relevant stimulus dimension (i.e. grouping) compared to the dimension that was not relevant for the task. A similar task-dependent effect

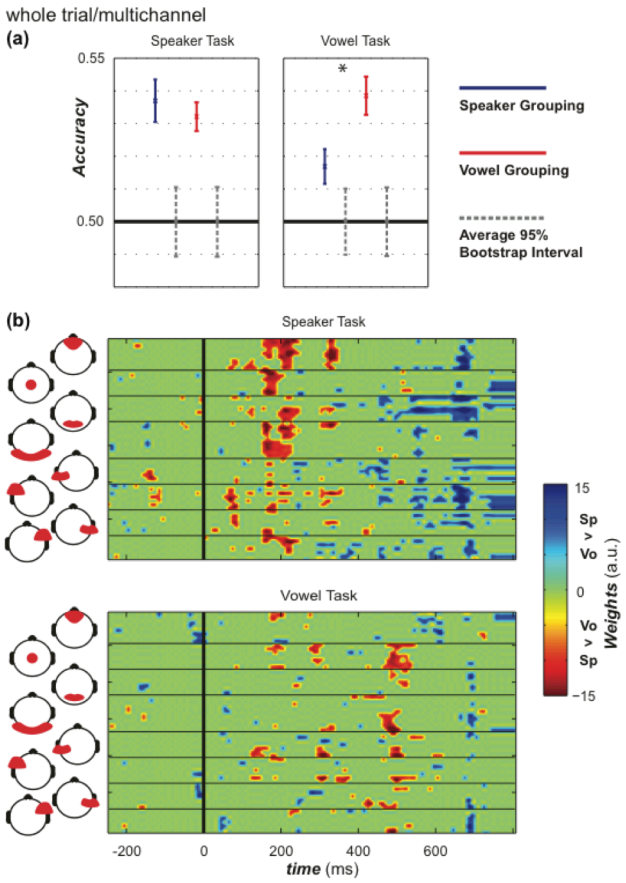



**Figure 6. Results of the Whole Trial/Multichannel Analysis.** (a) Classification performances using all sampling points and channels are presented for the speaker (left) and vowel task (right panel). Bars denote the average and SEM of classification performance across subjects. (b) Weight differences for all channels are presented by blue and red colors indicating larger weights for the speaker and vowel grouping, respectively.

was also found when the whole set of data was analyzed by means of the *multichannel* analysis.

# Discussion

## Pattern recognition and EEG Data

We have illustrated different strategies for analyzing EEG data using a pattern recognition algorithm. We have shown that it is feasible to distinguish experimental conditions above chance level, at the fine-grained level of speaker

and vowel identity. Although low, our single-trial classification accuracies were significant even at a single subject level, which indicate that –- although noisy – EEG single trial responses carry information on the neural processing of individual speech sounds. Significance was assessed with a resampling approach (permutation testing) that detects systematic classification biases and provides an empirical estimation of the chance level. In particular, our classification performances were lower than those typically reported in EEG-based BCI experiments (e.g. ~70-90% for motor imagery based BCIs [Lotte *et al.*, 2007]). There may be several reasons for this. First, our experimental paradigm included stimuli and conditions that result in largely similar EEG signals (low CNR) compared to BCI paradigms that are constructed to maximize response differences between classes. Another reason may be the data processing scheme prior to classification. Compared to our approach, analyses in BCI experiments often employ more sophisticated preprocessing and feature selection techniques (e.g. Laplacian filtering, common spatial patterns, genetic algorithms [Bashashati *et al.*, 2007; van Gerven *et al.*, 2009]). In this study, preprocessing of data included standard filtering and ICA but no further feature selection and enhancement. Univariate or multivariate feature selection algorithms – which may lead to considerable increases of accuracy values especially in high-dimensional cases – were not applied. Especially in cases of classifications with many features *(whole-trial/multichannel*), wrapper methods such as Recursive Feature Elimination (Guyon *et al.*, 2002; De Martino *et al.*, 2008) should be beneficial for both to obtain higher accuracies and select informative features.

With regard the to classification algorithm, we selected a GNB classifier that may be seen as a 'pseudo' multivariate approach, which has the advantage of providing interpretable weights (similar to *t* statistics) as it assumes independency among features (i.e. diagonal covariance matrix). Other machine-learning techniques such as linear discriminant analysis (e.g. Duda *et al.*, 2001) and support vector machines (Vapnik, 1995), which take feature correlations into account, have been previously applied to EEG datasets in the context of BCI (e.g. Bashashati *et al.*, 2007; Lotte *et al.*, 2007). The use of these or other classifiers may lead to higher accuracies compared to GNB-based classification, but this may come at the cost of the interpretability of the results and increase of computation time.

In this study, trials were classified based on EEG time courses. However, representing trials by means of event-related (de)synchronization (Pfurtscheller and Lopes da Silva, 1999) or measures of coherence and synchrony (e.g. Besserve *et al.*, 2007; Bonte *et al.*, 2009; Varela *et al.*, 2001) and classifying those may provide complementary and more detailed information (e.g. phase estimates, band-pass filtered signals and wavelet coefficients have been employed by Luo and Poeppel

[2007], Kerlin *et al.* [2010], and Rieger *et al.* [2008], respectively). Finally, our analyses could only detect effects that were strictly time-locked to the stimulus. Thus, single trials of the same condition that differed in terms of latencies could not be accurately classified. As this may be a relevant aspect in EEG/MEG, it is desirable - for future extensions of the proposed method - to include classification schemes that account for possible latency differences across trials.

**Types of Classification**

The focus of the present study was on examining how grouping of features in the temporal and spatial (channel) domain influences the results of EEG classification analyses (see Table 1). With regard to the temporal domain, a choice needs to be made between a hypothesis-driven analysis limited to a few temporal windows of interest (i.e. *predefined windows approach* which is closest to conventional ERP analyses) and a data-driven analysis with feature sets consisting of signal amplitude at all time points within a trial. This *whole-trial* approach promises to be more sensitive as several ERP components may contribute to distinguishing between two experimental conditions (Blankertz *et al.*, 2011). As a possible alternative, we also examined a shifting-window approach with multiple sequential classifications that - compared to the whole trial approach -possesses the advantage of assessing information content over time. Regarding the spatial domain, the choice is between performing multiple classifications channel-by-channel and a single classification using all channels simultaneously.

In general, a multichannel and whole-trial approach seems desirable as it does not rely on previous assumptions and enables the pattern recognition algorithm to fully exploit information contained in both the topographic and temporal

Table 1. Qualitative Comparison of Classification Approaches

| | **Predefined Windows** | | **Shifting Windows** | | **Whole Trial** | |
|---|---|---|---|---|---|---|
| | Single Channel | Multiple Channels | Single Channel | Multiple Channels | Single Channel | Multiple Channels |
| A-priori selection | Yes | Yes | No | No | No | No |
| Time course | - | - | Accuracies | Accuracies | Weights | Weights |
| Topography | Accuracies | Weights | Accuracies | Weights | Accuracies | Weights |
| Amount of tests | #channels by #windows (305)[a] | #windows (5) | #channels by #windows (6161) | #windows (101) | #channels (61) | 1 (1) |

[a]The numbers in brackets denote the amount of statistical tests required in this particular study with 61 channels, 5 predefined time windows, and 101 shifting windows.

distribution of signal amplitudes. Furthermore, such a data-driven approach relies on a single classification, thus avoiding the problem of multiple comparisons, which applies – at different extents – to all other combinations (see Table 1: *Amount of tests*).

Results of our analyses, however, highlighted several aspects that need to be considered when using this type of approach. When using all available features in a single classification (*whole-trial/multi-channel*), detection of both informative time windows and topographies is based on feature weights whereas a single accuracy value describes the overall information content. As illustrated in Figure 6, this approach detected the general effect on accuracy of the *vowel task* for the classification of vowels compared to speakers (Fig. 6.a, right panel) but failed to detect the expected opposite modulation for the speaker task (Fig. 6.a, left panel). Furthermore, the interpretation (and statistical testing) of weights to derive topographical and temporal information is not straightforward (Fig. 6.b). Especially when the number of features is very large, estimates of weights may be noisy. In case of SVM or LDA –based classification, additional issues may arise. For instance, Blankertz *et al.* (2011) describe a hypothetical case where high weights (as determined by a LDA classifier) are associated with one channel that does not contribute any class-related information. At the cost of increasing the number of classifications, the number of features can also be reduced by using a *whole trial/single channel* approach (Fig. 5). The analysis resulted in neurophysiologically plausible accuracy-based topographic maps that clearly highlight the task dependence of the informative neuronal sources but could only roughly indicate which intervals are relevant (Fig. 5.b).

Our results for the *shifting window* approaches (Fig. 4) indicate that these are the most appropriate for tracing information content over time. In fact, both single and multi-channel analyses were able to detect – without prior hypotheses on the temporal windows - the early and task-independent processing of vowels, which becomes maximal at ~200ms (corresponding to P2). This is in accordance with the idea that an early stimulus-driven analysis processes – by default – acoustic features which are informative of speech content, like first or second formant frequencies (e.g. Obleser *et al.*, 2004; Bonte *et al.* 2009). Although less significantly, our results additionally show that speaker identity information is present at similar latencies indicating bottom-up processing of speaker-relevant acoustic features, like fundamental frequency and timbre (Belin *et al.*, 2004; Charest *et al.*, 2009; Bonte *et al.*, 2009).

Later task dependent processing (~280ms), expressed by enhanced classification performance (*single channel* analysis) or higher weights (*multichannel*

approach) for the task-relevant dimension of stimuli, was found to occur mostly at right (speakers) or left lateralized (vowels) channels, which is in accordance with earlier studies (Belin and Zatorre, 2003; Formisano *et al.*, 2008; Hickok and Poeppel, 2007; van Kriegstein and Giraud, 2004). Additionally, a late task-dependent effect between 450 and 700ms after sound onset was detected that is most likely related to the memory maintenance of the relevant information for performing correctly the one-back task.

In the case of the *single channel/shifting window* analysis the amount of multiple testing is highest and statistical testing would require a proper correction. Using the Bonferroni approach is known to result in over-conservative corrections. A proper correction requires an empirical estimate of the likelihood that $k$ consecutive windows are significant by chance, which in turn requires permutation testing for each channel and time window. The computational load for this is very high, however it is becoming tractable thanks to the increasing availability of parallel processing. The number of classifications is greatly reduced with a multiple channel approach, which thus seems the most viable choice for tracing the temporal profile of information content also because the number of features considered in each classification is not excessively large (corresponds to the number of channels). As a consequence of the reduced number of tests, early and late effects on speaker and vowel grouping could be still detected after correcting for multiple testing (FDR) for the *multichannel* but not the *single channel* approach (Fig. 4).

## Conclusions

We have illustrated different ways of analyzing EEG data by means of a pattern classification algorithm. Outcomes of the analyses show that grouping or separating available features (channels, time windows) helps highlighting different aspects of information content in the data. Because of the high temporal resolution of EEG (and MEG) a shifting window approach with sequential multi-channel classifications proved to be the most valuable as it allows tracing the temporal evolution of stimulus and task-related neural information processing.

**Acknowledgements**

Financial support by the Netherlands Organization for Scientific Research, Innovative Research Incentives Scheme VENI Grant 451-07-002 (MB) and VIDI Grant 452-04-330 (EF) is gratefully acknowledged. We thank Giancarlo Valente for comments and discussions.

## References

Åberg, M. C., Wessberg, J., 2007. Evolutionary optimization of classifiers and features for single-trial EEG Discrimination. Biomed Eng Online 6, 32.

Bashashati, A., Fatourechi, M., Ward, R. K., Birch, G. E., 2007. A survey of signal processing algorithms in brain–computer interfaces based on electrical brain signals. J Neural Eng 4, R32-57.

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, 2105-2109.

Benjamini, Y., Hochberg, Y., 1995. Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57, 289-300.

Besserve, M., Jerbi, K., Laurent, F., Baillet, S., Martinerie, J., Garnero, L., 2007. Classification methods for ongoing EEG and MEG signals. Biol Res 40, 415-437.

Birbaumer, N., 2006. Breaking the silence: Brain-computer interfaces (BCI) for communication and motor control. Psychophysiol 43, 517-532.

Bishop, C. M., 2007. Pattern Recognition and Machine Learning (Information Science and Statistics) (2nd ed.). Springer, New York.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., Müller, K.-R., 2011. Single-trial analysis and classification of ERP components - a tutorial. NeuroImage 56, 814-25.

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.

Charest, I., Pernet, C. R., Rousselet, G. a, Quiñones, I., Latinus, M., Fillion-Bilodeau, S., et al., 2009. Electrophysiological evidence for an early processing of human voices. BMC Neurosci 10, 127.

Darlington, R. B., Hayes, a F., 2000. Combining independent p values: extensions of the Stouffer and binomial methods. Psychol Methods 5, 496-515.

Duda, R. O., Hart, P. E., Stork, D. G. 2001. Pattern Classification: Pattern Classification (2nd ed.). John Wiley & Sons, New York.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322, 970-973.

van Gerven, M., Farquhar, J., Schaefer, R., Vlek, R., Geuze, J., Nijholt, A., Ramsey, N., et al., 2009. The brain–computer interface cycle. J Neural Eng 6, 041001.

Golland, P., Fischl, B., 2003. Permutation tests for classification: Towards statistical significance in image-based studies. In C. Taylor & J. Noble (Eds.), Information Processing in Medical Imaging (Vol. 2732, pp. 330-341). Springer, Berlin/Heidelberg.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V., 2002. Gene Selection for Cancer Classification using Support Vector Machines. Mach Learn 46, 389-422.

Haynes, J., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat Rev Neurosci 7, 523-534.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat Rev Neurosci 8, 393-402.

Kerlin, J. R., Shahin, A. J., Miller, L. M., 2010. Attentional gain control of ongoing cortical speech representations in a "cocktail party". J Neurosci 30, 620-628.

Kriegstein, K. V., Giraud, A., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22, 948-955.

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., Arnaldi, B., 2007. A review of classification algorithms for EEG-based brain–computer interfaces. J Neural Eng 4, R1-13.

Luo, H., Poeppel, D., 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54, 1001-1010.

Makeig, S., Westerfield, M., Jung, T., Enghoff, S., Townsend, J., Courchesne, E., Sejnowski, T. J., 2002. Dynamic brain sources of visual evoked responses. Science 295, 690-694.

Mitchell, T. (1997). Machine Learning (1st ed.). Mcgraw-Hill.

Obleser, J., Elbert, T., Eulitz, C., 2004. Attentional influences on functional mapping of speech sounds in human auditory cortex. BMC Neurosci 5, 24.

Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: A tutorial overview. NeuroImage 45, S199-S209.

Pfurtscheller, G., Lopes da Silva, F. H., 1999. Event-related EEG/MEG synchronization and desynchronization: basic principles. Clin Neurophysiol 110, 1842-1857.

Rieger, J. W., Reichert, C., Gegenfurtner, K. R., Noesselt, T., Braun, C., Heinze, H., Kruse, R., et al., 2008. Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. NeuroImage 42, 1056-1068.

Schad, A., Schindler, K., Schelter, B., Maiwald, T., Brandt, A., Timmer, J., Schulze-Bonhage, A., 2008. Application of a multivariate seizure detection and prediction method to non-invasive and intracranial long-term EEG recordings. Clin Neurophysiol 119, 197-211.

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory (1st ed.). Springer.

Varela, F., Lachaux, J., Rodriguez, E., Martinerie, J., 2001. The brainweb: Phase synchronization and large-scale integration. Nat Rev Neurosci 2, 229-239.

Wolpaw, J. R., Birbaumer, N., McFarland, D. J., Pfurtscheller, G., Vaughan, T. M., 2002. Brain-computer interfaces for communication and control. Clin Neurophysiol 113, 767-791.

# Chapter 3

# Multiclass fMRI Data Decoding and Visualization Using Supervised Self-Organizing Maps

*Corresponding manuscript (under review):*

Hausfeld, L., Valente, G. & Formisano, E.
Multiclass fMRI Data Decoding and Visualizations using Supervised Self-Organizing Maps.

## Abstract

When multivariate pattern decoding is applied to fMRI studies entailing more than two experimental conditions, a most common approach is to transform the multiclass classification problem into a series of binary problems. Furthermore, for decoding analyses, classification accuracy is often the only outcome reported although the topology of activation patterns in the high-dimensional features space may provide additional insights into underlying brain representations. Here we propose to decode and visualize voxel patterns of fMRI datasets consisting of multiple conditions with a *supervised* variant of self-organizing maps (SSOMs). Using simulations and real fMRI data, we evaluated the performance of our SSOM-based approach. Specifically, the analysis of simulated fMRI data with varying signal-to-noise and contrast-to-noise ratio suggested that SSOMs perform better than a *k*-nearest-neighbor classifier for medium and large numbers of features (i.e. 250 to 1000 or more voxels) and similar to support vector machines (SVMs) for small and medium numbers of features (i.e. 100 to 600 voxels). However, for a larger number of features (>800 voxels), SSOMs performed worse than SVMs. When applied to a challenging 3-class fMRI classification problem with datasets collected to examine the neural representation of three human voices at individual speaker level, the SSOM-based algorithm was able to decode speaker identity from auditory cortical activation patterns. Classification performances were similar between SSOMs and other decoding algorithms; however, the ability to visualize decoding models and underlying data topology of SSOMs promotes a more comprehensive understanding of classification outcomes. We further illustrated this visualization ability of SSOMs with a re-analysis of a dataset examining the representation of visual categories in the ventral visual cortex (Haxby et al., 2001). This analysis showed that SSOMs could retrieve and visualize topography and neighborhood relations of the brain representation of eight visual categories. We conclude that SSOMs are particularly suited for decoding datasets consisting of more than two classes and are optimally combined with approaches that reduce the number of voxels used for classification (e.g. region-of-interest or searchlight approaches).


*Keywords — fMRI; decoding; multiclass classification; self-organizing maps*

## Introduction

During the past years, the traditional voxel-wise analysis of fMRI data has been complemented by so-called Multi-Voxel Pattern Analysis (MVPA; e.g. Haynes and Rees, 2006; Norman et al., 2006). In contrast to the conventional activation-based statistical analysis, which examines each voxel separately for activation differences between experimental conditions using, for example, the General Linear Model (GLM; Friston, 1995), MVPA represents an information-based analysis (Kriegeskorte et al., 2006) that exploits information contained in spatial (i.e. multi-voxel) activation patterns.

Various decoding algorithms have been proposed in MVPA fMRI studies. These algorithms differ in terms of their complexity and include correlation-based approaches (Haxby et al., 2001), Gaussian-naïve Bayes classifiers (Mitchell et al., 2004), linear discriminant analysis (Cox and Savoy, 2003, Kriegeskorte et al. 2006), linear and non-linear support vector machines (SVMs; Cox and Savoy, 2003; LaConte et al., 2005; Mourão-Miranda et al., 2005) and sparse logistic regression (SLR; Miyawaki et al., 2008; Ryali et al., 2010; Yamashita et al., 2008). More recently, the combination of classification algorithms and voxel selection strategies (e.g. De Martino et al., 2008; Langs et al., 2011; Yamashita et al., 2008) and use of multiple classifiers has been proposed (e.g. Kuncheva and Rodríguez, 2010).

In this paper, we present an fMRI decoding approach based upon self-organizing maps (SOMs). SOMs were developed to visualize high-dimensional data by converting the topology of data points into simple geometrical relationships on a two-dimensional grid (Kohonen, 2001). By preserving only the most important topological relationships, this algorithm abstracts from high-dimensional input and provides insight into the underlying data structure. These properties rendered it an important tool for data exploration in various domains. *Supervised* SOMs (SSOMs) inherit these characteristics and extend the SOMs such that they can classify unseen samples. The decoding of fMRI data with SSOMs differs with respect to other approaches in two relevant aspects.

First, SSOMs are not restricted to binary comparisons and thus allow for inherent multiclass decoding (i.e. when the fMRI measurements include more than two conditions). Many studies that analyzed fMRI data by means of MVPA employed experimental designs with two conditions or translated multiclass problems into a series of binary comparisons (one-versus-one or one-versus-all schema) paired with post-hoc processing (e.g. majority voting) to determine the predicted class (e.g. Beauchamp et al., 2009; Cox and Savoy, 2003; Ethofer et al., 2009; Kamitani and Tong, 2005; Mourão-Miranda et al., 2006; Reddy et al., 2010;

Swisher et al., 2010; Walther et al., 2009). One reason is that one of the most applied classification algorithms in MVPA fMRI applications, the SVM, is in its basic form restricted to two class problems (but see Crammer and Singer [2002] and Weston and Watkins [1999] for multiclass SVM formulations and Martínez-Ramón et al. [2006] for decoding of fMRI data using multiclass SVM). So far, only few fMRI studies applied MVPA in a multiclass setting without employing binary comparisons and necessary post-processing. In these studies decoding was performed with naïve Bayes (Brouwer and Heeger, 2009; Mitchell et al., 2004) or neural network classifiers (Hanson et al., 2004; Polyn et al., 2005).

Second, SSOMs offer the possibility to visualize the underlying distribution of fMRI activity patterns used to train the decoding model. These visualizations can reveal the topology of the underlying activity patterns in high-dimensional feature (voxel) space. When analyzing fMRI activation patterns by MVPA, the experimenter is interested in the degree of separation between classes in feature space. In the case of two classes, the interpretation of classification results is straightforward, i.e. the accuracy of predictions describes the degree of separation of the response patterns. However, for classification of response patterns with more than two classes, focusing on overall classification accuracy or performances of many binary comparisons provides only a partial, non-intuitive view on class distributions in high-dimensional voxel space. Thus, when performing multiclass classification an additional post-processing step is usually required to infer and visualize underlying class distributions (or topology), e.g. via confusion matrices. For example, Abdi and colleagues (2009) visualized class distributions in a multiclass setting by first computing pairwise comparisons between classes and visualizing, based on these outcomes, the underlying topology using Principal Component Analysis (PCA) and a bootstrapped Multidimensional Scaling (MDS). It should be stressed that, in these cases, class distributions are deduced using the classification performance of the decoding algorithm. Conversely, SSOMs visualize the data directly and thus do not require additional post processing. Another approach that allows visualizing data in fMRI is representational similarity analysis (RSA; Kriegeskorte et al., 2008). In RSA, representational dissimilarity matrices (RDMs) are calculated using pairwise distances (e.g. correlation distance) between activity patterns evoked by distinct stimulus classes. Subsequently, topology of activity patterns can be visualized by MDS of the dissimilarity matrices. A notable distinction compared to SSOMs is that RDMs are calculated from the entire feature set with equal weighting, whereas SSOMs-based topologies rely on the optimized combination of features, which results from learning of the stimulus labels. Thus, SSOM-based visualization may lead to a better abstraction of high-dimensional activity patterns.

Below we illustrate how SSOMs can be used in the context of fMRI decoding to perform multiclass classification and visualize class topology. We present results on simulated fMRI data to illustrate the validity and usefulness of SSOM classification and visualization in the decoding of single data sets as well as in analyses involving multiple cross-validation splits and/or multiple subjects. Furthermore, the proposed method is applied to the analysis of two fMRI studies. First, we examined the performances of our SSOM-based method in a challenging case of brain-based decoding of speaker identity using new fMRI data collected while participants listened to short non-linguistic vocalizations from three speakers (*dataset 1*, 3-class problem). Second, we re-analyzed publicly available fMRI data from the Haxby et al. (2001) study, where participants were presented with pictures of eight different object categories (Haxby et al., 2001; *dataset 2,* 8-class problem). Many other classification strategies have already been tested using this dataset, which thus provides a good benchmark for evaluating our method. Furthermore, the rich variety of visual categories employed allows showcasing the added value of visualizing the class topologies using SSOMs. For both simulated and real fMRI data, we compared the classification performances of SSOMs to linear and non-linear SVMs and a *k*NN classifier.

# Methods

### Self-Organizing Maps (SOMs)

Self-organizing maps (also called Kohonen maps or networks) are a special type of neural network that was first proposed by (Kohonen, 1982). Several properties have rendered it an important tool for exploration, visualization and abstraction of high-dimensional data (Kohonen, 2001). The SOM typically consists of a two-dimensional rectangular grid of nodes or units (in the following nodes and units are exchangeable terms) each associated with a model of the high-dimensional input data. The weights of these models are iteratively adapted during a learning process, which changes models to optimally span the range of input data. SOM nodes with their associated weights organize such that similar patterns in the high-dimensional space are grouped in clusters using a non-linear competitive learning strategy (Kohonen, 2001). More specifically, nodes that are close to each other in the two-dimensional SOM after training represent patterns of the input space that are similar (i.e. have a small distance). Furthermore, regions in high-dimensional feature space that are more densely populated are represented by more nodes compared to sparse regions with few data points. These properties arise from the learning process that occurs in three main stages (Kalteh et al., 2008). First, the nodes are initialized to span the range of input values for each

dimension. Second, the competitive, unsupervised learning scheme assigns each input pattern to the closest node, i.e. the *best matching unit* (BMU), using Euclidean distance in most implementations. Then, the weight vectors of the BMU and neighboring units defined by a neighborhood function are updated to better match the input pattern (i.e. these associated models are shifted towards the input pattern). These training steps are repeated until a specified number of iterations or a convergence criterion is met (e.g. Cheng, 1997; Kohonen, 2001). In a third step, the resulting nodes and attached models can be visualized using several strategies (e.g. Vesanto, 1999).

Formally, a SOM consists of a rectangular two-dimensional grid with $U$ units. Each unit $i$ is described by a $N$-dimensional model or weight vector $\mathbf{m}_i = [m_{i1},\ldots, m_{iN}]$ where $N$ is the number of input features. The basic organization of SOMs is usually a rectangular or hexagonal lattice (i.e. SOM nodes have four or six neighbors, respectively). We used a hexagonal lattice because they are preferable for visualization purposes (Kohonen, 2001; Vesanto, 1999). For SOM learning, training samples $\mathbf{x}_k = [x_{k1},\ldots, x_{kN}]$ ($k = 1,\ldots, K$) are iteratively presented and the best-matching unit (BMU) $\mathbf{m}_{BMU}$ is selected according to smallest distance, i.e.

$$\|\boldsymbol{x}_k - \boldsymbol{m}_{BMU}\| = \min_i(\|\boldsymbol{x}_k - \boldsymbol{m}_i\|), \qquad (1)$$

where || denotes Euclidean distance. In the following, weights of map units are modified with the following update rule:

$$\boldsymbol{m}_{t+1} = \boldsymbol{m}_t + \alpha_t h_{BMU}(r_t)\|\boldsymbol{x}_k - \boldsymbol{m}_i\|, \qquad (2)$$

where $t$ denotes the learning iteration, $\alpha_t$ the learning rate and $h_{BMU}(r_t)$ the Gaussian neighborhood kernel around winning unit $\mathbf{m}_{BMU}$ with radius $r_t$. After weight adaption, BMUs are redefined with the new unit-specific weight vectors. Both learning rate and radius of the neighborhood are decreasing functions over time. This learning leads to an early stage that sets the general layout of the map by allowing large adjustments and a fine-tuning stage with small changes. In this paper we used the batch computation of the SOM and not the stepwise, recursive algorithm as it is faster and more robust (e.g. Kohonen, 2001). The MATLAB-based SOM-toolbox (http://www.cis.hut.fi/somtoolbox/) was used for SSOM training.

**Supervised SOMs (SSOMs)**

Kohonen (2001) outlined how unsupervised SOMs could be modified to predict unseen instances (see Wongravee et al. [2010] and Xiao et al. [2005, 2006] for applications of this type of SSOM). The basic idea is to append a vector that specifies class membership to the SOM input. In this way, the training incorporates both the measured data (multivoxel patterns) as well as the class label

for model learning simultaneously, i.e. the algorithm does not differentiate between labels and input data. Melssen and co-workers (2006) proposed two alternatives for a supervised SOM algorithm. Both of these approaches consist of two separate SOMs, one representing the input data (multivoxel patterns in this paper) and the other reflecting the assigned class labels. During the training process a shared BMU or separate BMUs for the two maps are determined by taking distances in both SOMs into account. The weight vectors of SOM units for both maps are updated as in usual SOMs. In this paper, we apply the supervised SOM as proposed by Kohonen (2001) mainly for lower computational costs.

The extension suggested by Kohonen (2001) is based on a modified input vector $\mathbf{x}^*_k = [\mathbf{x}_k\ \mathbf{c}_k]$ for model training that results from concatenating input trials $\mathbf{x}_k$ and a $C$-dimensional ($C$ denotes the number of classes) class-vector $\mathbf{c}_k = [c_{k1},\ldots, c_{kC}]$ where $c_{ki} = 1$ if trial $k$ belongs to class $i$ and $c_{kj} = 0$ ($j \neq i$) otherwise (*1-of-K coding*; Bishop, 2006). Similarly, a vector $\mathbf{v}_i = [v_{i1},\ldots, v_{iC}]$ is appended to the weight vectors of SSOM-units $\mathbf{m}_i$ to form $\mathbf{m}^*_i = [\mathbf{m}_i\ \mathbf{v}_i]$ with $N+C$ elements. After SSOM training, map units are ascribed to one class by inspecting the last $C$ elements of the map weight vectors $\mathbf{v}_i$: the index with the largest value determines the label of map unit $\mathbf{m}_i$. In general, class-vectors $\mathbf{c}_k$ have norm $\tau$ (e.g. $\mathbf{c}_k = [0\ \tau\ 0]$ would be the class-vector coding the second class out of three). Large $\tau$ lead to better class separation of the supervised SOM but simultaneously increase the risk that SSOMs reflect the 'artificial' concatenated inputs $\mathbf{x}^*_k$ rather than the original inputs $\mathbf{x}_k$, which might lead to poor generalization performance. The parameters $\tau$ and $U$ for the two fMRI datasets were set according to results from the respective simulated datasets ($\tau = 0.2$ and $U = 64$ for dataset 1; $\tau = 1.0$ and $U = 100$ for dataset 2) and not optimized using measured data. The chosen parameters provided a good trade-off between data-driven organization and supervision leading to reasonable degree of clustering necessary for generalization and visualization (see supplemental material S1 for choices of $\tau$ and $U$). SOM units were initialized with weights that corresponded to eigenvectors of the largest $U$ eigenvalues of the training data.

