CSLT TECHNICAL REPORT-20160007 [Thursday 8th December, 2016]

Distraction Detection Using Sparse Discriminative Analysis

Dong Wang

Correspondence: wang-dong99@mails.tsinghua.edu.cn
Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Car driving is a serious task and any distraction may lead to catastrophic consequence. It is therefore very desirable to detect drivers' mental distraction and produce alter message when necessary. Psychologists have found that human's mental status can be featured by some representative signals of human brain, and machine learning methods can be employed to discriminate a driver's mental status (normal or distracted) based on these signals. This paper contributes in two directions for distraction detection: (1) we found using sparse models can help discover the most significant signals and prevent from over-fitting; (2) we found some simple signal processing techniques can find more powerful features than those conventionally used by psychologists.

Keywords: sparse discriminative analysis; distraction detection

1 Introduction

Driver distraction is one of the leading causes of traffic crashes. The weight of driver distraction on accident statistics varies depending on the criteria used to attribute distraction [1, 2, 3, 4, 5]. A naturalistic driving study "100 car study" released by the National Highway Traffic Safety Administration (NHTSA) reported that failures of attention contributed to 78% of all crashes and 65% of all near-crashes [6]. Visual, auditory, biomechanical, and cognitive distraction are four major types of driver distraction. In comparison to the extensive body of research examining the attention principles that govern distraction by external stimuli (the first three categories), the topic of cognitive distraction has been relatively understudied. Cognitive distraction occurs when the driver's mind is not focusing attention on the driving task [7, 8, 9, 10]. Because the mental state of drivers is not observable, no simple measure can index cognitive distraction precisely. Currently, researchers use subjective report measures, driver physiological measures, driver physical measures, driving performance measures, and hybrid measures to assess driver distraction [11]. Various physiological and biological measures such as electroencephalography (EEG), electrooculogram (EOG), heart rate variability (HRV), functional magnetic resonance imaging (fMRI), functional near infrared imaging (fNIR), and galvanic skin response have been employed to detect cognitive state changes [12, 13, 14]. However, EEG is the only physiological signal that has been shown to accurately reflect subtle shifts in alertness and attention that can be identified and quantified on a second-by-second time-frame [15]. Compared to other physiological and biological indices, EEG is a more direct and accurate technique to indicate when a driver is thinking something unrelated to the driving tasks. Several

ratios of EEG power bands (e.g., alpha, beta, theta, etc.) have been found to correlate with cognitive distraction [14, 16, 17, 18, 19, 20]. For example, Cunningham, et al. [17] recorded the EEG activities from seven electrode sites and computed three different band ratios: beta/alpha, beta/theta, and beta/(alpha+theta) at each site. The authors observed significant increases from pre- to post- TUTs in two power band ratios, beta/(alpha+theta) and beta/alpha from all parietal lobe sites (Pz, P3, and P4). Although EEG has been found to correlate with cognitive distraction in a variety of tasks (e.g., vigilance, memory, reading comprehension, auditory oddball, and signal detection), none of the existing researches involves a driving task. Many technologies have been adopted to mitigate the effects of distraction. One promising strategy involves developing algorithms/models to differentiate the driver's distracted state from normal driving conditions, and then using them to adapt the in-vehicle technologies to detect driver distraction in real time. Machine learning technology provides several algorithms of searching large volumes of data for unknown patterns. It has been successfully applied to capture the differences in driving behaviors and ocular activities when people drive normally and when they are distracted. Different training methods, model characteristics, and feature selection criteria were presented and compared [21, 22, 23, 24, 25]. For example, Liang and her colleagues applied support vector machines (SVMs) [25] and Bayesian networks [24] to develop a real-time approach for detecting cognitive distraction using drivers' eye movements and driving performance data. In these studies, data for training the models were collected using a static driving simulator, with real human subjects performing a specific secondary task while driving. The objective of the present study was to apply machine learning techniques to develop a real-time approach for the detection of cognitive distraction, using drivers' EEG data.

2 Preliminary

2.1 Challenge with distraction detection

Distraction detection can be cast to a problem of mental status classification, where normal and distracted status are treated as two status that need to be classified frame by frame. This classification task, as shown in [24], can be simply conducted by an off-the-shelf machine learning tool. However in practice, it is not such easy. There are at least two problems preventing from a blind application of machine learning tools: firstly, the training data for each subject is often very limited, so there is a serious over-fitting problem; second, the signals are from multiple channels, which means, on one hand, the information received is redundant; and on the other hand, many task-irrelevant signals are received. How to extract the most representative information that is more related to mental status and that is robust against noise (sensory noise, other psychological factors, etc.) is highly important. Traditional approaches use psychological knowledge to extract informative features [17]; however, this psychologically-derived features can only solve the one-channel problem; for multiple channels, the channel redundancy still exists. In this paper, we propose to select features (from multi-channel input) by sparse models, which automatically discovers the most related features by looking at their discriminative power on the task in hand. Moreover, we assume psychological-driven features may have lost some important information so are probably suboptimal.

