Salmi, Juha; Koistinen, Olli-Pekka; Glerean, Enrico; Jylänki, Pasi; Vehtari, Aki; Jääskeläinen, Iiro; Mäkelä, Sasu; Nummenmaa, Lauri; Nummi-Kuisma, Katarina; Nummi, Ilari; Sams, Mikko

Distributed neural signatures of natural audiovisual speech and music in the human auditory cortex

Published in:
NeuroImage

DOI:
10.1016/j.neuroimage.2016.12.005

Published: 15/08/2017

Document Version
Peer reviewed version

Please cite the original version:
Salmi, J., Koistinen, O-P., Glerean, E., Jylänki, P., Vehtari, A., Jääskeläinen, I., ... Sams, M. (2017). Distributed neural signatures of natural audiovisual speech and music in the human auditory cortex. NeuroImage, 157, 108-117. https://doi.org/10.1016/j.neuroimage.2016.12.005

Distributed neural signatures of natural audiovisual speech and music in the human auditory cortex

Juha Salmi, Olli-Pekka Koistinen, Enrico Glerean, Pasi Jylänki, Aki Vehtari, Iiro P. Jääskeläinen, Sasu Mäkelä, Lauri Nummenmaa, Katarina Nummi-Kuisma, Ilari Nummi, Mikko Sams

www.elsevier.com

PII: S1053-8119(16)30712-1
DOI: http://dx.doi.org/10.1016/j.neuroimage.2016.12.005
Reference: YNIMG13626

To appear in: *NeuroImage*

Received date: 11 April 2016
Revised date: 2 November 2016
Accepted date: 3 December 2016

Cite this article as: Juha Salmi, Olli-Pekka Koistinen, Enrico Glerean, Pas Jylänki, Aki Vehtari, Iiro P. Jääskeläinen, Sasu Mäkelä, Lauri Nummenmaa, Katarina Nummi-Kuisma, Ilari Nummi and Mikko Sams, Distributed neura signatures of natural audiovisual speech and music in the human auditory cortex *NeuroImage,* http://dx.doi.org/10.1016/j.neuroimage.2016.12.005

Distributed neural signatures of natural audiovisual speech and music in the human auditory cortex

Juha Salmi[1,2,3*], Olli-Pekka Koistinen[1*], Enrico Glerean[1], Pasi Jylänki[1], Aki Vehtari[1], Iiro P. Jääskeläinen[1], Sasu Mäkelä[1], Lauri Nummenmaa[1,4], Katarina Nummi-Kuisma[5], Ilari Nummi[1], Mikko Sams[1]

[1]Department of Neuroscience and Biomedical Engineering (NBE), School of Science, Aalto University, Finland

[2]Advanced Magnetic Imaging (AMI) Centre, School of Science, Aalto University, Finland

[3]Institute of Behavioural Sciences, Division of Cognitive and Neuropsychology, University of Helsinki, Finland

[4]Turku PET Centre, University of Turku, Finland

[5]DocMus Unit, Sibelius Academy, Helsinki, Finland

Corresponding author: Name: Mikko Sams: Department of Neuroscience and Biomedical Engineering (NBE), School of Science, Aalto University, Finland, P.O. BOX 12200, 00076 AALTO, FINLAND. Tel.: +358 50 5215739. Mikko.Sams@aalto.fi

Abstract

During a conversation or when listening to music, auditory and visual information are combined automatically into audiovisual objects. However, it is still poorly understood how specific type of visual information shapes neural processing of sounds in lifelike stimulus environments. Here we applied multi-voxel pattern analysis to investigate how naturally matching visual input modulates

supratemporal cortex activity during processing of naturalistic acoustic speech, singing and instrumental music. Bayesian logistic regression classifiers with sparsity-promoting priors were trained to predict whether the stimulus was audiovisual or auditory, and whether it contained piano playing, speech, or singing. The predictive performances of the classifiers were tested by leaving one participant at a time for testing and training the model using the remaining 15 participants. The signature patterns associated with unimodal auditory stimuli encompassed distributed locations mostly in the middle and superior temporal gyrus (STG/MTG). A pattern regression analysis, based on a continuous acoustic model, revealed that activity in some of these MTG and STG areas were associated with acoustic features present in speech and music stimuli. Concurrent visual stimulus modulated activity in bilateral MTG (speech), lateral aspect of right anterior STG (singing), and bilateral parietal opercular cortex (piano). Our results suggest that specific supratemporal brain areas are involved in processing complex natural speech, singing, and piano playing, and other brain areas located in anterior (facial speech) and posterior (music-related hand actions) supratemporal cortex are influenced by related visual information. Those anterior and posterior supratemporal areas have been linked to stimulus identification and sensory-motor integration, respectively.

Graphical abstract

Keywords: audiovisual, fMRI, multi-voxel pattern analysis, music, speech

1. Introduction

Our brain integrates auditory and visual information automatically into audiovisual objects. Concordant visual information enhances auditory perception. For instance, viewing concurrent

visual speech improves the accuracy of temporal discrimination of the acoustic speech (Vroomen & Stekelenburg 2011). Relatively little is known about audiovisual processing of music, but apparently matching visual information adds to perception of instrumental music, yet in a way that is distinct from audiovisual speech perception (see Saldana and Rosenblum 1993, Vatakis and Spence 2006).

1.1. Brain areas involved in speech vs. music

In order to discover the effect of visual stimulation on processing of auditory information in supratemporal auditory cortex, we first have to characterize the areas involved in processing of unimodal auditory stimuli. Music and speech share, for instance, requirement for fine-grained pitch discrimination (Zatorre and Baum 2012), periodic patterns (Patel 2003a) and even higher order structures (Patel 2003b). A few studies have revealed reliable intrahemispheric regional dissociation in cortical processing of complex music and speech features: speech-related spectral irregularity of sounds activates temporal cortex areas, mainly in the middle temporal gyrus (MTG), that are more anterior-lateral to those activated by music-related temporal regularity (Tervaniemi et al. 2006, Santoro et al. 2014). Recent studies utilizing multivariate pattern analysis (MVPA) have detailed different stages in processing unimodal speech and music (Abrams et al. 2011, Norman-Haignere et al. 2015, Rogalsky et al. 2011, Ryali et al. 2010). For instance, Abrams et al. (2011) suggested that unimodal speech and music involve largely the same temporal structure, but distinct spatial patterns to these stimuli can be classified in the inferior frontal gyrus, posterior and anterior superior temporal gyrus (STGp/a) and MTG, and auditory brainstem.

1.2 Brain areas involved in audiovisual modulations

Specific types of concurrent visual input modulate auditory processing in distributed temporal-cortical areas overlapping with those involved in unimodal auditory processing (Kayser et al. 2007). Integration of face and voice (for a review see Campanella and Belin 2007, Yovel and Belin 2013),

and audiovisual action processing (for a review see Hein and Knight 2008) are examples of sensory-integration processes that have been widely studied. Visual input modulates activity in multiple areas, including the primary auditory cortex (Sams et al. 1991, Foxe et al. 2002, Pekkola et al. 2005, Kayser et al. 2005) as well as anterior and posterior temporal lobe areas (von Kriegstein et al. 2005, Pekkola et al. 2006, Campanella and Belin 2007, Perrodin et al. 2014). The role of anterior MTG in coupling the face and voice information, in particular, has been demonstrated in several studies (see Campanella and Belin 2007, Yovel and Belin 2013).

Accumulating evidence suggests that audiovisual modulations are largely based on modulation of temporal processing, not changes in the overall response amplitudes (Allman et al. 2008, Iurilli et al. 2012, Lakatos et al. 2007, 2009). For instance, when monkeys are presented with naturalistic sounds accompanied with matching visual stimulus, firing rate of the neurons in the auditory cortex and inter-trial variability of the activation is decreased (Dahl et al. 2010, Kayser et al. 2010).

1.3 Multi-voxel pattern analysis

While the conventional mass-univariate general linear model (GLM) approach is straightforward to implement in studies examining regional activity evoked by isolated stimulus features, it is more problematic when overlapping stimulus features activate distinct multivariate patterns of neural activity within a given region (Ben-Yakov et al. 2012, see also Henson 2006). Multi-voxel pattern analysis (MVPA) represents an opposite way of modeling, trying to predict stimulus categories using an entire hemodynamic activation pattern, without being restricted to an assumption of certain predefined response function or stimulus model (Norman et al. 2006, Pereira et al. 2008, Mur et al. 2009). By enabling classification of complex stimulus-specific activation patterns even in the absence of regional amplitude changes, MVPA provides a powerful new approach to investigate the mechanisms of audiovisual integration (Pooresmaeli et al. 2014, Gentile et al. 2015, Li et al. 2015, Rohe & Noppeney 2015). For instance, Li et al. (2015) recently found distributed content-specific

(male vs. female, crying vs. laughing) supratemporal activations during audiovisual perception of faces and voices during selective attention to particular features. The effects of matching visual input on processing music and speech, however, remain unclear.