For the prediction of testing trials, the elements of the weight vectors containing class information are detached and unseen instances $\mathbf{x}_{test}$ (i.e. without class vector $\mathbf{c}_{test}$) are presented. In contrast to the usual approach in which a trial is classified according to the label of the BMU, we employed the 10 best-matching units (10-BMUs) to accumulate evidence for classification (see Fig. S3 for results of $k$-BMUs). As SOMs reflect the topology underlying the input data, using $k$-BMUs ($k$>1) can be seen as taking the neighborhood of the BMU into account, which leads to more robust classification performances (e.g. Haufeld et al., 2012;

Silva and Del-Moral-Hernandez, 2011). In particular, we computed a classification index

$$CI_c = \sum_{i=1\ldots10} (\mathrm{v}_{ic} \cdot \exp(-\|\boldsymbol{x}_{BMU} - \boldsymbol{x}_{BMUi}\|^2) \ , \qquad (3)$$

to obtain evidence that trial $\mathbf{x}_{test}$ belongs to class $c$ ($v_{ic}$ denotes the class specific certainty of $i$th best-matching unit and $\mathbf{x}_{\mathrm{BMU}i}$ is the model of the $i$th best-matching unit). The supervised SOM predicts an unseen trial according to the class obtaining largest *CI*.

**Multi-Split and Multi-Subject SSOMs.**

One main advantage of using SSOMs is that these offer opportunities to visualize decoding models (see Fig. 1B). In this study, we visualized the decoding model, i.e. the SSOM, and results by projecting the weight vectors $\mathbf{m}_i$ onto two dimensions using principal component analysis (PCA; see e.g. Vesanto [1999] and Xiao and colleagues [2005] for other types of visualization). We chose the PCA-based approach because it is a simple and comprehensive linear approach that, in most cases, provided similar visualizations compared to non-linear mappings like multi-dimensional scaling (MDS).

SSOMs are trained for each cross-validation split and each of these single-split SSOM possesses its own topology. In order to have an understanding of the common topology across cross-validations it is necessary to create a *multi-split SSOM,* i.e. a representation that generalizes from SSOMs of single splits. Similarly, to identify the common topology of SSOMs across single subjects requires establishing a *multi-subject SSOM*. Here, we suggest abstracting from single to multiple SSOM properties (i.e. weight vectors, label certainties, and training/testing trial occupation) with a two-step procedure: First, single-split or subject-specific response patterns are transformed into a common space and, second, an optimal correspondence between the nodes of single SSOMs is found.

*SSOM Alignment.* The first obstacle to create general SSOMs is the fact that underlying activation patterns might not be based on the same voxels (e.g. in this study voxel selection depends on properties derived from the training set and not a region-of-interest [ROI] or searchlight approach for which the same voxels are used across splits). To resolve this alignment problem we follow an approach recently proposed by Haxby and colleagues (2011) that was developed to align response patterns of different participants by creating a common space. Here, a procrustean transformation is applied to align the centroids of all nodes with the same label of single SSOMs to find a transformation that maps SSOM nodes into a common space. The parameters for the linear transformations (i.e. rotation and translation; we did not include a scaling parameter as data were already

normalized) are estimated in a three-step process (cf. Haxby et al., 2011). First, the single SSOM weight vectors are aligned to an iteratively updated reference to form an *initial reference* SSOM. A randomly chosen single SSOM served as the first reference. Consecutively, this reference was updated by the average of the old reference and the newly aligned SSOM. The reference obtained in the end of this stage formed the *initial reference*. For the second processing stage, all single SSOMs were aligned to the *initial reference* SSOM and the average of the aligned single SSOMs formed the *final reference*. As a last step, single SSOMs were aligned to the *final reference*.

The estimated parameters of the final procrustean transformation for each SSOM were used to convert the weight vectors of single SSOMs to a common space. Note that for cases in which the same voxels are selected across splits the alignment via procrustean transformation is not needed and can be skipped.

*Node Matching*. In this step, the aim is to find a matching of nodes across different splits (or subjects) to form the *general* SSOM. Once the matching is found, node properties (i.e. coordinates, node labeling, training- and testing trial occupation) are averaged accordingly. In order to find a good matching of nodes we defined an error function and used the simulated annealing algorithm (SA; Kirkpatrick et al., 1983) to approximate its global minimum. The error function $E$ we aimed to minimize was

$$E = \omega_1 E_1 + \omega_2 E_2 + \omega_3 E_3, \qquad (5)$$

where $E_1$, $E_2$ and $E_3$ denoted distance, connection and correspondence error, respectively (see supplemental material S2). For SA all errors were rescaled to lie between 0 and 1 and the parameters of the error function were $\omega_1 = 0.7$, $\omega_2 = 0.2$, and $\omega_1 = 0.1$. We chose this error weighting to base the matching mostly on the obtained SSOM node weights and less so on node labels. In order to ensure a visually comprehensive connection layout, we introduced the connection error, however, with small weighting.

In each step of the SA, we perturbed two randomly selected nodes of one randomly chosen SSOM. A node matching with a lower energy state was always accepted whereas for a higher energy state the new matching was accepted with probability $p = exp[(E_{old} - E_{new})/T_{iter}]$, where $T_{iter}$ decreases with iterations ($T_0 = 10$, $T_{iter} = \gamma^{iter-1} \cdot T_0$ and $\gamma = 0.9$ for dataset 1; $T_0 = 1$, $T_{iter} = \gamma^{iter-1} \cdot T_0$ and $\gamma = 0.75$ for dataset 2). The result of this procedure was a node matching between SSOMs that minimized the energy function (Eq. 5) and a general SSOM could be created. To determine node labels and training/testing

trial node occupation we averaged the single SSOM counterparts. This general SSOM can be visualized similarly to single SSOMs (see above).

**SSOM Visualization**

For visualizations of SSOMs, we projected the trained SSOMs onto the first two principal components using customized functions provided by the MATLAB-based SOM toolbox (see 2.1). In addition, for dataset 2, class-specific maps were derived by computing a weighted average over the (linear) models of all nodes of the class considered. For weighting, we ranked the nodes with respect to their distance from the center of mass of the respective class (i.e. nodes weighting was computed according to $(Y - rank_{dist} + 1)/Y$ where Y denotes the number of nodes of the respective category). Note that the maps in Figure 12 do not show the average activation for each category but – owing to the SSOM algorithm – they display the set of features (voxels) that discriminate one category optimally from the other categories.

**2.5 fMRI Datasets**

*Simulated fMRI Data.* A 3-class fMRI dataset including 3 runs was simulated following procedures described in (De Martino et al., 2008). For each class we simulated 30 trials using SNRs (i.e. signal amplitude/standard deviation of noise) of 0.35 and 0.5 (referred to as *low* and *high* SNR, respectively) and CNRs (amplitude difference between conditions/noise standard deviation) of 0.15, 0.25, and 0.35 (referred to as *low*, *medium* and *high* CNR, respectively). SNR and CNR values were randomly assigned independently to each voxel with a standard deviation of 0.1 (SNR) and 0.01 (CNR). The underlying anatomy and region definitions were based on a real dataset employing auditory stimuli (see 2.6.2). For generating spatio-temporal activation patterns, boxcar time-courses were convolved with hemodynamic responses that varied across voxels with respect to its time to positive peak (drawn from normal distribution with $\mu$ = 4s and $\sigma$ = 0.5s). Subsequently, we added temporally auto-correlated noise to obtain the simulated signal with an autocorrelation of $\rho \sim N(0.5, 0.1)$ for each voxel (cf. De Martino et al., 2008). Finally, voxel time-courses were sampled at a repetition time (TR) of 2.6s that was also used for acquiring the auditory fMRI dataset (see 2.6.2) resulting in ~180 TRs per run (trials lasted between 5 and 7 TRs). We simulated two regions in the auditory cortex with different response properties. Voxels of the first region (432 voxels) responded to all classes but with larger amplitude to one of the *C* classes. In the second region (1453 voxels), voxels responded to sounds but did not differentiate between classes. These responsive regions were embedded within a dataset of temporally correlated noise including a total of 16,505 voxels (cortex mask).

In a second set of simulations that we used for estimation of parameters for the dataset 2, specifications of the above-mentioned simulations remained the same but the number of classes was set to 8 (corresponding to the fMRI data) and 12 trials per class were simulated. We defined one region (523 voxels) that differentiated classes with an SNR of 0.5 and CNR of 0.25.

*Speaker Identification Study*. To examine the classification approach with a challenging dataset (i.e. difficult to decode), we acquired fMRI data while 5 participants performed a speaker identification task. Vocalized sounds (< 1s) were played during a 1.2s period of silence between image acquisitions in a slow event-related design (5-7 TRs between single trials) and subjects were asked to indicate the speaker identity after sound presentation. Functional runs were collected at 3 Tesla (Allegra, Siemens) and consisted of 18 slices positioned parallel to the Sylvian fissure obtained with a T2-weighted gradient echo, EPI sequence (TR 2.6s, TE 30ms, TA 2.4s; voxel size $2\times2\times2mm^3$). Anatomical images were obtained using a high resolution ($1\times1\times1mm^3$), T1-weighted MPRAGE sequence. Presented sounds were 30 different short non-linguistic vocalizations (e.g. "aww", "uuh") of three speakers (1 female, 2 males referred to as *f*, *m1*, *m2*). Before scanning, participants were familiarized with the voice identities with a short practice session consisting of 20 vocalizations not presented during fMRI data acquisition.

*Visual Object Categories Study*. We reanalyzed a publicly available data set from the Haxby et al. (2001) study. In this study, participants viewed pictures of faces, cats, houses, chairs, scissors, shoes, bottles, and scrambled version thereof. Provided data were converted to BrainVoyager QX format (v2.4, Brain Innovation). We restricted our analysis to the Region of Interest (ROI) defined by the mask included in the dataset covering the ventral temporal object-selective cortex. Thus, neither univariate nor multivariate voxel selection was necessary. For more specifications on the experiment and data the reader is referred to the original study (Haxby et al., 2001).

**Data Processing and Analysis**

*Preprocessing*. The preprocessing of fMRI data consisted of slice-scan-time correction, motion correction, transformation into Talairach space, temporal high-pass filtering (0.005 Hz) including removal of linear trends and spatial smoothing with a Gaussian kernel (2mm FWHM) using BrainVoyager QX.

*Cross-Validation and Voxel Selection*. For cross-validation of the first dataset and corresponding simulations, we divided each run into two half-runs (the first half contained trials 1 to 15 and the second trials 16 to 30 of the respective run; conditions were balanced over half-runs) and performed 6-fold cross-validations

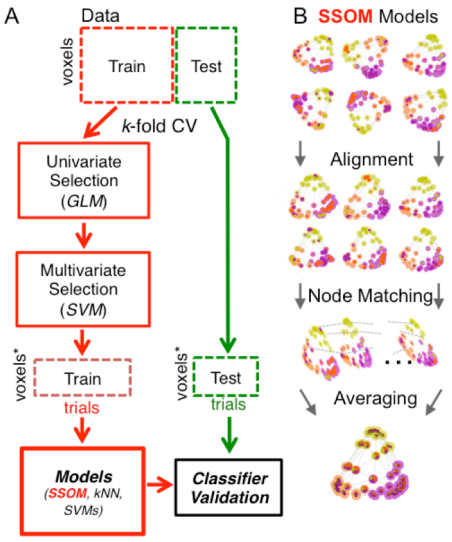



**Figure 1. Flowchart of Data Analysis and Creation of General SSOMs.** Panel A presents the main processing components employed in this study. After data acquisition and preprocessing, the dataset was divided according to a k-fold cross-validation scheme into an independent training and testing set. A univariate selection of features (i.e. voxels) via a GLM and subsequent multivariate selection using an SVM-based approach was performed on the training data (please note that the voxel selection steps were only used for analyzing dataset 1). The set of voxels surviving feature selection were used to construct the testing set. Next, different decoding algorithms were applied to create classification models that were validated using the validation set. Panel B shows the different processing steps to derive a general SSOM. Models of SSOMs (obtained in the Classification Models step in panel A) were aligned to be represented in a common high-dimensional space with procrustes transformation. Then, simulated annealing was employed to match nodes of single SSOMs that, in the end, were used to construct the general SSOM by averaging node membership and testing trial occupation.

to assess classification performance. For model training we defined datasets consisting of 5 of the 6 half-runs. The remaining half-run was used to determine how well the trained model was able to generalize. The second dataset (visual object categories) was analyzed in a leave-run-out scheme. This resulted in a 12-fold cross-validation for five subjects and 11-fold cross-validation for one subject.

For the first dataset we performed a series of feature selection steps (see Fig. 1A). We limited the analysis on an anatomical mask covering auditory responsive regions of temporal cortex and reduced the number of voxels with a GLM, which was computed using the training set by selecting the strongest responding 2000 voxels across conditions. An ensemble feature selection method (e.g. Abeel et al., 2010) was used to define feature sets with different numbers of voxels. This

feature selection approach relies on bootstrapped aggregation of SVM-based feature ranking. In this work, for each of 25 bootstrap samples, features surviving univariate selection were ranked according to model weights of a linear SVM ($C_{svm}$ = 1; one-versus-one scheme) and ranks across bootstrap samples were averaged to obtain the final ranking. We created 12 differently sized feature sets by removing iteratively the lowest ranked 20% voxels starting with 1000. This resulted in feature sets of 1000, 800, 640, 512, 410, 328, 262, 210, 168, 134, 107, and 86 voxels.

For both sets of data the single-trial response (in percent signal change) was fitted to a hemodynamic response model for each voxel. The obtained β values indicating response amplitude were used as features representing single trials. We normalized features of both training and test trials with an inter-quartile-range (IQR) normalization (median and 1st and 3rd quartile were estimated using training data): $x_i^{IQR} = 1.35 * (x_i - Q_{50})/(Q_{75} - Q_{25})$, where $x_i$ is the estimated β value of the $i$th trial and $Q_{50}$, $Q_{25}$ and $Q_{75}$ are the median, first and third quartile, respectively. Compared to z-score normalization, using the median and IQR for normalization is more robust to outliers. The scaling factor assures that z-score and IQR normalization is comparable for normally distributed data. For simulations we decoded the three simulated classes and, similarly for the real dataset, we classified the identity of the three speakers (f, m1, m2), whose vocalizations were presented to the subjects. We compared outcomes of SSOMs with a $k$NN classifier ($k$ = {1, 2, 3, 5, 10, 15, 20}), a linear SVM ($C_{svm}$ = 1), and a non-linear SVM (RBF-kernel; grid search of soft-margin and kernel parameter $C_{svm}$ [$2^{-6}$, $2^{-4}$,…, $2^{4}$] and $\sigma_{svm}$ [$2^{-7}$, $2^{-5}$,…, $2^{3}$]). Parameters for $k$NN and non-linear SVMs were optimized in 5-fold cross-validation. For SVM classification, the multiclass classification problem was transformed into binary classification using the one-versus-one scheme. We used the Spider toolbox (www.kyb.tuebingen.mpg.de/bs/people/spider) for SVM (linear and non-linear) and the BioInformatics toolbox for $k$NN classification (www.mathworks.nl/products/bioinfo/). Generalization performance was assessed with accuracy (i.e. the number of correct predictions divided by the total number of testing trials) and true positive rates (TPR) for each class to assess class-specific performance.

*Statistical Testing.* For simulations, we tested whether performances with discriminative voxels were higher than simulations without discriminative voxels using a Monte Carlo approximate permutation test (e.g. Good, 2000). We created 1999 randomly shuffled resamplings of the 20 simulations with and 10 simulations without discriminative voxels. The difference between the two types of simulations of the true groups was compared to the empirical distribution created by the

resampling procedure. The significance was calculated by dividing the count of instances in the permutation distribution that was larger than the actual difference (one count was added to both numerator and denominator). The 95% confidence interval of the estimated p-value is given by $\hat{p} \pm 1.96 * \sqrt{\hat{p}(1-\hat{p})/n}$, where $\hat{p}$ is the estimated p-value and $n$ the number of permutations. Comparisons between classification algorithms were done similarly. However, in contrast to the above-mentioned procedure, we randomly changed the sign of the elements of the difference vector between two classification approaches 1999 times and compared the actual average difference to the differences obtained with the Monte Carlo procedure.

For real data we tested – for each voxel selection level - whether classification performance was better than expected by chance by permuting class labels at a single subject level. Due to high computational costs (see Table S.1) the number of permutations was limited to 99. However, we also examined in detail a selected voxel selection level, for which we performed 999 permutations at single subject level. Corresponding single-subject accuracies and their confidence interval were derived as described for the simulations. For group statistics we employed an exact permutation test. Specifically, we determined the difference between the accuracy with true labels and the average permutation accuracies (observed 95% CI of average permutation accuracy for individual subjects 0.33±.098). The amount of occurrences of larger differences in the permutation distribution divided by the number of permutations denoted the significance. Comparisons between classification algorithms were performed with the same exact permutation test by computing accuracy differences between two classification approaches.

## Results and Discussion

### Simulations

*SSOM properties and visualization.* Figure 2 shows a SSOM obtained for one data split (low SNR, medium CNR; 262 voxels). The maps visualize the model as follows. First, weight vectors $\mathbf{m}_i$ of SSOM-units are projected onto two dimensions using PCA. Each node of the SSOM has a colored shading according to the class it represents (determined by the winning index of the class-defining vector $\mathbf{v}_i$). It can be seen that the SSOM grid reflects the properties of the 3-class dataset by forming a triangular-like shape with a clustering of nodes at the tips that represent one class. The shape indicates that, after training of the SSOM algorithm, nodes representing voxel activation patterns of the training set cluster at three regions in the high-dimensional input space. The occurrence of three clusters

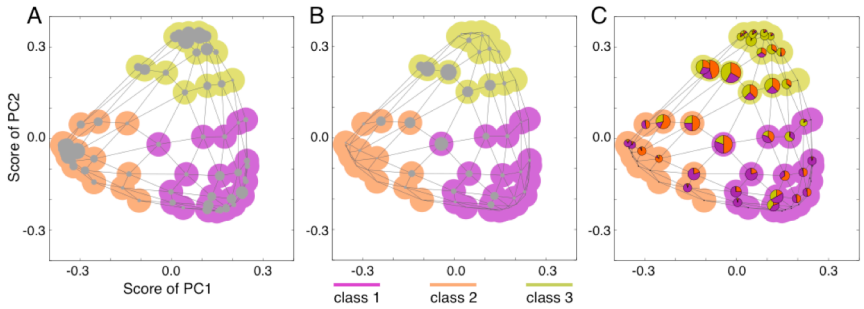



**Figure 2. Visualization of Single-Split SSOM.** The shading of nodes indicates which of the three classes the node represents. The radius of the inner grey circles denotes the number of training (A) and testing trials (B) for which the node was among the 10-BMUs. The pie charts in (C) show the class membership of testing trials of the respective units (the radii of pie charts denotes as in B the number of testing trials). See text for details.

is expected because input vectors contain class information leading to the class-distinguishing SSOM. To investigate the model's properties, we computed the 10-BMUs for training and testing trials. The radii of the grey circles in figure 2A (2B) indicate for each node the amount of trials that were best described by the node weight vectors $\mathbf{m}_i$ (i.e. the number of occurrences that this node belonged to the 10 BMUs). It can be seen that the model succeeds in describing training trials: nodes at the tips of the triangle are more often among the BMUs (Fig. 2A). In contrast, testing trials occupy more often nodes in the center of the triangle revealing that the model cannot differentiate as well between classes for the testing trials compared to trials in the training set (Fig. 2B). To provide a visual account of the generalization performance of the SSOM, pie charts describe to which class the testing trials belong (Fig. 2C). Here, colors indicate the proportion of classes of testing trials whereas the radius reflects the amount of trials falling onto single nodes (same as in Fig. 2B). It can be seen that nodes in the center of the triangle show similar proportions of testing trials for all classes. Nodes at the tips, however, do not attract many testing trials but are more specific with respect to classes. For this model it can be seen that testing trials falling on nodes at tips are most often consistent with the node's class label, thereby suggesting above chance classification (classification accuracy for this split: 0.40). Note that examining the performance of SSOMs on test trials is essential for a correct interpretation of topologies. In fact, due to supervised nature of the training, a clustering of nodes from the same class occurs to some extent also for random data. However, this is meaningful only if it generalizes to new trials. Hence, visualizations should only be consulted when classification performance is above chance level (similar to MVPA

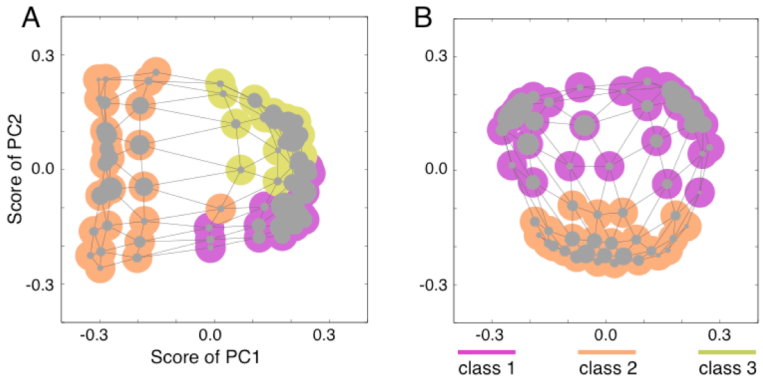



**Figure 3. Examples of Information Provided by SSOM Topology.** SSOMs show the same information as in Fig. 1A but for two special cases. In (A) the simulated 3-class dataset contains two classes that have similar activation patterns. Panel (B) shows a case in which binary classification was carried but the simulated data actually contained three classes.

studies presenting discriminative maps). In such cases, the SSOMs visualization can provide interesting insights into the decoding process. For example, when performing 3-class classification and the underlying representation of two classes is similar, SSOMs would show smaller distances for these classes in comparison to the other class (see Fig. 3A for an example). Similarly, strong evidence for separating one class into two classes would lead to subclusters of nodes within a class (Fig. 3B). In addition, other measures besides classification accuracy can be conveyed 'at a glance' (e.g. misclassification for specific classes).

To create a general SSOM from single-split SSOMs (Fig. 4 shows projections of SSOMs for different splits) we applied the alignment and node-matching procedure (see 2.3) that assures that class centroids overlap in common space. This linear transformation is then applied to produce a general SSOM. Note that for visualization purposes we did not present the classes of testing trials as pie charts for single-split SSOMs but their class-specificity, which we defined as the ratio of the classification index (cf. Eq.3) for the class with most and medium number of trials.

The center of figure 4 shows the corresponding general SSOM derived from single-split SSOMs. The nodes that share class labels cluster together forming triangular shapes in the PCA mapping. Similar to properties observed in single SSOMs, nodes between class clusters possess a higher attraction for testing trials, however, with less class-specificity. Nodes at the tips attract fewer testing trials but are more class-specific. It can be seen that for some of these nodes, the majority of

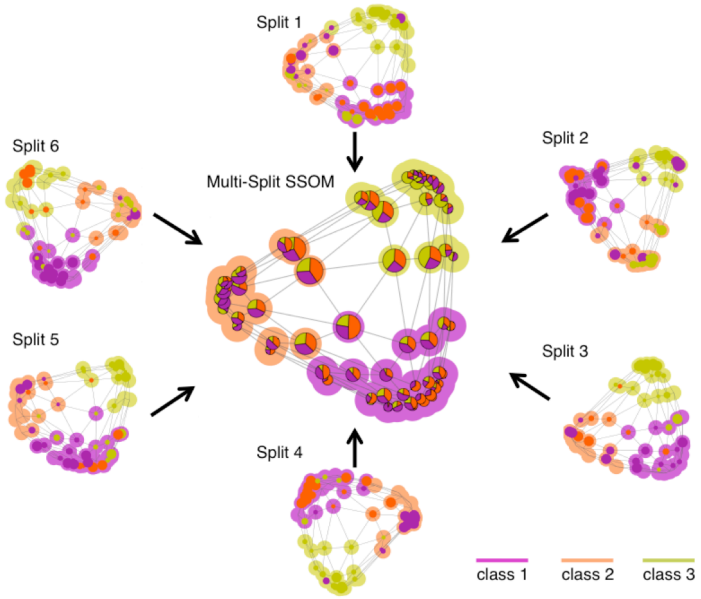



**Figure 4. Visualizations of Single-Split and Multi-Split SSOMs.** Shadings of nodes denote the three associated classes and inner circles of single-split SSOMs reflect the class specificity of testing trials for which nodes were BMUs (diameter scales with class specificity and color denotes the prevalent label of testing trials). Pie charts in the multi-split SSOM show class membership of testing trials and radii denote the number of training trials. Scales of principal components are equal to figure 1. The general SSOM was enlarged by a factor of 2.0 compared to single-split SSOMs.

testing trials did not belong to the class attributed to the node according the SSOM algorithm. In those cases this indicates misclassification and, more specifically, which classes are being confused. In the results for the particular simulation presented in figure 4, a faithful classification of the yellow class can be observed whereas the purple and orange one are often confused. This is also reflected in TPRs (yellow: 0.60, purple: 0.30, orange: 0.33; classification accuracy: 0.41).

*Performance of SSOMs compared to other decoding approaches*. For statistical evaluation of performances, we created 20 simulations with different noise patterns and randomized voxel's class preference in the discriminative region (as described in 2.5). In addition, we created 10 simulations in which we replaced voxels in discriminative region with voxels that were still responsive but no longer discriminative (i.e. these simulations did not contain any information about conditions).

As can be seen in figure 5 all decoding algorithms were able to classify class labels successfully independent of SNR for medium and high CNR (summed accuracies across selection levels, $p$ = .005±.0031 [significance estimate and 95% confidence interval thereof] for all decoding algorithms) and for high SNR and low CNR ($p$ < .005±.0031 for all decoding algorithms). We found that classification accuracy increased with a larger number of features.

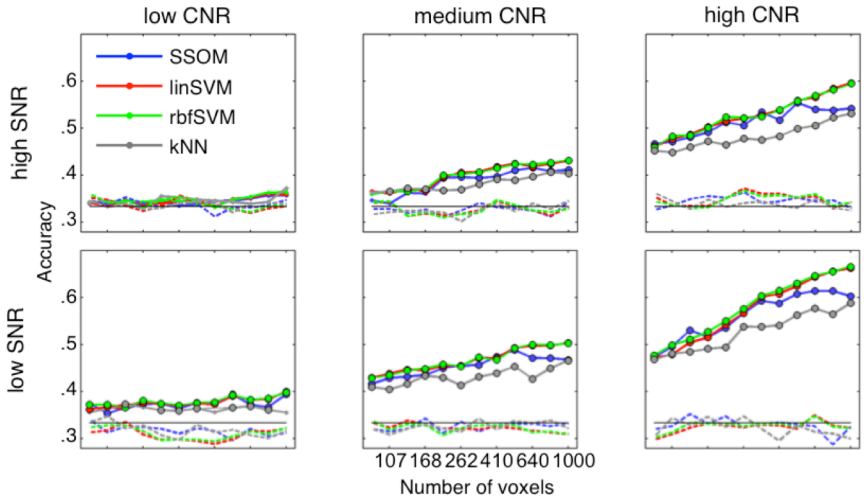



**Figure 5. Decoding Accuracy of Simulated Data for Different SNR and CNR.** Solid lines denote decoding performance for simulations with informative voxels and dotted lines show performance without informative but responsive voxels. Note that results depict an average decoding performance across simulations with independent noise patterns and randomly assigned voxel specificity in the discriminative region. Markers with black edge indicate classification performances above chance (p < .05; exact permutation test).

Focusing on SSOMs, we found that they perform better compared to *k*NN classification for simulations with medium CNR and high SNR (summed accuracies across selection levels, $p$ = .0215±.0064) and high CNR and low and high SNR ($p$ = .011±.0046). Level-specific comparisons showed that SSOMs could better discriminate classes compared to *k*NN classification for voxel selections consisting of medium-sized feature sets (i.e. 328-640 voxels; $p$ = .0555±.0100) for three SNR-CNR combinations (SNR/CNR: low/high, high/medium, and high/high). This suggests that although SSOMs and *k*NN classification share properties like computing distances to training samples (*k*NN) and nodes (SSOM) to predict test instances, SSOMs outperform *k*NN classification. One possible explanation is that in contrast to *k*NN classification, SSOM abstracts from individual training patterns to form SSOM nodes. This

might lead to more robust generalization due to less sensitivity to noise or outlier samples.

