Applied Informatics
a SpringerOpen Journal

RESEARCH Open Access

Cross-participant modelling based on joint or disjoint feature selection: an fMRI conceptual decoding study

Hiroyuki Akama[1*], Brian Murphy[2,3], Miao Mei Lei[1] and Massimo Poesio[4,5]

* Correspondence:
akama.h.aa@m.titech.ac.jp
[1]Graduate School of Decision Science and Technology, Tokyo Institute of Technology, W9-10, 2-12-1, O-okayama, Meguro-ku, Tokyo 152-8552, Japan
Full list of author information is available at the end of the article

Abstract

Multivariate classification techniques have proven to be powerful tools for distinguishing experimental conditions in single sessions of functional magnetic resonance imaging (fMRI) data. But they are vulnerable to a considerable penalty in classification accuracy when applied across sessions or participants, calling into question the degree to which fine-grained encodings are shared across subjects. Here, we introduce joint learning techniques, where feature selection is carried out using a held-out subset of a target dataset, before training a linear classifier on a source dataset. Single trials of functional MRI data from a covert property generation task are classified with regularized regression techniques to predict the semantic class of stimuli. With our selection techniques (joint ranking feature selection (JRFS) and disjoint feature selection (DJFS)), classification performance during cross-session prediction improved greatly, relative to feature selection on the source session data only. Compared with JRFS, DJFS showed significant improvements for cross-participant classification. And when using a groupwise training, DJFS approached the accuracies seen for prediction across different sessions from the same participant. Comparing several feature selection strategies, we found that a simple univariate ANOVA selection technique or a minimal searchlight (one voxel in size) is appropriate, compared with larger searchlights.

Keywords: FMRI; MVPA; Machine learning; Feature selection; Cross-session; Cross-subject

Background

The general linear model (GLM) is a univariate analysis aiming to detect global activations over contiguous voxels, which exhibits a groupwise significant signal change between conditions. As this method is based on the smoothing of activation values, it is not designed to detect information encoded locally as fine patterns across individual voxels. Multivariate pattern analysis (MVPA) is now widely used in cognitive neuroscience to predict (decode) physiological or psychological states encoded in the brain, without assumptions of spatial smoothness or contiguity (Haxby et al. 2001; Cox and Savoy 2003; Mitchell et al. 2004). Of special interest to the authors, it has been successfully applied to the study of semantics, and linguistics more generally (Wang et al. 2003; Mitchell et al. 2008; Pereira et al. 2010; Pereira et al. 2011; Huth et al. 2012).

In machine learning analyses, such as MVPA, performance usually improves as the amount of training data increases. However, in brain decoding studies, it is often found

that within-session analyses (with cross-validation, to avoid double dipping (Kriegeskorte et al 2009)) attain higher classification performance than cross-session analyses (e.g. classifying data from a held-out session, after training on one or more separate sessions) involving larger quantities of training data. The reasons for this performance penalty are not entirely understood, though many candidates have been identified, including inaccuracies in registration to an atlas or movement correction, differences in overall brain shape and local folding patterns of sulci and gyri, and genuine differences in functional patterns whether temporary (e.g. due to caffeine) or more enduring (e.g. due to different functional localisation).

Recently, Haxby et al. (2012) proposed a method termed 'hyperalignment' which eliminates the penalty in classification accuracy across subjects. First, a supervised feature mapping approach (a specialized feature selection) is used across subjects, based on data gathered during the viewing of a rich audiovisual stimulus (a film). Without using any spatial constraints, sets of voxels are identified across subjects, which collectively exhibit similar functional sensitivity across the time course of the fMRI data. The training data is labelled, in the sense that the fMRI recording is temporally aligned to the film, and there is a direct equivalence between the time points across subjects. After this feature selection/mapping stage, different data from the same pairs (or set) of subjects are used for cross-subject learning (e.g. training on labelled data from participant A and testing on similarly labelled data from participant B).

Here, we propose methods for cross-session classification which differ from Haxby et al. (2012) principally in that we preserve the conventional cross-subject spatial constraints, assuming functional equivalence between co-registered points in a shared atlas space. Additionally, all testing and training data, from all sessions, are gathered during the same behavioural paradigm. We term the session used for classifier training the source session, S, and the session which we want to test the target, T. The target data is additionally partitioned into a portion used for feature selection only (T1) and a portion which is held out during all stages of training for validation (T2).

Our methods involve strategies for feature selection: the procedure for choosing the most sensitive and informative voxels to feed into a machine learning classifier. In joint ranking feature selection (JRFS), one of two approaches we present, conventional univariate feature-selection strategies are used (ANOVA and searchlight) based on data from *both* the source dataset (S) and on a partition of the target dataset (T1). The other approach (disjoint feature selection (DJFS)) uses the target dataset partition only (T1). In both variants, we assume spatial equivalence across sessions, and training of our linear models is carried out using only the source data S, and testing using the held-out target data T2. The voxels identified during feature selection are akin to a region of interest, and the subset which is subsequently used by the trained linear classifiers are distributed regions with shared local coding patterns across S and T.

As some data from the target session is needed before model training, this approach is appropriate for cognitive neuroscience studies which study shared functional activations among an experimental group of subjects, rather than diagnostic applications. Assuming that the application of this method is validated by successful cross-session or cross-subject classification of unseen trials (where the benchmark for success is the level of classification accuracy seen for within-subject analyses), the regions identified during the feature selection stage can be said to contain areas which share systematic patterns of local coding across those sessions or subjects.

We hope these methods will prove useful in analysing fine-grained differences in brain states, and in this paper, we describe their application to fMRI data recorded during a covert property generation task (cf. Mitchell et al. 2008). Our general research interest is in determining the extent to which the distributed and overlapping coding patterns elicited by conceptual stimuli (Haxby et al. 2001; Pulvermüller 2005) are shared across languages. Thus, to demonstrate these methods, we use fMRI data from a covert property generation task involving language switching by bilingual participants (either early acquisition Korean-Chinese bilinguals or late acquisition Chinese-Japanese bilinguals). Stimuli consist of an image accompanied by its captioned name in one of the two languages the subject speaks, and the task is to covertly produce in the other language.

In the next section, we describe the Methods used in the experiment and its analysis. In the results, we show that our DJFS and JRFS methods for data partitioning and feature selection result in considerable, significant improvements of classification performance for cross-session and cross-subject prediction, and approach the benchmark levels found for within-subject MVPA analyses.

Methods

Overview

This experiment is a partial replication of the experiments described in Akama et al. (2012) and Mitchell et al. (2008) on which that study was based. The subjects were asked to silently rehearse semantic properties on presentation of a conceptual stimulus (an image with paired caption), in a slow event-related design, while we scanned a coarse whole-brain image ($3 \times 3 \times 6$ mm) at a short TR of 1 s. Early- and late bilingual participants took part in two separate language-switched sessions: in one session, they had to silently produce in their first language in response to stimuli captioned in their second language, and in another session, production was in the second language and captions are in the first language.

MVPA analyses first established a within-session benchmark with conventional cross-validated classification. Then, several variations were attempted, where one partition of the target data was used for feature selection (either using a simple ANOVA selector or a searchlight of varying radius), and we examined the effect these had on success of cross-session and cross-subject classification over the other held-out partition of the target dataset.

Participants

A total of 15 graduate students from universities in Tokyo participated in the current experiment. Fourteen participants completed the study, as one dropped out after the first session for personal reasons, and this participant (l_P8) was excluded from the analyses reported here. There were four males and ten females, with a mean age of 24 years (range 22 to 28 years) and mean education of 18 years (range 16 to 20 years). All participants were native speakers of Mandarin Chinese who grew up in China and were also proficient in a second language. One subgroup was composed of early Korean-Chinese bilinguals (denoted e_P < *number* > in this paper: seven persons, one male, six females), and the other of late Chinese-Japanese bilinguals (denoted l_P < *number*>: eight persons, three

male, five female). See the 'Appendix' section for further details on the selection of participants and testing of their language fluency.