1.4 The aim of the present study

We applied Bayesian logistic regression to classify transient temporal cortex activity patterns measured during audiovisual and auditory speech, singing, and piano playing. The analysis was based on probabilistic classification models that attach a given activation pattern to the most probable one of two or three classes based on linear combinations of the voxel activations, where the signs and absolute values of the voxel coefficients represent the contribution of each voxel to the classification task. By visualizing the posterior probability distributions of the coefficients as brain maps, we expected to reveal neural systems discriminating between audiovisual vs. auditory conditions or between auditory speech, singing and piano playing, likely being represented in complex spatial patterns in distributed neuronal networks (see Abrams et al. 2011, Norman-Heignere et al. 2015, Rogalsky et al. 2011, Ryali et al. 2010 for unimodal studies and Li et al. 2015, Vetter et al. 2014 for audiovisual studies). In order to address this specific research question, we selected a method that is, unlike often used searchlight MVPA approaches (see Mur et al. 2009), able to detect sparse patterns associated with activity in widely distributed brain networks. To promote sparsity in the posterior solution, the voxel coefficients were given short-tailed Laplace priors, which should improve both the generalizability and interpretability of the solution (Williams 1995). The performance of the classification models was tested by a cross-validation across 16 participants.

The data were acquired in an fMRI experiment, where participants watched and listened to audiovisual and purely auditory versions of songs that were either spoken, sung, or played with a piano. The visual input in the speech and singing conditions was the face of the speaker/singer, and

in the piano conditions participant saw the players finger movements on a keyboard. Singing condition that contained the acoustic structure of music, but had the same voice and mostly similar visual information as in speech condition, was expected to provide additional information about the effects of the visual input type (facial processing in singing vs. hand action in piano playing) and specific spectrotemporal characteristics of music (tone vs. voice) on auditory processing. By using spoken lyrics of the songs in the speech condition we were able to control for semantic and syntactic structures, as well as tempo. The trade-off was that the stimulus was not the most common type of narrative speech but more like listening to poetry reading. As many previous studies (Beauchamp et al. 2004, Romanski & Hwang 2012, Wayne & Johnsrude 2012, Conrad et al. 2013, Li et al. 2015), we used complex naturalistic stimuli in order to activate widespread temporal cortex areas associated with audiovisual processing. Such complex stimulation is important also, because it includes nuanced spectro-temporal features that are critical in discriminating between real-life music and speech. Half of the trials contained synchronous matching auditory and visual stimuli, and the other half only auditory stimuli that were identical to those in audiovisual stimuli. Identical auditory stimuli thus canceled acoustic differences related to differences between audiovisual vs. auditory speech, singing, and piano conditions.

We had two predictions: 1) Coherent visual input mostly amplifies processing within the set of brain areas dedicated to processing auditorily presented speech and music, 2) or there are distinct brain areas that specifically contribute to multimodal integration, not involved in auditory processing per se. Furthermore, we expected that visual stimulation containing facial movements would modulate the activity in anterior MTG (Campanella and Belin 2007, Yovel and Belin 2013), and that viewing visual hand actions (piano) would, in turn, modulate the activity in the dorsal auditory stream involved in spatial processing and sensorimotor integration (Rauschecker 2011).

2. Materials & Methods

2.1. Participants

We studied 16 healthy participants (6 females; 1 left handed; age range 21–40 years, M_{age} = 28 years, SD_{age} = 2.6 years) with no neurological or psychiatric illnesses or contraindications for functional magnetic resonance imaging, and with normal vision and hearing. All were native Finnish speakers. Seven participants reported music as their hobby, five had experience in playing a musical instrument, and three had studied music theory (15, 10, and 3 years). The study was approved by the Ethical Committee of Hospital District of Helsinki and Uusimaa, and was conducted in accordance with the Declaration of Helsinki. All subjects were compensated for their time and travel costs, and they signed ethics-committee-approved, informed consent forms.

2.2. Stimuli and experimental procedure

To construct the audiovisual stimulus set, we initially selected 18 popular songs, which were played with piano by a professional musician and recorded. Acoustic features of the recordings were analyzed with MIR toolbox (Latrillot et al. 2007). Final songs were chosen based on high variation in acoustic features (sound energy in time, event density, tempo, pulse clarity, acoustic roughness, pitch variability, musical mode, and mode variability). The final stimuli were recordings of three popular songs, Jingle Bells (J. Pierpont; duration 77 s), Those Were the Days (B. Fomin; duration 126 s), and Summertime (G. Gershwin; duration 106 s). High acoustic variability and high number of volumes (mean stimulus duration 103 s) for each stimulus were assumed to provide sufficient stimulus-response mapping. That is, by selecting songs with high acoustic variability (representative collection of features) we expected to activate widespread neuronal populations and to further increase the variability in feature-specific brain responses, which is important in pattern analysis. In contrast with less-than-20-second blocks often used in studies examining regional activity during presentation of repetitive and isolated stimuli, we used quite long stimuli. Repetition

suppression was expected to play a minor role here due to continuous changes in stimulus features and their dynamics (see Grill-Spector et al. 2006). Importantly, our analysis was not based on signal onset amplitudes, which provide the largest effects in regional analysis of isolated stimuli. Instead, pattern analysis uses regularities in signal time series that actually increase during presentation of prolonged naturalistic stimulus (e.g., Yeo et al. 2007). Earlier studies have also shown that the feature selectivity of auditory cortical neurons remains high during prolonged naturalistic stimulation (e.g., Mukamel et al. 2005), and also several other prior studies have demonstrated that BOLD signal collected during viewing of naturalistic stimuli is reliable and it does contain sufficient information about stimulus-specific activations (e.g., speech, music, faces, colors, stimulus movements) even when the same stimulus is never repeated (e.g., Alluri et al. 2010, Burunat et al. 2016, Farbood et al. 2016, Huth et al. 2016, Lahnakoski et al. 2012, see Hasson et al. 2010 for a review). All songs were recorded in three different ways i) played with piano (Piano), ii) sung by one voice a cappella (Singing), or iii) spoken as normal speech (Speech) keeping the same tempo as when they were sung or played. With the aim of having the Piano stimuli as comparable as possible, the piano part had a melody line similar to the sung condition as well as accompanying harmony. Singing and Speech were performed with Finnish lyrics. After the experiment, each participant evaluated the familiarity and pleasantness of the music and lyrics, on a scale from one to seven.

Piano was recorded using one binaural stereo microphone (OKM Technik by Soundman) at the height of the pianist's head inside the grand piano and one room microphone (AKG C-1000) on top of the piano. Voice recordings were done with the same microphones positioned in front of the singer.

The microphones were connected to an M-AUDIO firewire soundcard, and the acoustic signal was sampled at 44100 Hz with 16-bit precision. A high-definition video was recorded during performance (Canon HD camera). The pianist's hand movements were recorded from above.

Singing and Speech were recorded synchronized to the piano tempo by simultaneously listening to the piano recordings. During Singing and Speech, the video camera was directed to the actors face. For a playback, the sound intensities of Piano, Singing, and Speech were digitally equalized over the whole piece, and the sound quality was improved by reducing background hiss and mild compressing, using Logic Pro (Apple). During the experiment, each stimulus was presented with (Audiovisual) or without (Auditory) the corresponding video stream, resulting in a total of 3 [(Piano, Singing, Speech) x 2 (Audiovisual, Auditory)] stimulus categories and a total of 18 stimuli (3 songs per stimulus category).

In order to isolate the brain responses associated with specific acoustic features, we extracted time series of two acoustic features over sliding temporal windows of 500 ms from the stimuli used in the experiment. These features were pulse clarity (temporal regularity) and spectral entropy (spectral irregularity). 'Speechness' is described by high values of acoustic spectral irregularity compared to the piano sounds. On the other hand, temporal regularity captures the sound 'musicness', due to the regularity in the musical notes and their attacks, compared to speech where different consonants can alter the sense of rhythm. The time series were then downsampled to TR resolution and convolved with the canonical hemodynamic response function. Other timbral features such as brightness or spectral centroid were relatively constant since there was only one type of sound per time (piano or voice).