Compared to both types of SVMs, the decoding accuracy of SSOMs was not significantly different (summed accuracies across selection levels, $p = .135 \pm .150$ for all SNR and CNR settings). Focusing on performances for single selection levels SVMs outperformed SSOMs only when the number of voxels in the feature set was large (i.e. 800 and 1000 voxels; $p = .045 \pm .0091$) for simulations with three SNR/CNR combinations (low SNR/high CNR, high SNR/medium CNR, and high SNR/high SNR).

To better understand the decoding results, we examined the sets of voxels selected by the ensemble feature selection. As can be seen in figure 6, the number of voxels that belonged to the region with discriminative (i.e. informative) voxels increased until the largest set of features (dashed lines). This indicates why decoding algorithms show a general increase in performance. However, the number of informative voxels decreased compared to the number of noise voxels (i.e. voxels that did not contain information about class identity) for increasing sets of features (solid lines). This might explain the performance plateau for SSOMs for high CNR cases as the larger number of noisy voxels entering the set of features counteracts the increase of informative voxels. SVMs are more robust to this increase of noisy voxels by their intrinsic weighting of features, which is in line

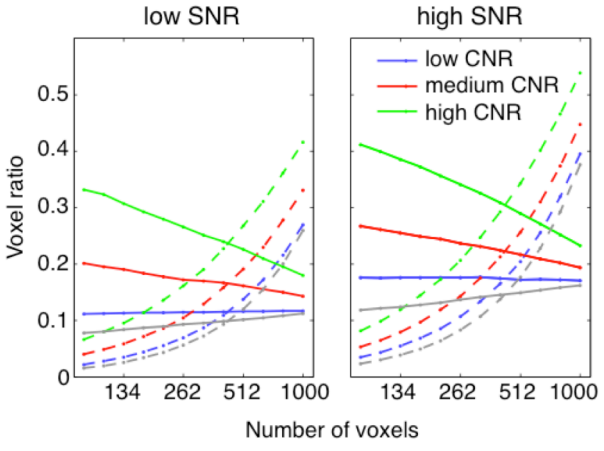



**Figure 6. Voxel Selection of Simulated Data for different SNR and CNR.** Solid lines depict the ratio of informative versus non-informative voxels among selected voxels. Dashed lines indicate the number of selected voxels within the discriminative region divided by total number of informative voxels. Grey lines show the respective results for simulations without informative but responsive voxels.

with the observed performance increase observed also for large sets of voxels. Thus, for cases with a large number of features (e.g. whole-brain classification), SVMs should be chosen over SSOMs.

**Speaker Identification Study**

*Voxel Selection*. Figure 7 shows the group map of the voxel ranking for the left and right hemispheres established by the ensemble feature selection procedure. We used cortex-based alignment (CBA) to transform single-subject selection maps into a common space (Goebel et al., 2006). For the group results, we computed maps that focus on the consistency of important regions across subjects. In particular, we selected the highest ranked 262 voxels for each subject and created a group map thereof that depicts for how many subjects a region was included in the combined univariate-multivariate selection procedure. Voxels selected with high consistency among subjects were found in left planum temporale (PT), bilateral middle superior temporal gyrus (STG)/Heschl's sulcus (HS), lateral Heschl's gyrus (HG), right posterior superior temporal sulcus (STS), and left anterior and posterior portions of STG. The right STS and HG/HS are in agreement with regions found to be involved in processing of voices (Belin et al., 2000; Moerel et al., 2012) and in the discrimination of individual speakers (Formisano et al., 2008). However, in contrast to Formisano et al. (2008) - where subjects were passively listening to the stimuli - we also observed left anterior STG as being important for speaker classification. This might be due to the active speaker discrimination task employed in our new experiment (see also Andics et al. [2010] that found the same region to be involved in an active speaker identification task).

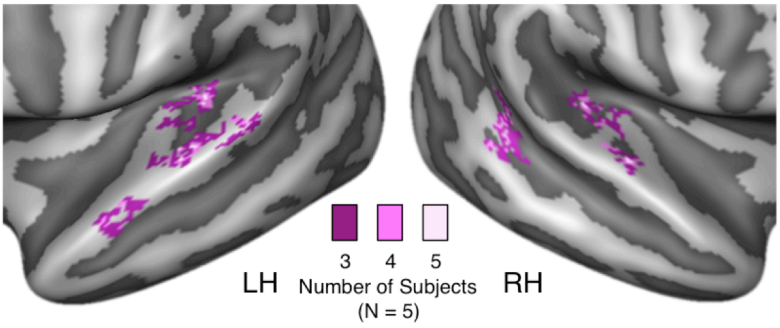



**Figure 7. Group Maps from the Ensemble Feature Selection Approach.** For each subject the most important 262 voxels were selected and projected onto the group-aligned cortex. The colors depict for how many subjects a region was included in the combined univariate-multivariate selection procedure.

*Performance of SSOMs compared to other decoding approaches.* We compared the SSOM decoding performance with those of *k*NN and SVM (linear and RBF-kernel) classification. Figure 8 shows the average decoding performance for the different voxel selections and classification algorithms. We found that all algorithms perform above chance level for medium numbers of voxels ($p < .05$).

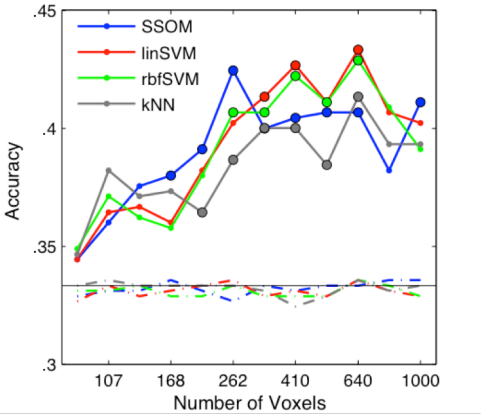



**Figure 8. Average Classification Accuracy for Real fMRI Data.** Solid lines show average accuracy and dashed lines denote empirical chance level computed by permuting class labels. Markers with black edge indicate significant classification performance on a group level ($p < .05$; exact permutation test).

Compared to *k*NN classification, SSOMs perform similarly (summed across selection levels, $p = .2903$) but with a better decoding accuracy for medium numbers of voxels. In particular, SSOMs were found to classify better compared to *k*NN classification for 210, 262 and 512 voxels ($p = .0323$). This finding is in line with the results of the simulations (see 3.1.2). In addition, SSOMs show similar performance to linear and non-linear SVMs for this set of data (summed across selection levels, $p > .40$). The increase of classification performance with larger voxel sets can be observed earlier in SSOMs and reaches its maximum at 262 voxels. In contrast, linear and non-linear SVMs show a more gradual increase and classify most accurately using 640 voxels. For single voxel selection levels, we found that SSOMs performed better compared to linear SVMs for 262 voxels ($p = .0323$) and worse than linear and non-linear SVMs for large number of voxels, i.e. 640 and 800 voxels ($p = .0323$).

In addition to decoding accuracy, we also examined class-specific outcomes. In figure 9 (upper left panel) we plotted for SSOMs the average TPR for each of the three voices (see table 1 for single subject results). We found that the female

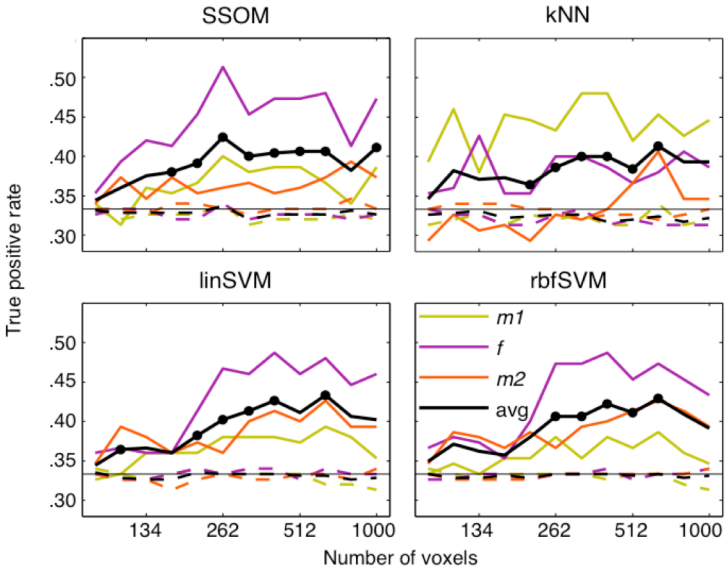



**Figure 9. Class-Specific TPR for Decoding Algorithms.** Colored solid lines denote single speaker TPRs (purple-f, yellow-m1, orange-m2), the black lines classification accuracy and dashed lines indicate TPR for permutations. Markers on the classification accuracy show decoding performance above chance.

speaker is decoded more truthfully compared to male speakers for most voxel selection levels. This is in line with the acoustic properties and corresponding behavioral results showing that the distinction between two male speakers is more challenging compared to detecting the female speaker (results not shown). With respect to TPRs the class-specific performances were found to be similar between SSOMs and SVMs. In contrast, class-specific performances obtained with *k*NN classification show a different pattern. Here, the first male speaker showed the highest TPR for most selection levels.

The decoding results for the real dataset suggest that SSOMs perform similar or slightly better compared to linear and non-linear SVMs for a small and medium number of voxels. For voxel sets with many voxels SVMs could discriminate better between classes probably due to their better robustness to noisy voxels. Compared to *k*NN classification we found that SSOMs had superior classification performance and that the TPRs of SSOMs for single classes matched the behavioral results in contrast to *k*NN classification. Thus, these results suggest that a classification with SSOMs is applicable for smaller regions when inherent multiclass classification is important.

*SSOM visualization*. Figure 10A shows the visualizations of general SSOMs (i.e. abstracting from single-split SSOMs) with 262 voxels for all subjects. For most subjects the nodes indicating the female voice attracted the testing trials of the very same class. In contrast, male voices were often confused among each other. For subject S4 and S5 (Fig. 10A, lower right SSOM) one can see a different pattern revealing faithful classification also for one male voice.

Figure 10B visualizes the group SSOM resulting from single-subject SSOMs. In the upper panel the average occupation of training trials is shown. It can be seen that in the center of class-specific clusters (i.e. at the tips of the triangle) the training trials belonged to the assigned node label. The further away from the class center a node is situated, the more training trials from the neighboring class cluster fall onto the node. This, as a proof of concept, indicates that the generalization procedure for SSOMs as proposed in section 2.3 seems to work well when applied twice (first to form the single-subject SSOMs and second to create the group SSOM). Focusing on testing trials (Fig. 10B lower panel) the group SSOM indicated that the dataset was challenging (small class-specificity) but in line with

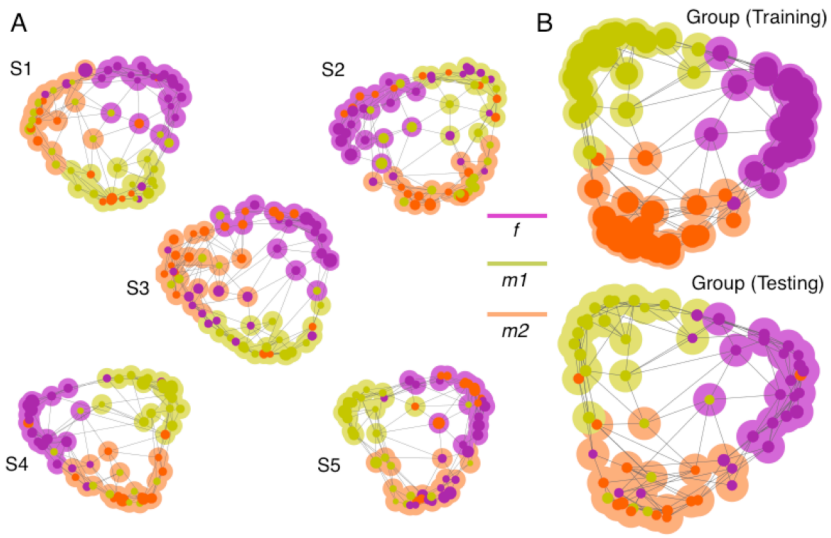



**Figure 10. Single Subject and Group SSOMs.** Speaker labels for nodes are color-coded (f-purple, m1-orange, m2-yellow). Panel (A) shows single-subject SSOMs with class-specificity of testing trials. In (B) the group SSOM with the class-specificity for training (upper panel) and testing trials (lower panel) is presented. All SSOMs depict the decoding model with 262 voxels. Note that for group SSOMs in (B) class-specificity expressed by circle diameter is scaled equally for training and testing trials whereas the scaling of class-specificity changes for single subject SSOMs in (A). Group SSOMs are enlarged by a factor of 1.5 compared to single-subject SSOMs.

classification accuracies and TPRs showed that all speakers are relatively well separated but especially in the case of one male speakers (m2) test trials were often mislabeled. With respect to SSOM topology, we did not observe differences in class relationships in the projections.

**Visual Object Categories Study**

*Performance of SSOMs compared to other decoding approaches*. For the data of the visual object categories study by Haxby and colleagues (2001), all decoding algorithms provided high classification performances clearly above theoretical chance level of 0.125 (see Fig. 11) using voxels in object-selective ventral temporal cortex. Results showed that SSOMs resulted in higher classification accuracy compared to *k*NN classification. In addition, linear SVMs - showing the highest performance among the tested algorithms – were found to have significantly higher classification accuracy compared to the other classification techniques.

Investigating TPRs for single classes shows similar patterns of decoding performance for all classification algorithms. These show - similar to results of the original paper (see supplemental material of Haxby et al. [2001]) - that faces, houses and scrambled pictures could be identified best, followed by chairs and cats and the lowest decoding accuracies for scissors, bottles and shoes (see Fig. 11).

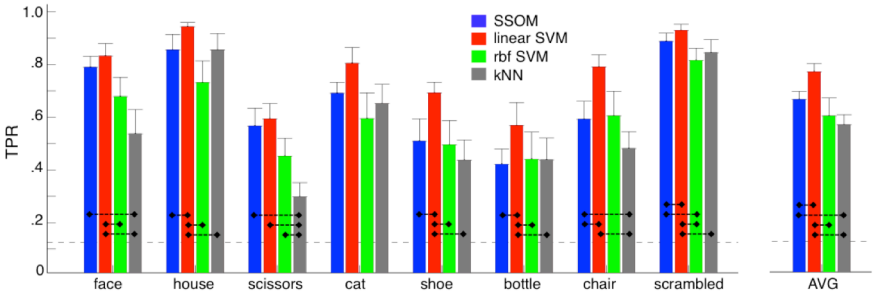



**Figure 11. Classification Performance for Visual Object Category Data.** Bars show the TPR for different classes and average classification accuracy. Results for the four classification algorithms employed in this study are presented in different bar colors (blue: SSOM, red: linear SVM, green: non-linear SVM, grey: kNN). Dashed lines denote significant differences between algorithms ($p < .05$; exact permutation test).

*SSOM Visualization*. Figure 12A shows the group SSOM of the visual object categories study. Apart from the observation that node classes and classes of testing trials were consistent (reflecting the high classification performance for this dataset), the SSOM visualization revealed several features of the dataset. First,

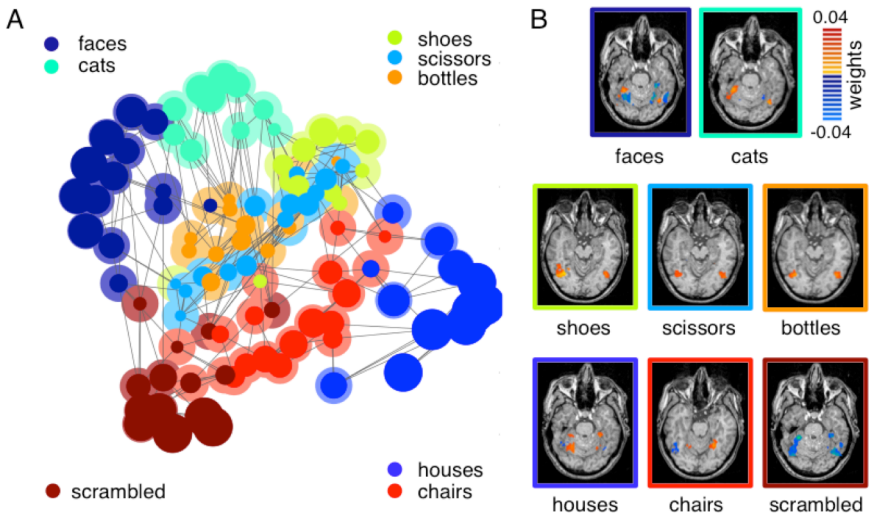



**Figure 12. SSOM and Category Maps for Visual Object Category Data.** Panel A shows the Group SSOM for the 8-class dataset projected using the first two principal components. Colors denote the different categories. Category maps of one exemplary subject are presented in panel B. These maps are constructed from a weighted average of nodes belonging to the same class and represent activation patterns.

scrambled objects formed a tight cluster separated from all other categories. The face and house categories appeared at opposite sides of the SSOM reflecting the highly differential activation patterns for these categories in object-selective ventral temporal cortex. Interestingly, the nodes of the cat category were close to the face category forming a larger cluster probably reflecting an animate/inanimate or a natural/man-made ordering principle for objects (see Kriegeskorte et al., 2008). The categories bottles, shoes and scissors showed overlapping or merging clusters that suggests more similar cortical activation patterns compared to other categories. Activation patterns used by the SSOM algorithm to classify new instances are presented in Figure 12B. These maps show the set of voxel that separate one category optimally from others.

## Conclusions

In this study, we applied supervised self-organizing maps to decode simulated and real fMRI datasets. The feasibility of SSOMs was shown by means of a simulated set of data. We found that SSOMs performed similarly to SVMs for small and medium numbers of voxels (i.e. 100-600 voxels) and were superior to

*k*NN classification especially for voxel sets of medium size. For large feature sets (> 800 voxels), SVMs outperformed SSOMs. For real data acquired from a challenging voice identification experiment, we showed that SSOMs performed similar to SVMs and *k*NN classification. We found that, as expected, the female speaker was easier to decode compared to the two male voices using activity patterns from auditory responsive regions of temporal cortex. Results for a dataset with known high classification performance showed that SSOMs were able to classify a dataset with 8 classes. It would be interesting to extend the comparison to other classification algorithms used in fMRI data analysis, especially inherently multivariate ones like Gaussian Processes (Rasmussen & Williams, 2006), sparse logistic regression (Yamashita et al., 2008; Ryali et al., 2010), Naïve Bayes (Brouwer & Heeger, 2008; Mitchell et al., 2004), Decision Tree classifiers (e.g. Kuncheva et al., 2010; Richiardi et al., 2011), or neural networks (Hanson et al., 2004; Polyn et al., 2005). Making use of the visualization possibilities from SOMs, we suggested one approach to plot the classification model and its generalization performance that provides information about underlying representations. To generalize from single SSOMs to a general SSOM we put forward an approach that first establishes a mapping of SSOMs into a common space and subsequently matches the SSOM-units. Group SSOMs were created and visualized by applying the summarizing approach twice (first, to create single-subject SSOMs form single splits and second, to generalize from single-subject SSOMs to a group SSOMs). SSOM visualizations for the two real datasets highlight the potential to depict the topology of the activation patterns underlying the successful decoding results.

## Acknowledgements

This work was supported by Maastricht University (LH, GV, EF) and by the Netherlands Organization for Scientific Research (NWO): VICI Grant 453-12-002 (EF). We thank Federico De Martino for comments on the article.

## References

Abdi, H., Dunlop, J.P., Williams, L.J., 2009. How to compute reliability estimates and display confidence and tolerance intervals for pattern classifiers using the Bootstrap and 3-way multidimensional scaling (DISTATIS). NeuroImage 45, 89–95.

Abeel, T., Helleputte, T., Van de Peer, Y., Dupont, P., Saeys, Y., 2010. Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. Bioinformatics 26, 392–398.

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

Beauchamp, M.S., LaConte, S., Yasar, N., 2009. Distributed representation of single touches in somatosensory and visual cortex. Human Brain Mapping 30, 3163–3171.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Bishop, C.M., 2006. Pattern recognition and machine learning. Springer Verlag.

Brouwer, G.J., Heeger, D.J., 2009. Decoding and reconstructing color from responses in human visual cortex. The Journal of Neuroscience 29, 13992–14003.

Cheng, Y., 1997. Convergence and ordering of Kohonen's batch Map. Neural Computation 9, 1667–1676.

Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. NeuroImage 19, 261–270.

Crammer, K., Singer, Y., 2002. On the learnability and design of output codes for multiclass problems. Machine Learning 47, 201–233.

De Martino, F., Valente, G., Staeren, N., Ashburner, J., Goebel, R., Formisano, E., 2008. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. NeuroImage 43, 44–58.

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P., 2009. Decoding of emotional information in voice-sensitive cortices. Current Biology 19, 1028–1033.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "What?" Brain-based decoding of human voice and speech. Science 322, 970–973.

Friston, K.J., 1995. Statistical parametric maps in functional imaging : A general linear approach. Human Brain Mapping 2, 189–210.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Human Brain Mapping 27, 392–401.

Good, P.I., 2000. Permutation tests. Springer Verlag.

Hanson, S.J., Matsuka, T., Haxby, J.V., 2004. Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? NeuroImage 23, 156–166.

Haufeld, L., Santoro, R., Valente, G., Formisano, E., 2012. Classification and visualization of multiclass fMRI data using supervised self-organizing maps. In: Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on, IEEE, pp. 65–68.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72, 404–416.

Haynes, J.-D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nature Reviews Neuroscience 7, 523–534.

Kalteh, A.M., Hjorth, P., Berndtsson, R., 2008. Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application. Environmental Modelling & Software 23, 835–845.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nature Neuroscience 8, 679–685.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by simulated annealing. Science 220, 671–680.

Kohonen, T., 1982. Analysis of a simple self-organizing process. Biological Cybernetics 44, 135–140.

Kohonen, T., 2001. Self-organizing maps. Springer Verlag.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proceedings of the National Academy of Sciences USA 103, 3863–3868.

Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience 2, 4.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J, Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126-1141.

Kuncheva, L.I., Rodríguez, J.J., 2010. Classifier ensembles for fMRI data analysis: an experiment. Magnetic Resonance Imaging 28, 583–593.

LaConte, S., Strother, S., Cherkassky, V., Anderson, J., Hu, X., 2005. Support vector machines for temporal classification of block design fMRI data. NeuroImage 26, 317–329.

Langs, G., Menze, B.H., Lashkari, D., Golland, P., 2011. Detecting stable distributed patterns of brain activation using Gini contrast. NeuroImage 56, 497–507.

Martínez-Ramón, M., Koltchinskii, V., Heileman, G.L., Posse, S., 2006. fMRI pattern classification using neuroanatomically constrained boosting. NeuroImage 31, 1129–1141.

Melssen, W., Wehrens, R., Buydens, L., 2006. Supervised Kohonen networks for classification problems. Chemometrics and Intelligent Laboratory Systems 83, 99–113.

Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to decode cognitive states from brain images. Machine Learning 57, 145–175.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.-A., Morito, Y., Tanabe, H.C., Sadato, N., Kamitani, Y., 2008. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron 60, 915–929.

Moerel, M., De Martino, F., Formisano, E., 2012. Processing of natural sounds in human auditory cortex: Tonotopy, spectral tuning, and relation to voice sensitivity. The Journal of Neuroscience 32, 14205–14216.

Mourão-Miranda, J., Bokde, A.L.W., Born, C., Hampel, H., Stetter, M., 2005. Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. NeuroImage 28, 980–995.

Mourão-Miranda, J., Reynaud, E., McGlone, F., Calvert, G., Brammer, M., 2006. The impact of temporal compression and space selection on SVM analysis of single-subject and multi-subject fMRI data. NeuroImage 33, 1055–1065.

Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends in Cognitive Sciences 10, 424–430.

Polyn, S.M., Natu, V.S., Cohen, J.D., Norman, K.A., 2005. Category-specific cortical activity precedes retrieval during memory search. Science 310, 1963–1966.

Rasmussen, C.E., Williams, C.K.I., 2006. Gaussian processes for machine learning. MIT Press.

Reddy, L., Tsuchiya, N., Serre, T., 2010. Reading the mind's eye: Decoding category information during mental imagery. NeuroImage 50, 818–825.

Richiardi, J., Eryilmaz, H., Schwartz, S., Vuilleumier, P., Van De Ville, D., 2011. Decoding brain states from fMRI connectivity graphs. NeuroImage 56, 616–626.

Ryali, S., Supekar, K., Abrams, D.A., Menon, V., 2010. Sparse logistic regression for whole-brain classification of fMRI data. NeuroImage 51, 752–764.

Silva, L.A., Del-Moral-Hernandez, E., 2011. A SOM combined with KNN for classification task. In: Presented at the Neural Networks (IJCNN), The 2011 International Joint Conference on, IEEE, pp. 2368–2373.

Swisher, J.D., Gatenby, J.C., Gore, J.C., Wolfe, B.A., Moon, C.-H., Kim, S.-G., Tong, F., 2010. Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. The Journal of Neuroscience 30, 325–330.

Vesanto, J., 1999. SOM-based data visualization methods. Intelligent Data Analysis 3, 111–126.

Walther, D.B., Caddigan, E., Fei-Fei, L., Beck, D.M., 2009. Natural scene categories revealed in distributed patterns of activity in the human brain. The Journal of Neuroscience 29, 10573–10581.

Weston, J., Watkins, C., 1999. Support vector machines for multi-class pattern recognition. In: Proceedings of the Seventh European Symposium On Artificial Neural Networks.

Wongravee, K., Lloyd, G.R., Silwood, C.J., Grootveld, M., Brereton, R.G., 2010. Supervised self organizing maps for classification and determination of potentially discriminatory variables: Illustrated by application to nuclear magnetic resonance metabolomic profiling. Analytical Chemistry 82, 628–638.

Xiao, Y.-D., Clauset, A., Harris, R., Bayram, E., Santago, P., Schmitt, J.D., 2005. Supervised self-organizing maps in drug discovery. 1. Robust behavior with overdetermined data sets. Journal of Chemical Information and Modelling 45, 1749–1758.

Xiao, Y.-D., Harris, R., Bayram, E., Santago, P., Schmitt, J.D., 2006. Supervised self-organizing maps in drug discovery. 2. Improvements in descriptor selection and model validation. Journal of Chemical Information and Modelling 46, 137–144.

Yamashita, O., Sato, M.-A., Yoshioka, T., Tong, F., Kamitani, Y., 2008. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. NeuroImage 42, 1414–1429.

# Supplemental Material

## Parameter Selection for SSOMs

The performance and properties of the SSOM are influenced by the parameter setting of the algorithm. These parameters include the number of map units $U$, the parameter $\tau$ describing the amount of supervision and the number of nearest nodes $k$ used to compute the prediction.

As a general rule, we favored medium numbers of map units ($U$; to allow for enough insights about clustering with visualizations still being comprehensive) and small values of supervision weighting ($\tau$; to assure that SSOMs reflect the underlying data and not the concatenated input vector of data and class membership). To estimate the impact of parameter choices we computed outcome measures for different combinations of parameters on the simulated fMRI datasets

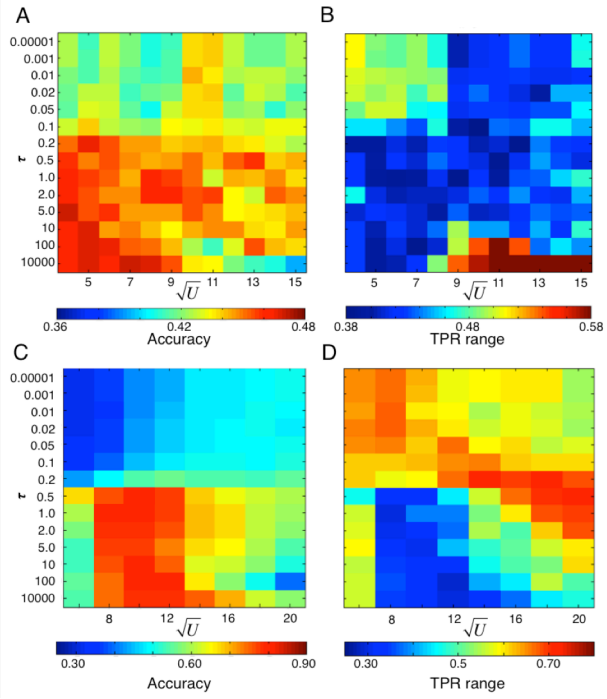



**Figure S1. Influence of Map Size and Supervision Weighting on Accuracy and Range of TPRs of SSOMs.** Panel (A) shows the classification accuracy and (B) the range of TPRs for combinations of different map sizes (U) and values of supervision weighting (τ) for the 3-class simulation. Similarly, panel C and D present classification accuracy and TPR range for 8-class simulation, respectively. Note that map size is indicated by its side length, i.e. square root of the total number of maps units.

(see 2.5.1). Focusing on the 3-class dataset, we found a region of parameter combinations of high classification performance for map sizes between 16 and 64 units and supervision parameter $\tau$ between 0.2 and 10000 (Fig. S1A). The lowest range of TPRs was observed for a supervision weighting between 0.2 and 10 (Fig. S1B). These two outcomes motivated our choice of parameters ($U$ = 64, $\tau$ = 0.2) because we aimed to have a medium number of map units $U$ and the smallest supervision weighting $\tau$ with high classification performance and small TPR range. With a similar reasoning we chose the parameters for the 8-class dataset ($U$ = 100, $\tau$ = 1.0; see panels C and D of Fig. S1).