Materials

The experimental paradigm used here is based on Mitchell et al. (2008) and Akama et al. (2012). A total of 40 contrast-normalized grey-scale photographs were used in the present study. All pictures were chosen from a set of stimuli previously used for predicting EEG activation patterns (Murphy et al. 2009; Murphy et al. 2011). There were 20 different images for each of two categories of stimuli, tools and mammals, and we prepared three versions of each image with the name of the object given in Chinese, Japanese or Korean as a caption. E-Prime 2.0-Standard software was used to present the stimuli and guaranteed synchronization with the fMRI scanner. In each run, all 40 images were presented in random order on the back-projected screen. Each session included 6 runs, to give a total of 240 image presentation trials. The list of 40 concepts is given below (written in Chinese, Japanese and Korean), and such concepts are illustrated in the Additional file 1 ('Materials') with the corresponding pictures:

Mammals: anteater (食蚁兽, アリクイ, 개미핥기), armadillo (穿山甲, アルマジロ, 아르마딜로), beaver (河狸, ビーバー, 비버), camel (骆驼, ラクダ, 낙타), deer (鹿, シカ, 사슴), elephant (大象, ゾウ,코끼리), fox (狐狸, キツネ, 여우), giraffe (长颈鹿, キリン, 기린), gorilla (大猩猩, ゴリラ, 고릴라), hare (野兔, 野ウサギ, 토끼), hedgehog (刺猬, ハリネズミ, 고슴도치), hippopotamus (河马, カバ, 하마), kangaroo (袋鼠, カンガルー, 캥거루), koala (考拉熊, コアラ, 코알라), mole (地鼠, モグラ, 두더지), monkey (猴子, サル, 원숭이), panda (熊猫, パンダ, 참대곰), rhinoceros (犀牛, サイ, 코뿔소), skunk (臭鼬鼠, スカンク, 스컹크), zebra (斑马, シマウマ, 얼룩말). Tools: Allen key (六角匙, 六角レンチ, 앨런 볼트용 렌치), axe (斧头, 斧, 도끼), chainsaw (链锯, チェーンソー, 동력 사슬톱), craft knife (工艺刀, クラフトナイフ, 다용도 칼), file (锉刀, ヤスリ, 줄), hammer (铁锤, ハンマー, 망치), nail (钉子, 釘, 못), paint roller (油漆滚筒, 塗量ローラー, 페인트 롤러), trowel (抹泥刀, 移植ゴテ, 모종삽), pliers (钳子, ペンチ, 펜치), plunger (活塞, パイプ吸引具, 플런저), power drill (电钻, 電気ドリル, 동력 천공기), rake (耙子, 熊手, 갈퀴), saw (锯子, ノコギリ, 톱), scraper (刮刀, こすり落とし用ヘラ, 긁어내는 도구), scissors (剪刀, ハサミ, 가위), screw (螺丝钉, ネジ, 나사), sickle (镰刀, 鎌, 낫), spanner (扳手, スパナ, 스패너), tape measure (卷尺, 巻き尺, 줄자).

The orthographic form of the stimuli varies: from pure ideograms (Chinese), an alphabetic language (Korean Hangul), and a combination of ideograms (Kanji) and syllabary (Katakana) in Japanese[a].

Task design

A slow event-related design was used. The participants attended two separate scanning sessions carried out on two different days separated by at least 1 week, alternating the languages for stimuli and task. In each session, the stimuli were presented with captions in one language, and the participant was asked to complete property generation in the other language in which they had fluency. For example, Korean-Chinese bilinguals were asked to participate in one K → C session (stimulus captions in Korean, property generation in Chinese) and in a C → K session on a different day (Figure 1).

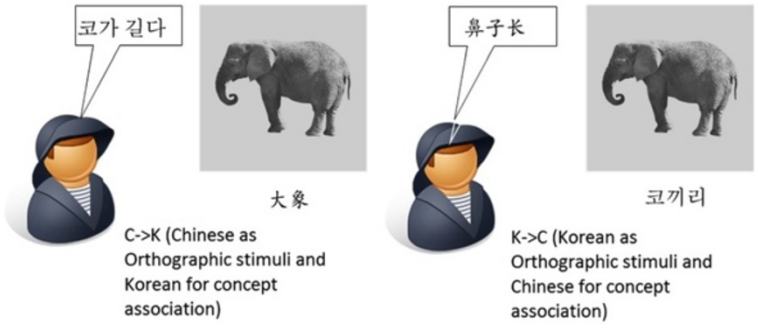

Figure 1 Schematic of stimulus and task in both sessions for Korean-Chinese early bilingual speakers. 'Language A → Language B' signifies one trial in which a participant was presented with stimuli captions in language A and was asked to perform covert property generation in language B by thinking the appropriate features of the corresponding concept. In this example, a bilingual participant retrieves the words 'a long nose' as the important feature of an 'elephant' by simultaneously performing a language switch in one trial.

Similarly, each Chinese-Japanese subject completed both a C → J session and a J → C session. The order of these two sessions was alternated across subjects.

Each session had 6 repeated runs for a total of 240 trials. In each trial, each concept was presented for 3 s followed by a fixation cross for 7 s. There were six additional presentations of a fixation cross of 40 s each, distributed just after each run. During the 3-s stimulus presentation, the participants were asked to do a silent property generation task, thinking of appropriate features of the corresponding concept in the required language, and this was followed by a fixation cross presentation time of 7 s during which participants were asked to fixate their eyes on the cross silently and no response was required (Figure 2). See the Appendix for further details on the choice of behavioural task and pre-session training that participants performed.

Multivariate pattern analyses

Pre-processing of the fMRI data was performed with Statistical Parametric Mapping software (SPM8, Wellcome Department of Cognitive Neurology, London, UK). The data were motion corrected, co-registered to the anatomical images, segmented to identify grey matter and normalized into standard Montreal Neurological Institute (MNI) space at a resliced voxel size of 3 × 3 × 6 mm.

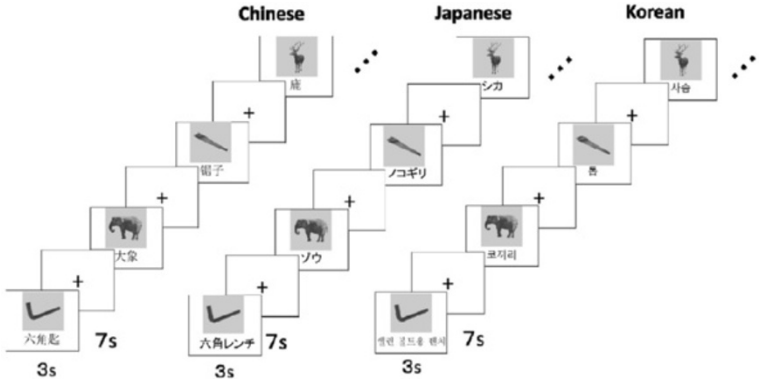

Figure 2 Experimental stimuli in each of three caption language conditions.

The MVPA analyses used PyMVPA 2.0 (http://www.pymvpa.org/), a Python package developed to run machine-learning programs applied to neurological data. With the exception of the novel data partitioning and feature selection strategies presented here, the classification methods and associated parameter settings were adopted unchanged from Akama et al. (2012). The realigned, co-registered, segmented and normalized (but unsmoothed) images of each participant in each session were used to train classifiers for voxel pattern discrimination, allowing us to discriminate the semantic category (animal or tool) of individual trials. To approximate the hemodynamic response, we used a boxcar average over trial volumes, taking an onset delay of 4 s and a boxcar width of 4 s (in Akama et al. (2012), we show that this can give excellent results, comparable with canonical models of the HRF).

We used an *L2*-norm Penalized Logistic Regression classifier (PLR) as implemented in the PyMVPA package1. In any logistic regression (see e.g. Hastie et al, 2009, chapter 4), a β-weighted linear combination of the explanatory variables X is used to estimate the response variable y (1), embedded within a sigmoid logistic function (2). In our case the explanatory variables are the fMRI data, and the response variable is the class {animal, tool}. The logistic function models this two class variable as {0,1}, mapping the real numbers to the interval (0,1), centred around a decision boundary of 0.5.

$$\hat{y} = S(\beta X) \tag{1}$$

$$S(t) = \frac{1}{1 + e^{-t}} \tag{2}$$

$$\underset{\beta}{\overset{argmin}{}} \left\{ \|y - S(\beta X)\|^2 + \lambda \|\beta\|^2 \right\} \tag{3}$$

The *L2* penalty (or Ridge) regularization term avoids overfitting, thereby dealing both with the high dimensionality and redundancy in fMRI data. Optimization of the fit is by gradient descent, simultaneously minimizing the sum of squares of the β weights, and the modelling error, which is the sum of squared errors between the actual classes and modelled classes (3).

For the regularisation tuning parameter λ, we used the default value of 1.0, as was used successfully in Akama et al. (2012). This value could be further optimised by nested cross-validation. However, our aim here is not to show which MVPA strategy gives the maximum classification performance (indeed, one might try other classifiers). Rather, we want to demonstrate the effect of feature selection strategies given a fixed classifier setting.

The reported accuracies measure the proportion of trials for which a trained classifier could correctly determine semantic class. In within-subject analyses, we used sixfold cross-validation, over folds consisting of interleaved trials, and the reported accuracies are means over each of the six test folds. In the JRFS and DJFS settings (see next subsection), the target data set was partitioned into two portions (T1 and T2), and each was tested separately: T1 was tested after T2 had been used during feature selection and vice versa. Reported accuracies are the mean of the result of these two computations. Additionally, cross-session analyses were carried out in two modes: using only one source dataset for training and using the group of all other participants' data for training (to examine the effect on classification accuracy of having larger numbers of trials, from a broader sample of subjects).

Joint ranking feature selection and disjoint feature selection

For JRFS, we split the dataset of the target session T into two halves. Each of these (T1 and T2) in turn was used together with the whole source session dataset S for feature selection. Training of the PLR classifier then proceeded using S only, and testing was performed on the unseen partition of T. In all cases, the top ranked 5,000 voxels were selected as input to the classifier.