During fMRI, the 18 stimuli were presented in an order that was counterbalanced between different stimulus categories (Auditory vs. Audiovisual, and Speech vs. Singing vs. Piano). Participants were instructed to actively attend to the stimuli during the experiment. In order to have the setup as naturalistic as possible, we did not include any active task during fMRI. Stimuli were separated by 5-s breaks. This type of presentation was selected to reduce the effect of possible carry-over effects between subsequent stimuli. The audio was played to the subjects in the MRI scanner with an UNIDES ADU2a audio system (Unides Design, Helsinki, Finland) via plastic tubes through porous

EAR-tip (Etymotic Research, ER3, IL, USA) earplugs. The video was projected on a semi-transparent screen behind the participant's head using a 3-micromirror data projector (Christie X3, Christie Digital Systems Ltd., Mönchengladbach, Germany). The distance to the screen was 34 cm via a mirror located above their eyes (visual angle 12˚, binocular view width 24 cm). After the experiment, participants were interviewed regarding their behavior in the scanner and to approve that they listened and watched attentively to all stimuli and stayed alert during the scan. Post-experimental ratings were collected outside the scanner in order to keep the length of the experiment reasonable and to keep the stimulation as naturalistic as possible.

2.3. MRI data acquisition and preprocessing

MR imaging was performed with a 3.0 T GE Signa Excite MRI scanner (GE Medical Systems, USA) using a quadrature 16-channel head coil. Whole-brain data were acquired with T2* weighted echo-planar imaging (EPI), sensitive to the blood oxygenation dependent (BOLD) contrast using the following imaging parameters: 29 axial slices, slice thickness 4 mm, 1-mm gap between slices, in-plane resolution 3.4 mm x 3.4 mm, voxel matrix 64 x 64, TR = 2000 ms, TE 32 ms, flip angle = 90°, ascending interleaved acquisition. Altogether 1160 functional volumes were acquired continuously during the experiment. T1-weighted inversion recovery spin-echo volume was acquired for anatomical alignment (TE 1.9 ms, TR 9 ms, flip angle 15°). The T1 image acquisition used the same slice prescription as the functional image acquisition, except for a denser in-plane resolution (in-plane resolution 1 mm x 1 mm, matrix 256 x 256) and thinner slices (1 mm, no gap).

fMRI data was preprocessed with the Functional Magnetic Resonance Imaging of the Brain Centre's (FMRIB) software library (FSL, release 4.1.6 www.fmrib.ox.ac.uk/fsl, Smith et al. 2004). To allow for the initial stabilization of the fMRI signal, first 5 volumes of each session were excluded from the analysis (during this time a blank screen was presented). The data were motion corrected (McFlirt), and non-brain matter was removed (BET). The data were co-registered

(FLIRT) first to anatomical image allowing 9 DOF and then registered to MNI152 standard space (Montreal Neurological Institute) allowing 9 DOF. The data were spatially smoothed with a Gaussian kernel of 6 mm (FWHM) to decrease spatial noise in the statistical analysis (see Op de Beeck 2010 for spatial filtering when using MVPA) and high-pass filtered with 100-s cutoff.

For MVPA, an area in the bilateral temporal cortex that covered the parietal operculum cortex (POC), planum temporale (PT), Heschl's gyri (HG), planum polare (PP), posterior superior temporal gyrus (STGp), anterior superior temporal gyrus (STGa), posterior middle temporal gyrus (MTGp), and anterior middle temporal gyrus (MTGa) was defined in the Harvard-Oxford cortical template. The data set used in the MVPA included a total of 2875 features, each representing one voxel in this area covering all the template subregions. The above-mentioned anatomical regions were used in describing and discussing the results (Figures 2–4, Supplementary Figure 1).

Prior to MVPA, the data samples were standardized by dividing each individual voxel time series by its standard deviation and setting its mean to zero. Each of the six categories (Audiovisual and Auditory Speech, Singing, and Piano), contained data (samples) from 157 EPI volumes (39 samples for Jingle Bells, 64 for Those Were the Days, and 54 for Summertime) from each of the 16 subjects. Altogether there were thus 157 x 16 = 2512 samples per category. MVPA was performed for this time series data while treating each sample as a separate observation. Hence, the total number of observations in the MVPA was 15072 (6 categories x 2512 observations per category).

2.4. Multi-voxel pattern analysis

Our MVPA analysis was based on Bayesian treatment of logistic regression classifiers that attach a given transient activation pattern to the more probable one of two stimulus classes ($c = \pm 1$) according to a linear combination of the voxel activations x weighted by the unknown voxel coefficients w. In the logit model, this linear combination is transformed into a class probability by

the logistic activation function, $\Pr(c=+1)=l^{-1}(w^T x)=\frac{1}{1+e^{-w^T x}}$, so that a positive value is transformed into a class probability greater than 0.5 and negative value into a class probability less than 0.5 ($\Pr(c=-1)=1-\Pr(c=+1)$). Thus, positive voxel coefficients represent sensitivity to the positive stimulus class and negative coefficients to the negative stimulus class.

In the Bayesian treatment, the voxel coefficients were given independent Laplace priors $p(w_j|\lambda)=\frac{1}{2\lambda}e^{\frac{|w_j|}{\lambda}}$ with a constant scale hyperparameter λ, in order to promote sparsity in the posterior distribution and hence improve both generalizability and interpretability of the solution (Williams 1995). The short-tailed Laplace prior does not enforce coefficients to zero, but suppresses the absolute values of irrelevant voxels so that the amount of voxels regarded significant decreases, when compared to a model using a Gaussian prior distribution. The multivariate posterior distribution of the coefficients was approximated using an expectation propagation algorithm (van Gerven et al. 2010, Minka 2001) implemented in the FieldTrip toolbox (Oostenveld et al. 2011).

For the final models trained using the data of all 16 participants, the scale hyperparameter of the Laplace prior was optimized (candidate values $\lambda=10^k$, where $k\in\{-6,-5,-4,-3,-2\}$) by maximizing the mean log predictive probability (MLPP) obtained in a leave-one-out cross-validation across participants (one participant at a time was left out from the training data set and the model was trained using the remaining 15 participants) and averaged over all seven binary classification tasks (Lamnisos et al. 2012). The posterior distribution of the voxel coefficients was visualized by presenting the marginal posterior probabilities for positive sign of each coefficient as a brain map, which we call a signature pattern. Hence, the probability scores in the signature pattern maps reflect the relative contribution of each voxel to the classification (Figures 2 and 4) or linear regression (Figure 3). The significances ($p < 0.05$) of the resulting voxel scores were tested by retraining the classifiers 100 times using datasets, where the class labels of the observations of one

subject were randomly permuted and the same label order was used for all other subjects (Pesarin 2001). The significance thresholds were obtained by gathering together all the 100 x 2875 retrained voxel values and taking the 95th percentile. Thus, the thresholds apply to single voxels, but they have not been corrected for multiple comparisons. The maximum statistics approach that could have been used (see Nichols & Holmes 2001) in order to correct for multiple comparisons would be overtly conservative in this type of region-of-interest-based analysis. Similar statistical testing was conducted for each classifier (see Figures 2–4).

The predictive classification accuracies were tested by a nested cross-validation procedure across the 16 participants, where one participant at a time was left out for testing and the model was trained using only the remaining 15 participants. The scale hyperparameter was selected separately for each cross-validation fold by maximizing the mean log predictive probability (MLPP) obtained in an inner cross-validation across the remaining 15 participants (one participant at a time was again left out for testing and the model was trained using only the remaining 14 participants) and averaged over all seven binary classification tasks. The significance ($p < 0.05$) of each classification accuracy was tested by repeating the cross-validation (with the same scale hyperparameter value as used for the final models) using datasets, where the class labels of the observations of one subject were randomly permuted and the same label order was used for all other subjects. The empirical chance level was obtained by taking the 95th percentile of the obtained classification accuracies.

Four separate binary classifiers were trained to discriminate between Auditory and Audiovisual stimuli, both separately for each stimulus type (Piano, Singing, and Speech), as well as for all Auditory versus Audiovisual stimuli together. In the signature patterns of these symmetrical classifiers, the marginal posterior probabilities for positivity of the voxel coefficients were scaled from 0…1 to -1…1, so that a value near -1 indicates high probability for the coefficient to be negative (i.e., the voxel is, with a high posterior probability, more sensitive to the negative stimulus class than to the positive stimulus class).