To quantify the amount of clustering we calculated the average silhouette index (SI; Rousseeuw, 1987) for combinations of map size $U$ and supervision $\tau$ (Fig. S2) using the underlying SSOM models. Interestingly, we found two different states of SI. For the 3-class dataset we observed one low level of SI for values of $\tau < 0.2$ indicating poor clustering and a state with higher SI for $\tau \geq 0.2$ that suggest faithful clustering. This supports our choice of $\tau$ = 0.2 and indicates why

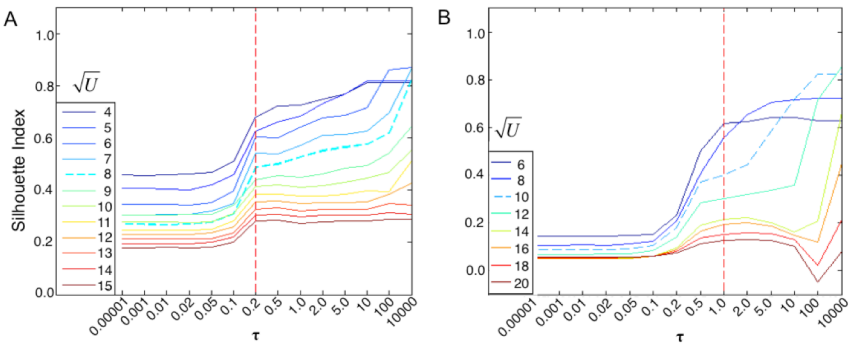



**Figure S2. Degree of Clustering as Assessed by Silhouette Index (SI) for SSOMs with Different Numbers of Nodes.** Panel A shows SI for the 3-class simulations and B for 8-class simulations. Large SI denote that data are appropriately clustered. Line colors denote SI for SSOMs with different numbers of nodes. The dashed red line indicates the chosen supervision parameter τ and the dashed turquoise line the chosen number of nodes U. Note that map size is indicated by its side length, i.e. square root of the total number of maps units.

classification performance increases in particular from $\tau$ = 0.1 to $\tau$ = 0.2. For the 8-class dataset a state with low clustering was found for $\tau < 0.5$ and the high state for $\tau \leq 0.5$.

Similar to our earlier findings (Haufeld et al., 2012) we observed that increasing the number of nodes for predictions leads to more robust and accurate

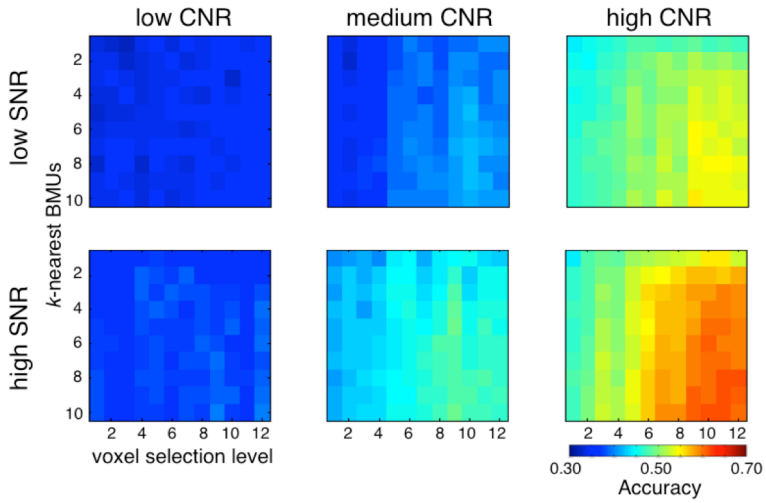



**Figure S3. Influences of k-BMUs and Voxel Selection on Classification Accuracy.** Larger voxel sets led to higher accuracies. Similarly, classifying with more BMUs led to increases in decoding performance. Larger voxel sets led to higher accuracies. Similarly, classifying with more BMUs led to increases in decoding performance.

performance (Fig. S3). Classification performance in this 3-class problem increased both with larger voxel selections and the number of nearest nodes used for class prediction. For the chosen $k$, i.e. $k = 10$, we computed two different outcome measures dependent on map size $U$ and supervision parameter $\tau$: classification accuracy and the range of TPR (i.e. the difference in TPRs between classes with highest and lowest TPR; the fact that the simulations were created with equal information for all 3 classes entails that models reflecting the data faithfully should result in a small range of TPRs).

**Error definitions for Simulated Annealing**

In order to form the multi-split and multi-subject SSOM, it is necessary to determine a good matching of nodes across different splits and subjects. Our approach was to define an error or energy function and to employ simulated annealing (SA; Kirkpatrick et al., 1983) to determine its global minimum. The error function $E$ we aimed to minimize consisted of three errors (i.e. distance, connection and correspondence error):

$$E = \omega_1 E_1 + \omega_2 E_2 + \omega_3 E_3, \qquad (S1)$$

where $\omega_1 = 0.7$, $\omega_2 = 0.2$ and $\omega_1 = 0.1$.

The *distance error* $E_1$ denotes the average Euclidean distance between nodes for all pairwise comparisons of corresponding nodes of $SSOM_p$ and $SSOM_q$ (Eq. S2), namely:

$$E_1 = \sum_{p,q\ (SSOM)} \sum_{n\ (Node)} \left\| m_n^{(p)} - m_n^{(q)} \right\|^2, p > q, \tag{S2}$$

where $m_n^{(p)}$ denotes the *n*-th node of the *p*-th SSOM. The *connection error* $E_2$ was introduced to penalize visually complex node connectivity layouts by punishing connections between nodes of different classes within each SSOM. The error is calculated as the sum of the number of neighbors with different class labels normalized by the total number of neighbors across SSOMs and nodes (Eq. S3).

$$E_2 = \sum_{SSOM} \sum_{n(Node)} \frac{\sum_{j \in \mathcal{N}_n} g_j^{(n)}}{|\mathcal{N}_n|}, \tag{S3}$$

where $\mathcal{N}_n$is the set of neighboring nodes of node *n*, $|\mathcal{N}_n|$ its cardinality and $g_j^{(n)}$ is 0 if node *j* belongs to the same class as node *n*.

$$g_j^{(n)} = \begin{cases} 1, & c_j \neq c_n \\ 0, & c_j = c_n \end{cases}. \tag{S3.1}$$

The *correspondence error* $E_3$ favors nodes with the same label across different SSOMs (Eq. S4). It yields small values when the class labels across SSOMs are similar:

$$E_3 = \sum_{n\ (Node)} N_{SSOM} - \max\left(O(n,1), \ldots, O(n,c)\right), \tag{S4}$$

where $O(n,j)$ denotes in how many SSOMs node *n* represented class *j*.

**Computational Costs**

As an estimate of the complexity of the employed processing steps (cf. Fig. 1) we measured their duration for one participant of the speaker identification study. Costs for univariate and multivariate selection were 2.12s and 14.74s, respectively. Table S1 shows that computational costs for SSOM training were small compared to, for example, training of rbf SVM and increase with the number of features included in the training set. However, these costs were small compared to the Node Matching procedure via SA.

Table S1. Computational Costs for Classification and SSOM Visualization in Seconds.

| | Number voxels | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Classification** | **86** | **107** | **134** | **168** | **210** | **262** | **328** | **410** | **512** | **640** | **800** | **1000** |
| SSOM | 0,12 | 0,15 | 0,18 | 0,23 | 0,31 | 0,42 | 0,61 | 0,88 | 1,53 | 1,99 | 3,02 | 5,30 |
| linear SVM | 0,09 | 0,05 | 0,05 | 0,04 | 0,04 | 0,04 | 0,05 | 0,05 | 0,05 | 0,05 | 0,05 | 0,06 |
| rbf SVM | 12,41 | 12,65 | 12,79 | 13,08 | 12,60 | 12,35 | 13,10 | 13,52 | 14,11 | 14,76 | 15,63 | 16,51 |
| kNN | 0,08 | 0,05 | 0,06 | 0,07 | 0,06 | 0,07 | 0,12 | 0,14 | 0,17 | 0,14 | 0,26 | 0,21 |
| **Visualization** | | | | | | | | | | | | |
| Alignment | 0,11 | 0,18 | 0,28 | 0,47 | 0,82 | 1,44 | 2,69 | 6,75 | 28,09 | 26,52 | 66,97 | 155,58 |
| Node Matching | 235 | 229 | 243 | 313 | 293 | 439 | 372 | 479 | 520 | 697 | 1021 | 1531 |

## References

Haufeld, L., Santoro, R., Valente, G., Formisano, E., 2012. Classification and Visualization of Multiclass fMRI Data Using Supervised Self-Organizing Maps. Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on, pp. 65–68.

Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P., 1983. Optimization by Simulated Annealing. Science 220, 671–680.

Rousseeuw, P.J., 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. J Comput Appl Math 20, 53–65.

# Chapter 4

# Task-Dependent Decoding of Speaker and Vowel Identity in Superior Temporal Cortex

*Corresponding manuscript (under review):*

Bonte, M., Hausfeld, L., Scharke, W., Valente, G. & Formisano, E. Task Dependent Decoding of Speaker and Vowel Identity in Superior Temporal Cortex.

## Abstract

Selective attention to relevant sound properties is essential for every day listening situations. It enables the formation of different perceptual representations of the same acoustic input and is at the basis of flexible and goal-dependent behavior. Here we investigate the role of the human auditory cortex in forming behavior-dependent representations of sounds. We use single-trial functional magnetic resonance imaging (fMRI) and analyze cortical responses collected while subjects listen to the same speech sounds (vowels /a/, /i/ and /u/) spoken by different speakers (boy, girl, male) and perform a delayed-match-to-sample task on either speech sound or speaker identity. Univariate analyses show a task-specific activation increase of the right superior temporal gyrus/sulcus (STG/STS) during speaker categorization and of the right posterior temporal cortex during vowel categorization. Beyond regional differences in activation levels, multivariate classification of single trial responses demonstrates that the success with which single speakers and vowels can be decoded from auditory cortical activation patterns depends on task demands and subject's behavioral performance. Speaker/vowel classification relied on distinct but overlapping regions across the (right) mid-anterior STG/STS (speakers) and bilateral mid-posterior STG/STS (vowels), as well as the superior temporal plane including Heschl's Gyrus/Sulcus. The task dependency of speaker/vowel classification demonstrates that the informative fMRI response patterns reflect the top-down enhancement of behaviorally relevant sound representations. Furthermore, our findings suggest that successful selection and processing of task-relevant sound properties relies on the joint encoding of information across early and higher-order regions of the auditory cortex.

## Introduction

In natural listening situations, we are surprisingly efficient in selecting, processing and grouping relevant acoustic elements of a sound while ignoring other elements of the same sound and the possible interference of background noise. This processing enables deriving distinct perceptual representations from the same acoustic input and is at the basis of adaptive and goal-oriented behavior. Whether and how the auditory cortex contributes to the formation of these representations remains largely unknown. In ferrets, neurons in the primary auditory cortex (PAC) have been shown to selectively tune their receptive field properties to behaviorally relevant auditory features (Fritz et al., 2003; Atiani et al., 2009), which suggests that goal-dependent sound representations may emerge already in PAC. In humans, fMRI responses in posterior auditory cortical regions were shown to become right lateralized during a pitch categorization task and left lateralized during a duration categorization task using the same frequency modulated tones (Brechmann and Scheich 2005), which suggests that modulatory and task-dependent effects are strongest in non-primary sub-regions within the auditory cortex.

Selective processing and grouping of specific acoustic elements is also pertinent to the extraction of different types of information from complex and socially relevant signals such as speech. For example, extracting phonemic categories requires a grouping of auditory features along the relevant dimension (e.g. formants of a vowel), independently of variations in other dimensions (e.g. fundamental frequency [F0] of a speaker's voice). Similarly, recognizing a voice requires extracting speaker specific acoustic characteristics (e.g. F0, timbre), independently of phonemic content. Task-dependent perceptual representations of multidimensional speech stimuli may emerge in specialized higher-order modules in the posterior superior temporal cortex for speech content (von Kriegstein et al., 2010; Mesgarani and Chang, 2012) and in the right anterior superior temporal sulcus for speaker identity (von Kriegstein et al., 2003). Alternatively, these representations may rely on spatially distributed general purpose auditory mechanisms involving also early auditory areas (Formisano et al., 2008; Kilian-Hütten et al., 2011a). Within such a distributed system, speech or speaker categories may emerge via task-dependent temporal binding of the responses of multiple (and spatially distant) neuronal populations, each one encoding for relevant acoustic features (Bonte et al., 2009).

The present fMRI study investigates the role of early and higher-order auditory cortex in forming goal-dependent representations of speech. Previous fMRI studies have investigated task-dependent speech processing by analyzing regional

changes in averaged activity across different experimental conditions. Here we apply multivariate pattern recognition techniques to single-trial fMRI responses and examine how task demands influence the spatial pattern of neural responses to individual sounds. We ask our subjects to perform delayed-match-to-sample tasks on either speaker or vowel identity and decode the neural representation of individual vowels or speakers in these two task contexts. Furthermore, we study the specific contribution to speaker identification of higher-order voice-selective areas by performing region-of-interest-based analyses using independently-acquired voice localizer data (Belin et al., 2000).

## Methods

### Participants

Ten healthy native Dutch adults (mean age 24.1 ± 2.4 years; 6 females; 9 right-handed) gave their written informed consent and participated in the study. Handedness was assessed by a handedness questionnaire adapted from Annett (1979). None of the participants had a history of neurological abnormalities and all had normal hearing as assessed with a pure tone audiogram (detection thresholds of frequencies from 250-8000 Hz at 0-20 decibels). Participants received a monetary reward for participation. Approval for the study was granted by the Ethical Committee of the Faculty of Psychology and Neuroscience at Maastricht University.

### Stimuli

Stimuli were speech sounds consisting of three natural Dutch vowels (/a/, /i/, and /u/) spoken by three native Dutch speakers (sp1: 9-year-old boy, sp2: 9-year-old girl, and sp3: adult male). To introduce acoustic variability typical of natural speech perception, for each vowel and for each speaker we included two different tokens. For instance, condition 'a-sp1' included two different utterances of the vowel /a/ spoken by speaker 1 (Fig. 1). We used children voices in addition to an adult voice because a shorter version of the experiment was used in a subsequent developmental fMRI study. Furthermore, this allowed investigating the recognition of children voices that, unlike adult voices, are not readily distinguished based on F0 and whose identification additionally relies on formant frequencies (Bennet and Weinberg, 1979; Perry et al., 2001). Stimuli were digitized at a sampling rate of 44.1 kHz, D/A converted with 16 bit resolution, band pass filtered (80 Hz to 10,5 kHz), downsampled to 22.05 kHz, and edited with PRAAT-software (Boersma and Weenink, 2002). Stimulus length was equalized to 350 ms (original range 258 to 364 ms), by using PSOLA (100-400 Hz as extrema for the F0 contour). We

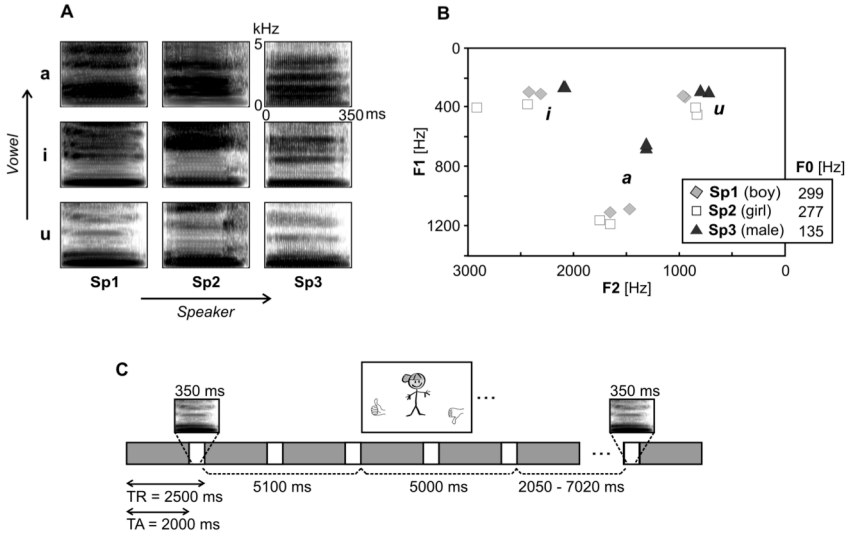



**Figure 1. Stimuli and Design.** A. Spectrograms of one exemplar of each of the 9 speech conditions. Stimuli consisted of three vowels (/a/, /i/, /u/) pronounced by three speakers (sp1:boy, sp2:girl, sp3:male). B. F1/F2 formant values for all stimuli (2 utterances per vowel for each speaker) and mean ± SD fundamental frequency (F0) values for each of the three speakers. C. Schematic overview of an experimental trial and the fMRI stimulation protocol including a black-and-white version of the 'boy' decision picture. Decision pictures consisted of cartoons of a boy, a girl or a man (speaker task), or the letter combinations 'aa', 'ie' and 'oe', corresponding to the pronunciation of the 3 Dutch vowels (vowel task).

carefully checked our stimuli for possible alterations in F0 after length equalization and did not find any detectable changes. Sound intensity level was numerically equalized across stimuli by matching peak amplitudes. To avoid acoustic transients (clicks) that would be created by a sharp cut-off, stimuli were faded with 100 ms exponential onset and offset ramps.

## Experimental design and procedure

We investigated task dependent processing of speaker and vowel identity by comparing the processing of the 9 speech conditions (a-sp1, a-sp2, a-sp3, i-sp1, i-sp2, i-sp3, u-sp1, u-sp2, u-sp3) during the performance of delayed-match-to-sample tasks on either speaker or vowel identity (Fig. 1C). Both tasks consisted of (1) the presentation of one of the speech stimuli (350 ms), followed by (2) a decision picture presented at the center of the screen, 5.1 seconds after speech stimulus offset, followed by (3) a match/mismatch response of the participant, indicated by pressing a response button with the right index or middle finger respectively. During the speaker task, decision pictures consisted of cartoons of a

boy (see Fig. 1C for a black-and-white version), a girl or a man. During the vowel task, decision pictures consisted of the letter combinations 'aa', 'ie' and 'oe', corresponding to the pronunciation of the 3 Dutch vowels. Decision pictures remained on screen until the button press or for a maximum time of 5 seconds. The sequence of speech stimuli was pseudo-randomized to avoid immediate repetitions of the same speech condition (e.g. a-sp1). Half of the trials included matching and the other half mismatching pictures, presented in a pseudo-randomized order, balanced per task, across experimental runs, and for each of the 9 speech conditions.

All subjects participated in two fMRI sessions with a between session break of 1 to maximally 10 days. At the start of the first session participants were familiarized with the three voices and performed practice trials to make sure both speaker and vowel tasks were understood and the three speakers and vowels were recognized correctly. The practice trials were repeated at the start of the second session. Both fMRI sessions consisted of three experimental runs, each run consisting of 4 alternations of the speaker and vowel tasks (run 1, 3 and 5: speaker task – vowel task –vowel task –speaker task; run 2, 4 and 6: vowel task – speaker task –speaker task –vowel task). We used 12 different sequences of speech stimuli, each of them occurring once in the speaker and once in the vowel task, across different fMRI sessions or (in two cases) in the first and third run of a session. In total, each run included 21 trials per task and 2 or 3 presentations of each of the 9 speech conditions. Across both fMRI sessions, each of the 9 speech conditions was presented 14 times per task.

**Functional MRI measurement**

Brain Imaging was performed with a Siemens Allegra 3 Tesla scanner (head setup) at the Maastricht Brain Imaging Center. During both fMRI sessions three 12 minute functional runs were collected (3 mm × 3 mm × 3 mm) using a standard echoplanar-imaging (EPI) sequence (repetition time [TR] = 2500 ms, acquisition time [TA] = 2000 ms, field of view [FOV] = 192 mm x 192 mm, matrix size = 64 x 64, echo time [TE] = 32 ms). Each volume consisted of 33 slices (distance factor 10%), covering the whole brain, except the most superior part of the posterior parietal cortex in some participants. Speech stimuli were presented binaurally at a comfortable listening level via MR compatible headphones, in the 500-ms silent gap between two volume acquisitions (Fig. 1C). According to a slow event-related design, the average inter-trial-interval between two speech stimuli was 15 seconds (range 12.5 to 17.5 seconds). Decision pictures were presented 5.1 seconds after the offset of the speech stimuli to allow a clear estimation of the auditory activation before the onset of visual and response-

related activity. During both experimental sessions a high-resolution structural scan (1 mm × 1 mm × 1 mm) was collected using a T1-weighted three-dimensional ADNI sequence ([TR] = 2050 ms, [TE] = 2.6 ms, 192 sagittal slices).

In the second session, an additional 12.5 minute voice localizer run was collected using the same EPI sequence and slice positioning of the main experiment, but with a TR of 3.0 seconds, leaving 1 second of silence for sound presentation. The voice localizer run consisted of 24 stimulation blocks (18 seconds / 6 volumes per block, one sound per volume) alternated with 12 seconds rest (4 volumes). During the stimulation blocks, participants listened to either (1) vocal sounds (including 7 non-speech sounds and 5 meaningless speech sounds), (2) other natural categories (musical instruments, environmental and animal sounds), both adapted from Belin et al. (2000), or (3) amplitude modulated (8 Hz) tones ranging from 0.3 to 3 kHz.

**fMRI pre-processing**

Functional and anatomical data were first analyzed using BrainVoyager QX 2.6 (Brain Innovation). Pre-processing of functional data included slice scan-time correction (using sinc interpolation), high-pass temporal filtering to remove nonlinear drifts of five or less cycles per time course, 3-dimensional motion correction, co-registration to individual structural images and normalization of anatomical and functional data to Talairach space (Goebel et al., 2006). All participants minimized head movements to maximally 2 mm in any direction. For univariate analysis functional data were spatially smoothed with a Gaussian kernel of 4 mm × 4 mm × 4 mm FWHM. Multivariate analysis was performed on unsmoothed functional data. Based on the high-resolution anatomical scans, individual cortical surfaces were reconstructed from gray–white matter segmentations. An anatomically aligned group-average cortical surface representation was obtained by aligning the individual cortical surfaces using a moving target-group average approach based on curvature information (cortex-based alignment, Goebel et al., 2006).

**Univariate fMRI analysis**

In order to map fMRI signal time courses from volume space to surface space, values located between the grey/white matter boundary and up to 4 mm into grey matter towards the pial surface were sampled with trilinear interpolation and averaged, resulting in a single value for each vertex of a cortex mesh. Random effects (RFX) general linear model (GLM) analysis was performed on time course data sampled on individual cortical surface meshes, aligned to the cortical group surface mesh using cortex-based alignment. The GLM model included one

predictor per stimulus condition (convolved with a double gamma hemodynamic response function) as well as confound predictors including each participant's motion correction parameters. Functional contrast maps (t-statistics) were calculated to assess sound-evoked fMRI responses during the speaker and vowel tasks (all sounds speaker task > baseline; all sounds vowel task > baseline). Direct task contrasts were analyzed for speaker task specific activity ((speaker task > vowel task) & (speaker task + vowel task > baseline)) and vowel task specific activity ((vowel task > speaker task) & (speaker task + vowel task > baseline)). Univariate stimulus effects were analyzed for each of the three speakers, independently of which vowel they pronounced (e.g. a-sp1 + i-sp1 + u-sp1 > (a-sp2 + i-sp2 + u-sp2 + a-sp3 + i-sp3 + u-sp3)/2) and for each of the three vowels, independently of who pronounced the vowel (e.g. a-sp1 + a-sp2 + a-sp3 > (i-sp1 + i-sp2 + i-sp3 + u-sp1 + u-sp2 + u-sp3)/2). Stimulus effects were analyzed in both the speaker and in the vowel task condition. All functional contrast maps were corrected for multiple comparisons by applying a cluster-size threshold with an initial voxel-level threshold of $p=0.01$ (overall activity) or $p=0.05$ (task and stimulus contrasts) and submitting the maps to a whole-brain correction criterion based on the estimate of the map's spatial smoothness (Forman et al., 1995; Goebel et al., 2006).

**Multivariate fMRI Analysis**

Multivoxel patterns of sound-evoked fMRI responses were analyzed by applying a machine learning algorithm (support vector machine, SVM; Vapnik, 1995) in two functional regions of interest (ROIs) based on each single subject's voice-localizer data. The first ROI included all auditory responsive voxels in the superior temporal cortex (STC). The STC ROI was defined from the independent localizer data by calculating for each subject a functional contrast map (voices + other + tones > silent baseline), applying a false discovery rate (FDR) correction for multiple comparisons (at $p<0.05$) and taking the intersection of this functional contrast map with an anatomical STC mask. The same anatomical STC mask was applied across subjects and in all subjects included the superior temporal plane, STG and STS as well as all superior temporal activity to voices, other natural categories and tones. The second ROI comprised the individually determined voice selective STC voxels also based on the independent localizer data. Voice ROIs were defined as regions showing significantly stronger activity to voices as compared to both other sound categories and tones (voices > (other + tones)/2). To prevent large between-subjects differences in the size of the voice ROIs, the exact statistical threshold was set on an individual basis (Frost et al., 2012; Bonte et al., 2013).

*Classification procedure.* Preprocessed functional time series were first divided into "trials" (one trial per sound presentation). Testing and validation sets were created using a 14-fold cross-validation procedure in which one trial out of 14 was left out for every condition. As input to the classifiers (features) we used beta estimates of the fitted double-gamma hemodynamic response, which were computed for single trials and voxels. For trial estimations we considered one TR before sound onset and the first two TRs following sound onset. The beta values were normalized across trials for each voxel using interquartile-range normalization (median, 1st and 3rd quartile were estimated using training trials):

$$x_i^{IQR} = 1.35 * \frac{x_i - Q_2}{Q_3 - Q_1}$$

where $x_i$ is the beta value of the *i*th trial and $Q_2$, $Q_1$ and $Q_3$ are the median, first and third quartile, respectively. This normalization is less sensitive to outliers than z-scores and - due to the scaling factor - provides comparable results when data are normally distributed. The voxels that were used to discriminate different speakers or vowels were specified by the ROIs as defined above. For classification we employed the SVM algorithm (soft margin parameter $C = 1$) as implemented in the spider toolbox (http://people.kyb.tuebingen.mpg.de/spider/). The three-class problem (classification of 3 speakers/vowels) was transformed into binary classifications using a one-versus-one scheme (i.e. sp1 vs. sp2, sp1 vs. sp3, sp2 vs. sp3 for speakers and /a/ vs. /i/, /a/ vs. /u/, /i/ vs. /u/ for vowels). In this approach multiclass classification is based on classifying pairs of conditions and the prediction for a test trial is determined by the condition that the binary classifiers predict most often. When one trial was equally often assigned to two classes, the class with the highest score of the classifier was chosen as the predicted one. Speaker classification was performed by grouping the trials of the three speakers regardless of vowels (e.g. sp1 = a-sp1 + i-sp1 + u-sp1). Vowel classification was performed by grouping the trials of the three vowels regardless of speakers (e.g. /a/ = a-sp1 + a-sp2 + a-sp3). For each of the three binary classifications per task, model-weights were used to indicate the importance for single voxels (see below, 'Mapping of Informative Regions'). Classification performance was reported in terms of overall accuracy, i.e. the number of correct predictions across speakers/vowels divided by the total number of speaker/vowel test trials.

*Statistical Testing.* To test whether classification values were significantly above chance, we performed the same multivoxel pattern analysis as described above with randomly shuffled condition labels per subject (number of permutations = 99). On a group level we performed a random-effects analysis using an exact

permutation test (number of permutations = 1022; Good, 2000), and comparing the single-subject accuracy of speaker/vowel classification with the average permutation accuracy of the respective subject. Single-subject and group significance levels were estimated by counting the number of permutations where the accuracy was larger than the actual classification accuracy and then dividing by the number of permutations (one count was added to both numerator and denominator for a more robust estimate of the significance value). To investigate the task-dependence of speaker/vowel classification we performed repeated measures ANOVAs on the ranked single-subject accuracies and assessed interaction effects between task and speaker/vowel classification accuracy.

*Mapping of Informative Regions.* Discriminative maps of locations that contributed most to the classification of the speakers/vowels were determined within the STC ROI. For each binary comparison, weights were linearized by ranking the absolute values. In a next step, we averaged the maps of binary comparisons to create a rank map for the multiclass classification. Single-subject maps were created by averaging the maps across cross-validations. These maps were projected onto the cortical surfaces of the individual subjects and subsequently projected on the group-averaged and cortex-based aligned cortex mesh. Inter-individual consistency maps were created by indicating for each vertex the number of subjects for which this vertex was within the fourth quartile of the SVM ranking (i.e. among the highest ranked 25%).

*Self-Organizing Maps.* For visualization of informative activation patterns we used self-organizing maps (SOMs; Kohonen, 2001; Formisano et al., 2008). For the speaker and vowel task we selected the 15 most informative voxels for single trial classification of speakers and vowels respectively. We concatenated the normalized vectors for all subjects and trained a rectangular SOM with 4x5 units with hexagonal connections using the MatLab based SOM toolbox (http://www.cis.hut.fi/somtoolbox/). We visualized the SOMs by showing the first two principal components of the high-dimensional model of SOM units. For both the speaker and the vowel task, the SOMs were trained using the average response patterns of the nine stimulus conditions (a-sp1, i-sp1, …, u-sp3). After training the 'best-matching units' (BMUs) for single trials were computed using Euclidean distance. Then, we labeled each SOM unit with the stimulus condition label for which this unit was most often the best matching one. The selectivity for each unit was determined by dividing the number of trials of the winning class by the total number of trials for which this unit was the BMU.