The DJFS method only differs in that the selector is applied to the half partition of the target data T only. Otherwise, the same set of settings and analyses were used.

For both, we first used ANOVA as the selector and then tried a cross-validated searchlight with several radius settings beyond the central voxel ($r = \{0, 1, 2, 3\}$). The searchlight (Kriegeskorte and Bandettini 2007) varies from ANOVA in two interesting ways. First, it is a cross-validated selection method and so an extra level of validation underlies the ranking it supplies. Second, with the searchlight, we can vary the radius of the globe of voxels considered and investigate the extent to which coding regularities are local.

Results and discussion

Classification within single-participant sessions

The classification result within a single-participant session is established here to be used as a comparison benchmark for cross-session predictions. Classification accuracy was measured by the proportion of single trials whose semantic category (animal or tool) was successfully determined. The two classes are balanced, so chance performance was 50%, and accuracies above 55.8% were significant (at $p < 0.05$, binomial test over independent trials, chance 50%, $n = 240$). This chance level was identical to the result of a permutation test with random labelling.

As shown in Figure 3, for experiment I (Korean-Chinese bilinguals), the classification accuracy was well over this threshold for all sessions, both in the Korean-to-Chinese language-switching condition (K → C, L1 → L2; image captions in Korean, covert generation in Chinese) and also in the Chinese-to-Korean (C → K, L2 → L1) condition. In the K → C condition, the mean accuracy was 93.4% (SD = 4%), ranging from 70.7% to

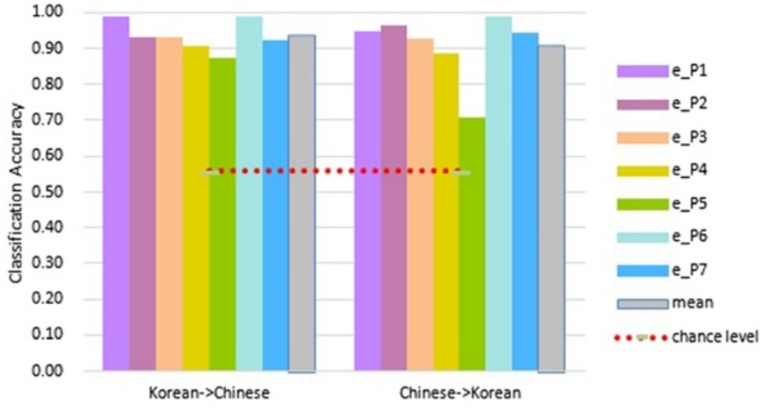

Figure 3 Accuracies for within-single-participant session classification in experiment I (Korean-Chinese early bilinguals).

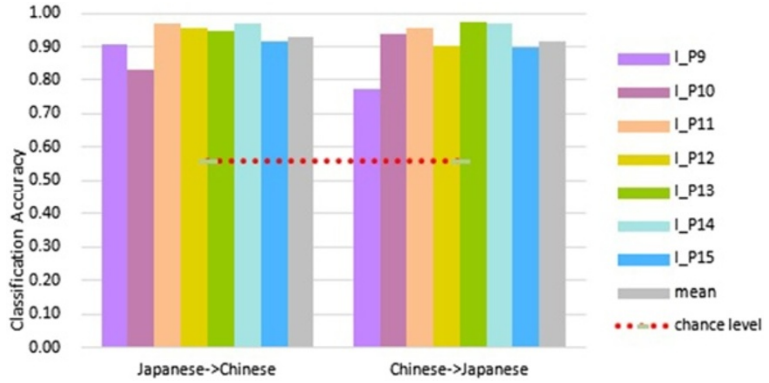

Figure 4 Accuracies for within-single-participant session classification in experiment II (Chinese-Japanese late bilinguals).

98.8%. In the C → K condition, the mean accuracy was 90.8% (SD = 9%), ranging from 77.5% to 96.7%.

Figure 4 shows the equivalent results for experiment II (Chinese-Japanese second language learners), where all individual accuracies were again well beyond the significance threshold (55.8%). In the C → J (L1 → L2) condition (captions in Chinese, covert task in Japanese), the mean accuracy was 91.6% (SD = 7%) ranging from 77.5% to 97.5%. In the J → C (L2 → > L1) condition, the mean accuracy was 92.7% (SD = 5%), ranging from 83.3% to 96.7%.

Within-participant cross-session classification

Here, we see whether category-specific activation patterns are shared between different sessions in the same participant. The PLR classifier with ANOVA feature selection was trained on all 240 trials from one session and then tested directly to discriminate among animal and tool presentations, on the 240 trials from the same participant's other experimental session.

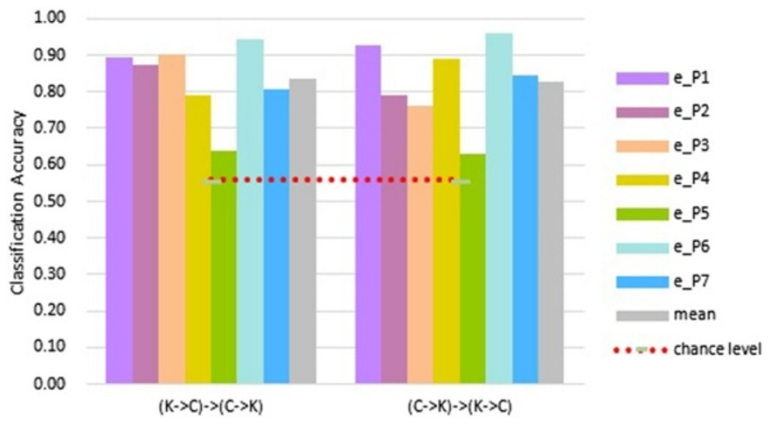

Figure 5 Accuracies for within-participant cross-session classification in experiment I (Korean-Chinese early bilinguals).

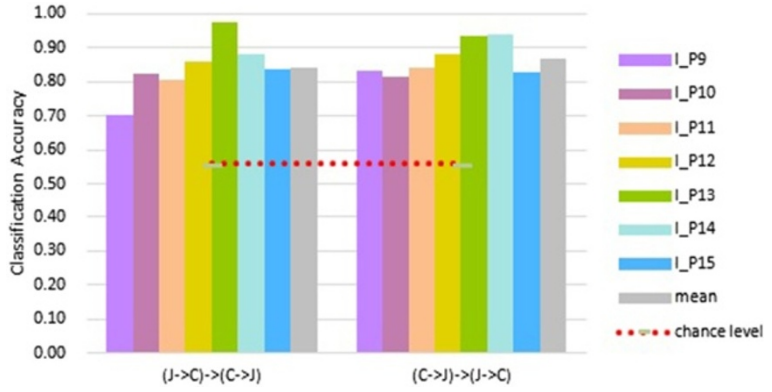

Figure 6 Accuracies for within-participant cross-session classification in experiment II (Chinese-Japanese late bilinguals).

Compared to within-session classification, the cross-session prediction was slightly less successful, as demonstrated in Figures 5 and 6. Still, all results were significantly above chance at the same threshold of 55.8%.

For experiment I (early Korean-Chinese bilingual participants), the (K → C) → (C → K) analysis (training on data from the Korean caption/Chinese production session; testing on data from the Chinese caption/Korean production session) achieved a mean accuracy over seven participants of 83.6% (SD = 10%), ranging from 63.7% to 94.2% (Figure 5, left panel). In the other direction, (C → K) → (K → C), the classification accuracy was also significant, with a mean accuracy of 82.9% (SD = 11%), ranging from 62.9% to 95.8% (Figure 5, right panel). For the late Chinese-Japanese bilingual group (Figure 6), the mean classification accuracy was 84.0% (SD = 8%) in the (J → C) → (C → J) analysis (range 70.4% to 97.5%) and 86.7% (SD = 5%) in the (C → J) → (J → C) analysis (range 81.2% to 93.3%).

Cross-participant classification: pairwise and groupwise

Here, we performed a similar analysis to that done in the previous section (PLR classifier, ANOVA feature selection, different training and testing sessions) but classified the data from one participant after training on the data of different participants. We first do this training on the data from single-participant sessions and then training on whole groups of participants.

In Figure 7, we present the classification accuracy when doing ANOVA feature selection and training on one dataset (see *y*-axis) and testing on another (*x*-axis). For comparison, on alternating cells just off the diagonal, we have shown the within-participant cross-session accuracies (as already shown in Figures 5 and 6).

The mean classification accuracy for the other cells in Figure 7 (where training and testing data come from different participants) is 64.6%. This is considerably lower than that seen for within-session and within-participant analyses, but 69.0% of the individual test/training pairs were above the significance threshold of 61.3% (binomial test with Bonferroni correction, $n = 168$).