To conduct a three-class classification between Piano, Singing, and Speech, we trained three more binary classifiers using only the Auditory stimuli: Piano vs. Singing/Speech, Singing vs. Piano/Speech, and Speech vs. Piano/Singing. The signature patterns of these classifiers were presented together in one brain map using the normal probability scale, and the predictive classification accuracy of the three-class classifier was determined by choosing the most probable stimulus type based on the class probabilities of the three binary classifiers. We also trained one vs. one classifiers for the three stimulus types in order to make sure that none of them biases the results based on the one vs. two classifiers.

Finally, a pattern regression analysis with a Gaussian noise term was used to examine the linear effects of 'musicness' across auditory and audiovisual conditions. A similar analysis was also performed for 'speechness'. The noise variance and the Laplace prior scale hyperparameter were optimized by minimizing the cross-validated mean squared error. To take into account the possible overlap of the patterns of 'musicness' and 'speechness', 'musicness' was used as an additional regressor when modelling 'speechness' and vice versa (Valente et al. 2014). This analysis was performed in order to interpret which patterns in the classification analyses follow the acoustic features of sounds and which are likely to reflect "higher level processes".

Additional GLM analysis was conducted to demonstrate that the differences between the task conditions are not observed in the mean regional signals. This analysis was performed using fMRIB Improved Linear Model (FILM). Regressors were derived from the onset timings and durations of the same stimuli that were included in the MVPA. Hence, the time series data was the same in the GLM and MVPA analyses. Hemodynamic responses to each of the six stimulus conditions were modeled using gamma function and its temporal derivatives. The high-pass filter applied to the model was the same that was applied to the data. Pause periods served as a baseline in the model. The same one vs. one contrasts that were studied in MVPA were analyzed with GLM. Statistical thresholds for the resulting voxel maps were inferred using permutation-testing (5000 permutations)

tool implemented in FSL (Randomise). Thresholding was conducted by using Threshold-Free Cluster Enhancement option.

3. Results

3.1. Classification accuracies

MVPA of the activity in the temporal cortex areas of both hemispheres was successful in classifying the brain activity patterns into Auditory Speech vs. Singing vs. Piano and into Audiovisual vs. Auditory stimulus classes (Figure 1).

3.2. Unimodal auditory classifiers

Figure 2b shows the voxels that formed the signature patterns discriminating between Auditory Speech, Singing, and Piano in the temporal cortex area included in MVPA (see Figure 2a). Voxels contributing to these signature patterns were distributed bilaterally over wide areas in both auditory cortices, forming intermixed clusters continuing from one labeled brain region to another. A large area in the right hemisphere, including areas in STGa, STGp, and MTGa, contributed significantly in discriminating Piano from Speech and Singing. A set of left-hemisphere areas also contributed to the discrimination, but the spatial organization of these areas was different than in the right temporal cortex. Areas discriminating Singing or Speech from two other stimulus types were distributed all over the left and right temporal cortices. Voxels located primarily in left STGa and right MTGp discriminated Singing from Speech and Piano. Distributed signature patterns including areas in left MTGp and right PP discriminated Speech from Piano and Singing. The results of one vs. one classifiers (Piano vs. Speech, Piano vs. Singing, Singing vs. Speech) were consistent with the results based on the one vs. two classifiers.

3.3. Pattern regression analysis with a continuous acoustic model

The signature patterns of 'musicness' in the linear regression analysis were observed mainly in bilateral left STGp, right STGp, and left HG/PT/POC (Figure 3). The signature patterns of 'speechness' were observed mainly in left STGa/p, and bilateral MTGp (Figure 3).

3.4. Audiovisual vs. auditory classifiers

The signature patterns of the four Audiovisual vs. Auditory classifiers are visualized in Figure 4 (see Figure 2a for the names of the subregions and Table 2 for the local maxima). Visual information affected brain activity in multiple areas. These areas included early auditory areas in HG, as well as higher-level auditory areas, for instance, in STG, MTG, and POC. When using the data of all stimulus types together, the Audiovisual vs. Auditory signature pattern showed most significant effects in bilateral STGp, right MTGp, and left POC (Figure 4, AV vs. A All). When using the data of Piano stimulus type alone, the most significant AV-related effects were found in bilateral POC (Figure 4, AV vs. A Piano). In the case of Singing, the most significant effects were found in STGa, especially in the right hemisphere (Figure 4, AV vs. A Singing), and in the case of Speech, bilaterally in MTGa/p (Figure 4, AV vs. A Speech).

3.5. Results of the GLM analysis

GLM analysis contrasting singing vs. speech, and singing vs. piano produced widespread activity in the left STGp, left anterior planum temporale, and right MTGp. In addition, singing vs. speech showed activity in the right STG, left HG and PP, and singing vs. piano in the bilateral MTG. GLM analysis did not reveal significant differences between piano vs. speech (Supplementary Figure 1).

Furthermore, GLM analysis did not reliably discriminate between the Audiovisual vs. Auditory stimuli. The only significant effect associated with modulation caused by visual information was observed in the right MTGp for audiovisual vs. unimodal auditory speech (Supplementary Figure 1).

3.6. Subjective ratings of the stimuli

The obtained familiarity rating values were 5.3 ± 1.01 (mean ± SD) for music and 4.4 ± 1.08 for lyrics, confirming that the songs were familiar as expected. Familiarity with music theory and years with music as a hobby correlated positively with subjectively rated familiarity of the music ($r = 0.5$, $p < 0.05$ and $r = 0.7$, $p < 0.01$, respectively). However, neither of these variables was associated with the MVPA classification accuracy. Classification accuracy in distinguishing between Piano vs. Singing was, however, correlated with subjectively evaluated pleasantness of Piano vs. Singing ($r = 0.53$, $p < 0.05$). That is, the more pleasant the stimulus, the higher the classification accuracy. The pleasantness ratings showed differences between Piano, Speech, and Singing: Piano was estimated more pleasant than Singing ($t = 3.92$, $p < 0.0001$) or Speech ($r = 5.53$, $p < 0.0001$), and Singing was estimated more pleasant than Speech ($t = 2.32$, $p < 0.05$). However, pleasantness did not correlate with accuracy of the audiovisual vs. auditory classifiers.

3.7 Additional MVPA's

To reveal the possible overlap in the brain activity associated with Piano vs. Singing vs. Speech and stimulus valence (see Section 3.6 for the results of the valence ratings), we performed an MVPA (regression model) between fMRI activation and the valence ratings. The coefficient of determination (the square of the correlation coefficient between predicted and true valence ratings) for the model was only 2%, and the histogram of voxel coefficients was near the one obtained by randomized data. We thus conclude that temporal cortex signature patterns are not reliably linked with valence ratings.

In addition to the auditory three-class classification between Piano vs. Singing vs. Speech, we conducted a similar three-class classification using only the audiovisual stimuli. The obtained classification accuracy was 67%, which was approximately 11 %-units higher than when using only the auditory stimuli. A permutation test confirmed that this difference was statistically significant (p

< 0.01). We also conducted an additional cross-validation test, where we used the auditory data of 15 participants for training and the audiovisual data of the remaining subject for testing. The obtained three-class classification accuracy was 51%, which was clearly higher than the empirical chance level (36 %, $p < 0.05$), but 6 %-units lower than when testing with auditory data. Also this difference was confirmed statistically significant ($p < 0.01$) in a permutation test.

Finally, in order to make sure that important features were not lost when selecting an approach utilizing the sparsity-promoting Laplace prior, we ran similar analyses using a Gaussian prior. The classification accuracies were only slightly lower than with Laplace prior, and also the signature patterns were comparable, even if the amount of voxels considered significant was about 10% higher.

4. Discussion

In the present study, we characterized signature patterns of supratemporal cortex activity associated with naturalistic audiovisual and auditory speech, singing, and instrumental piano perception. Bayesian logistic regression analysis successfully discriminated between activation patterns elicited by auditory speech, singing, and piano playing (Figures 1 and 2). In addition, we found that matching visual input modulated activity patterns in widely distributed temporal cortex areas, which were distinct from the areas contributing to the classification of unimodal auditory stimuli (Figure 4). Hence, the brain networks processing different auditory features and those involved in audiovisual processing are both specific to speech and music but distinct from each other. Specific brain areas contributing to audiovisual signals may provide important information for the basis for understanding how our brain encodes complex audiovisual objects. The reported significance maps represent the shared information in brain networks involved in processing of audiovisual speech and music. However, it should be noted that the presented approach aiming in controlling for the

acoustic variability is not suitable for estimating the predictive performance of a method to be trained with a new data with other types of stimuli.