*Regression Analysis.* We employed a generalized linear model with a logit link function (McCulloch and Searle, 2001) to test whether behavioral accuracy of

speaker/vowel identification could be predicted from speaker/vowel classification accuracy within the STC and Voice ROIs. In this regression analysis, the log-odds ratio of the behavioral performances (modeled with a binomial distribution) was fitted using a design matrix consisting of a constant (intercept) and a predictor based on the fMRI classification accuracy. To assess whether a predictor (beta coefficient) was significantly different from 0, z-scores were computed for each predictor by dividing the corresponding beta coefficient by its standard error. The overall fit of the regression model was assessed using a $\chi^2$ - test of residual deviances (with low $\chi^2$ and $p_{fit} > 0.05$ indicating a good model fit).

## Results

### Behavioral results

All participants correctly identified each of the three speakers and vowels during a practice session outside the scanner. During the fMRI experiment, all participants performed well-above chance level (50%) during the delayed match-to-sample speaker and vowel identity tasks, although they had more difficulty to identify the children as compared to the adult speaker. That is, during the speaker task, percentage correct answers was (mean (SD)): boy 88.8 (7.6)%; girl 83.3 (9.3)%; man 98.3 (3.0)%. These differences led to a significant main effect of speaker $F(2,18)=13.0$; $p=0.000$, and pair wise differences between identification accuracies for the boy and the man ($t(9)=3.9$; $p=0.004$), the girl and the man ($t(9)=4.4$; $p=0.002$), but not for the boy and the girl ($t(9)=-1.8$; n.s.). During the vowel task, percentage correct answers corresponded to: vowel /a/ 99.8 (0.8)%; vowel /i/ 99.8 (0.8)%; vowel /u/ 98.8 (2.0)%, without significant differences between vowels.

### Univariate responses during the Speaker and Vowel task

During both the speaker and the vowel task, sounds evoked significant blood-oxygen-level dependent (BOLD) responses in a wide expanse of the superior temporal cortex, including early auditory areas (Heschl's Gyrus/ Heschl's Sulcus), the planum temporale and extending along the superior temporal gyrus (STG), superior temporal sulcus (STS) and middle temporal gyrus (MTG) (Fig. 2). Outside the temporal lobe, the medial prefrontal cortex was activated during both tasks whereas the bilateral posterior STS/MTG, and the right superior frontal gyrus were significantly activated only during the vowel task. Because the GLM focused on modeling of sound-evoked BOLD responses prior to both the presentation of the decision picture and the subsequent motor response, our maps did not show significant activation in visual or motor areas.

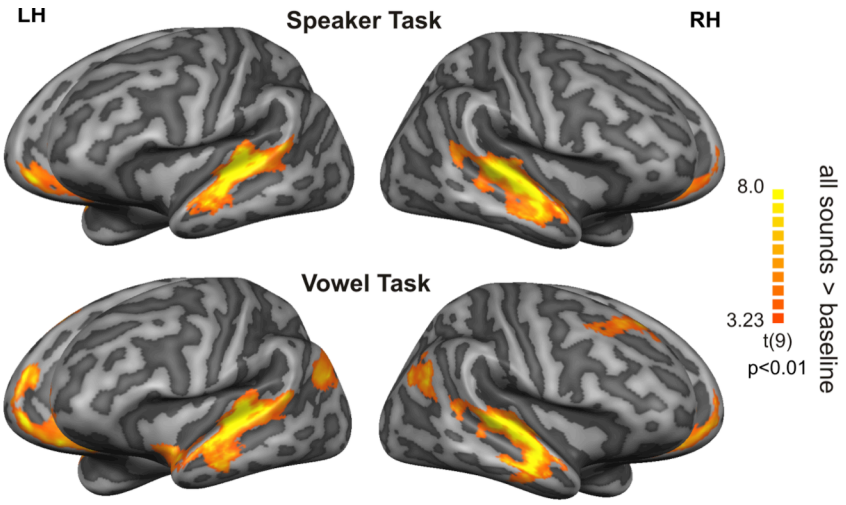



**Figure 2. Speech Sound Processing during Speaker and Vowel Tasks.** Functional contrast maps (t-statistics) illustrating the overall pattern of cortical responses during performance of the speaker task (speaker task > baseline) and vowel task (vowel task > baseline). Maps are visualized on inflated representations of the left (LH) and right (RH) hemispheres (light gray: gyri and dark gray: sulci), resulting from the realignment of the cortices of the 10 subjects. The maps are corrected for multiple comparisons by applying a cluster-size correction at $p<0.01$.

Results further suggested task specific activations for the speaker (Fig. 3 – blue colors) and the vowel (Fig. 3 – red colors) task. Although most of these activations were symmetrical across hemispheres, only two right hemispheric clusters survived cluster-size multiple comparisons correction. A cluster in the right middle STG/STS showed enhanced activity during the speaker as compared to the vowel task, whereas a cluster in the right posterior STS/MTG was more active during the vowel as compared to the speaker task (Fig. 3 – highlighted clusters and BOLD time-courses). Exclusion of the fMRI responses to sounds that were followed (after 3-5 TR) by an incorrect response did not change these task effects.

Analysis of stimulus effects did not show systematic univariate stimulus differences. Speaker and vowel stimuli did not show any significant activation differences along the task relevant dimension (boy, girl or man during the speaker task; /a/, /i/ or /u/ during the vowel task). Along the task-irrelevant dimension two stimulus contrasts did reach significance. During the speaker task the vowel /u/ elicited significantly stronger activity as compared to both other vowels in a region on the left mid to anterior STG. During the vowel task, the adult voice elicited significantly stronger activity as compared to both children voices in

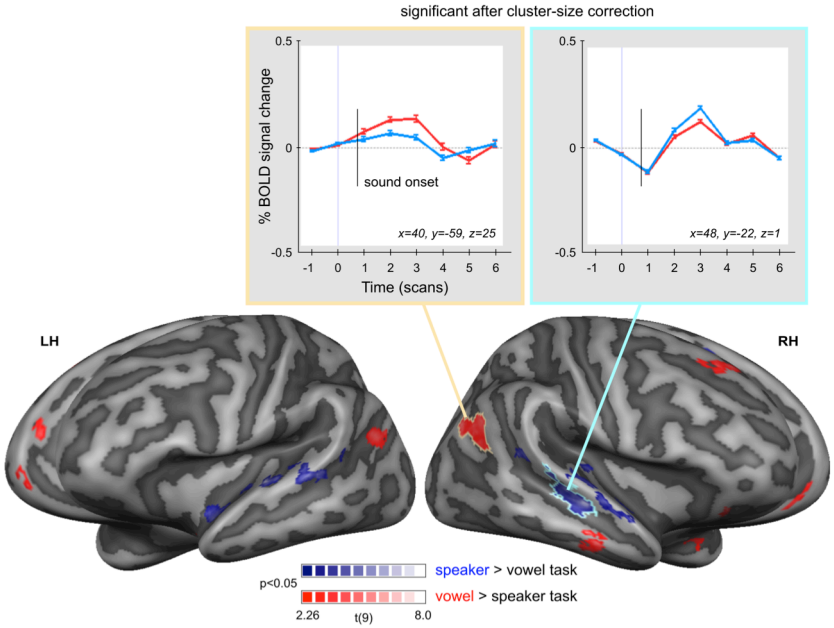



**Figure 3. Univariate Speaker and Vowel Task Effects.** Functional contrast maps (t-statistics) of task effects are shown for the speaker (blue colors) and vowel (red colors) tasks. Maps are visualized on inflated and aligned group-averaged representations of the left (LH) and right (RH) hemispheres. The maps show uncorrected activation clusters (> 9mm2). Two right hemispheric clusters survived multiple comparisons correction (cluster-size threshold at $p<0.05$): a mid STG/STS cluster that showed larger activity during the speaker task and a posterior STS/MTG cluster that showed larger activity during the vowel task. The time-course of task-related activity in both clusters is illustrated by plotting BOLD percentage signal change with respect to volume acquisitions (TR resolution). Talairach coordinates (x,y,z) refer to the center of gravity of the two clusters.

bilateral clusters on the temporal plane. No significant stimulus effects were found for the other two speakers or vowels.

### Task-dependent decoding of Speaker and Vowel identity

Beyond regional differences in overall activation levels, we investigated the task-dependent representations of individual speaker and vowel stimuli with a machine learning classification algorithm (SVM). In a first analysis, classification was performed within individually determined regions of auditory responsive superior temporal cortex (STC ROI, see Methods). The algorithm's success in speaker/vowel discrimination was strongly modulated by task demands. That is, in the STC region (Fig. 4A – middle panel), speaker stimuli were successfully discriminated based on fMRI responses obtained during the speaker but not

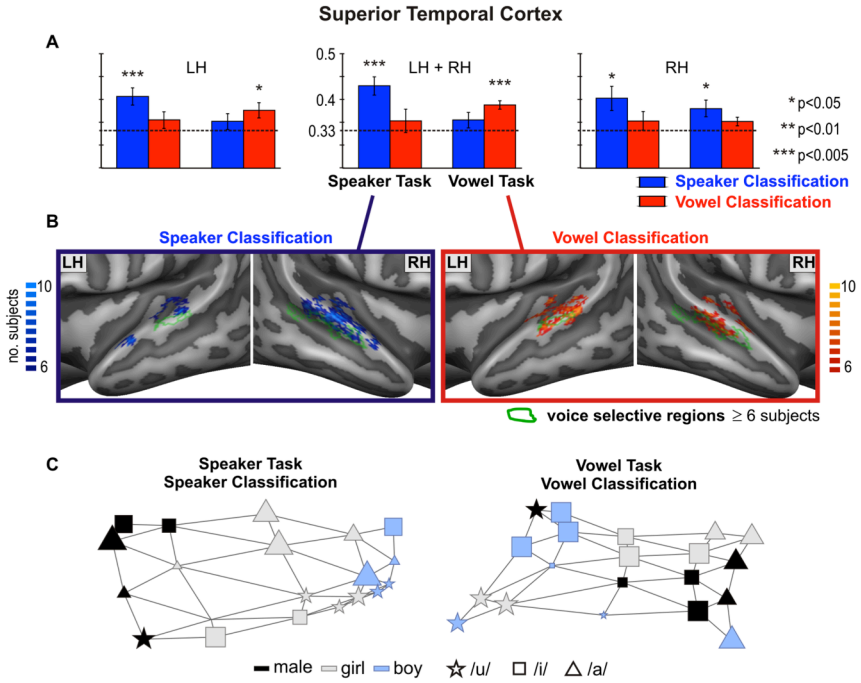



**Figure 4. Task dependent Classification of Speakers and Vowels.** A. Group averaged classification accuracies for speakers and vowels in the superior temporal cortex (STC ROI), during the speaker and vowel task. LH = left hemisphere, RH = right hemisphere. Statistical significance was determined with respect to empirical (permutation-based) chance level (dotted lines, range: 0.331-0.336, mean: 0.333, i.e. corresponding to theoretical chance level). B. Discriminative maps are illustrated for speaker classification during the speaker task and for vowel classification during the vowel task. These binary maps show for how many subjects a voxel was among the 25% most discriminative voxels and are visualized on inflated and aligned group-averaged representations of the temporal cortex. C. Self-organizing map (SOM) displays illustrating brain-based representation of speakers during the speaker task and of vowels during the vowel task. The maps are based on the 15 most discriminative speaker/vowel STC voxels across the 10 participants. The colors (speakers) and symbols (vowels) show which stimulus condition was assigned to a unit. The size of the unit indicates how often the stimulus condition was assigned to that unit (unit selectivity, see Methods).

during the vowel task, while vowel stimuli were successfully discriminated based on fMRI responses obtained during the vowel but not during the speaker task. This task-dependent decoding success was confirmed by a significant stimulus-by-task interaction for ranked single-subject accuracies ($F(1,9)=5.67$; $p=0.041$).

To assess the spatial layout and consistency across subjects of discriminative voxels underlying this task-dependent speaker and vowel classification, we constructed binary discriminative maps (Fig. 4B). These maps illustrate for how

many subjects a voxel was among the 25% most discriminative voxels. Speaker discriminative voxels (blue colors) clustered on the temporal plane, along Heschl's Gyrus/Heschl's Sulcus, and, especially in the right hemisphere, along the mid to anterior STG/STS. Vowel discriminative voxels (red colors) were distributed more bilaterally than those of speakers, and clustered on the temporal plane, along Heschl's Gyrus/Heschl's Sulcus, and on the mid to posterior STG/STS.

We visualized the spatial proximity and grouping of discriminative voxels contributing most to speaker and vowel classification using self-organizing maps (SOMs) (Fig. 4C). As expected from the significant classification accuracies in these conditions, the SOM-based two-dimensional displays showed vowel-invariant speaker grouping during the speaker task and speaker-invariant vowel grouping during the vowel task. Visual inspection of the spatial proximity of the individual speakers and vowels further indicates that speaker representations are ordered according to the average F0 of the speaker's voices (i.e. from left to right: male (135 Hz), girl (277 Hz) and boy (299 Hz)), while vowels are ordered according to their combined F1 and F2 values (i.e. from left to right: /u/, /i/ and /a/, following the diagonal of their representation in F1/F2 space, see Fig. 1B).

*Hemispheric lateralization.* Possible differences in lateralization were assessed by inspecting classification accuracies separately for the left and right STC (Fig. 4A – left and right panels). Results showed accurate speaker discrimination during the speaker task in both the left and right STC, and also during the vowel task in the right STC. Instead, accurate vowel discrimination only occurred during the vowel task and only in the left STC. The decoding accuracies in the separate left and right hemisphere STC ROIs did not show significant interaction effects.

*Contribution of Voice selective regions.* In a further analysis, classification was performed within individually determined regions of voice selectivity (Voice ROI, see Methods and Fig. 5AB) as well as within the STC ROI after subtracting the voice ROI (Fig. 5C). When restricting speaker/vowel classification to voxels in the bilateral Voice ROI (Fig. 5B), speakers could be discriminated above chance during the speaker task and vowels could be discriminated above chance during the vowel task, but there was no significant stimulus-by-task interaction ($F(1,9)=1.5$; n.s.). The same pattern of results was obtained when classifying speakers/vowels within the larger STC ROI after subtracting the Voice ROI (Fig. 5C), this time accompanied by a significant stimulus-by-task interaction ($F(1,9)=13.0$; $p=0.006$). Possible differences in lateralization were assessed by inspecting classification accuracies separately for the left and right hemisphere Voice and STC-Voice ROIs (Fig. 5BC – left and right panels). Within each of the

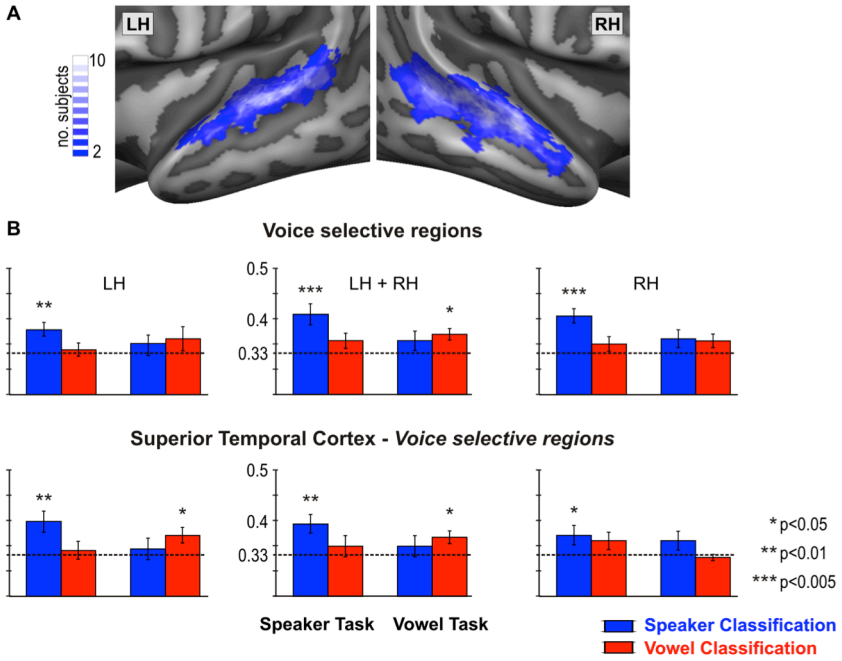



**Figure 5. Contribution of Voice Selective Regions.** A. Probabilistic maps illustrating spatial overlap of the individually determined Voice selective regions (Voice ROI). The maps are visualized on inflated and aligned group-averaged representations of the temporal cortex and show 20 to 100% subject overlap (n=2 to n=10). LH= left hemisphere, RH= right hemisphere. B. Group averaged classification accuracies for speakers and vowels during the speaker and vowel task in the individually determined Voice selective regions. Statistical significance was determined with respect to empirical (permutation-based) chance level (dotted lines, range: 0.330-0.334, mean = 0.333). C. Classification accuracies within the superior temporal cortex (STC ROI) after subtracting the Voice selective regions (Voice ROI).

four hemisphere specific ROIs, classification accuracies showed above chance classification of speakers during the speaker task. Only voxels in the left STC-Voice ROI also discriminated vowels during the vowel task, leading to a significant stimulus-by-task interaction ($F(1,9)=4.9$; $p=0.054$). Together, these results confirm an important role of the temporal voice areas (Belin et al., 2000) in the neural representation of speaker identity. However, they also show that parts of the STC that do not belong to these category selective regions are informative of speaker (and vowel) identity.

*Relation with Behavioral performance.* We used regression analysis to investigate whether individual differences in the accuracy of speaker classification predicts individual differences in behavior (accuracy of speaker identification). We concentrated on the speaker task because behavioral performance was close-to-

ceiling on the vowel task. Regression analysis showed that classification performance in the left STC could significantly predict participants' behavioral accuracy (model fit: $\chi^2(8)=14.80$, $p_{fit}=0.063$; predictor: $z=3.71$, $p=0.0002$; Fig. 6). In none of the other ROIs, the relation between classification and behavioral accuracy reached significance.

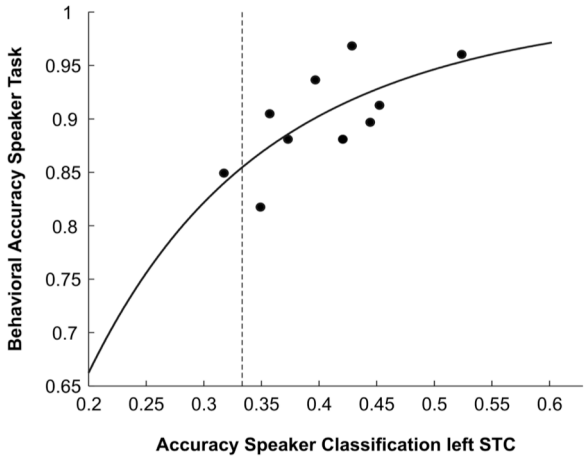



**Figure 6. Relation between Behavioral and Classification Accuracy.** Regression plot illustrating the relation between participant's behavioral accuracy of speaker identification and corresponding speaker classification accuracy in the left superior temporal cortex (left STC ROI). The vertical dotted line reflects the group averaged empirical chance level for classification accuracy.

## Discussion

We investigated single-trial fMRI responses measured while participants categorized the same natural speech sounds according to speaker or vowel identity. The task dependency of the speaker and vowel decoding accuracy demonstrates that the fMRI response patterns in auditory cortex (and, possibly, the underlying neural representations) reflect the top-down enhancement of behaviorally relevant sound representations. Furthermore, our findings highlight the role of early – together with higher order - auditory regions in the formation and maintenance of these representations.

To investigate the task-dependent categorization of sounds, we used delayed match-to-sample tasks that require the extraction and maintenance of either speaker or vowel information for several seconds until the presentation of a decision picture. Univariate analysis of sound-evoked responses showed extensive

and largely overlapping activation bilaterally in superior temporal cortex in both task contexts, reflecting sensory/perceptual analysis of the speech sounds (Belin et al., 2000; Binder et al., 2000; Scott et al., 2000). Both tasks also activated medial prefrontal regions, probably reflecting cognitive aspects of the tasks including maintenance of speaker/vowel identity in short-term memory and/or activation of task-relevant stimulus-response mappings (Duncan and Owen, 2000; Euston et al., 2012). Our analysis did not show other regions often implicated in speech perception, such as the left inferior frontal cortex (Hickock and Poeppel, 2007; Rauschecker and Scott, 2009). This may relate to specific stimulus and/or task demands. In particular, left inferior frontal activity is often observed during effortful lexical-semantic analysis of e.g. vocoded or spectrally rotated speech (Davis and Johnsrude, 2003; Eisner et al., 2010; Obleser and Kotz, 2010) and may contribute to decoding of ambiguous consonant-vowel stimuli (Lee et al., 2012).

Besides this network of largely overlapping brain activations, task-specific effects were observed in two regions. First, the right middle STG/STS, as well as smaller sub-threshold bilateral STG/STS and temporal plane clusters, showed stronger activation during the speaker task. This speaker task modulation confirms and extends previous reports of the involvement of these superior temporal regions in the passive and/or active processing of human voices (Belin et al., 2000; von Kriegstein et al., 2003; Andics et al., 2010; Moerel et al., 2012; Bonte et al., 2013; Latinus et al., 2013). Second, the right posterior STS/MTG showed stronger activation during the vowel task. Although this region is not typically involved in speech sound processing, it overlaps with an extended region in the inferior parietal lobe that has been related to the processing of learned audio-visual relations (Killian-Hütten et al., 2011b; Naumer et al., 2009). It can be speculated that the observed activation of the posterior STS/MTG during the vowel task relates to the nature of our delayed match-to-sample task which required matching of vowel sounds to their well-known visual counterparts (letters).

Beyond regional differences in overall activation levels, our multivariate decoding results demonstrate that distinct but overlapping response patterns across early and higher-order auditory cortex entail abstract goal-dependent representations of speech stimuli. Speaker discrimination most consistently relied on voxels clustering in early auditory regions (Heschl's gyrus/Heschl's Sulcus), on the temporal plane, as well as in regions along the mid-to-anterior (right) STG/STS that overlap with the superior temporal voice areas (Fig. 6A; Belin et al., 2000; Moerel et al., 2012; Bonte et al., 2013; Latinus et al., 2013) and with right STG/STS regions recruited during voice recognition tasks (von Kriegstein et al., 2003; Lattner et al., 2005; Andics et al., 2010). Vowels could be significantly decoded from voxels clustering in similar early auditory regions (Heschl's

gyrus/Heschl's Sulcus), on the temporal plane, as well as in bilateral regions along the mid-to-posterior STG/STS that have been related to the processing of isolated phonemes (Jäncke et al., 2002; van Atteveldt et al., 2004; Obleser and Eisner, 2009; Kilian-Huetten et al., 2011a) and to the processing of speech spoken by different speakers (von Kriegstein et al., 2010; Mesgarani and Chang, 2012). Our task-dependent speaker/vowel discriminative maps bear striking resemblance to those previously obtained when decoding speaker/vowel identity from fMRI responses to adult voices in a passive listening paradigm (Formisano et al., 2008). This similarity holds despite the use of different stimuli (children versus adult voices), a different experimental design (active tasks versus passive listening), scanning parameters (e.g. resolution of 3 mm isotropic versus 2 mm isotropic) and decoding methods (e.g. ROI-based feature selection versus recursive feature elimination). One interesting exception involves the left anterior STG that only contributed to speaker decoding in the present study. While this region may also show voice selectivity when using a voice localizer, this selectivity is found less consistently across subjects as compared to the right anterior STG/STS (in our localizer it showed voice selectivity in 7 out of 10 subjects, see also Moerel et al., 2012). Although the specific role of the left anterior STG remains to be determined, it may entail the active selection and perceptual representation of children and/or adult voices in the presence of task-irrelevant (verbal) information.

The task dependence of speaker/vowel classification accuracies suggests that the observed auditory cortical responses patterns reflect the perceptual categorization of sounds along the task-relevant stimulus dimension. Furthermore – in left STC - the speaker decoding accuracy significantly predicted participant's identification accuracy, which emphasizes the behavioral relevance of these patterns. In a pervious study (Andics et al., 2010), voice identification performances correlated significantly with activation changes in clusters of the left and right STG/STS. The left focus in the present study may be due to two factors. First, our multivariate analysis allowed relating behavioral performances to direct measures of identity information in the fMRI response patterns. This may reflect more closely the neural encoding of speaker identity as compared to activation level differences. Second, the behavioral variability in our study was mainly driven by the children voices, which - unlike adult voices - are not readily distinguished based on F0 (Murry and Singh, 1980; Baumann and Belin, 2010). Their identification may require the use of more subtle differences, e.g. in formant frequencies (Bennet and Weinberg, 1979; Perry et al., 2001) that may be processed in the left STC.

Whereas our findings confirm the involvement of voice/speech selective superior temporal regions, they show that auditory cortical maps of speaker and

vowel identity are not limited to these higher-order regions. Instead these categorical speech maps extend 'backwards' to regions that are assumed to restrict themselves to sensory processing of individual acoustic-phonetic speech features (Hickock and Poeppel, 2007; Rauschecker and Scott, 2009). The role of early, in addition to higher-order, auditory cortex in the task-dependent encoding of sound is consistent with evidence from animal electrophysiology (Fritz et al., 2003; Atiani et al., 2009). Furthermore, recent human fMRI-decoding studies show that similarly distributed superior temporal cortical patterns predict the abstract categorical representation of natural sounds (Formisano et al., 2008; Staeren et al., 2009), the subjective perceptual interpretation of ambiguous speech syllables (Kilian-Hütten et al., 2011a), and sound category learning (Ley et al., 2012). Within such a distributed system, task-dependent grouping of relevant speech features may emerge via transient phase alignment of neuronal responses in multiple non-adjacent cortical patches, each encoding for one or more of these speech features. Indeed, using EEG, we previously observed a task-dependent temporal alignment of oscillatory responses to individual speakers/vowels, that starts around 200 ms after stimulus onset and follows an initial analysis of acoustic-phonetic stimulus differences (Bonte et al., 2009). Furthermore, single-trial decoding of the same EEG data demonstrated task-independent classification of both speaker and vowel identity in early time windows, followed by sustained and task-dependent classification of speakers during the speaker task and of vowels during the vowel task (Hausfeld et al., 2012). Because the BOLD signal integrates neural processing over longer time-scales, these stronger and later task-dependent stimulus modulations may be most prominent in our findings, suppressing short-lived and earlier stimulus-driven processes. In fact, when presenting the same speech conditions in the context of a passive listening paradigm both vowels and speakers could be decoded (Formisano et al., 2008). Note, however, that in the present study the decoding accuracy in the task irrelevant dimension – although weaker - show above-chance (non significant) trends. It is thus possible that when increasing the spatial resolution (e.g. Formisano et al., 2008) and/or functional contrast to noise ratio (e.g. using higher magnetic fields), also the weaker signals along the task-irrelevant dimension may become significantly decodable.

The present study measured top-down modulation of fMRI responses in healthy adults to three vowels and three speakers that were presented in isolation in order to obtain distinct neural activation patterns. Extension of these results to attention-dependent processing of words or concatenation of words in streams of longer speech segments and in varying acoustic conditions (e.g. noisy environments), provides a compelling challenge and will contribute to a general brain-based decoder of sounds in the context of real-life situations. Furthermore,

extension to different age groups and subject populations may reveal relevant aspects of learning and plasticity in auditory cortical representations during normal and anomalous development.

**Acknowledgements**

Work supported by the Netherlands Organization for Scientific Research (NWO): VENI-grant no. 451-07-002 to MB. We thank Annemarie Graus for assistance in data acquisition.

**References**

Andics A, McQueen JM, Petersson KM, Gal V, Rudas G, Vidnyanszky Z (2010) Neural mechanisms for voice recognition. Neuroimage 52:1528-1540.

Annett M (1979) Family handedness in three generations predicted by the right shift theory. Ann Hum Genet 42:479-491.

Atiani S, Elhilali M, David SV, Fritz JB, Shamma SA (2009) Task difficulty and performance induce diverse adaptive patterns in gain and shape of primary auditory cortical receptive fields. Neuron 61:467-480.

Baumann O, Belin P (2010) Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol Res 74:110-120.

Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B (2000) Voice-selective areas in human auditory cortex. Nature 403:309-312.

Bennett S, Weinberg B (1979) Acoustic correlates of perceived sexual identity in preadolescent children's voices. J Acoust Soc Am 66:989-1000.

Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cereb Cortex 10:512-528.

Boersma P, Weenink D (2002) Praat 4.0: a system for doing phonetics with the computer [Computer software]. Amsterdam, The Netherlands: Universiteit van Amsterdam.

Bonte M, Frost MA, Rutten S, Ley A, Formisano E, Goebel R (2013) Development from childhood to adulthood increases morphological and functional inter-individual variability in the right superior temporal cortex. Neuroimage 83:739-750.

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.

Brechmann A, Scheich H (2005) Hemispheric shifts of sound representation in auditory cortex with conceptual listening. Cereb Cortex 15:578-587.

Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. J Neurosci 23:3423-3431.

Duncan J, Owen AM (2000) Common regions of the human frontal lobe recruited by diverse cognitive demands. Trends Neurosci 23:475-483.

Eisner F, McGettigan C, Faulkner A, Rosen S, Scott SK (2010) Inferior frontal gyrus activation predicts individual differences in perceptual learning of cochlear-implant simulations. J Neurosci 30:7179-7186.

Euston DR, Gruber AJ, McNaughton BL (2012) The role of medial prefrontal cortex in memory and decision making. Neuron 76:1057-1070.

Forman SD, Cohen JD, Fitzgerald M, Eddy WF, Mintun MA, Noll DC (1995) Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. Magn Reson Med 33:636-647.

Formisano E, De Martino F, Bonte M, Goebel R (2008) "Who" is saying "what"? Brain-based decoding of human voice and speech. Science 322:970-973.

Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6:1216-1223.

Frost MA, Goebel R (2012) Measuring structural-functional correspondence: spatial variability of specialised brain regions after macro-anatomical alignment. NeuroImage 59:1369-1381.

Goebel R, Esposito F, Formisano E (2006) Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum Brain Mapp 27:392-401.

Good P (2000) Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer.

Hausfeld L, De Martino F, Bonte M, Formisano E (2012) Pattern analysis of EEG responses to speech and voice: influence of feature grouping. Neuroimage 59:3641-3651.

Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nat Rev Neurosci 8:393-402.

Jäncke L, Wustenberg T, Scheich H, Heinze HJ (2002) Phonetic perception and the temporal cortex. Neuroimage 15:733-746.

Kilian-Hutten N, Valente G, Vroomen J, Formisano E (2011a) Auditory cortex encodes the perceptual interpretation of ambiguous sound. J Neurosci 31:1715-1720.

Kilian-Hutten N, Vroomen J, Formisano E (2011b) Brain activation during audiovisual exposure anticipates future perception of ambiguous speech. Neuroimage 57:1601-1607.

Kohonen T (2001) Self-Organizing Maps. Berlin: Springer-Verlag.

Latinus M, McAleer P, Bestelmeyer PE, Belin P (2013) Norm-based coding of voice identity in human auditory cortex. Curr Biol 23:1075-1080.

Lattner S, Meyer ME, Friederici AD (2005) Voice perception: Sex, pitch, and the right hemisphere. Hum Brain Mapp 24:11-20.

Lee YS, Turkeltaub P, Granger R, Raizada RD (2012) Categorical speech processing in Broca's area: an fMRI study using multivariate pattern-based analysis. J Neurosci 32:3942-3948.

Ley A, Vroomen J, Hausfeld L, Valente G, De Weerd P, Formisano E (2012) Learning of new sound categories shapes neural response patterns in human auditory cortex. J Neurosci 32:13273-13280.

McCulloch CE, Searle SR (2001) Generalized, Linear, and Mixed Models. New York: John Wiley and Sons.

Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233-236.

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.

Murry T, Singh S (1980) Multidimensional analysis of male and female voices. J Acoust Soc Am 68:1294-1300.

Naumer MJ, Doehrmann O, Muller NG, Muckli L, Kaiser J, Hein G (2009) Cortical plasticity of audio-visual object representations. Cereb Cortex 19:1641-1653.

Obleser J, Eisner F (2009) Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn Sci 13:14-19.

Obleser J, Kotz SA (2010) Expectancy constraints in degraded speech modulate the language comprehension network. Cereb Cortex 20:633-640.

Perry TL, Ohde RN, Ashmead DH (2001) The acoustic bases for gender identification from children's voices. J Acoust Soc Am 109:2988-2998.

Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718-724.

Scott SK, Blank CC, Rosen S, Wise RJ (2000) Identification of a pathway for intelligible speech in the left temporal lobe. Brain 123 Pt 12:2400-2406.

Staeren N, Renvall H, De Martino F, Goebel R, Formisano E (2009) Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol 19:498-502.

van Atteveldt N, Formisano E, Goebel R, Blomert L (2004) Integration of letters and speech sounds in the human brain. Neuron 43:271-282.

Vapnik VN (1995) The Nature of Statistical Learning Theory. New York: Springer-Verlag.

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

von Kriegstein K, Smith DR, Patterson RD, Kiebel SJ, Griffiths TD (2010) How the human brain recognizes speech in the context of changing speakers. J Neurosci 30:629-638.

# Chapter 5

# Auditory Cortical Representations during Speaker Identification in Noisy Auditory Scenes

*Corresponding manuscript (in preparation):*

Hausfeld, L., & Formisano, E.
Auditory Cortical Representations during Speaker Identification in Noisy Auditory Scenes.

## Abstract

Listeners effortlessly identify familiar voices regardless of speech content and in the presence of background noise. Using functional MRI, we examined the neural basis of speaker identification in the context of auditory scenes. To this end, participants listened to various short, non-linguistic vocalizations from three different speakers presented without noise or with interfering sounds (white noise or environmental noise). Then, we investigated whether activation patterns in auditory cortex (AC) could be used to decode speaker identity under the different listening conditions. Our results showed that speaker identity could be decoded above chance for sounds without background noise and with white noise but not within natural scenes. Furthermore, we found that activation patterns estimated in silence could predict speaker identity from patterns evoked by sounds in white noise and vice versa. The successful classifications for silent and white noise cross-decoding suggest that neuronal populations in AC (including early auditory areas) represent individual voices at a level of sound processing which is not purely acoustic. Specifically, we take the informative fMRI response patterns to reflect neuronal representations of a speaker voice which are robust to large acoustic variations (as those observed in different non-linguistic utterances) and to - certain extent - to interfering noise. Finally, we speculate that the differences we observe in the representation of voices in natural scenes reflect substantially different neuronal processing related to mechanisms of selective attention required to segregate two natural auditory objects in a scene.

# Introduction

Human listeners possess the remarkable skill to recognize a relevant sound without seeming effort even in the presence of loud and interfering background conditions. Difficult to model and replicate in artificial systems (e.g. Kinnunen and Li, 2010), this perceptual capability also holds when the goal is to identify a familiar voice that is mixed to concurrent sounds like other voices and environmental noise (Cherry, 1953). The aim of the present study is to investigate – using fMRI and multivariate decoding – the neural underpinnings of speaker identification in noisy auditory scenes.

## Voice and Voice-Identity Processing

In the past years, several studies examined the neural processing of voices by comparing brain responses to human vocal sounds and other sound categories using fMRI (Belin et al., 2002; 2000) and EEG (Charest et al., 2009; Grossmann et al., n.d.). FMRI studies identified regions with greater activation for voices in bilateral middle and posterior superior temporal gyrus (STG) and sulcus (STS) and right anterior STG/STS (Belin et al., 2000; Bonte et al., chapter 4; Charest et al., 2012; Ethofer et al., 2009; 2013; Latinus et al., 2011; 2013). While it has been established that these regions - often referred to as temporal voice areas (TVAs) – are involved in the processing of human voices, their "division of labor" in processing specific features of the vocal signal is less clear (but see Warren et al. [2006] for a proposal). Furthermore, it remains an open question *if*, *how* (i.e. based on which acoustic features) and *where* (i.e. in which of the regions) the identity of a speaker is derived within this temporal network.

Psychoacoustical investigations of auditory processing suggest that relevant cues for speaker identification are the fundamental frequency (F0 or pitch), which is determined by glottal folds of the speakers, and higher formant frequencies (F1-F4) that are shaped by vocal tract properties of speakers (Baumann and Belin, 2010; Lavner et al., 2000). It should be noted, however, that - even for short and stable stimuli like vowels - the relevance of a specific cue for speaker identification might differ depending on the speakers' voice characteristics and task context (Lavner et al., 2000). Beyond vowels, longer speech sounds (e.g. words or sentences) offer other cues for reliable speaker identification like formant dynamics and prosody (see for example Dellwo et al. [2007] and Hill [2007] for an overview). In line with the diversity of the possible speaker identity cues, recent behavioral and imaging evidence supports the hypothesis that voices are encoded in a flexible multidimensional representational space, which may be based on

relative (i.e. prototype based) rather than absolute and feature–inherent properties (Bruckert et al., 2010; Latinus and Belin, 2011; Lavner et al., 2001; Papcun, 1989).

From a neuroscientific and modeling perspective, it is of great relevance understanding how speaker identity is derived in the brain. A common model assumes that the individual cues are processed in hierarchical and specialized processing channels, culminating in one region where speaker identity is represented (Andics et al., 2010; Belin et al., 2002; Belin and Zatorre, 2003; Imaizumi et al., 1997; Latinus et al., 2011; Nakamura et al., 2001). Alternatively, the different cues may be extracted in the auditory cortex via general-purpose auditory processing and brought together to form a speaker identity by distributed and adaptive neural mechanisms for binding (Bonte et al., 2009; Formisano et al., 2008, in press). The first model is supported by evidence from PET (Nakamura et al., 2001) and fMRI using stimulus adaptation (Andics et al., 2010; Belin and Zatorre, 2003; Latinus et al., 2011) and task modulations (Kriegstein et al., 2003). This line of evidence points to the right anterior STG as the region where speaker identity is represented. The second model is supported by recent findings from our group, which suggest that the neural representation of a speaker's identity emerges from the encoding of information occurring in early (bilateral lateral Heschl's gyrus [HG] and sulculs [HS]) as well as in higher-level auditory cortical regions (i.e. TVAs). The fine-grained temporal realignment of the neural response patterns has been put forward as a flexible mechanism that binds together the activity of neuronal populations in spatially remote auditory regions (Bonte et al., 2009; Hausfeld et al., 2012a). Further support that processing of voices and the formation of speaker identity may rely on general-purpose auditory processing stems from evidence showing that: (1) carefully controlling for spectral-temporal acoustic differences between voices and other categories greatly reduces the response differences in the auditory cortex and in the TVAs (Staeren et al., 2009) and (2) that TVAs present greater sensitivity for acoustic features which are typical of human voices (e.g. frequencies < 1000 Hz) even when they are stimulated using non-vocal stimuli and (Moerel et al., 2012).

**Voice processing in noisy scenes**

The present study contributes to this debate on voice processing by examining the representation of speaker identity in noisy scenes. To date, only few fMRI studies investigated the processing of voices in the presence of interfering background (e.g. Binder et al. [2004] and Bishop and Miller [2009]). These studies, however, focused on intelligibility and comprehension of speech segments and did not specifically investigate the neural processing and representations for speaker identification.

Recently, studies using magnetoencephalography (MEG) or electrocorticographic recordings (ECoG) investigated speech processing in multi-talker environments. They found that delta and gamma frequency phase and power of signals recorded in auditory cortex track attended speech (and speaker) but less so ignored speech (Ding and Simon, 2012; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013). In contrast to the studies above that presented long speech streams (sentences), the present study focuses on short non-linguistic vocalizations. In addition, by using fMRI to measure cortical responses we cover the entire auditory responsive areas within temporal cortex whereas ECoG electrodes are restricted to small regions within auditory cortex. We presented subjects with various types of short vocalizations and asked them to perform a speaker identification task while measuring their fMRI activation. The voiced sounds were presented in silence but also with concurrent white and natural noise. We made use of a decoding analysis that has been successfully employed in auditory experiments to examine speech content and speaker identity processing (Bonte et al., chapter 4; Formisano et al., 2008) perceptual interpretations of ambiguous speech (Kilian-Hütten et al., 2011) and representations of sound categories (Staeren et al., 2009) and their formation during learning (Ley et al., 2012). For decoding of speaker identity in noise, we applied a novel fMRI-decoding approach based on supervised self-organizing maps (SSOMs) that allows multiclass decoding and offers an intuitive visual account of classification results and categorical representation (Hausfeld et al., 2012b, chapter 3).

## Methods

### Participants

Five right-handed female participants (29.0 years ± 2.91) employed by Maastricht University participated in the study. All subjects reported no history of hearing loss and deficits or neurological abnormalities, signed a written consent and received a monetary reward after participation. The study was approved by the ethics committee of the Faculty of Psychology and Neuroscience at Maastricht University.

### Stimuli

Stimuli were created using vocal sounds and artificial or natural background. The experiment consisted of a 3 x 3 stimulus design with speaker and background as factors (see Fig. 1). Voice stimuli were recorded from three Dutch speakers (*f*: female, F0 = 246±43Hz [mean ± standard deviation; see 2.4]; *m1:* male, F0 = 172±42Hz; *m2*: male, F0 = 114±24Hz) and consisted of 50 short non-linguistic

vocalizations (e.g. "aww", "uuh"). Stimuli were digitized at 44.1kHz, D/A converted with a resolution of 16bit, band-pass filtered (50Hz–10.75kHz) and downsampled to 22.05kHz. The length of vocalizations varied between 600 and 1000ms (median: 905ms, quartile range: 801-966ms). Backgrounds were either Gaussian white noise or natural auditory scenes (e.g. recordings of rain, cars, cafeteria noise). Natural scenes had a sampling rate of 44.1kHz and 16bit resolution and were downsampled to 22.05kHz. The backgrounds scenes were 1100ms long and cut from a long recording randomly for each of the stimuli. Artificial (i.e. Gaussian white noise) noise was presented at one of five SNRs (6, 0, -3, -6, -9dB) with respect to root mean square (RMS) amplitude. Natural background was presented with equal RMS compared to vocalizations, i.e. 0dB SNR.

To avoid acoustic transients, voices and background stimuli faded linearly with 20ms onset and offset ramps. The stimuli for the three experimental conditions (silence, artificial or natural noise background; abbreviated as *SIL*, *WN*, *NN*, respectively) were created by embedding voice stimuli in 1100ms length segments of silence (SIL), artificial noise (*WN*) or natural noise (*NN*) (see Fig. 1). The voice always started at 50ms and vocal stimuli were as long as noise stimuli (i.e. 1100ms).

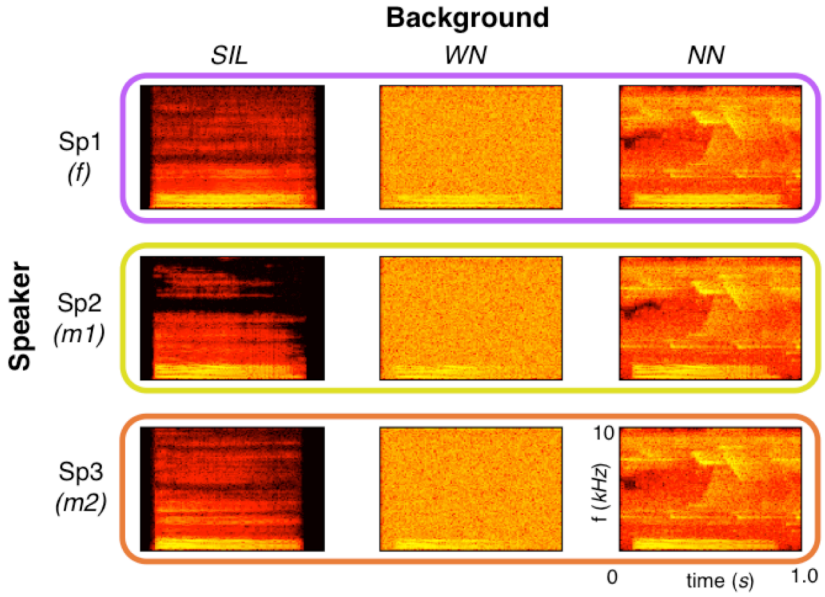



**Figure 1. Design of Stimuli.** In this study, each sound contained a vocalization of one of three voices (f-female, m1/m2-male) and belonged to one of three background conditions: no background (SIL), white noise (WN) or natural noise (NN). Spectrograms for one exemplary vocalization are shown for all speaker by background combinations.

**Stimuli Analysis**

The use of natural stimuli in this experiment (utterances and background) allows presenting subjects to sounds that are similar to everyday life experiences. However, compared to artificial sounds (e.g. tones, broadband noise), properties of natural stimuli are more difficult to control. Thus, we performed a careful analysis of the stimuli characteristics.

We analyzed the recorded vocalizations with respect to F0 or pitch and higher formants (F1-F4) by extracting F0 and formant contours for each stimulus without background noise using Praat v5.3.39 (Boersma, 2001). To retrieve the F0 contours, we used standard Praat parameters with the exception of F0 range: For the female voice we extracted the pitch within a range of 100 to 600Hz, for speaker m1 within [75 400] Hz and for speaker m2 within [60 300] Hz. Subsequently, the contours for single stimuli were summarized using their average and standard deviation (see Table 1 for a summary and Table S1 for properties of single stimuli). We also extracted formant contours (F1-F4) with the suggested parameter setting (the number of formants was set to 4 with maximum frequencies of 6500Hz and 5500 Hz for the female and male voices, respectively). We describe formants by their average frequency over time and computed formant dispersion (FD) as the average frequency difference between formants (see Table 1).

In addition, we analyzed the spectral content of the stimuli both in silence and within the two types of background noise by computing the power spectral density (PSD) with reference to the white noise stimulus of 0dB (dB WN). All analyses were performed using a log-frequency scale using 174 bins between 25 and 10000 Hz. In addition to comparison of spectral power between speakers within each spectral bin (Wilcoxon rank-sum test for all combinations of speaker comparisons), we performed multivariate speaker classification from the stimuli in the three background conditions. We used a linear SVM ($C$ = 1, one vs. one scheme, 6-fold cross-validation) and the PSD estimates (174 bins) as input features to the classifier.

**Experimental Design**

During the fMRI measurements (see below), subjects performed a speaker identification task. To learn the voices and become familiar with the task, participants underwent a practice session outside of the scanner. To familiarize with the voices subjects first listened to one-minute excerpts of reading from each of the speakers. During this initial phase, subjects were shown sketches of faces,

**Table 1. Acoustic Parameters of Sound Stimuli.** Parameters are presented for the three speakers and sounds presented during fMRI data acquisition and behavioral training.

| | fMRI | | | behavioral training | | |
|---|---|---|---|---|---|---|
| | *f* | *m1* | *m2* | *f* | *m1* | *m2* |
| F0 (Hz) | 247 ± 8 | 170 ± 7 | 114 ± 4 | 244 ± 10 | 177 ± 11 | 115 ± 6 |
| Mod F0 (Hz) | 49 ± 5 | 36 ± 4 | 19 ± 3 | 53 ± 6 | 36 ± 5 | 16 ± 3 |
| F1 (Hz) | 725 ± 24 | 662 ± 29 | 584 ± 28 | 719 ± 45 | 671 ± 38 | 597 ± 30 |
| F2 (Hz) | 1993 ± 73 | 1458 ± 57 | 1695 ± 61 | 1825 ± 90 | 1504 ± 75 | 1506 ± 75 |
| F3 (Hz) | 3273 ± 48 | 2794 ± 30 | 2815 ± 47 | 3203 ± 68 | 2815 ± 43 | 2783 ± 39 |
| F4 (Hz) | 4269 ± 38 | 3732 ± 31 | 3843 ± 50 | 4310 ± 57 | 3808 ± 54 | 3831 ± 44 |
| FD (Hz) | 554 ± 23 | 412 ± 18 | 515 ± 20 | 495 ± 30 | 426 ± 25 | 455 ± 26 |
| Dur (ms) | 814 ± 23 | 891 ± 16 | 894 ± 15 | 877 ± 19 | 895 ± 20 | 911 ± 22 |

[1] F0 – average fundamental frequency (± SEM) over time; [2] stdF0 – average standard deviation of F0 over time; [3] F1/F2/F3/F4 - the average first to fourth formant frequencies over time; [4] FD –average formant dispersion (summed differences of adjacent formant frequencies) across stimuli; [5] Dur – the average duration of vocalizations across stimuli.

which were associated with the voices. Then, participants performed a speaker identification task using a subset of stimuli (180 stimuli, i.e. 20 stimuli for each speaker/background combination). They were asked to identify speakers by pressing one of three buttons; the button configuration was indicated by three face sketches at the bottom of the screen and changed between blocks. First, subjects received feedback on their performance after each trial for three blocks. When subjects were able to correctly identify 80% of the trials (average true-positive rate [TPR] for the male voices), they continued with a test blocks where they did the same task but without receiving feedback on their performance trial-by-trial (after one block their average performance was presented). Subjects underwent up to three test blocks in case the 80% criterion had not been met. Participants that performed below the criterion after the last test block were not considered for the fMRI scanning session. During scanning, subjects performed three blocks of the same identification task and no feedback was given. For the fMRI session, stimuli not used in the practice session (270 stimuli, i.e. 30 voice stimuli per speaker/background combination) were presented. Similar to the behavioral training, button configurations changed after each block within the fMRI session to remove the confounding correlation between activities related to button presses and speaker presentation.

**MRI Data Acquisition**

Brain imaging was performed with a 3T Siemens Allegra scanner (head setup) at the Maastricht Brain Imaging Center (Maastricht, The Netherlands). For each subject, three functional runs [mean length: 565 volumes] were collected using a standard echo-planar-imaging (EPI) sequence and high-spatial resolution acquisitions ($2 \times 2 \times 2mm^3$) (repetition time [TR] = 2.6s; acquisition time [TA] = 1.4s, field of view [FOV] = 192 mm x 192mm, matrix size = 128 x 128, echo time [TE] = 30ms). Each volume consisted of 18 slices placed parallel to the Sylvian fissure, covering the temporal and adjacent peri-Sylvian cortex. During measurements, subjects listened - at a comfortable listening level - to the stimuli that were presented binaurally in the 1200ms silent gap between volume acquisitions via MR compatible ear-buds (Sensimetrics S14, Sensimetrics Corporation). The presentation of the stimuli in the silent gap resulted in a clear separation between the acoustic stimuli and the scanner noise and ensured a clear perception of the stimuli. We employed a slow event-related design with an average inter-trial-interval of 15.6s (range 13 – 18.2s; i.e. 5-7 TRs). The sequence of stimuli was randomized for each subject such that each of the three functional runs included ten trials per each stimulus condition (90 trials/run) - resulting in a total of thirty trials per stimulus condition. Additionally, different white noise conditions were balanced across speakers (i.e. for each speaker 6 trials were selected for each of the 6 SNRs [1 in each half-run]) and SNR levels were randomly assigned to different utterances to avoid SNR by utterance interactions. Between the first two functional runs, anatomical images covering the whole brain were obtained with a $1 \times 1 \times 1$ $mm^3$ resolution using T1-weighted 3D ADNI sequence [TR: 2050ms; TE: 2.6ms; 192 sagittal slices]. The slow event-related stimulation scheme provided the possibility of estimating the time-courses of the hemodynamic responses to single presentations of sounds, required for our data analysis.

**Data Processing and Analysis**

*Preprocessing*. FMRI data preprocessing was done with BrainVoyager QX (v2.4, Brain Innovation) and consisted of slice-scan-time correction, motion correction, transformation into Talairach space, temporal high-pass filtering (0.005 Hz) including removal of linear trends and minimal spatial smoothing (Gaussian filter of 2mm FWHM).

*Univariate Analysis*. For univariate analysis of fMRI data, a general linear model (GLM; Friston, 1995) was computed by fitting the blood oxygen level-dependent (BOLD) response with the predicted time series for the three voice or noise classes independent of noise or voice class, respectively. The hemodynamic

response was modeled by convolving the predicted time courses with a canonical (double gamma) hemodynamic response function. We performed a group analysis (fixed-effects [FFX]) of the pairwise contrasts for voice and noise classes. Results were corrected for multiple comparisons by false discovery rate (FDR; Genovese et al., 2002).

*Multivariate Analysis.* In this study, we aimed to decode speaker identities (i.e. based on a labeled activation patterns we tested whether these could be used to predict the speaker in independent trials). In particular, we performed two types classification that differed with respect to the trials used as part of the training and evaluation set. The *within-background* classification was done within one background condition, i.e. both training and testing sets were restricted to examples of the same experimental condition. To test whether models of one condition were able to distinguish speakers in another background condition we also classified voices in an *across-background* scheme, in which case training and testing sets were from different background conditions. This type of decoding aimed at identifying brain activity patterns whose information is robust to background variations.

For these classification analyses, we made use of a recent decoding approach – supervised SOMs (Hausfeld et al., 2012b, chapter 3) – which provide the advantage of handling the present multiclass (three class) problem without the use of binary comparisons. In the following, a short description of the decoding algorithm is provided (for more information see Hausfeld et al., chapter 3).

*Self-Organizing Maps (SOMs).* A SOM (Kohonen, 2001) is a neural network that consists of a rectangular two-dimensional grid with $U$ units. Each unit $i$ is described by a $N$-dimensional weight vector $\mathbf{m}_i = [m_{i1},\ldots, m_{iN}]$ where N is the number of input features. We set the amount of map units $U$ was set to 64. After map units were initialized with random weights within the range of training samples training samples $\mathbf{x}_k = [x_{k1},\ldots, x_{kN}]$ ($k = 1,\ldots, K$) were iteratively presented and the best-matching unit (BMU) $\mathbf{m}_{BMU}$ was selected according to

$$\|\boldsymbol{x}_k - \boldsymbol{m}_{BMU}\| = \min_i(\|\boldsymbol{x}_k - \boldsymbol{m}_i\|), \quad (1)$$

where || denotes the Euclidean distance. In the following, weights of map units were modified with the following update rule:

$$\boldsymbol{m}_{t+1} = \boldsymbol{m}_t + \alpha_t h_{BMU}\, r(t)\, \|\boldsymbol{x}_k - \boldsymbol{m}_i\|, \quad (2)$$

where $t$ denotes the learning iteration, $\alpha_t$ the learning rate and $h_{BMU}(r_t)$ the neighborhood kernel of winning unit $\mathbf{m}_{BMU}$ with radius $r(t)$. Both learning rate and radius of the neighborhood are decreasing functions over time. This results in an early stage that sets the general layout of the map and a subsequent fine-tuning

stage. For SSOM training, we used the MatLab-based SOM-toolbox (http://www.cis.hut.fi/projects/somtoolbox/).

*Supervised SOMs (SSOMs).* One extension of SOMs for supervised learning is based on a modified input vector $\mathbf{x}^*_k = [\mathbf{x}_k\ \mathbf{c}_k]$ for model training that results from concatenating input trials $\mathbf{x}_k$ and a C-dimensional (C denotes the number of classes) class-vector $\mathbf{c}_k = [c_{k1},\ldots, c_{kC}]$ where $c_{ki} = 1$ if trial k belongs to class i and $c_{kj} = 0$ ($j \neq i$) otherwise. Similarly, a vector $\mathbf{v}_i = [v_{i1},\ldots, v_{iC}]$ is appended to the weight vectors of SSOM units $\mathbf{m}_i$ to form $\mathbf{m}^*_i = [\mathbf{m}_i\ \mathbf{v}_i]$ with N+C elements. After SSOM training, map units are ascribed to one class by inspecting the last C elements of the map weight vectors: the index with the largest value determines the label of map unit $\mathbf{m}_i$. Vectors $\mathbf{c}_k$ were not of unit length but of length □ = 0.2, which provided a good compromise between class separation and data fidelity (see Hausfeld et al., chapter 3). For the prediction of testing trials, we computed a measure of the 10 best-matching units (10-*BMU*) to accumulate evidence for classification, which has been found to lead to more robust classification performance (Hausfeld 2012b, Silva and Del-Moral-Hernandez, 2011). In particular, we computed a classification index

$$CI_c = \sum_{i=1\ldots10}(\mathrm{v}_{ic} \cdot \exp(-\|BMU - BMU_i\|^2)\,, \qquad (3)$$

to obtain evidence that trial $\mathbf{x}_{test}$ belongs to class $c$ ($v_{ic}$ denotes the class specific certainty of $i$th best-matching unit and $\mathrm{BMU}_i$ is the $i$th closest map unit). The supervised SOM predicts an unseen trial according to the class obtaining largest *CI*. For more details on how SSOMs were implemented and visualized, the reader is referred to Hausfeld and colleagues (chapter 3).

*Cross-Validation.* The performance of *within-background* classification was assessed using 6-fold cross-validations. Across classes performance was expressed by classification accuracy and the sensitivity index *d'* indicated class-specific performances (*d'* = *z*[true positive rate] – *z*[false positive rate]). For model training we defined datasets consisting of 5 of the 6 half-runs. The remaining half-run was used to determine how well the trained model was able to generalize.

For classification that involved different conditions for training and testing (i.e. *across-background* classification) the same half-run cross-validation scheme was used for training but the total number of trials remained in the test set to obtain a more reliable estimate of the accuracy.

*Voxel Selection and Features.* First, we limited the decoding analysis to an anatomical mask covering auditory responsive regions of temporal cortex. As a feature reduction step, we further reduced the number of voxels with a GLM. This was computed based on the training set and the strongest responding 2000 voxels

were selected. For each voxel, a (double-gamma) hemodynamic response model was fitted to each single-trial response (in percent signal change). The obtained □ values were used as input features for the classifier. We normalized features across trials using inter-quartile-range normalization where median and first and third quartiles were estimated using training trials (cf. Hausfeld et al., chapter 3).

To define feature sets with different numbers of voxels we used an ensemble feature selection method (e.g. Abeel *et al.*, 2010). For each of 25 bootstrap samples, features surviving univariate selection were ranked according to model weights of linear SVM ($C_{svm}$ = 1; one-versus-one scheme). The final ranking was obtained by averaging the ranks across bootstrap samples. We created 12 differently sized feature sets by removing iteratively the lowest ranked 20% starting with 1000 voxels, which resulted in sets of 1000, 800, 640, 512, 410, 328, 262, 210, 168, 134, 107, and 86 voxels.