In Figure 8, the results are shown where ANOVA feature selection and training are performed over the data from 13 participants (26 sessions, 6,240 stimulus trials in

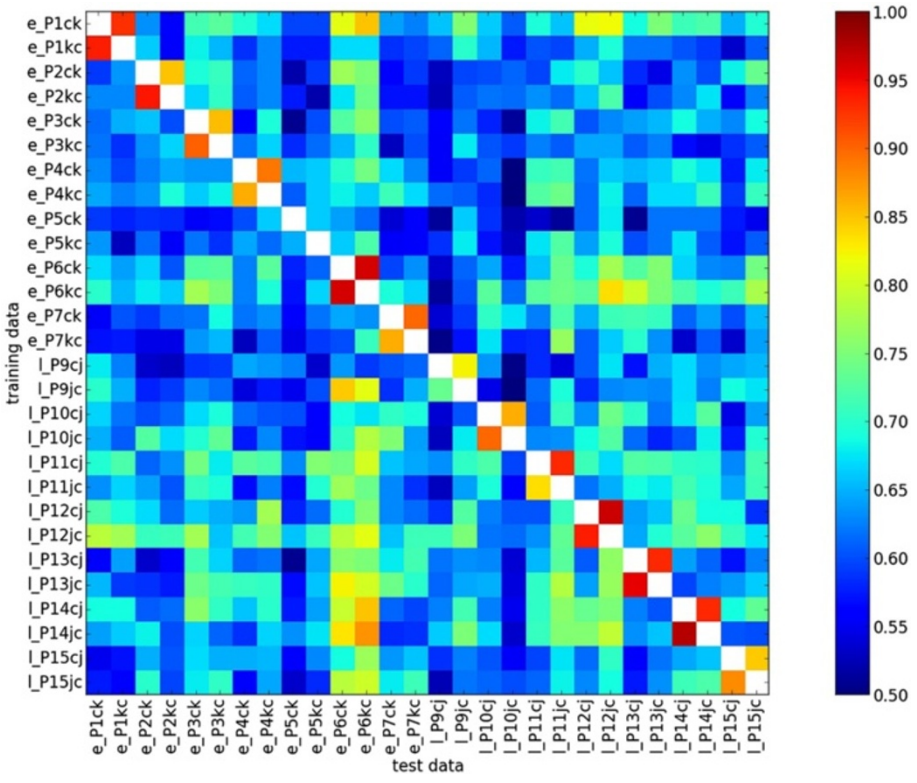

Figure 7 Pairwise cross-participant classification accuracy (all experimental sessions). The *y*-axis indicates the training dataset, and the *x*-axis the test dataset.

total), while trials from the remaining held-out participant (2 sessions) are classified. Of the 28 classification analyses, 92.9% reach a significance threshold of 60% (with Bonferroni correction, $n = 28$). We see a clear improvement of mean classification accuracies, which reach 74.5% compared to 64.6% in the previous analysis that used only one session at a time for training, rather than 26 here. But it is still considerably lower than the mean accuracy of 83.2% seen for cross-session analyses using only a single training session from the same participant (Figures 5 and 6; $t = 7.7$, $p < 2.8 \times 10^{-8}$).

JRFS cross-participant classification: pairwise and groupwise

In the last section, we saw that there was a clear performance penalty for training across participants relative to cross-validated testing/training from a single participant, even when dramatically increasing the amount of training data by including multiple sessions. Here, we introduce a joint feature selection strategy to try to address that. Up until now, our feature selections have used only the same *source* data as is used in training. Here, feature selection is performed jointly using all of the source data and one half of the data from the *target* session dataset at a time. Training is still executed using the source data only, and the testing uses the other held-out half of the target set that did not contribute to feature selection. In each analysis, this process is performed twice (i.e. twofold), so that each half of the target dataset can be tested separately, and the accuracies given are the mean of those two separate accuracies.

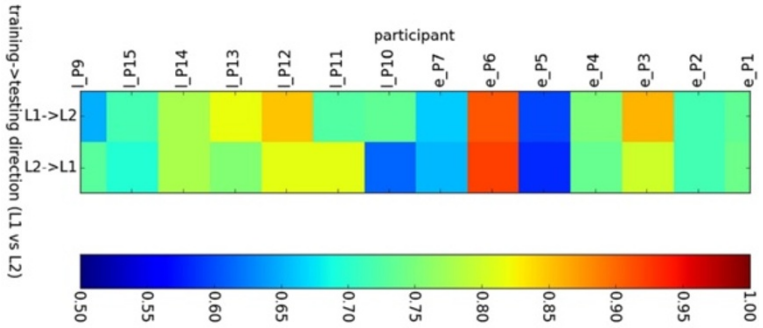

Figure 8 Groupwise cross-participant classification accuracy (all experimental sessions). The *x*-axis indicates the test participant, and the *y*-axis the language direction for that session: 'L1' and 'L2' means 'first native language' and 'second native language', respectively. L1 is Korean for the early bilinguals (with a participant identifier of 'e_P*') and Chinese for the late bilinguals (with a participant identifier of 'l_P*'), whereas L2 is Chinese for the early bilinguals (with a participant identifier of e_P*) and Japanese for the late bilinguals (with a participant identifier of l_P*).

Figure 9 shows the cross-session pairwise classification accuracies, as in Figure 7, but now using the modified JRFS strategy. Considering the cross-participant analyses only (off-diagonal cells in the plot), 89.7% of them were above the significance threshold of 61.3% (binomial test, $p < 0.05$, with a Bonferroni correction of $n = 168$). The mean accuracy rate was 71.3%, which was an average improvement of 6.7% points, in comparison to the pure cross-participant modelling (feature selection and training on one participant, testing on another; Figure 7). While there was a strong correlation of

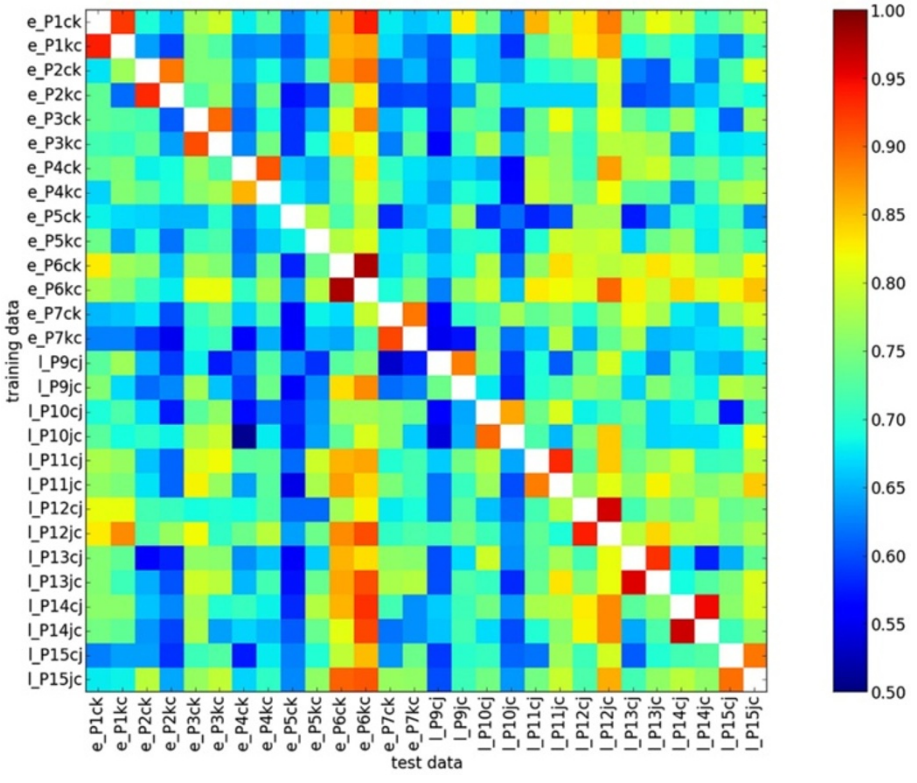

Figure 9 JRFS pairwise cross-participant classification accuracy (all experimental sessions). The *y*-axis indicates the training dataset, and the *x*-axis the test dataset.

$r = 0.78$ between the conventional and JRFS accuracies, the group-level difference was highly significant ($t = -36.69$, $p < 7.79 \times 10^{-82}$).

Figure 10 shows the results of the JRFS group classification accuracies (corresponding to Figure 8 which used a conventional source-dataset-only feature selection). The mean classification accuracy was 80.0%, significantly higher than that seen with conventional feature selection (74.5%, $t = -5.76$, $p < 3.97 \times 10^{-6}$) and increased by 8.7% points compared to the single- or cross-participant analyses in Figure 9.

Here, all the sessions exceed the significance threshold of 60% (with Bonferroni correction, $n = 28$). What is notable here is that the modelling for the participant 'e_P5', whose dataset recorded the worst accuracy in the within-participant prediction, was considerably improved as a result of the feature co-selection technique. The precision rates of the Korean-to-Chinese and the Chinese-to-Korean predictions in this particular case increased by virtue of JRFS from 59.6% to 66.2% and from 58.3% to 70.8%, respectively.