4.1 The influence of visual speech and music on supratemporal activity

Our study using pattern analysis for an fMRI data recorded during complex naturalistic stimulation characterizes the relative contribution and content-specificity of multiple superior temporal cortex areas in audiovisual processing (Figure 4). In general, these findings accord with prior work reporting that distributed brain areas, including early auditory cortex in HG, as well as higher order areas such as STG and MTG, participate in audiovisual speech processing (see Campanella and Belin 2007, Vroomen and Baart 2012, Erickson et al. 2014). Moreover, our results conform to the results of a recent MVPA study suggesting that concurrent visual speech modifies content-specific MTG areas during listening to dynamic auditory input (Li et al. 2015).

We further demonstrated that POC specifically contributes to audiovisual processing of music. More specifically, POC activity was associated with a condition, which contained hand actions related to piano playing (Figure 4). This result agrees with previous research suggesting that activity in POC is modulated by both auditory and visual motion input (Pavani et al. 2002, Krumbholz et al. 2005, Antal et al. 2008). Furthermore, in keeping with earlier findings (Erickson et al. 2014), the visual input associated with facial speech (here also singing), in turn, showed strongest modulatory effects in more ventral temporal cortex areas (Figure 4). That is, visual information modulated the dorsal areas only when it included hand actions. The distinction between dorsal and ventral areas modulated by visual input in our study accords with the proposed distinct processing streams for spatial processing and action perception vs. identification of auditory objects (Rauschecker and Tian 2000, Rauschecker 2011, DeWitt and Rauschecker 2012). That is, the speech-related ventral temporal areas might use visual information in order to facilitate language recognition (Campanella and Belin 2007), and dorsal temporal cortex areas involved in sensory-motor integration might, for

instance, improve the accuracy of temporal discrimination (Vatakis and Spence 2006). It is well known that the auditory and visual dorsal and ventral streams are overlapping in the inferior temporal and posterior parietal cortex (see Goodale and Milner 1992, Rauschecker and Tian 2000). However, the evidence of specific effects of visual information on the auditory pathways at early processing stages has been lacking. Hence, the present results indicate that matching audiovisual input may enhance the distribution of the processing into specialized where/how and what processing streams in temporal cortex areas where the auditory and visual input are combined.

While listening to singing, the visual input had the strongest modulatory effect on right STGa, and during speech perception on bilateral MTGp/MTGa (Figure 4). The anterior ventral temporal cortex areas modulated by visual singing and speech in our experiment are involved in multiple functions (for a review, see Rauschecker 2011, DeWitt and Rauschecker 2012) such as coupling between face and voice (von Kriegstein et al. 2005, Campanella and Belin 2007, Perrodin et al. 2014). Our results agree with the recent proposal that specialization to speech and processing of temporally prolonged stimuli involve ventral auditory stream areas including STG and MTG (see De Witt and Rauschecker 2012). In previous studies, it has been suggested that both the dorsal and ventral auditory streams are affected by auditory predictive coding (see Hickock 2012 for a review). It is possible that predictive coding, i.e., the comparison of higher level (audiovisual) predictions and lower level (auditory) signals mediated by backward and forward connections, is also the mechanism for audiovisual integration. Previous studies have suggested that visual predictive coding enhances detection of location and biological movement in the dorsal stream (Stekelenburg and Vroomen 2012), while in the ventral stream it may support speech recognition accuracy (Peelle and Sommers 2015).

Altogether our results concerning distinct effects of audiovisual speech and singing imply that processing facial information that complements auditory information is not focused to a specialized area (von Kriegstein et al. 2005, Campanella and Belin 2007), but affects processing in multiple

areas, likely depending on the nature of the acoustic input and/or temporal characteristics of the visually presented facial stimulus. Speech and singing share a lot of information (e.g., speaker's voice and tempo, and to a large extent also the characteristics of facial movements). Therefore the comparison of these conditions was specifically expected to reflect integration of specific acoustic and visual information characteristic for speech and music. As these effects were observed clearly in other areas than those discriminating between auditory speech vs. music, or 'speechness' and 'musicness' modeled as separate signals, we expect that these areas in particular are involved in processing visual information and integrating it with acoustic information (see Tervaniemi et al. 2006, Santoro et al. 2014 for auditory studies). In contrast to MVPA, GLM analysis showed significant regional modulation associated with visual input only in the audiovisual vs. auditory speech contrast. This activity was observed in the right MTGp (Supplementary Figure 1), an area that was also observed in MVPA analysis (Figure 4).

4.2. Unimodal auditory signature patterns

The music-related auditory signature patterns were focused on STG and the speech-related pattern to more anterior temporal cortex areas, particularly MTG (Figures 2 and 3). These findings are well in agreement with previous research reporting regional effects, both a study using complex stimuli (Santoro et al. 2014) as well as another study using more isolated but acoustically matching instrumental sounds and spoken words (Tervaniemi et al. 2006). In our study, the results of the classification analysis were highly consistent with the results of the regression analysis based on the linear effects of 'musicness' (low spectral entropy) and 'speechness' (low pulse clarity) derived from the acoustic features (comparison of Figures 2 and 3). The overlap between these results in right STG, right MTG, and left STG for piano condition and 'musicness', and in left STG, left MTGp, right MTG, and bilateral POC for speech and singing conditions and 'speechness' suggests that the acoustic features explained some of the differences in brain activity between auditory conditions in these areas. High performance of the auditory classifier when tested with audiovisual data suggests

that the class-information in the auditory activation patterns is preserved when the visual input is added to the stimulus.

4.3 Utilizing MVPA and complex naturalistic stimulation in brain research

During the recent years, the use of MVPA in the analysis of auditory fMRI data has rapidly increased (e.g., Formisano et al. 2008, Staeren et al. 2009, Ryali et al. 2010, Abrams et al. 2011, Kilian-Hutten et al. 2011, Lee et al. 2011, Linke et al. 2011, Rogalsky et al. 2011, Ley et al. 2012). The Bayesian MVPA approach provides an under-exploited means to examine distributed activity patterns in naturalistic paradigms that are difficult to model with rigid stimulus functions. By avoiding the stimulus model and gaining increased sensitivity from the pattern information, MVPA appears to be well suited for examining the distinctions between activation patterns involved in processing continuous stimulation such as audiovisual speech and music. Hence, MVPA may provide a novel approach to examine brain function during processing of complex naturalistic signals (see Hari and Kujala 2009, Hasson et al. 2010, Hasson and Honey 2012).

4.4. Limitations of the study

We used complex naturalistic stimulation in order to examine the effects of visual speech and music on auditory processing. While there are significant advantages in using complex stimulation and multivariate methods in examining the basis of audiovisual processing in the brain, there are some trade-offs related to this approach. Firstly, even though the stimulus features would be extracted in detail, the possible interactions of the complex features are difficult to fully account for in the model. Therefore, complementary studies utilizing more reduced stimuli are useful in confirming the role of specific stimulus feature combinations. Secondly, in real-life conditions people rarely have specific task to selectively process particular stimulus contents. In a non-forced task it may be more difficult to interpret what kind of goal-directed processes are involved in processing the stimulus. Thirdly, it is possible that the familiar stimuli were associated with covert activation due

to melody (for speech) or speech (for piano playing). However, our aim was to reduce inter-individual variance in the responses and minimize learning effects using repetitions of analogical familiar stimulus. Anyway, the possible covert activations should not affect our main results that are based on comparisons of identical auditory stimuli. Fourthly, when using naturalistic stimuli, for instance, the effects of familiarity of the stimulus type (e.g., seeing hands of a piano player or hearing a spoken song or a poet), arousal level or specific types of emotions raised by particular stimuli are difficult to control for and related differences across the conditions might affect the activity patterns. However, it should be noted that in the present MVPA results the probability estimates were equally distributed between the positive and negative classes, and the global familiarity or arousal effects across the conditions should be neglected in the analysis. Moreover, in the main analyses (audiovisual vs. auditory conditions) identical auditory stimuli were used, which canceled the differences between auditory stimuli. Fifthly, inter-individual variability of several temporal cortex areas (Morosan et al. 2001, Baumann et al. 2013, Pernet et al. 2015) is likely to decrease the accuracy of inter-subject classification. In future studies it would be important to complement this analysis by conducting a study in which a greater variety of stimulus sequences would be presented to individual participants in repeated scans and the classifiers would be trained to predict unforeseen stimuli within the same participants. Finally, even though spoken lyrics are acoustically comparable to normal speech it is possible that this type of "poetry-like" stimuli are processed differently than other types of speech passages, such as listening to a conversation.