*Relevance Maps*. To define cortical locations that were important for speaker classification, we used cortex-based alignment (CBA) to transform single-subject selection maps into a common space (Goebel et al., 2006). First, we selected the highest ranked 262 voxels for each subject and created a group map that depicts for how many subjects a vertex was included in the voxel selection procedure. These maps indicate for each vertex the consistency across participants, i.e. areas are color-coded by the number of subjects for which these were important for speaker identification (≥3/5 subjects).

**Statistical Testing of Decoding Results**

We estimated whether classification outcomes were significantly better than chance by using a permutation test (Stelzer et al., 2012). For this permutation approach, we created 99 permutations for each subject. Subsequently, we compared the average classification performance with the real labels to an empirical distribution that was created by drawing with replacement for each subject 99,999 times an accuracy outcome from the single subject permutations and computing their average. To compute the significance level, we counted the number of instances of the permutations that were larger than the accuracy with true labels and divided this value by the number of permutations (one instance was added to both numerator and denominator for a more robust estimate of the significance level).

# Results

## Analysis of Stimuli Properties

The extraction of F0 contours (see example in Fig. 2A and Table 1; Table S1 contains properties of single stimuli) of the vocalizations revealed that the average F0 differs between the three speakers (Fig. 2B, two-sample t-test; f - m1: $t(98) = 8.605$, $p < 10^{-12}$; f - m2: $t(98) = 18.743$, $p < 10^{-33}$; m1 - m2: $t(98) = 8.495$, $p < 10^{-12}$; Wilcoxon rank-sum test; f - m1: $z = 6.663$, $p < 10^{-10}$; f - m2: $z = 8.517$, $p < 10^{-16}$; m1 - m2: $z = 6.877$, $p < 10^{-11}$). Compared to vowels, which have been presented in previous studies (e.g. Formisano et al., 2008) and were recorded here for comparison purposes, we found larger deviation of F0 contours for the stimuli used in this experiment (std $F0_{vowel}$ = 13.26Hz, std $F0_{exp}$ = 34.76Hz; $t_{163} = 3.25$, $p < .001$).

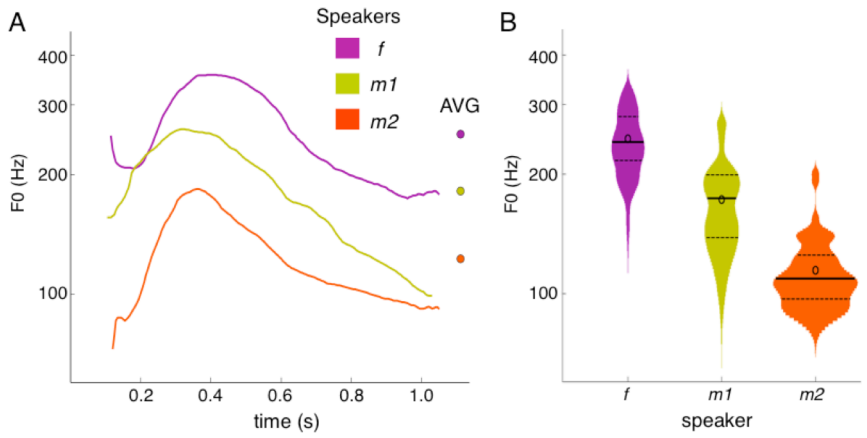



**Figure 2. Fundamental Frequency of Vocalizations.** Panel A shows for one stimulus ("uuh") the F0 contours and their average. Colors denote F0 contours and average for the three speakers. The distribution of average F0 across stimuli is depicted in Panel B. Circles denote the average F0 for speakers and solid and dashed lines denote the median and quartiles, respectively.

The analysis of the frequency spectra for the different voices indicated in which frequency bands the spectral energy differed across speakers. Figure 3A shows the median and 95% confidence interval of spectral power for vocalizations without background. It can be seen that the spectral power significantly differs (Wilcoxon rank-sum test) in three intervals: [60 160Hz], [850 1500Hz] and [3500 7000Hz]. Whereas the first and third interval showed differences between all three voices, the second interval was characterized by lower power for speaker m2. Figures 3B

and 3C show the average spectrum and variation across stimuli for the stimuli obtained by mixing the vocalizations with white and natural noise, respectively. The changes in frequency spectra of the stimuli reflected the properties of the noise, with flat frequency amplitude for white noise and with the 1/f characteristic profile for natural noise. Adding noise to vocalizations decreased the acoustic differences between the stimuli. Spectral differences between stimuli were narrower (compared to silence) for the low frequency interval ([90 150Hz]) both in the white noise and natural noise condition. No amplitude differences between stimuli were found in medium and high frequencies, which indicates that artificial and natural background stimuli masked amplitude differences for frequencies above 250 Hz in this particular set of stimuli.

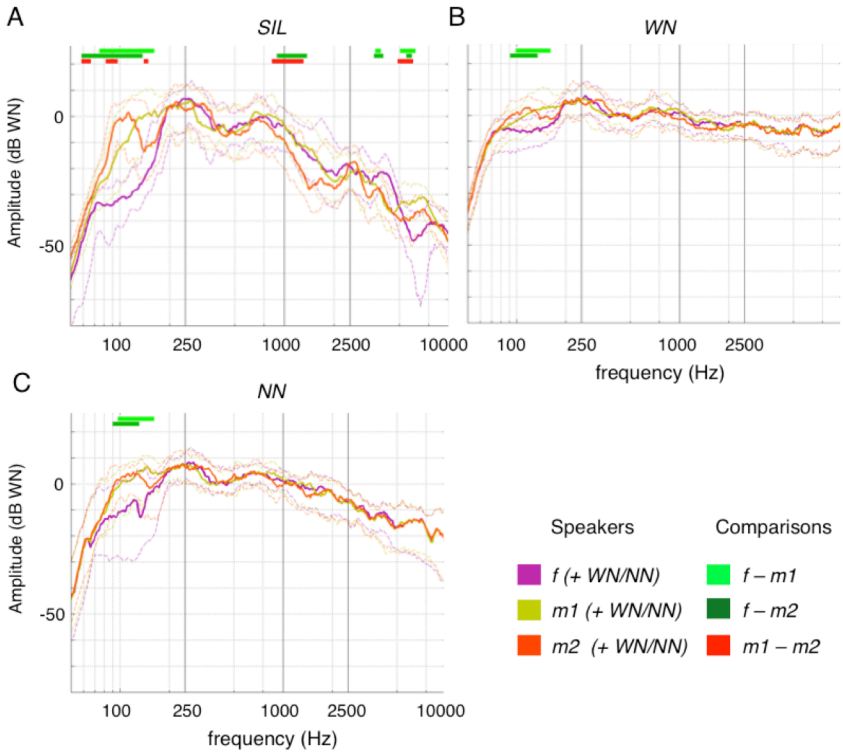



**Figure 3. Sound Spectra for each Speaker and Background Condition.** Average and quartiles of spectra for the three speakers (color) are depicted by solid and dashed lines, respectively. Thick lines above spectra show for which frequencies pairs of speakers have different amplitude. Panel A shows spectra for sounds without background, B for sounds with white noise (WN) and C for sounds with natural background (NN). Amplitude is expressed in dB WN (see text for details).

**Stimulus Classification**

The analysis above considers the different frequency range independently of each other. We therefore conducted a multivariate classification analysis of the stimulus properties which takes into account the entire PSD of the stimuli. As expected, the speaker classification based on the stimuli's PSD (see Fig. 4A) showed better voice identification within silence (accuracy: .844; *d'*: 4.23, 2.25, 2.35) compared to white noise (accuracy: .677; *d'*: 2.20, 1.31, 1.76) and natural noise (accuracy: .603; *d'*: 1.47, 1.28, 1.22).

The analysis of the SVM-weights showed that - in the silence condition - the classification relied on amplitude differences in the F0 range and frequencies above 2500Hz (Fig. 4B). Both of these frequency ranges match the ones detected in the previous analysis (see *Analysis of Stimulus Properties*). Interestingly, the middle range of frequencies that showed differences between speakers in the previous analysis was not important for speaker classification, which might be due to the fact that PSDs in this range are highly correlated to those in the higher frequencies that were weighted high in the classification. The analysis of the SVM-weights for the white noise condition indicated that frequencies in F0 range and between 800 and 1000Hz were important for differentiating speakers whereas high frequencies were weighted low for speaker classification. This outcome reflected the frequency regions of the previous analysis. For stimuli within natural noise, weights modulations were much smaller compared to the former classification models, which might be due to a high variability of weights between splits introduced by the acoustic variations of the natural background. However, the weight pattern

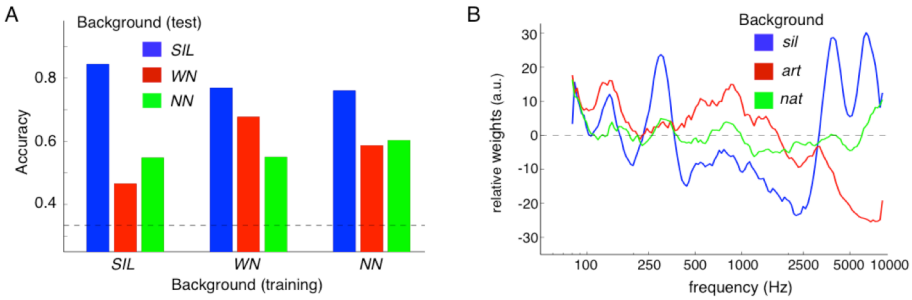



**Figure 4. Decoding Results for Spectral Features.** In panel A bars show the accuracy to classify speakers based on spectral amplitude for all combinations of background conditions for training and test sets. Panel B depicts the importance of spectral features for decoding models. Colored lines show the weights of linear classification models for speaker identity decoding. Weights are averaged across binary comparisons and centered on zero.

was found to follow a mixture of weight modulations of models for stimuli in silence and white noise.

We also tested whether the obtained models could be used to classify speakers in other background conditions. The results (Fig. 4A) showed that models trained in either noise condition could classify stimuli without noise (WN to SIL: accuracy .769, *d'*: 2.68, 1.85, 2.40; NN to SIL: accuracy .760, *d'*: 2.48, 2.07, 1.95). Interestingly, the performance was higher compared to stimuli that belonged to the training condition. In addition to that, the results showed that the SVM model obtained from stimuli without noise was worst to generalize to the other background conditions (SIL to WN: accuracy .465, *d'*: 0.03, 1.18, 0.76; SIL to NN: accuracy .548, *d'*: 1.40, 1.12, 0.99). This indicated that the silence model weighted features that helped to differentiate speakers, which were distorted by the added background noise (e.g. high frequency regions). Conversely, models obtained with white or natural noise were able to generalize to the other noise condition (WN to NN: accuracy .550, *d'*: 0.98, 1.12, 0.87; NN to WN: accuracy .587, *d'*: 1.01, 1.37, 1.26).

**Behavioral Results**

The analysis of reaction times (RT; see Table 2 middle column) in correct trials showed a main effect of speaker ($F_{2,8}$ = 67.280, $p$ < .001) and marginally significant interaction ($F_{4,16}$ = 3.016, $p$ = .050). RTs were faster for the female compared to the male speakers (m1: $t_4$ = -16.029, $p$ < .001; m2: $t_4$ = -6.498, $p$ = .009). A repeated measures ANOVA of response sensitivity (*d'*; see Table 2 right column) revealed a significant main effect of speaker ($F_{2,8}$ = 19.275, $p$ < .001), a main effect of background ($F_{2,8}$ = 6.099, $p$ = .025) but no interaction ($F_{4,16}$ = 1.599, $p$ = .223). Post-hoc tests (Bonferroni-corrected for multiple comparisons) showed more accurate response for the female compared to the two male speakers (m1: $t_4$ = 4.53, $p$ = .032; m2: $t_4$ = 4.26, $p$ = .039) but no significant differences between

**Table 2. Behavioral Results of the Speaker Identification Task.** Average and standard error of the mean are shown for all background-speaker combinations. RT: reaction time; *d'*: sensitivity index computed by *d'* = *z*(true positive rate) - *z*(false positive rate).

| | RT (ms) | | | d' | | |
|---|---|---|---|---|---|---|
| | *SIL* | *WN* | *NN* | *SIL* | *WN* | *NN* |
| *f* | 719 ± 50 | 727 ± 68 | 729 ± 45 | 4.42 ± 0.11 | 3.39 ± 0.16 | 3.77 ± 0.19 |
| *m1* | 959 ± 36 | 998 ± 49 | 1057 ± 45 | 2.59 ± 0.50 | 1.96 ± 0.28 | 2.04 ± 0.34 |
| *m2* | 961 ± 53 | 938 ± 27 | 898 ± 62 | 2.56 ± 0.48 | 2.41 ± 0.31 | 2.24 ± 0.29 |

background conditions (responses in the silent condition showed a tendency to be more accurate compared to WN and NN; $p$ = .057 and $p$ = .0435, respectively [uncorrected]).

**Univariate fMRI Results**

Vocalizations evoked significant fMRI responses in regions consistent with previous studies (Fig. 5A), including auditory regions in the Heschl's gyrus, multiple regions in the planum temporale (PT), in the superior temporal gyrus and sulcus (STG and STS, respectively), as well as in middle temporal gyrus, insular cortex and angular gyrus (Davis and Johnsrude, 2003; Hickok and Poeppel, 2007) (FDR-corrected: $q$ = .01, cluster-size threshold [CS]: 15). In order to test the effect of adding background to the stimuli we contrasted the silence condition with white and natural noise (Fig. 5B). The natural noise condition evoked higher activity compared to the silence condition in lateral Heschl's sulcus (left and right) and right PT ($q$ = .10, CS: 15; green color). Comparing the silent with artificial noise condition, we found higher activity for the white noise condition in left HS ($q$ =

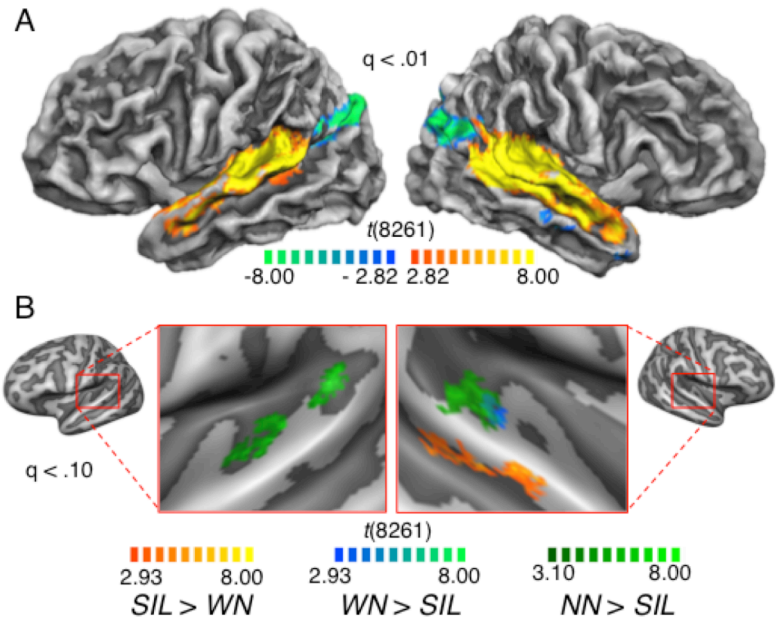



**Figure 5. Results of Univariate Analysis.** Panel A shows overall auditory cortical activation in response to the auditory stimuli as estimated with univariate FFX GLM (FDR-corrected: $q$ < .01; cluster-size threshold: 15). Panel B shows activation differences between background conditions (FFX GLM; $q$ < .10; cluster-size threshold: 15). All other contrasts for backgrounds and voices did not yield significant results.

.10, CS: 15; blue color). In addition we found that right middle STS and STG were more responsive to voices in silence than artificial noise conditions ($q$ = .10, CS: 15; orange color). The comparison of artificial and natural noise revealed higher activity did not show different activation ($q$ = .10, CS: 15).

Pairwise contrasts between different speakers - across or within background conditions – did not reveal any significant difference in BOLD response amplitude ($q$ = .10, cluster-size threshold: 10).

**Multivariate fMRI Results**

*Within-Background Decoding of Speaker Identity*. We first determined whether it was possible to decode speakers from the fMRI responses measured in the silence, WN and NN condition (Fig. 6, left, middle and right line pairs). Our decoding results showed that speaker decoding was possible when vocalizations were presented without interfering background noise (classification accuracy: .436, p = .027) and - to smaller extent - with WN background (classification accuracy: .416; p = .077). For stimuli that contained natural scenes as background, classifications results were not better than empirical chance level (classification accuracy: .411, p = .212).These findings suggest that activity patterns evoked by a wide range of vocalizations can be used to decode speakers. This holds for the silence and white noise condition (only marginally significant) but not for vocalizations accompanied by a natural background.

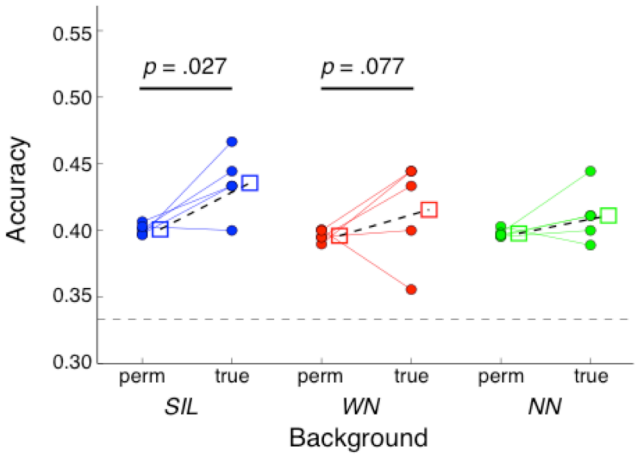



**Figure 6. Within-Background Decoding of Speaker Identity.** Filled circles show classification accuracy of single subjects for permuted labels (i.e. empirical chance level) and performance for real labels. Squares denote average decoding accuracy.

*Across-Background Decoding of Speaker Identity*. In the previous analysis, brain-based decoding of speakers was done within one type of background conditions, i.e. the decoding analysis was restricted to examples of the same experimental condition for training and testing. However, in the across-background decoding we tested whether fMRI responses measured under a specific background condition could be used to differentiate speakers in the remaining experimental conditions. Successful decoding would indicate that the response patterns reflect the representation of features of voices that are robust to changes in the acoustic background.

For this analysis (Fig. 7) we found that models based on stimuli containing only vocalizations and no background could decode speakers mixed to white noise but not to natural noise (WN: accuracy = .379, $p$ = .025; NN: accuracy = .353, $p$ = .854). Models based on stimuli containing artificial noise showed reliable speaker identification for vocalizations without noise but not within natural background noise (SIL: accuracy = .390, $p$ = .002; NN: accuracy = .371, $p$ = .210). Models trained on activation patterns evoked by vocalizations within natural noise could neither generalize to stimuli with artificial nor to the silence condition (SIL: accuracy = .349, $p$ = .934; WN: accuracy = .358, $p$ = .659).

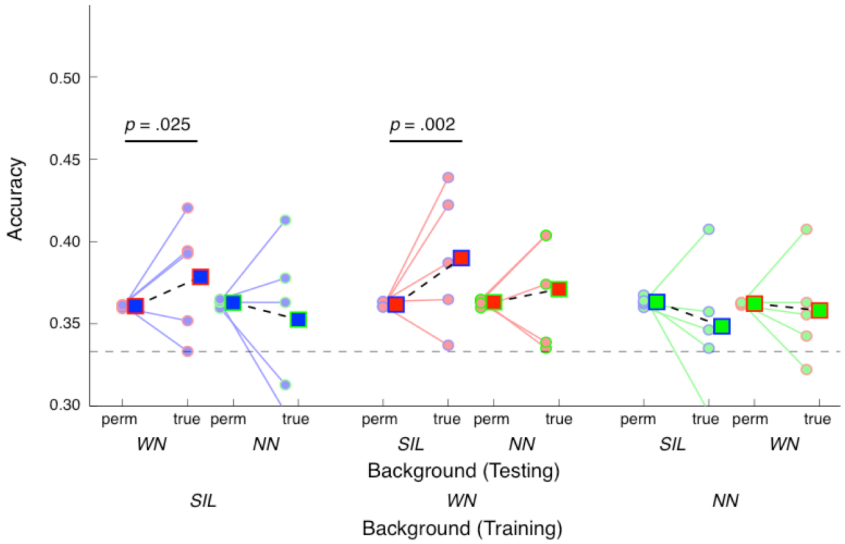



**Figure 7. Across-Background Decoding of Speaker Identity.** Filled circles (frame color denotes testing and fill color the training set) show classification accuracy of single subjects for permuted labels (i.e. empirical chance level) and performance for real labels. Squares denote average decoding accuracy.

*Voxel Selection*. Figure 8 outlines regions that were found to be important for speaker discrimination in the different background conditions. Areas important for speaker identity decoding in silence were localized in bilateral antero-lateral portions of HG/HS, posterior medial HG/HS, right posterior STS and left posterior STG and anterior portion of middle STG. Regions contributing to speaker decoding in white noise were the similar to the ones silence except that the left anterior portion of middle STG was absent. Models in natural noise relied on activation in right PT and antero-lateral HG and left HG/HS, middle and posterior STG.

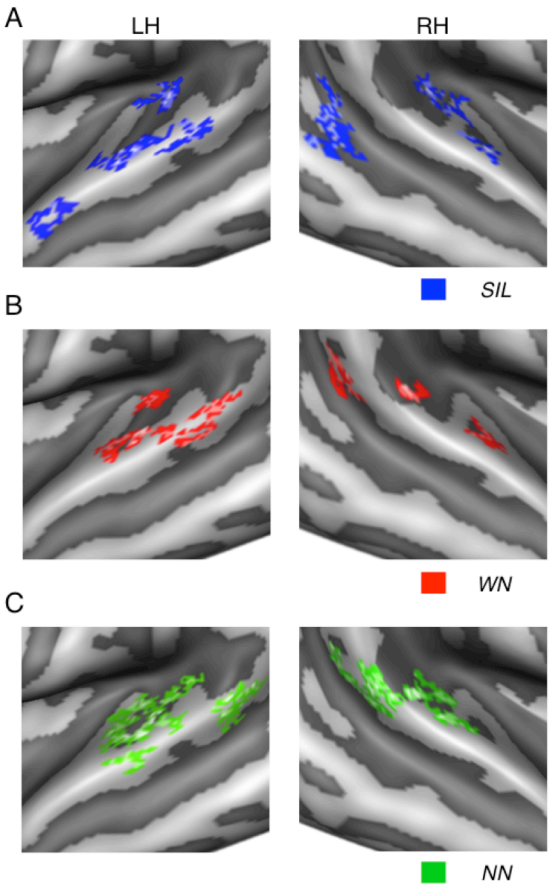



**Figure 8. Consistency Maps of Classification Models for Speaker Identity Decoding.** Colored areas depict areas that were important for at least 3 out of subjects for speaker decoding. Panel A shows areas contributing to speaker decoding for sounds without background, B for sounds with white noise (WN) and C for sounds with natural background (NN) (see text for details).

## Discussion

The goal of the current study was to investigate the representation of speaker identity in the human auditory cortex in the context of noisy auditory scenes. To this end, we used a novel MVPA approach to decode the identity of the speaker from the fMRI responses to a variety of short dynamic vocalizations from three speakers, presented in silence or mixed to artificial or natural background. Our results showed that – based on activation patterns in superior temporal cortex - we could decode speaker identity when sounds were presented in silence and, to a lesser extent, white noise but not when they were mixed to natural noise. Moreover, we found that activity patterns in silence or white noise could be utilized to cross-decode the identity of speakers in the other background condition, i.e. white noise and silence, respectively.

### Speaker Identity Decoding from Diverse and Dynamic Vocalizations

Our results showed that speakers could be identified across a wide variety of vocalizations in the silence condition using activity patterns in auditory cortex. Together with the observation that responses to individual speakers did not differ at level of single voxels. This finding is in agreement with the idea that the auditory cortex forms a model of a speaker's voice that generalizes across a wide range of utterances. These decoding results extend findings from Formisano and colleagues (2008) that classified the identity of speakers using vowel stimuli. In contrast to vowels, the stimuli used here were characterized by a larger variability of temporal modulations both for F0 contours and for higher formants. This implies that the informative spatial patterns used for decoding cannot solely reflect the responses to a stationary pattern of fundamental frequency and formants. The response patterns may rather reflect a speaker representation - possibly derived from the processing of these acoustic parameters - but is at the same time robust to their variations across different utterances by the same speaker.

### Speaker Decoding of Vocalizations in White Noise

In addition to successful decoding of speakers without interfering noise, we found a trend for information about speaker identity in activation patterns evoked by sounds with added white noise. The acoustic analysis of stimuli in white noise showed that spectral amplitude differences informative of speaker identity were mostly in the low frequencies (F0 range), whereas the contribution in the higher frequency ranges (present in the original utterances) were masked by the addition of white noise. This suggests that successful speaker decoding for vocalizations in white noise relies mostly on a representation derived from a speaker characteristic F0. Further support for this interpretation arises from the finding that a region in

left middle STG, which we interpret to process specifically speaker-related information in higher frequency bands (> 2500Hz), did not seem to contribute to speaker decoding as opposed to models established with sounds in silence (see *voxel selection*).

**Across-Background Decoding of Sounds in Silence and White Noise**

Remarkably, our most significant decoding compared to chance was found for models trained with white noise stimuli applied to classify speaker identity in the silence condition. Moreover, successful decoding was as well possible for the opposite case in which models trained with sounds in silence could predict the speaker of vocalizations within white noise. This reciprocity indicates a similarity between the evoked activation patterns in these two experimental conditions and again suggests that the speaker-specific information derives from some type of processing and transformation of the F0 frequency range.

**Decoding of Speakers in Natural Noise at Chance Level**

In contrast to stimuli in silence and white noise we could not decode the speaker's identity when presented within natural scenes. Based on the analyses of the acoustics of the stimuli (see Fig. 3C) and of the behavioral performances of the listener, this null result would not be expected (see Table 2). Apart from the possible methodological causes (e.g. insufficient sensitivity, choice of decoding algorithm, voxel selection approach, randomized natural scenes), there might be two interesting explanations. First, it might be that – despite the similarity of the average acoustic properties between WN and NN - the larger acoustic variability of natural scenes (10 different ones) may have led to a higher variability in the measured activation patterns. In turn, may have influenced the quality of the training and the performance of the brain-based classifier. Another possible explanation for the negative outcome is that natural noise - as opposed to white noise background - consisted of real-world auditory objects. Compared to white noise, the natural noise condition entails auditory scenes incorporating two or more natural objects. Although this is not evident from our behavioral results, such a setting might require the involvement top-down cortical mechanisms of attentional and stream selection, which may differ scene-by-scene. Similar to the effects of acoustic variability, this may result in auditory cortical responses that differ substantially trial-by-trial within the NN condition and from those evoked by vocalizations without background or within white noise.

**Regions Involved in Speaker Identity Decoding**

MVPA as employed here tests for information about experimental conditions in distributed activation patterns spread across the whole cortical surface covered

by fMRI measurements. However, we found localized patches on bilateral temporal cortex that contributed considerably more than other regions to successful speaker identity decoding. In the following, we will discuss these clusters in the light of previous studies concerned with voice and speech processing. For silence and white noise models, we found three clusters in the right hemisphere (posterior STG/STS to anterolateral HG and medial HS/PT) and three in the left hemisphere (posterior STG, anterolateral HG, medial HS) to be important for speaker identification. In addition to that, a cluster in left anterior STG contributed to speaker identity decoding for models based on vocalizations in silence but not white noise. As voices could not be distinguished within natural background, we will not look more specifically on classification models of this condition.

Recent findings showed that bilateral anterolateral HG seems to be involved in particular in processing of voice pitch (Griffiths et al., 2001; Jamison et al., 2006; Kriegstein et al., 2007; 2010; Patterson et al., 2002; Penagos et al., 2004) and pitch in complex non-voice stimuli (Ley et al., 2012). This is in accordance with the stimulus analysis and models of voice decoding with sounds that showed that voice pitch could be used to differentiate between speakers (see Figs. 2 and 3) in all background conditions.

A large cluster that contained information about speaker identity was found in right posterior STS/STG, which has been suggested to reflect acoustic similarity of voices (Andics et al., 2010) and to be involved in the extraction of speaker-related vocal tract parameters (e.g. formants; Kriegstein et al., 2010). This is in line with the classification of vocalization spectra that revealed that frequencies in the range of first, third and fourth formants were important for speaker identification for the silence (F3, F4) and white noise condition (F1). Formisano and colleagues (2008) found this cluster to contain information in both speaker and vowel classification, which is in agreement with the high entanglement of speaker-related vocal tract parameters (formant patterns) and speech content (vowel) (e.g. Turner et al., 2009).