DJFS cross-participant classification: pairwise and groupwise

In this section, the results of DJFS are reported in comparison of those of JRFS. For DJFS, the voxels are selected only from one half at a time of the dataset of a target subject T, model training is made on the whole dataset of a source subject S, and testing is executed on the held-out half of the dataset T. Figure 11 shows the results of the pairwise DJFS classification, which outperformed all the between-subject classification techniques. The mean classification accuracy was 75.7%, which was significantly higher than that of the JRFS (71.3%), with an improvement of 4.3% points ($t = -24.87$, $p < 3.2172 \times 10^{-99}$). Note that 94.8% of the subject combinations were above the significance threshold of 61.3% (binomial test, $p < 0.05$, with Bonferroni correction of $n = 168$).

The results of the DJFS groupwise classification accuracies also ranked top among all the groupwise modelling instances with the same configuration. The mean accuracy was 82.0%, an increase of 2% points compared to that of the JRFS groupwise classification ($t = -3.08$, $p < 0.00471$). The classification accuracy of all the sessions was greater than the significance threshold of 60% (with Bonferroni correction, $n = 28$) (Figure 12).

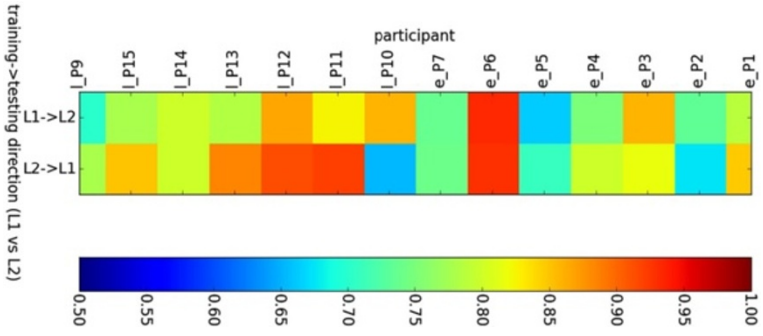

Figure 10 Groupwise cross-participant classification accuracy (all experimental sessions using the JRFS). This figure is to be compared with Figure 8.

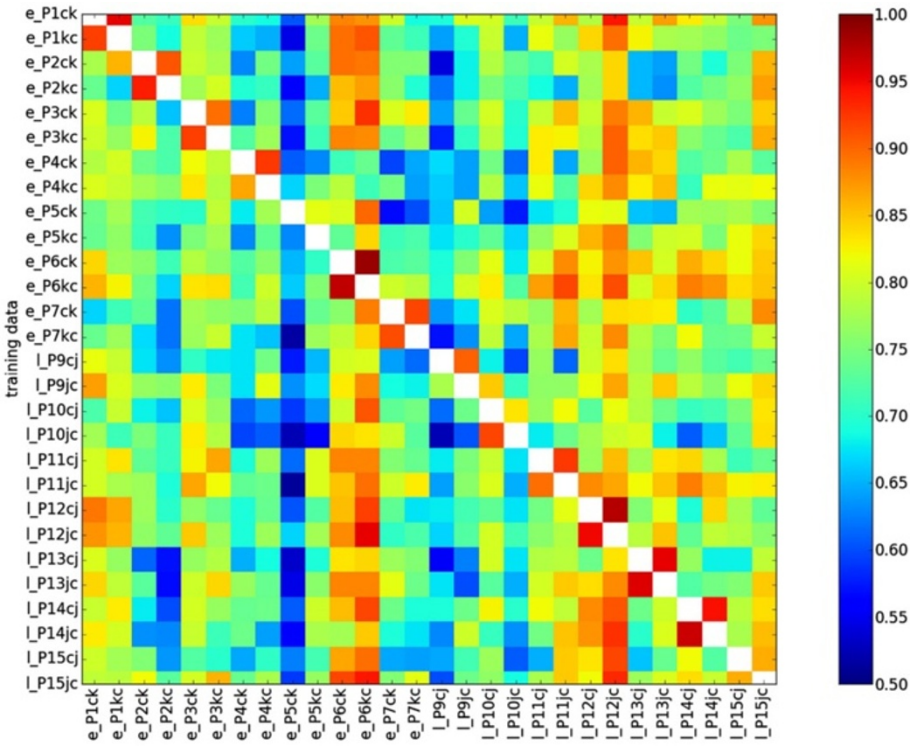

Figure 11 DJFS pairwise cross-participant classification accuracy, all experimental sessions. The *y*-axis indicates the training dataset, and the *x*-axis the test dataset.

JRFS and DJFS using a searchlight selector

It turns out that the JRFS and DJFS using a cross-validated searchlight were similarly effective to the ANOVA-based JRFS and DJFS described above. After removing all the results of the within-single-participant session prediction as before, we calculated (1) the mean and (2) the standard deviation of the classification accuracy and (3) the proportion of significant combination session patterns (accuracy larger than 61.3% at $p < 0.05$). For a JRFS searchlight selector using radii = 0, 1, 2, or 3, these statistics were {72.1%, 0.079, 91.1%}, {69.6%, 0.076, 87.4%}, {67.7%, 0.076, 79.3%} and {66.3%, 0.076, 73.6%}, respectively. The corresponding number for DJFS was consistently superior, recording {76.3%, 0.082, 95.3%}, {72.6%, 0.081, 91.9%}, {69.7%, 0.076, 86.9%} and {68.0%, 0.074, 83.1%}, respectively.

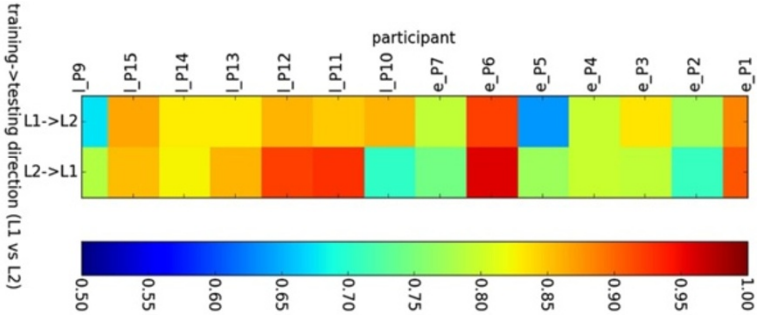

Figure 12 Group-wise cross-participant classification accuracy, all experimental sessions using the DJFS. This figure is to be compared with Figures 8 and 10.

Figures 13 and 14 represent the results of the cross-participant prediction based on the JRFS and DJFS with the searchlight.

The pattern of DJFS outperforming JRFS from the last analysis was seen again at all searchlight size settings. Contrary to our expectations, there was no significant improvement in performance using a searchlight selector over an ANOVA selector, despite a small apparent advantage for searchlights with radius of 0 (i.e. volume of a single voxel). There were significant disadvantages in using larger searchlights, relative to the ANOVA selector (multiple-comparison Bonferroni test executed posterior to a one-way ANOVA).

Summary of results

Figure 15 summarizes the results of all analyses, showing the mean classification accuracy over all datasets for each feature selection and data partitioning strategy examined. The results clearly illustrate the established effects of a cross-session penalty in classification accuracy, and in the groupwise results, an advantage as the number of training sessions and trials increases due to an improvement of signal-to-noise ratio and a broader sampling of the population of trials and subjects. Our feature selection and partitioning strategies both outperform the conventional methods, and DJFS has an advantage over JRFS, approaching the benchmark levels of within-subject analysis. In terms of the feature selector, cross-validated searchlight results are slightly higher than ANOVA, but not significantly, and only for a minimal searchlight size of 1 voxel.