4.5. Conclusions

This study revealed neural signature patterns associated with naturalistic speech and music perception. Additional matching visual input modulated activation in temporal cortex areas that were distinct from those segregating between speech and music within the acoustic domain. The results confirm that visual input modulates activity in distributed areas in the temporal cortex, specific to the stimulus type (speech, singing, piano playing). We suggest involvement of two

mechanisms and brain networks in audiovisual processing of naturalistic speech and music: 1) Coupling of face-voice information (audiovisual speech and singing) occurs in ventral temporal cortex, in areas more accurately determined by spectro-temporal characteristics of the input (speech or music). 2) Integration of visuomotor/spatial information (audiovisual piano playing) occurs in the dorsal temporal cortex areas apparently involved in merging auditory signals with other, perhaps higher-level, stimulus contents (see Rauschecker 2011).

None of the authors declare a conflict of interest

Acknowledgements:

The study was supported by the Academy of Finland (National Centres of Excellence program 2006–2011, grants #129670, #130412, #138145, #259752), European Research Council (Starting Grant #313000 to LN), the aivoAALTO project grant from the Aalto University. Special thanks to Ms Marita Kattelus for her help in collecting the MRI data,

References

Abrams DA, Bhatara A, Ryali S, Balaban E, Levitin DJ, Menon V (2011) Decoding temporal structure in music and speech relies on shared brain resources but elicits different fine-scale spatial patterns. Cereb Cortex 21:1507-1518.

Albert JH, Chib S (1993) Bayesian analysis of binary and polychotomous response data. J Am Stat Assoc. 88:669-679.

Allman BL, Bittencourt-Navarrete RE, Keniston LP, Medina AE, Wang MY, Meredith MA (2008) Do cross-modal projections always result in multisensory integration? Cereb Cortex 18: 2066-2076.

Alluri V, Toiviainen P, Jääskeläinen IP, Glerean E, Sams M, Brattico E. Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. Neuroimage 59:3677-89.

Antal A, Baudewig J, Paulus W, Dechent P (2008) The posterior cingulate cortex and planum temporale/parietal operculum are activated by coherent visual motion. Vis Neurosci. 25:17-26.

Balk MH, Ojanen V, Pekkola J, Autti T, Sams M, Jääskeläinen IP (2010) Synchrony of audio-visual speech stimuli modulates left superior temporal sulcus. Neuroreport 23:822-826.

Bartels A, Zeki S (2004) Functional brain mapping during free viewing of natural scenes. Hum Brain Mapp. 21:75-85.

Beauchamp MS, Argall BD, Bodurka J, Duyn JH, Martin A (2004) Unraveling multisensory integration: patchy organization within human STS multisensory cortex. Nat Neurosci. 7:1190-1192.

Burunat I, Toiviainen P, Alluri V, Bogert B, Ristaniemi T, Sams M, Brattico E (2016) The reliability of continuous brain responses during naturalistic listening to music. Neuroimage 124:224-31.

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

Conrad V, Kleiner M, Bartels A, Hartcher O'Brien J, Bülthoff HH, Noppeney U (2013). Naturalistic stimulus structure determines the integration of audiovisual looming signals in binocular rivalry. PLoS One 27:8.

Dahl CD, Logothetis NK, Kayser C (2010). Modulation of visual responses in the superior temporal sulcus by audio-visual congruency. Front Integr Neurosci. 4:10.

DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A 109:E505-514.

Engel AK, Senkowski D, Schneider TR (2012) In: Murray MM, Wallace MT, editors. The Neural Bases of Multisensory Processes. Boca Raton (FL): CRC Press.

Erickson LC, Heeg E, Rauschecker JP, Turkeltaub PE (2014) An ALE meta-analysis on the audiovisual integration of speech signals. Hum Brain Mapp. 4 (epublished ahead of print).

Farbood MM, Heeger DJ, Marcus G, Hasson U, Lerner Y (2015) The neural processing of hierarchical structure in music and speech at different timescales. Front Neurosci. 12:157.

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.

Gentile G, Björnsdotter M, Petkova VI, Abdulkarim Z, Ehrsson HH. (2015) Patterns of neural activity in the human ventral premotor cortex reflect a whole-body multisensory percept. Neuroimage. 109:328-340.

Ghazanfar AA, Takahashi DY (2014) The evolution of speech: vision, rhythm, cooperation. Trends Cogn Sci. 18. pii: S1364-6613, 00150-00158.

Goodale MA, Milner AD (1992) Separate visual pathways for perception and action. Trends Neurosci. 15:20-25.

Grill-Spector K1, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. Trends Cogn Sci. 10:14-23.Hari R, Kujala MV (2009) Brain basis of human social interaction: from concepts to brain imaging. Physiol Rev. 89:453-479.

Hasson U, Honey CJ (2012) Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. Neuroimage. 62:1272-1278.

Hasson U, Malach R, Heeger DJ (2010) Reliability of cortical activity during natural stimulation. Trends Cogn Sci. 14:40-48.

Hein G, Knight RT (2008) Superior temporal sulcus – It's my area: or is it? J Cogn Neurosci. 20:2125-2136.

Hickok G (2012) The cortical organization of speech processing: feedback control and predictive coding in the context of a dual-stream model. J Commun Disord. 45:393-402.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453-8.

Iurilli G, Ghezzi D, Olcese U, Lassi G, Nazzaro C, Tonini R, Tucci V, Benfenati F, Medini, P (2012) Sound-driven synaptic inhibition in primary visual cortex. Neuron 73:814-828.

Kayser C, Logothetis NK, Panzeri S (2010) Visual enhancement of the information representation in auditory cortex. Curr Biol. 20:19-24.

Kayser C, Petkov CI, Augath M, Logothetis NK (2007) Functional imaging reveals visual modulation of specific fields in auditory cortex. J Neurosci. 27:1824-1835.

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

Krumbholz K, Schönwiesner M, von Cramon DY, Rübsamen R, Shah NJ, Zilles K, Fink GR (2005) Representation of interaural temporal information from left and right auditory space in the human planum temporale and inferior parietal lobe. Cereb Cortex 15:317-324.

Lahnakoski JM, Glerean E, Salmi J, Jääskeläinen IP, Sams M, Hari R, Nummenmaa L (2012) Naturalistic fMRI mapping reveals superior temporal sulcus as the hub for the distributed brain network for social perception. 13:233.

Lamnisos D, Griffin JE, Steel MFJ (2012) Cross-validation prior choice in Bayesian probit regression with many covariates. Stat Comput 22:359-373.

Lakatos P, Chen CM, O'Connell MN, Mills A, Schroeder CE (2007) Neuronal oscillations and multisensory interaction in primary auditory cortex. Neuron 53, 279-292.

Lakatos P, O'Connell MN, Barczak A, Mills A, Javitt DC, Schroeder CE (2009) The leading sense: supramodal control of neurophysiological context by attention. Neuron 64, 419-430.

Lee YS, Janata P, Frost C, Hanke M, Granger R (2011) Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. Neuroimage 57:293-300.

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.

Li Y, Long J, Huang B, Yu T, Wu W, Liu Y, Liang C, Sun P (2015) Crossmodal integration enhances neural representation of task-relevant features in audiovisual face perception. Cereb Cortex. 25:384-395.

Linke AC, Vicente-Grabovetsky A, Cusack R (2011) Stimulus-specific suppression preserves information in auditory short-term memory. Proc Natl Acad Sci U S A 108:12961-12966.

Minka, T (2001) Expectation propagation for approximate Bayesian inference. In: Breese, J., Koller, D (Eds.), Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann.

Mukamel R1, Gelbard H, Arieli A, Hasson U, Fried I, Malach R (2005) Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science. 309:951–4.

Mur M, Bandettini PA, Kriegeskorte N (2009) Revealing representational content with pattern-information fMRI – an introductory guide. Soc Cogn Affect Neurosci 4:101-109.

Neal RM (2003) Slice sampling. The Annals of Statistics 31:705-741.

Nichols TE, Holmes AP (2001) Nonparametric Permutation Tests For Functional Neuroimaging: A Primer with Examples. Human Brain Mapping 15:1–25

Norman KA, Polyn SM, Detre GJ, Haxby JV (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 10:424-430.

Norman-Haignere S, Kanwisher NG McDermott JH (2015) Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. 16:1281-1296.

Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: Open source software for

advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. 2011:156869.

Op de Beeck HP (2010). Probing the mysterious underpinnings of multi-voxel fMRI analyses. Neuroimage 50:567-571.

Patel AD (2003a) Rhythm in language and music: parallels and differences. Ann N Y Acad Sci. 999:140-143.

Patel AD (2003b) Language, music, syntax and the brain. Nat Neurosci. 6:674-681.

Pavani F, Macaluso E, Warren JD, Driver J, Griffiths TD (2002) A common cortical substrate activated by horizontal and vertical sound movement in the human brain. Curr Biol. 12:1584-1590.

Peelle JE, Sommers MS (2015) Prediction and constraint in audiovisual speech perception. Cortex 68:169-181.

Pekkola J, Ojanen V, Autti T, Jääskeläinen IP, Möttönen R, Sams M (2006) Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. Hum Brain Mapp. 27:471-477.

Pereira F, Mitchell T, Botvinick M (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45:S199-209.

Perrodin C, Kayser C, Logothetis NK, Petkov CI (2014) Auditory and visual modulation of temporal lobe neurons in voice-sensitive and association cortices. J Neurosci. 34:2524-2537.

Pesarin F (2001) Multivariate permutation tests: with applications in biostatistics. John Wiley & Sons.

Pooresmaeili A, FitzGerald TH, Bach DR, Toelch U, Ostendorf F, Dolan RJ (2014) Cross-modal effects of value on perceptual acuity and stimulus encoding. Proc Natl Acad Sci U S A. 111:15244-15249.

Qi Y, Minka TP, Picard RW, Ghahramani Z (2004) Predictive automatic relevance determination

by expectation propagation. Published in: Brodley CE (ed.) Proceedings of the 21st International Conference on Machine Learning. Banff, Canada. New York, Accociation for Computing Machinery Inc.

Rauschecker JP (2011) An expanded role for the dorsal auditory pathway in sensorimotor control and integration. Hear Res. 271:16-25.

Rauschecker JP, Tian B (2000) Mechanisms and streams for processing of "what" and "where" in auditory cortex. Proc Natl Acad Sci U S A 97:11800-11806.

Rogalsky C, Rong F, Saberi K, Hickok G (2011) Functional anatomy of language and music perception: temporal and structural factors investigated using functional magnetic resonance imaging. J Neurosci. 31:3843-3852.

Rohe T, Noppeney U (2015). Cortical hierarchies perform Bayesian causal inference in multisensory perception. PLoS Biol. 24:e1002073.

Romanski LM1, Hwang J (2012). Timing of audiovisual inputs to the prefrontal cortex and multisensory integration. Neuroscience. 214:36-48.

Ryali S, Supekar K, Abrams DA, Menon V (2010) Sparse logistic regression for whole-brain classification of fMRI data. Neuroimage 51:752-764.

Saldana HM, Rosenblum LD (1993) Visual influences on auditory pluck and bow judgements. Percept Psychophys 54:406-416.

Sams M, Aulanko R, Hämäläinen M, Hari R, Lounasmaa OV, Lu ST, Simola J (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. Neurosci Lett. 12:141-145.

Santoro R, Moerel M, De Martino F, Goebel R, Ugurbil K, Yacoub E, Formisano E (2014) Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol. 10:e1003412.

Schön D, Gordon R, Campagne A, Magne C, Astesano C, Anton JL, Besson M (2010) Similar cerebral networks in language, music and song perception. Neuroimage 51:450-461.

Schönwiesner M, Zatorre R (2011) Cortical speech and music processes revealed by functional neuroimaging (pp. 657-676). In J.A. Winer, C.E. Schreiner (eds.), The Auditory Cortex, Springer, LLC.

Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TE, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy RK, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM (2004) Advances in functional and structural MR image analysis and implementation as FSL. Neuroimage 23 Suppl 1:S208-219.

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.

Stekelenburg JJ, Vroomen J (2012) Electrophysiological correlates of predictive coding of auditory location in the perception of natural audiovisual events Front Integr Neurosci 6:26.

Tervaniemi M, Szameitat AJ, Kruck S, Schroger E, Alter K, De Baene W, Friederici AD (2006) From air oscillations to music and speech: functional magnetic resonance imaging evidence for fine-tuned neural networks in audition. J Neurosci. 26:8647-8652.

van Gerven MA, Cseke B, de Lange FP, Heskes T (2010) Efficient Bayesian multivariate fMRI analysis using a sparsifying spatio-temporal prior. Neuroimage 50:150-161.

Valente G, Castellanos AL, Vanacore G, Formisano E (2014) Multivariate linear regression of high-dimensional fMRI data with multiple target variables. Human Brain Mapping 35:2163-2177.

Vatakis A, Spence C (2006) Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. Neurosci Lett. 393:40-44.

Vetter P, Smith FW, Muckli L (2014) Decoding sound and imagery content in early visual cortex. Curr Biol. 24:1256-1262.

von Kriegstein K, Kleinschmidt A, Sterzer P, Giraud AL (2005) Interaction of face and voice areas during speaker recognition. J Cogn Neurosci. 17:367-376.

Vroomen J, Baart M (2012) Phonetic recalibration in audiovisual speech. In: Murray MM, Wallace MT, editors. The Neural Bases of Multisensory Processes. Boca Raton (FL): CRC Press.

Vroomen J, Stekelenburg JJ (2011) Perception of intersensory synchrony in audiovisual speech: not that special. Cognition. 118:75-83.

Wayne RV, Johnsrude IS (2012) The role of visual speech information in supporting perceptual learning of degraded speech. J Exp Psychol Appl. 18:419-435.

Williams PM (1995). Bayesian regularization and pruning using a Laplace prior. Neural Computation. 7:117-143.

Yao H, Shi L, Han F, Gao H, Dan Y (2007) Rapid learning in cortical coding of visual scenes. Nat Neurosci. 10:772–8.

Yovel G, Belin P (2013). A unified coding strategy for processing faces and voices. Trends Cogn Sci. 17:263-271.

Zatorre RJ, Baum SR (2012) Musical melody and speech intonation: singing a different tune. PLoS Biol. 10:e1001372.

Table 1. Anatomical labels, cluster sizes (cs), probability scores (*p)*, and MNI-coordinates of local maxima in brain areas showing significant ($p < 0.05$) differences between three Auditory stimulus types (one vs. one).

Brain region	cs	*p*	X	Y	Z
Piano vs. others					
Right superior temporal gyrus, anterior	245	0.94	54	2	-16
Left middle temporal gyrus, anterior	91	0.84	-54	2	-28
Left middle temporal gyrus, posterior	56	0.81	-62	-38	0
Singing vs. others					
Left superior temporal gyrus, posterior	184	0.95	-66	-10	0
Right middle temporal gyrus, posterior	155	0.86	58	-34	0
Right parietal opercular cortex	24	0.75	58	-34	32
Left Heschl's gyrus	18	0.82	-42	-18	4
Speech vs. others					
Right middle temporal gyrus, anterior	128	0.82	58	-6	-32
Left middle temporal gyrus, anterior	116	0.85	-62	-10	-16
Left middle temporal gyrus, posterior	41	0.93	-66	-22	-8
Right insular cortex / Heschl's gyrus	26	0.84	42	-14	0

Table 2. Anatomical labels, cluster sizes (cs), probability scores (*p)*, and MNI-coordinates of local maxima in brain areas showing significant ($p < 0.05$) differences between Audiovisual vs. Auditory conditions.