For classification of speaker identity in silence, we found a cluster in left mid STG to contain information about speakers, which was absent in the white noise condition. We speculate that this cluster reflects the importance of higher frequencies (> 2500 Hz, including F3 and F4), as suggested by the speaker classification analysis condicted on the stimuli. First, amplitude differences in high frequencies did not differentiate speakers in the presence of white noise, which parallels the result that we do not find this region in the white noise condition. Second, Bonte and colleagues (chapter 4) find this cluster to be involved in

speaker classification for a set of speakers that contained two children voices that have been found to be distinguished in particular based on higher formants (e.g. Perry et al., 2001). Thus, we speculate that left mid STG might reflect processing of specifically fast changing features (F3, F4) that indicate speaker identity. The absence of this cluster in the study by Formisano and co-workers (2008) might be due to the passive paradigm (a similar cluster has been described by Bonte et al. [chapter 4] and Kriegstein et al. [2003] during an active speaker recognition task) or by the richer spectro-temporal structure of sounds presented in this study.

Interestingly, regions in middle and anterior STS/STG that have been associated with voice (Belin et al., 2000) and, specifically, voice identity processing (Andics et al., 2010; Belin et al., 2004; Belin and Zatorre, 2003; Bonte et al., chapter 4, Formisano et al., 2008; Kriegstein et al., 2003; Kriegstein and Giraud, 2004) did not contribute to speaker identity decoding in this study. A possible explanation for the negative outcome might be that these regions represent features that do not differentiate between speakers for the stimuli we employed. In contrast to other studies examining voice identity that either used vowels (Formisano et al., 2008; Bonte et al., chapter 4) or syllables (Andics et al., 2010; Belin and Zatorre, 2003), we used non-linguistic dynamic vocalizations. However, von Kriegstein (2003, 2004) employed stimuli that contained temporal modulations of fundamental and formant frequencies (spoken sentences) and showed that right anterior STG/STS responded stronger when participants attended voice identity compared to speech content. Note, however, that such results provide evidence that this region is involved in processing of non-verbal features of speech; however, they do not allow concluding that this region could also distinguish between speakers.

**Limitations and Future Directions**

Our experimental settings were restricted to three speakers, which raises the question whether our results would also hold for a larger set of talkers. Related to this limitation, Lavner and colleagues (2000) found that participants change the weighting of voice cues during speaker identification of vowel sounds depending on the speaker's voice characteristics. Nevertheless, the fundamental frequency and F1 and F3-F4 are important cues important cue to predict speaker similarity ratings in large sets of female and male speakers (Baumann and Belin, 2010; Lavner et al., 2000). Presenting a larger set of speakers in a similar paradigm with more participants is necessary to show that the results are valid in a more general setting.

Another limitation is that the significance of the effects relied on group statistics calculated on a small set of five subjects. Thus, it might be that possible

effects (e.g. successful decoding within stimuli with added natural background) were not detected because of low sensitivity of these analyses.

Finally, it could be possible case that our results were due to the specific SOM-based decoding approach we employed. However, we found similar results when using SVMs instead of SSOMs (outcomes not shown) to decode speaker identity. All these classification approaches, however, are based on *discriminative* models, which estimate the informative without using any information on the acoustic properties of the stimulus. It would be interesting, in future studies, to use a *generative* approach (e.g. encoding, Kay et al., 2008; Mitchell et al., 2008; Moerel et al., 2012; Naselaris et al., 2009) and test more directly the cortical representation of specific acoustic parameters of the stimuli and their contribution to speaker identification.

## Conclusions

Our results showed that speaker identity could be decoded above chance for sounds without background noise and with white noise but not within natural noise. Similarly we found that activation patterns in silence could predict speaker identity of patterns evoked by sounds in white noise and vice versa. Based on these findings, we suggest that neuronal populations in AC form representations of human voices, which are mostly based on the extraction of the speaker-characteristic pitch (F0) but are robust to the within-speaker variability across utterances

### Acknowledgements

This work was supported by Maastricht University (LH, EF) and by the Netherlands Organization for Scientific Research (NWO): VICI Grant 453-12-002 (EF). We thank Giancarlo Valente for comments on decoding and statistical testing.

### 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.

Baumann, O., Belin, P., 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychol Res 74, 110–120.

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.

Belin, P., Zatorre, R.J., Ahad, P., 2002. Human temporal-lobe response to vocal sounds. Brain Res Cogn Brain Res 13, 17–26.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B., 2000. Voice-selective areas in human auditory cortex. Nature 403, 309–312.

Binder, J.R., Liebenthal, E., Possing, E.T., Medler, D.A., Ward, B.D., 2004. Neural correlates of sensory and decision processes in auditory object identification. Nat Neurosci 7, 295–301.

Bishop, C.W., Miller, L.M., 2009. A multisensory cortical network for understanding speech in noise. J Cogn Neurosci 21, 1790–1804.

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

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.

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G.A., Kawahara, H., Belin, P., 2010. Vocal attractiveness increases by averaging. Curr Biol 20, 116–120.

Charest, I., Pernet, C., Latinus, M., Crabbe, F., Belin, P., 2012. Cerebral processing of voice gender studied using a continuous carryover fMRI design. Cereb Cortex 23, 958-966.

Charest, I., Pernet, C.R., Rousselet, G.A., Quiñones, I., Latinus, M., Fillion-Bilodeau, S., Chartrand, J.-P., Belin, P., 2009. Electrophysiological evidence for an early processing of human voices. BMC Neurosci 10, 127.

Cherry, E.C., 1953. Some experiments on the recognition of speech with one and with two ears. J Acoust Soc Am 25, 975–979.

Davis, M.H., Johnsrude, I.S., 2003. Hierarchical processing in spoken language comprehension. J Neurosci 23, 3423–3431.

Dellwo, V., Huckvale, M., Ashby, M., 2007. How is individuality expressed in voice? An introduction to speech production and description for speaker classification, In: Speaker Classification I, Lecture Notes in Computer Science. Springer.

Ding, N., Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A 109, 11854–11859.

Ethofer, T., Bretscher, J., Wiethoff, S., Bisch, J., Schlipf, S., Wildgruber, D., Kreifelts, B., 2013. Functional responses and structural connections of cortical areas for processing faces and voices in the superior temporal sulcus. NeuroImage 76, 45–56.

Ethofer, T., Van De Ville, D., Scherer, K., Vuilleumier, P., 2009. Decoding of emotional information in voice-sensitive cortices. Curr Biol 19, 1028–1033.

Formisano, E., De Martino, F., Bonte, M., Goebel, R., 2008. "Who" is saying "what?" Brain-based decoding of human voice and speech. Science 322, 970–973.

Formisano, E., Moerel, M., Bonte, M., (accepted). Functional MRI of the human auditory cortex, In: Functional MRI: From nuclear spins to brain function, Uluda□ , K. U□ urbil, K., and Berliner, L. J. (Eds.). Springer

Friston, K.J., 1995. Statistical parametric maps in functional imaging : A general linear approach. Hum Brain Mapp 2, 189–210.

Genovese, C., Lazar, N., Nichols, T., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15, 870-878.

Goebel, R., Esposito, F., Formisano, E., 2006. Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: From single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. Hum Brain Mapp 27, 392–401.

Griffiths, T.D., Uppenkamp, S., Johnsrude, I., Josephs, O., Patterson, R.D., 2001. Encoding of the temporal regularity of sound in the human brainstem. Nat Neurosci 4, 633–637.

Grossmann, T., Oberecker, R., Koch, S.P., Friederici, A.D., n.d. The developmental origins of voice processing in the human brain. Neuron 65, 852–858.

Hausfeld, L., De Martino, F., Bonte, M., Formisano, E., 2012a. Pattern analysis of EEG responses to speech and voice: Influence of feature grouping. NeuroImage 59, 3641–3651.

Hausfeld, L., Santoro, R., Valente, G., Formisano, E., 2012b. Classification and visualization of multiclass fMRI data using supervised self-organizing maps, in:. Presented at the Pattern Recognition in NeuroImaging (PRNI), 2012 International Workshop on, IEEE, pp. 65–68.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat Rev Neurosci 8, 393–402.

Hill, D.R., 2007. Speaker classification concepts: Past, present and future, in: Speaker Classification I, Lecture Notes in Computer Science. Springer.

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.

Jamison, H.L., Watkins, K.E., Bishop, D.V.M., Matthews, P.M., 2006. Hemispheric specialization for processing auditory nonspeech stimuli. Cereb Cortex 16, 1266–1275.

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

Kilian-Hütten, N., Valente, G., Vroomen, J., Formisano, E., 2011. Auditory cortex encodes the perceptual interpretation of ambiguous sound. J Neurosci 31, 1715–1720.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun 52, 12–40.

Kohonen, T., 2001. Self-organizing maps. Springer.

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.

Kriegstein, von, K., Giraud, A.-L., 2004. Distinct functional substrates along the right superior temporal sulcus for the processing of voices. NeuroImage 22, 948–955.

Kriegstein, von, K., Smith, D.R.R., Patterson, R.D., Ives, D.T., Griffiths, T.D., 2007. Neural representation of auditory size in the human voice and in sounds from other resonant sources. Curr Biol 17, 1123–1128.

Kriegstein, von, K., Smith, D.R.R., Patterson, R.D., Kiebel, S.J., Griffiths, T.D., 2010. How the human brain recognizes speech in the context of changing speakers. J Neurosci 30, 629–638.

Latinus, M., Belin, P., 2011. Anti-voice adaptation suggests prototype-based coding of voice identity. Front Psychol 2, 175.

Latinus, M., Crabbe, F., Belin, P., 2011. Learning-induced changes in the cerebral processing of voice identity. Cereb Cortex 21, 2820-2828.

Latinus, M., McAleer, P., Bestelmeyer, P.E.G., Belin, P., 2013. Norm-based coding of voice identity in human auditory cortex. Curr Biol 23, 1075–1080.

Lavner, Y., Gath, I., Rosenhouse, J., 2000. The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. Speech Commun 30, 9–26.

Lavner, Y., Rosenhouse, J., Gath, I., 2001. The prototype model in speaker identification by human listeners. J Speech Technol 4, 63–74–74.

Ley, A., Vroomen, J., Hausfeld, L., Valente, G., De Weerd, P., Formisano, E., 2012. Learning of new sound categories shapes neural response patterns in human auditory cortex. J Neurosci 32, 13273–13280.

Mesgarani, N., Chang, E.F., 2012. Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485, 233–236.

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., Prenger, R.J., Kay, K.N., Oliver, M., Gallant, J.L., 2009. Bayesian reconstruction of natural images from human brain activity. Neuron 63, 902–915.

Papcun, G., 1989. Long-term memory for unfamiliar voices. J Acoust Soc Am 85, 913–925.

Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., Griffiths, T.D., 2002. The processing of temporal pitch and melody information in auditory cortex. Neuron 36, 767–776.

Penagos, H., Melcher, J.R., Oxenham, A.J., 2004. A neural representation of pitch salience in nonprimary human auditory cortex revealed with functional magnetic resonance imaging. J Neurosci 24, 6810–6815.

Perry, T.L., Ohde, R.N., Ashmead, D.H., 2001. The acoustic bases for gender identification from children's voices. J Acoust Soc Am 109, 2988–2998.

Staeren, N., Renvall, H., De Martino, F., Goebel, R., Formisano, E., 2009. Sound categories are represented as distributed patterns in the human auditory cortex. Curr Biol 19, 498–502.

Stelzer, J., Chen, Y., Turner, R., 2012. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): Random permutations and cluster size control. NeuroImage.

Turner, R.E., Walters, T.C., Monaghan, J.J.M., Patterson, R.D., 2009. A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. J Acoust Soc Am 125, 2374–2386.

Warren, J.D., Scott, S.K., Price, C.J., Griffiths, T.D., 2006. Human brain mechanisms for the early analysis of voices. NeuroImage 31, 1389–1397.

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., Poeppel, D., Schroeder, C.E., 2013. Mechanisms underlying selective neuronal tracking of attended speech at a "Cocktail Party." Neuron 77, 980–991.

# 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.

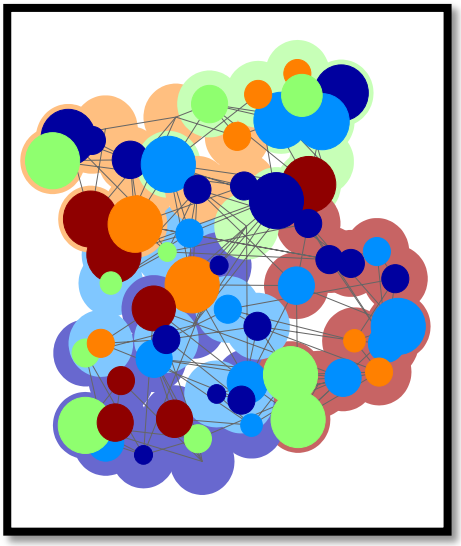
# Acknowledgements

Dear reader,

you wouldn't be able to read this personal and, thus, special piece of writing without help and support by many people. Without them, this thesis would not exist. Empirical scientific projects, their design, data acquisition, analysis, writing and ultimately publishing require, in my opinion, *always* a group effort (even more so in a multidisciplinary field like cognitive neuroscience). I tried to give credit to everyone who worked on and contributed (in a broad sense) to this PhD thesis: colleagues, friends at work and family.

Elia, you are the first person I'd like to acknowledge. You supervised me during the PhD the last four years, before that – together with Federico – during my Master's study and will continue on providing supervision during my post-doctorate. Thanks a lot for your invaluable help, precise comments, trust, long-lasting manuscript meetings, short advices, support and motivation when necessary, for sharing ideas, great sessions of brainstorming (which I loved a lot), for letting me work independently on weird and funny colorful grids, allowing me to follow many different directions and at the same time reminding me not to get lost in side tracks. I'm very happy that we will work together for the upcoming years and look forward to this new, exciting period that had a promising start already.

Thanks also to my co-supervisors Milene and Giancarlo. Giancarlo, my dearest *technical* supervisor, during these four years you were my invaluable source and authority when it came to data analysis. This included sharing programming insights, theories of statistics, simulated annealing and – for 90% of all time – discussions and hints concerning non-parametric approaches for statistical testing. I have a hard time thinking of one case for which we did not end up with my computer literally running for weeks collecting outcomes for *properly* randomized data. Thanks also for sharing 1) your seemingly never ceasing supply of snoepjes, speculaas and drops as well as 2) my enthusiasm over aesthetically pleasing, *abstract art*-like results (see right).



I also want to acknowledge my *voice* co-supervisor Milene. Besides contributing *essentially* to this thesis by providing superbly cleaned and corrected EEG data (chapter 2) and for offering me the possibility to work on the fMRI dataset by means of various analyses (chapter 4), you also helped a lot with your insights into voice processing discussing the outcomes of the 5th chapter. Writing this, I remember that during my Master studies you were the one person that provided me with much appreciated critical and detailed feedback on my writing and that during your course I reestablished my enthusiasm for EEG which will always remain one method of choice to me (despite the strong focus on fMRI around).

Federico, I think I wouldn't be at this stage and at this place without my internship under your daily supervision during my Master's. It did not only result in my first publication, but also provided the basic analysis framework (MVPlab) for upcoming projects reported in this thesis and studies of colleagues opting for a decoding approach. I'll keep your numerous BBQs, football invites and one particular concert in best memory! Thank you for letting me *park* my furniture in your apartment. I'm very much looking forward to learning from your high-field knowledge necessary for my future work.

Christl, Riny, Annemie and recently Eva, thank you for your help with all these forms, reimbursements and organizational things in general. I know it's not an easy job and I think you fulfill it tremendously.

Next I want to mention current and past members of our auditory group: Elia, Federico, Giancarlo, Milene, Lars, Noël, Julien, Nick, Anke, Roberta, Fren, Sanne, Kiki, Niels and Vittoria (it's probably a non-exhaustive list). I'm very thankful and enjoy the open and informal discussions, presentations and proposal meetings always providing critical and constructive feedback. Lars, it is impressive - and I can't stress enough how thankful I am - that you keep on organizing our meetings and that you always managed to revive the group and provide a new schedule. Thanks also for asking these nasty 'detail questions' that are in the end necessary for good experimental designs and conclusions!

I want to thank the group of people forming the Cognitive Neuroscience department. In the corridors and offices we inhabit and inhabited (the following is not a 'building effect') I found a very open, international, supportive, friendly and chatty environment, which makes this department an exceptional place to work in. Together I we could also spend some great time with you during conferences at places like Barcelona, Washington DC, Lausanne or London.

I enjoyed coming to work thanks to great office mates during my PhD years: Anniek, Gonny, Roberta, Franzi and Arne. Franzi, I appreciate you for your

opinions and I never want to miss our fun or more serious conversations about work, tea, private matters and *you know what*. I cherish these really a lot and I'm very glad we managed to stay in the same office and keep them up! Joao, Gojko, Martin, Marin, Britta, Joost, it was a great trip and idea to go running in Palma (although only four of us ended up really running). Kamil, thanks for initiating going to the movies and spending nice afternoons and evenings at fun places in Maastricht. The latter also holds for Joao, Rosanne, Alex, Marin, Matteo and Britta. Our weekend brunches, lunches or dinners were nice and had this lovable, lazy Sunday spirit. Thanks to Marin, Britta, Anke and Joao, for climbing/bouldering evenings. Kiki, during the last year we somehow always ended up talking a lot at every PhD party I can think of, we finally managed to meet for some pool and had a very nice day at PinkPop. I enjoyed this shared time very much. Next, I want to acknowledge the ones sharing rides to Maastricht during my time in Aachen: Anke, Valentin, Arne, Mario. You made the often dull commuting hours a lot easier to pass by. Franzi and Thomas, at your lovely place we had this combined dinner-tea tasting ☺ and a beautiful summer breakfast. I hope (and I'm confident) we manage to have more of these!

Cara Roberta, I have to admit I don't really know where to start. There are some facts that show we somewhat *met*: sharing one office for more than 4 years, working in the same group with the same people and the same supervisor, or me living at your place for some time. We watched numerous movies at your place (most often classic ones in black and white with some handsome tall guy being at least in his 40s) or at Lumière (then more modern movies but still classy). We had countless, joyful dinners at your place or somewhere in Maastricht (especially at restaurants serving sushi). During our work-related and private discussions I liked your refreshing, different take on things with which some problems disappeared without much further ado afterwards. There are many more points I could mention but I want to end with thanking you for being the *perfect* office mate and above all a great and trustworthy friend!

Felix and Gesa, we know each other since I arrived in Maastricht for studying here and you introduced me to this different environment (and studying here was different). We had walks to prepare for and relax before exams, numerous BBQs at Felix' and Thomas' place back then and you supported me at handball. I enjoyed our vacations or long weekends at the Baltic Sea, Dutch coast, Ostenfelde and - not a long time ago – our longer tour through France (à propos: I'll always owe you for Corse). You spoiled me with almost every dish that was prepared at your place and surprised me countless times with unexpected tastes. Gesa, I'll never forget my first *real* encounter with horses, me being overwhelmed and

feeling unable to cope at first: it was an impressive experience. Thank you also for your great choir performances at locations each with its own dignified spirit. Felix, if you don't knock at or just show up in our office for one or two days I'm getting worried ☺ Thank you for so many ideas, advices, explanations and discussions *über Gott und das Leben* (also literally). For me, you are one true scientist that avoids the trap of being impressed and convinced at the same time by some great figure of some fancy method, that remains curious and that possesses the sublime gift of expressing something seemingly complex in simple, clear cut terms.

Joao, we spend a lot of time together during the last year, at work and even more so during spare time. There were many nights at Zondag, Take 5, dinners at your place, the take-away Thai, this Indian restaurant, weekend lunches at coffee lovers and again Zondag. There was the intense time for Alex' movie when we had to cope with each other at least 8 hours a day (and night) for 3, 4 consecutive days shooting and editing ☺. We share a passion for sports (e.g. bouldering which we have to pick up again) and some inexplicable weakness for fancy analyses ending up in pretty figures. We discussed various important things which I won't mention right here. I love your thinking – and you convinced me by now – that everything nice has some Portuguese roots.

Anke, you already mentioned in your acknowledgements that we followed the same trajectory before ending up in Maastricht being PhD students (if I'm not mistaken we even attended once the same course in Osnabrück). Taken out of context the following might sound weird: *you* were the reason I decided to move to Maastricht! Providing context: indeed, without discussing future directions with you while being participant in your research, I never would have considered to follow the Master in Maastricht. Our friendship (that amongst many other things included bouldering, cooking, board games, wine, tea and coffee) changed during the last years in ways I didn't think of. To me it is still remarkable how you always understand and follow what I want to express while I'm having trouble saying it; just saying 'you know what I mean' or even a simple gesture or glance is enough. I'll never forget how we strongly supported, consulted and could rely on each other during troubled times. Thank you for being the true friend you are!

It's time to give credit to my Maastricht housemates who added to the multinational, casual and *fabulous* spirit accompanying the house at Hertogsingel (first Britta and Alex and now Anna and Gojko). During the last year I felt home in Maastricht and in particular this old and charming house which to a large extent was because of you. Some non-exhaustive list of events: BBQs and dinners at Britta's, discussions with Alex about various sorts of movies (from splatter to Nouvelle Vague to Hitchcock), evenings with nice wine going along with *deep* or

"deep" topics (Britta, Alex, Gojko, Anna), watching movies or boxing (Alex, Britta) and *finally* playing some board games (Gojko, Anna).

Roberta, Matteo and Tommasso, I want to let you know that I'll always keep my *Italian* time at your place in best memory and I'll always be grateful that you accommodated me without hesitating. I immediately felt (and still feel) welcome at your place where you enriched my life with great Italian dishes, espresso in the morning, great dinner and post-dinner discussions. Matteo, thanks a lot for your salmon and chicken dishes, for many caffè espressi in your office and letting me know that *jesus died for somebody's sins but not mine*. Tell me if I'm wrong, but during this time I think I even developed some sort of rudimentary and, unfortunately, still pre-existing form of Italian (correlating maybe with my official but only temporary Italian citizenship). I know that it's not easy to host one extra person for two months. Thank you for taking this risk and - something that came natural to you and probably went by unnoticed – *critical support* during this special time of mine.

Talking about accommodation… Anke and Sven, I don't know how often I made use of your offer to stay over in Aachen. Thanks for that, also and especially when I asked less than a day in advance. I loved my room ☺

Nick, you were a great colleague in Maastricht. We *ran* into each other on a weekly basis during football in Aachen (both of us having a more physical interpretation of playing although we easily could make use of our superior technical skills) and you even invited this funny bunch of people (Sven, Tim, Campmann, Thomas) to your place in New York (establishing the concept of the *philosophical fire escape*). Thank you for initiating great trips to the French Alps and Ardennes. All of these three travels were special and brilliant in their own way and I'm happy I had the chance to join every single one of them. Tom, Michelle, Job, I always enjoyed spending time with each of you (and all of you together) at Thembi's, Café Zuid or your places.

There have been more happenings and events that I missed and, more importantly, people I forgot to mention: Christianne, Judith, Peter, Rainer, Bernadette, Jan, Anna, Sanae, Sanne, Mehrdad, Inge, Katie, Valerie, Francesco, Martin, Joel, Judith and *many* others. You all make Maastricht and this department a great place for prospering high-quality research with impressive, sophisticated projects and at the same time extend it to a place with more than just professional encounters!

Anna, die meiste Zeit meines Doktoranden-Daseins haben wir gemeinsam verbracht – als Neuzugezogene in Aachen in einer unserer Wohnungen mit zwei weiteren kleinen Mitbewohnern. Wir haben viele schöne, rührende, traurige, intensive Erlebnisse geteilt. Wir haben voneinander gelernt, und Dinge von uns, und über uns erfahren, von denen ich weiß, dass sie beständig in unseren weiteren Leben nachhallen werden. Du hast mich in meiner Arbeit großartig unterstützt (dabei viel ausgehalten), mir geholfen Dinge anders zu sehen und (nicht zu unterschätzen) Arbeit Arbeit sein zu lassen. Ohne dich wäre der PhD gewiss anders verlaufen mit weniger schönen und nicht so schönen Ereignissen und vor allem weniger Wärme und Nähe. Ich werde unsere gemeinsame Zeit niemals missen und Dich *immer* in Erinnerung behalten. Für all deine Unterstützung, Hilfe und Verständnis während der schwierigen Zeiten des PhDs: danke! Für alles andere: DANKE!!!

Janni, mon petit frère, il mio fratellino: schön, dass du wieder in der *Nähe* bist und nicht mehr auf der anderen Seite des Atlantiks. Du warst natürlich immer ein *willkommener Grund* in die amerikanische Provinz (Roanoke, VA) oder die frankophone kanadische Metropole zu reisen (Montréal). Seltsamer- und glücklicherweise klappte das durch Konferenzen, Summer Schools und Urlaube häufiger als gedacht. Trotz der räumlichen Distanz seit Beginn des Studiums haben wir es geschafft, einiges an gemeinsamer Zeit zu finden. Gerade fallen mir Dinge ein wie das Jazz Festival, die Canadians, die Zeit in Konstanz zu deinem Einzug, einen nicht wirklich gelungenen Umzug meinerseits von Maastricht nach Aachen, Besuche deinerseits bei mir und die Woche Segeln in den Kykladen. Letzteres war ein toller Abschluss der freien Zeit nach Abgabe der Dissertation, an die ich mich gerne und lange erinnern werde (Gruß an Urs, Kati, Judith, Julia und Thomas). Zu guter Letzt: Ich finde es interessant, dass wir uns beruflich/thematisch annähern (ein wenig zumindest); ich bin gespannt wohin das bei dir führt und wünsche dir dabei viel Glück und Erfolg!

Sylvia und Thomas, vielen Dank für eure immerwährende, bedingungslose Unterstützung sowohl vor dem Studieren als auch während meiner Studentenzeit in Osnabrück und Maastricht und auch in den letzten 4 Doktorandenjahren in Maastricht! Danke für euren Rat bei wichtigen Entscheidungen und eure praktische Hilfe. Ich wusste immer, dass ich in *jeder* Angelegenheit auf euch zählen kann, was meinem Leben eine wichtige Stabilität gab. Auch wenn ich nun nach Bayern muss und nicht mehr ins heimelige Hannover, so freue ich mich trotz allem auf jeden Besuch und jedes Wiedersehen.

# Publications

**Hausfeld, L.**, De Martino, F., Bonte, M., Formisano, E., 2012. Pattern analysis of EEG responses to speech and voice: Influence of feature grouping. *Neuroimage* 59, 3641-3651.

Riecke, L., Vanbussel, M., **Hausfeld, L.**, Baskent, D., Formisano, E., Esposito, F., 2012. Hearing an illusory vowel in noise: Suppression of auditory cortical activity. *J Neurosci* 32, 8024-8034.

Ley, A., Vroomen, J., **Hausfeld, L.**, Valente, Gl., De Weerd, P., Formisano E., 2012. Learning of new sounds shapes neural response patterns in human auditory cortex. *J Neurosci* 32, 13273-13280.

**Hausfeld, L.**, Santoro, R., Valente, G., Formisano, E., 2012. Classification of multiclass fMRI data using supervised self-organizing maps. In: Pattern Recognition in Neuroimaging (PRNI), IEEE International Workshop on, 65-68.

Correia, J., Formisano, E., Valente, G., **Hausfeld, L.**, Jansma, B., Bonte, M. (*accepted*). Brain-based translation: fMRI decoding of spoken words in bilinguals reveals language-independent semantic representations in anterior temporal lobe. *J Neurosci.*

## Submitted manuscripts

Bonte, M., **Hausfeld, L.**, Scharke, W., Valente, G., Formisano, E. (*under review*). Task-dependent decoding of speaker and vowel identity in superior temporal cortex.

Eck, J., Kaas, A.L., Mulders, J.L., **Hausfeld, L.**, Kourtzi, Z., Goebel, R. (*under review*) The effect of task instruction on haptic processing: The neural underpinnings of roughness and spatial density perception.

**Hausfeld, L.**, Valente, G, Formisano, E.. (*under review*) Multiclass fMRI-data decoding and visualization using supervised self-organizing maps.

## Manuscripts in preparation

**Hausfeld, L.**, Formisano, E. Cortical representations during speaker identification in noisy auditory scenes in human auditory cortex.

Goffeaux, V., Duecker, F., **Hausfeld, L.**, Schiltz, C., Goebel, R. Orientation tuning for faces in the fusiform face area and Primary visual cortex.

Ley, A., **Hausfeld, L.**, Formisano, E., Vroomen, J. MVPA multimodal categories decoded in temporal cortex.

# Curriculum Vitae

Lars Hausfeld was born in Hannover, Germany, on November 21st 1983. He attended the Gymnasium Tellkampschule Hannover and completed his secondary education in 2003. After civilian service at Nordstadt Klinikum Hannover, Lars continued his education at the University of Osnabrück studying Cognitive Science. His bachelor thesis about the integration of audio-visual stimulation was based on empirical work under supervision of Prof. Peter König and Dr. Hans-Peter Frey. Lars obtained his Bachelor's degree in 2007. Subsequently, he enrolled in the Research Master program in Cognitive Neuroscience at Maastricht University. During his research internship, Lars applied single-trial decoding techniques to EEG measurements under supervision of Prof. Elia Formisano and Dr. Federico De Martino. For his thesis he was awarded the studentenprijs of the Stichting Wetenschapsbeoefening of Maastricht University. In 2009 he obtained his Master's degree, cum laude. Lars continued working at the Department of Cognitive Neuroscience as a PhD candidate under the supervision of Prof. Elia Formisano, Dr. Giancarlo Valente and Dr. Milene Bonte. Within the scope of his PhD project, he worked on the development and application of decoding methods for EEG and fMRI and investigated voice processing in the human brain. Since October 2013, Lars is working as post-doctoral fellow in the same research group.