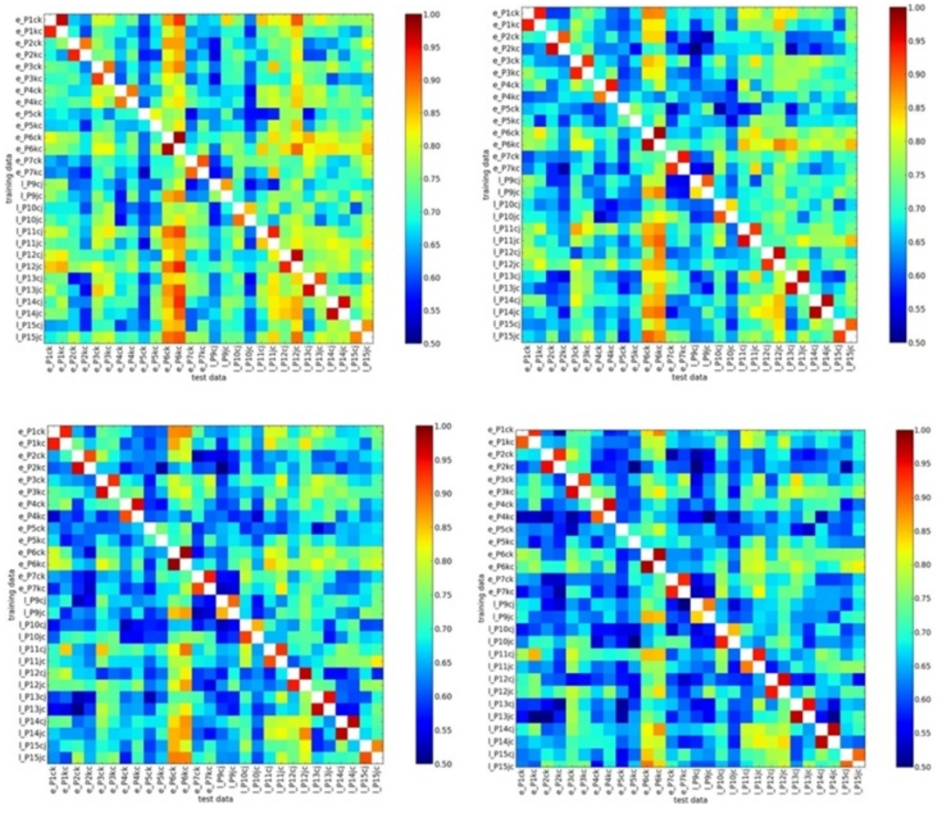

Figure 13 Results for cross-participant classification using JRFS and a searchlight selector. Searchlight radius varies: 0 (upper left), 1 (upper right), 2 (lower left) and 3 (lower right).

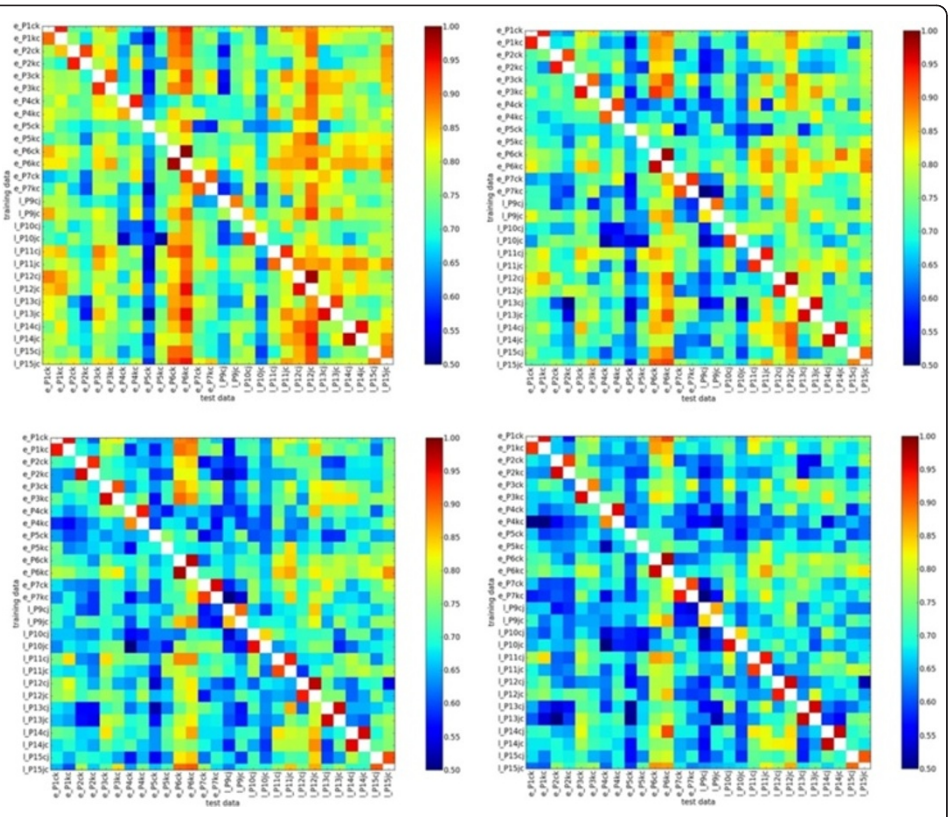

Figure 14 Results for cross-participant classification using DJFS and a searchlight selector. Searchlight radius varies: 0 (upper left), 1 (upper right), 2 (lower left) and 3 (lower right).

Conclusions

This paper presents MVPA techniques in which a feature selection was carried out using a subset of a target dataset T. Using one of two simple and computationally cheap partitioning methods (JRFS and DJFS), we alleviated the typical penalty in classification accuracy seen for cross-participant and cross-session studies. DJFS proved

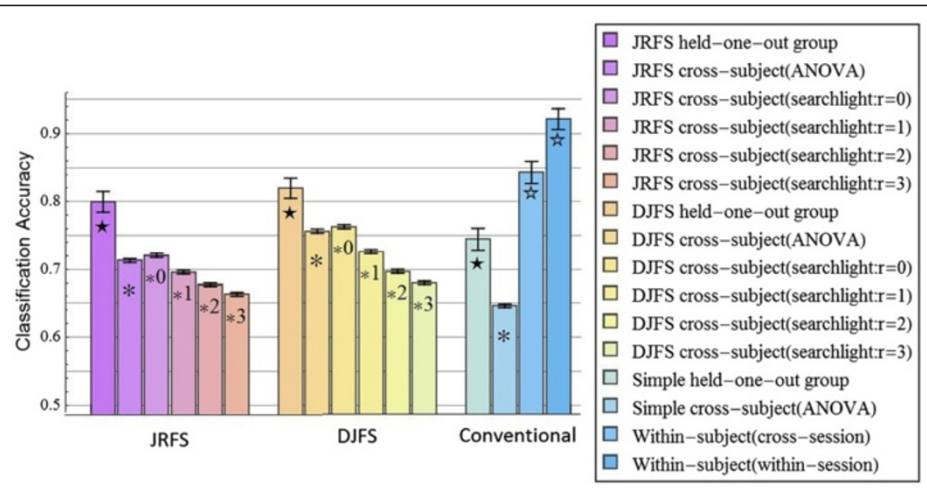

Figure 15 Summary bar plot for the performance achieved the JRFS and DJFS techniques. Error bars represent standard error. The meaning of the auxiliary symbols is as follows; filled star: held-one-out group, asterisk: cross-subject, empty star: within-subject.

the more effective of the two methods, and when applied with a minimal searchlight as a selector and after training on a group of participants, its performance on classification of trials from an unseen subject approached that seen for within-session analyses.

In some senses, the analyses that we propose might be seen as a "weaker" task than conventional cross-subject classification, but "harder" than within-participant cross-validated classification. However, the main value of the classification accuracy measures here are to validate the learning. The very competitive predictive accuracy attained by our final classifiers demonstrates that we have generalised across participants: after the feature-selection stage, we go on to learn linear weights over voxels, from a set of animal/tool-labelled trials from subject S only. Using those learned patterns from subject S, we can determine with high accuracy the stimulus type of unseen trials from subject T. While the classifier did see a different held-out set of trials from subject T, this was during feature selection only, and can be compared to the use of region-of-interest localizers. Crucially, it does not tell the classifier what information each voxel carries (i.e. whether a relative increase in BOLD activity is associated with animals or is associated with tools, and what linear combination of those voxels is most reliable in estimating trial type). That knowledge must be learned from subject S, and successful classification demonstrates cross-participant commonality of local coding patterns.

Figure 16 illustrates the effects of the different feature selection strategies. Among the widely distributed patterns, some tendencies emerge. For this pair of participants, the T session shows more sensitivity in frontal regions, and the S session more in posterior areas. JRFS extracts from a mix of regions across the two sessions (S-T2 and S-T1), pointing to areas of shared local coding patterns.

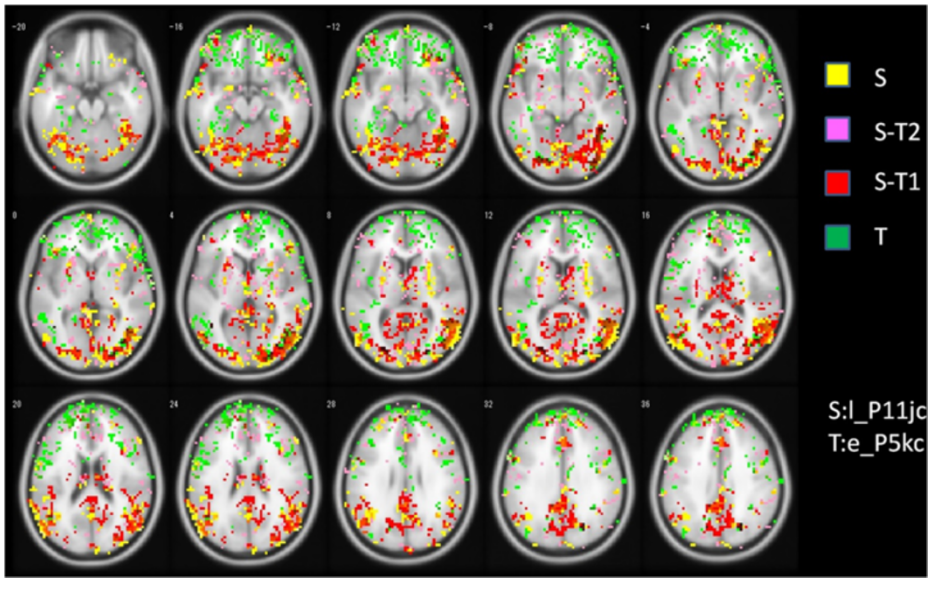

Figure 16 The top 5,000 voxels selected for an illustrative source/target pair visualized using the xjView toolbox. ANOVA-ranked selection is shown for within-subject cases (S:l_P11jc and T:e_P5kc) and both cross-subject cases (S-T1 and S-T2).