Brain region	cs	*p*	X	Y	Z
Audiovisual vs. Auditory Piano					
Superior temporal gyrus, anterior	217	0.7	18	64	30
Middle temporal gyrus, posterior	113	0.76	68	50	30
Heschl's gyrus	60	0.76	70	56	42
Parietal operculum cortex	23	0.86	70	42	52
Planum polare	16	0.82	68	66	34
Parietal operculum cortex	10	0.69	24	54	50
Auditory vs. Audiovisual Piano					
Superior temporal gyrus, posterior	139	0.7	68	40	40
Superior temporal gyrus, posterior	63	1	16	44	38
Middle temporal gyrus, posterior	20	0.61	18	52	30
Heschl's gyrus	15	0.47	68	60	34
Planum polare	14	0.5	24	56	34
Middle temporal gyrus, anterior	11	0.57	72	64	18
Audiovisual vs. Auditory Singing					
Planum polare	143	0.93	14	66	38
Planum temporale	120	0.77	58	46	42
Middle temporal gyrus, posterior	73	0.66	78	56	28
Middle temporal gyrus, posterior	43	0.57	10	48	30
Superior temporal gyrus, posterior	17	0.57	10	54	42

Auditory vs. Audiovisual Singing					
Planum polare	170	0.89	10	44	36
Superior temporal gyrus, anterior	36	0.83	76	64	38
Parietal operculum cortex	31	0.61	68	52	44
Middle temporal gyrus, posterior	25	0.51	72	44	28
Superior temporal gyrus, posterior	18	0.61	70	42	38
Middle temporal gyrus, anterior	15	0.66	72	64	24
Audiovisual vs. Auditory Speech					
Parietal operculum cortex	129	0.58	64	52	44
Superior temporal gyrus, posterior	85	0.66	12	52	38
Planum polare	26	0.53	20	66	30
Parietal operculum cortex	18	0.72	72	42	50
Parietal operculum cortex	13	0.44	20	48	54
Middle temporal gyrus, posterior	11	0.53	14	48	28
Superior temporal gyrus, anterior	11	0.74	76	62	38
Auditory vs. Audiovisual Speech					
Middle temporal gyrus, posterior	72	0.89	76	42	38
Middle temporal gyrus, posterior	66	0.97	18	44	36
Middle temporal gyrus, posterior	65	0.67	18	52	28
Parietal operculum cortex	50	0.68	76	52	46
Middle temporal gyrus, posterior	43	0.72	74	54	30
Middle temporal gyrus, anterior	19	0.54	72	60	20
Middle temporal gyrus, anterior	16	0.72	18	64	20

Figure 1. Cross-validated classification accuracies of the auditory Piano vs. Singing vs. Speech classifier (the three-class accuracies are specified class-wise below the overall accuracy) and the four Audiovisual vs. Auditory classifiers. The dashed lines indicate the empirical chance levels ($p < 0.05$) obtained in permutation tests.

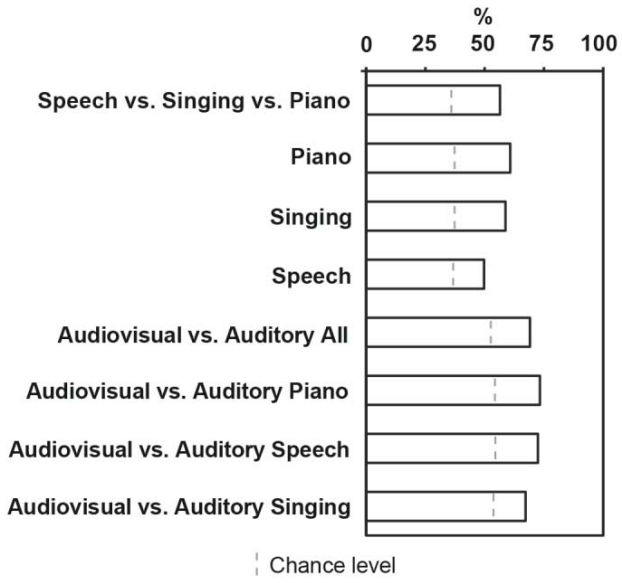
Figure 2. a) A temporal lobe area included in all analyses contained primary and secondary/association auditory cortical areas bilaterally. The figure also shows the borders of the specific subregions (PP, planum polare; HG, Heschl's gyrus; POC, parietal opercular cortex; PT, planum temporale; STGa, anterior superior temporal gyrus; STGp, posterior superior temporal gyrus; MTGa anterior middle temporal gyrus; MTGp, posterior middle temporal gyrus) based on the Harvard-Oxford atlas. b) The signature patterns associated with different stimulus types in the auditory Piano vs. Singing vs. Speech classification are visualized on a flattened temporal cortex map (thresholded at $p < 0.05$).

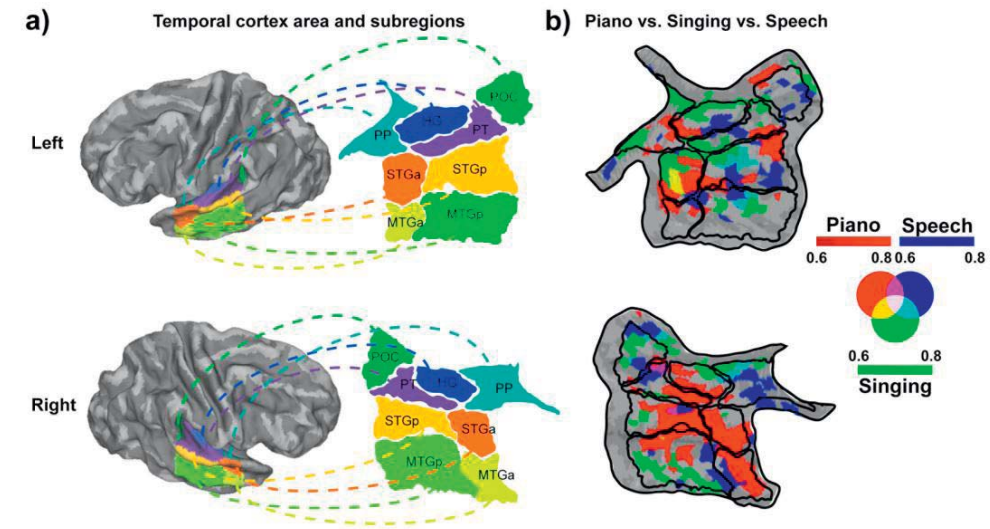
Figure 3. Signature patterns associated to 'musicness' and 'speechness' in a linear regression analysis. The included temporal lobe area is the same as in Figure 2, and the results are visualized on a similar flattened temporal cortex map (thresholded at $p < 0.05$).

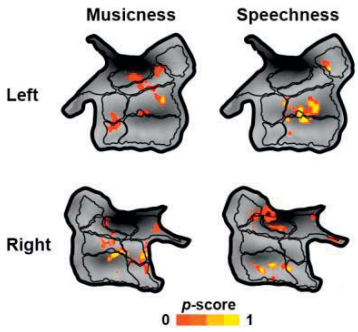
Figure 4. Signature patterns associated with Audiovisual vs. Auditory conditions are visualized on flattened temporal cortex maps (thresholded at $p < 0.05$). The included temporal lobe area is the same as in Figure 2. The probabilities are scaled from 0…1 to -1…1. AV, Audiovisual; A, Auditory.

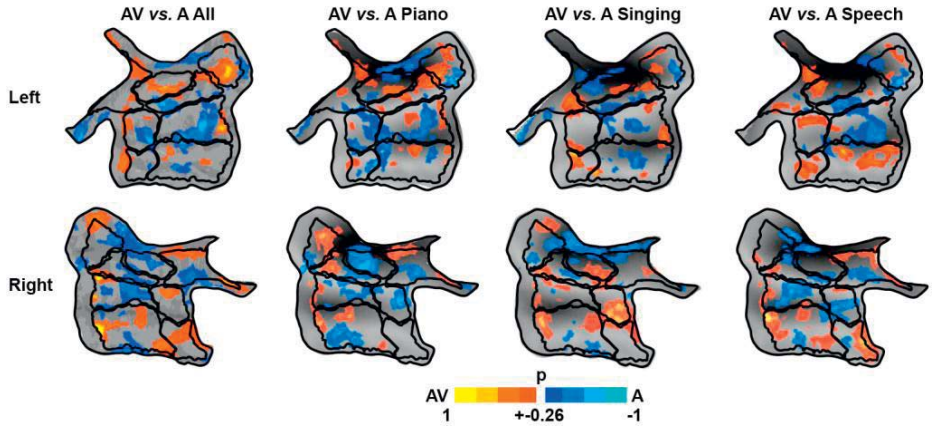
Highlights

- We used MVPA to study neural signatures of lifelike audiovisual speech and music.
- Audiovisual speech and audiovisual music modulated the activity in distinct supratemporal areas.
- Other brain areas were specific to corresponding unimodal auditory signals.
- Specific visual input may modulate the anterior and posterior auditory pathways.




a)
Temporal cortex area and subregions
Left
Right
POC
PP
HG
PT
STGa
STGp
MTGa
MTGp
b)
Piano vs. Singing vs. Speech
Piano
Speech
0.6
0.8
0.6
0.8
0.6
0.8
Singing


Musicness
Speechness
Left
Right
p-score
0
1


AV vs. A All
AV vs. A Piano
AV vs. A Singing
AV vs. A Speech
Left
Right
p
AV
A
1
+-0.26
-1