Situating our method relatively to Haxby et al. (2012), we largely alleviate the penalty in classification accuracy typically seen when classifying across individuals, relative to within-session classification, while using a technique which is straightforward and computationally cheap. In contrast to their analyses that used recordings during film viewing for the feature-selection/mapping stage, our method identifies the subset of voxels that are much more specific in their sensitivity to the experimental task at hand. In our continuing work, we will now use these methods to examine fine-grained conceptual representations across speakers of different languages and in sessions where the same speaker alternates between native languages.

Endnote

[a]The neural substrate of phonology is thought to be shared among these languages, together with English. In contrast, there is some evidence from orthographic priming effects (Weekes et al. 2005) of an alternative visual word form area located in the angular gyrus. However, these differences should be orthogonal to the conceptual distinctions that are the focus of this study.

Appendix

Participants

Early and late bilinguals (Paradis 2003) were chosen for the study, as the age of language acquisition is thought to effect cortical representations (see Perani and Abutalebi (2005) for a recent review). All participants were paid for their participation. All participants had normal or corrected-to-normal vision and were assessed to be strongly right-hand dominant, also reporting no left-handed immediate relatives. The participants were healthy, were not taking medication, and had no record of serious physical neurological or psychiatric illness. They gave and signed a written informed consent in accordance with the guidelines established by the Ethics Committee of the Graduate School of Decision Science and Technology at Tokyo Institute of Technology. All seven participants in the Korean-Chinese early bilingual group were ethnic Korean Chinese from the Yanbian Korean Autonomous Prefecture of Jilin Province or its neighbouring areas in China. Those participants acquired Korean from early childhood in their family and learned Chinese in school. On a five-point scale (1 = 'very non-proficient', 5 = 'very proficient'), the Korean-Chinese bilingual participants self-reported as 'very proficient' in their Chinese reading (mean = 4.71), listening (mean = 5.00), speaking (mean = 5.00) and writing ability (mean = 4.29). Similarly, their self-reports for Korean ranged from 4.71 (writing) to 5.00 (reading, speaking and listening). An independent-samples t test showed no significant differences between self-reports of L1 and L2 ability (reading t = 1.549, writing t =1.643, $p > 0.05$). For the Chinese-Japanese high proficiency second language learners group, we recruited eight late bilingual participants. These individuals had all passed the most advanced N1 level of the Japanese Language Proficiency Test (JLPT), indicating the ability to understand and use Japanese in a broad variety of circumstances (http://www.jlpt.jp/e/about/index.html). JLPT is the most widely recognized test for measuring abilities in the Japanese language and is administered by Japan Education Exchanges and Services. The eight Chinese-Japanese second language learners started learning Japanese at an average age of 17.6 (SD = 1.06, range 15 to 18)

and had studied Japanese for an average of 6.3 years (SD = 2.55, range 2.5 to 11 years), including classroom study. The participants were living in Japan at the time of the experiment. The mean length of their stay in Japan was 6.3 years (SD = 2.12, range 2.5 to 8 years). According to their self-reported questionnaires, the language more frequently used in their daily lives was Japanese. On average, the participants reported themselves as 'proficient' (mean = 4.13) in their Japanese reading and listening ability (mean = 4.13) and 'moderately proficient' in Japanese speech (mean = 3.75) and writing (mean = 3.75). In contrast, although they had lived in Japan for a long time, their ratings of Chinese language abilities were still all very high, ranging from 4.63 (writing in Chinese) to 4.88 (reading, listening and speaking in Chinese). A t test showed significant differences between Chinese and Japanese (listening $t = 2.898$, reading $t = 4.243$, speaking $t = 5.463$, writing $t = 3.564$; $p < 0.05$).

Behavioural task

To ensure that each participant had a consistent set of properties to think about during on-line tasks, the participants were asked to get acquainted with these stimuli and perform a property rehearsal task before the scanning session (Mitchell et al. 2008). Considering that Chinese-Japanese second language learners would possibly have an unbalanced degree of language proficiency, they were required to do sufficient self-preparation as well as to rehearse under supervision for at least an hour before each scanning session. As the task level was higher than that in a previous study (Akama et al. 2012), the offline rehearsal prior to each fMRI session was more intense than in the previous study. Note that these experiments were held exclusively in the orthographic condition. The framework of cross-language prediction is nearly symmetric across two sessions but not from the viewpoint of writing systems. However, Chinese and Korean characters engage the same visual word form area in proficient early Chinese-Korean bilinguals. So, the factors related to the orthographic differences might

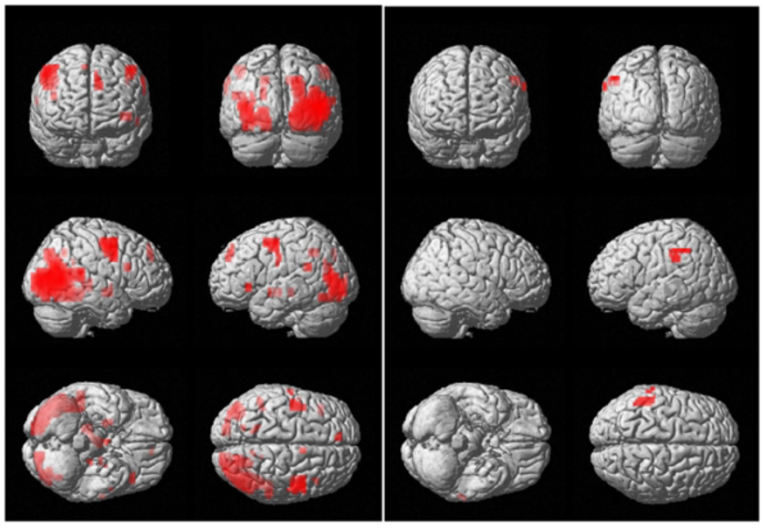

Figure 17 Averaged brain activations of mammal > tool and vice versa. (left) Averaged brain activations of mammal > tool of the random effects analysis (uncorrected, $p < 0.0005$). (right) Averaged brain activations of tool > mammal of the random effects analysis (uncorrected, $p < 0.0005$).

be considered as sufficiently controlled even without cutting off the pure visual area working independently of language processing. It is worth noting in passing that we used covert tasks in line with Mitchell et al. (2008) as well as conventional bilingual research (Kim et al. 1997; Hernandez et al. 2001; Chee et al. 2003; Wang et al. 2007). Such studies avoid overt speech tasks, as they focus on the neural response of the left frontal lobe in language-switching conditions, particularly the pre-central gyrus and Broca's area which can be affected by vocal behaviour.

GLM results of all the sessions

The data from the 14 early and late bilingual participants, totaling 28 sessions, were analysed with GLM procedure using SPM8. The figure (Figure 17) represents the result of the random effects group analysis ($p < 0.0005$ uncorrected) applied to the data of all our participants. According to the t contrast of mammal > tool in our study (the sign of inequality '>' means here a contrast direction), the mammal items of which intensity was significantly larger than the tool versus mammal comparisons

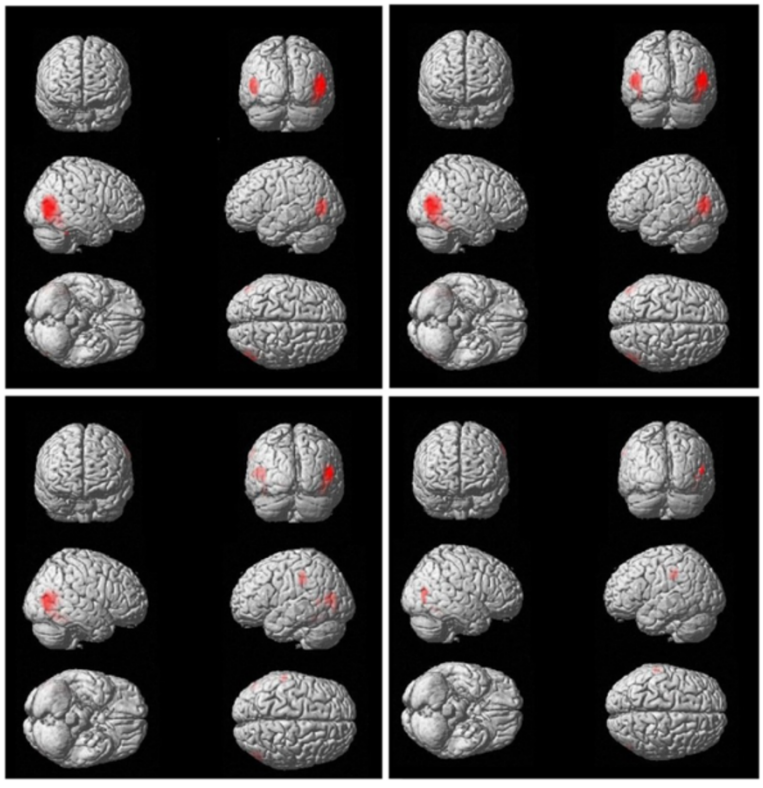

Figure 18 Result maps of mean searchlight. Radii = 3 (top left), 2 (top right), 1 (bottom left) and 0 (bottom right). When the radius is set to the maximum value of 3, the regions were strongly concentrated in the right middle temporal gyrus, right cerebellum and left middle occipital gyrus. When it is 2, relatively, a wide range of brain areas were sensitive; particularly activated were the right middle temporal gyrus, right inferior temporal gyrus, left middle occipital gyrus, left inferior temporal gyrus as well as left inferior occipital gyrus. With the radius narrowed down to 1, informative voxels were scattered in the right middle temporal gyrus, right fusiform gyrus, left middle occipital gyrus, left middle temporal gyrus, left inferior occipital gyrus and left supramarginal gyrus. And finally, when removing the adjacent voxels from the searchlight scope (radius = 0), brain informativity was shown in the right middle temporal gyrus, left superior frontal gyrus, right fusiform gyrus, right inferior temporal gyrus and left supramarginal gyrus.

showed a large area of strong activation distributed in the left and right middle temporal gyri, left and right middle occipital gyri, etc. On the other hand, some peaks for the tool > mammal activations could be found out in the left inferior parietal lobe and left supramarginal gyrus but without forming significant clusters of contiguous voxels. The important regions extracted by this univariate contrastive analysis accorded well with topographic patterns which have been established to show animal and tool specificity (Pulvermüller 2001; Binder et al. 2009; Akama et al. 2012).

Feature selection by searchlight

We provide here a close-up view of the brain map information to examine the methods of *voxel retrieval* based on the searchlight. The searchlight with the different radius values (0,1,2,3) computed the voxelwise mean accuracies across the sessions, and using the method of searchlight (Jimura et al. 2012), the z-scores were screened out with the threshold of 3.08 corresponding to the p value of 0.001 under the hypothesis of normal distribution (Figure 18). The local sensitivity maps thus created with respect to each radius were reprocessed by using the *xjview* toolbox to produce rendered images. The important finding here is that in larger searchlights, in which cross-subject decoding models become less accurate, significantly higher accuracy regions were concentrated in fewer but larger blocks, and regions of interest for the semantic contrasts in GLM such as the left supramarginal gyrus disappeared.

Additional file

Additional file 1: Materials.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The HA and LMM ran a series of fMRI experiments at Tokyo Tech, performed analysis of these datasets, and wrote draft versions of this paper. HA and BM developed novel analysis techniques together, and BM produced original scripts to automate analysis. BM was involved in the organization and proofreading of the paper. BM and MP provided training and supervision in the design and analysis of the fMRI experiments. All authors read and approved the final manuscript.

Acknowledgements

We are grateful to Jorge Jovicich (Co-Director, Functional Neuroimaging Laboratory in the Centre for Mind/Brain Sciences (CiMEC), University of Trento) and his staff; to Noboru Hidano (Professor of Economics, Graduate School of Decision Science and Technology, Tokyo Institute of Technology, Leader of the fMRI working group at Tokyo Tech), Koji Jimura (Researcher, Precision and Intelligence Laboratory, Tokyo Institute of Technology) and Li Na (former graduate school student at Tokyo Institute of Technology); and to Kai-Min Chang and Tim Keller (Machine Learning and Psychology Departments, respectively, Carnegie Mellon University) for their support and feedback on this research. The work described here was funded in part by grant 1R01HD075328 (CRCNS, National Institutes of Health, USA) and Kaken-Kiban (C) 23500171 (JSPS, Japan).

Author details

[1]Graduate School of Decision Science and Technology, Tokyo Institute of Technology, W9-10, 2-12-1, O-okayama, Meguro-ku, Tokyo 152-8552, Japan. [2]Knowledge & Data Engineering, EEECS, Queen's University Belfast, BT9 5BN Belfast, Northern Ireland. [3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA. [4]Centre for Mind/Brain Sciences, University of Trento, Rovereto 38068, Italy. [5]School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK.

Received: 21 November 2013 Accepted: 19 March 2014
Published: 2 September 2014

References

Akama H, Murphy B, Li N, Shimizu Y, Poesio M (2012) Decoding semantics across fMRI sessions with different stimulus modalities: a practical MVPA study. Front Neuroinform 6:24, doi:10.3389/fninf.2012.00024
Binder JR, Desai RH, Graves WW, Conant LL (2009) Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb Cortex 19:2767–2796
Chee MWL, Soon CS, Ling Lee H (2003) Common and segregated neuronal networks for different languages revealed using functional magnetic resonance adaptation. J Cogn Neurosci 15:85–97
Cox DD, Savoy RL (2003) Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19:261–270
Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning (Vol.2, No.1). New York: Springer
Haxby JV, Gobbini MI, Maura L, Ishai FA, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430
Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, Hanke M, Ramadge PJ (2012) A common, high-dimensional model of the representational space in human ventral temporal cortex. Neuron 72(2):404–16. doi:10.1016/j.neuron.2011.08.026
Hernandez AE, Dapretto M, Mazziotta J, Bookheimer S (2001) Language switching and language representation in Spanish–English Bilinguals: an fMRI study. Neuroimage 14:510–520
Huth AG, Nishimoto S, Vu AT, Gallant JL (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76(6):1210–1224
Jimura K, Russell AP (2012) Analyses of regional-average activation and multivoxel pattern information tell complementary stories, Neuropsychologia 50:544–552.
Kim KHS, Relkin NR, Lee K, Hirsch J (1997) Distinct cortical areas associated with native and second languages. Nature 388:171–174
Kriegeskorte N, Bandettini P (2007) Combining the tools: activation- and information-based fMRI analysis. Neuroimage 38:666–668
Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker C (2009) Circular analysis in systems neuroscience: the dangers of double dipping. Nat Neurosci 12:535–540
Mitchell TM, Hutchinson R, Niculescu RS, Pereira F, Wang X, Just M, Newman S (2004) Learning to decode cognitive states from brain images. Mach Learn 57:145–175
Mitchell T, Shinkareva S, Carlson A, Chang K, Malave V, Mason R, Just M (2008) Predicting human brain activity associated with the meanings of nouns. Science 320:1191–1195
Murphy B, Baroni M, Poesio M (2009) EEG responds to conceptual stimuli and corpus semantics. Proc ACL/EMNLP 2009:619–627
Murphy B, Poesio M, Bovolo F, Bruzzone L, Dalponte M, Lakany H (2011) EEG decoding of semantic category reveals distributed representations for single concepts. Brain Lang 117(1):12–22
Paradis M (2003) Differential use of cerebral mechanisms in bilinguals. In: Banich MT, Mack M (eds) Mind, brain, and language: multidisciplinary perspectives., pp 351–370. London: Lawrence Erlbaum
Perani D, Abutalebi J (2005) Neural basis of first and second language processing. Curr Opin Neurobiol 15:202–206
Pereira F, Botvinick M, Detre G (2010) Learning semantic features for fMRI data from definitional text. In: Proceedings of first workshop on computational neurolinguistics. NAACL HLT: Los Angeles, pp 1–9
Pereira F, Detre G, Botvinick M (2011) Generating text from functional brain images. Front Hum Neurosci 5:72. doi:10.3389/fnhum.2011.00072
Pulvermüller F (2001) Brain reflections of words and their meaning. Trends Cogn Sci 5:517–524
Pulvermüller F (2005) Brain mechanisms linking language and action. Nat Rev Neurosci 6(7):576–582
Wang X, Hutchinson R, Mitchell TM (2003) Training fMRI classifiers to discriminate cognitive states across multiple participants. Adv Neural Inf Process Syst 16:709–716
Wang Y, Xue G, Chen C, Xue F, Dong Q (2007) Neural bases of asymmetric language switching in second-language learners: an ER-fMRI study. Neuroimage 35:862–870
Weekes BS, de Zubicaray G, McMahon K, Eastburn M, Bryant M, Wang D (2005) Orthographic effects on picture naming in Chinese: a 4 T fMRI study. Brain Lang 95:14–15

doi:10.1186/2196-0089-1-1
Cite this article as: Akama *et al.*: **Cross-participant modelling based on joint or disjoint feature selection: an fMRI conceptual decoding study.** *Applied Informatics* 2014 **1**:1.