We study more general features derived from spectral processing, hoping that will involve more information that would be discovered and utilized by the sparse models. Our experiments demonstrated that the general features plus feature selection based on sparse models work much better than psychological features plus SVM, an approach that has been demonstrated successful in previous studies [25].

2.2 Sparse models

Imposing an l_1 or *lasso* penalty to achieve sparsity on features has been extensively studied in both regression [26, 27] and classification [28, 29, 30, 31]. By adding an l_1 regularization term to the original cost function, the coefficients of less important feature dimensions are effectively driven to zeros, leading to a natural and efficient feature selection approach which can be used for discovering the most important channels and dimensions in the data of mental distraction. A remarkable advantage of this sparsity-based approach is that the promising features are selected simultaneously as an entire group.

We investigate two sparse models in this paper: one is based on the simple linear discriminative model while the other is based on the SVM model. The former is simple and efficient, but the latter is more consistent with the detection component, if the classifier used in the detection is an SVM.

2.2.1 Sparse linear discriminative analysis (SDA)

Following the formulation of [28], let $X \in R^{N\times P}$ be a data matrix where N is the number of observations and P is the dimension of the feature vector; further let $Y \in \{0,1\}^{N\times K}$ be the class variables in which Y_{nk} is an indicator variable for which the n-th observation belongs to the k-th class. The optimal scoring criterion for LDA involves recasting the classification problem as a regression problem by turning the categorical target (class label) to a continuous target by multiplying a score vector θ_k. The objective function takes the following form [28]:

$$min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2\} \quad s.t. \quad \frac{1}{N}\theta_k^T Y^T Y\theta_k = 1, \;\; \theta_k^T Y^T Y\theta_l = 0 \;\; \forall l < k,$$

where θ_k is the K-dimensional score vector, and β_k is a P-dimensional vector of variable coefficients. Note that this is a sequential optimization problem where the 'discriminative directions' $\{\beta_k\}$ are attained one by one. To enforce sparsity in the discriminative directions, [28] appended an l_2 term and an l_1 term to the cost function, given by:

$$\begin{aligned} &min_{\beta_k,\theta_k}\{||Y\theta_k - X\beta_k||_2^2 + \gamma\beta_k^T\Omega\beta_k + \lambda||\beta_k||_1\} \\ &\quad s.t. \quad \frac{1}{N}\theta_k^T Y^T Y\theta_k = 1, \;\; \theta_k^T Y^T Y\theta_l = 0 \;\; \forall l < k, \end{aligned} \tag{1}$$

where Ω is a positive definite matrix to avoid singularity when the observations are mutually dependent or when the dimension is large, i.e., $P > N$, and λ and γ are non-negative hyperparameters. Note that the l_1 penalty introduced by the third

term in the above equation enforces sparsity on β_k, and more dimensions of β_k are driven to zeros with a larger λ [26].

In the case of a two-class classification problem, there is only one discriminative direction β. The optimization problem is then simplified as follows:

$$\begin{aligned} &min_{\beta,\theta}\{||Y\theta - X\beta||_2^2 + \gamma\beta^T\Omega\beta + \lambda||\beta||_1\} \\ &\quad s.t. \quad \frac{1}{N}\theta^T Y^T Y\theta = 1. \end{aligned} \tag{2}$$

Eliminating θ by a simple calculation leads to:

$$min_{\beta}\{||\hat{Y} - X\beta||_2^2 + \gamma\beta^T\Omega\beta + \lambda||\beta||_1\}, \tag{3}$$

where $\hat{Y}$ is the normalized class indicator matrix whose elements are given by:

$$\hat{Y}_{n,k} = \sqrt{\frac{N}{N_k}},$$

where N_k is the number of observations of the k-th class. We see that the optimization problem for the classification task equals to the optimization problem of a regression task in the case of two classes, which has been stated in [32]. Further, notice that Eq. (3) is an elastic net problem if $\Omega = I$, and a generalized elastic net problem for an arbitrary symmetric positive definite matrix Ω. This elastic net problem can be solved by the algorithm proposed by [27].

Once the optimal β is obtained, for a new observation $x \in R^P$, a simple classification can be conducted by setting a threshold on $\beta^T x$. In this work, however, we treat the SDA as a keyword selector instead of a classifier. First, notice that β is sparse, which indicates that only a fraction of the dimensions of X contributes to the decision. We therefore select the features (words) whose corresponding coefficients in β are not zero as keywords; these keywords are then used to build a new low-dimensional text feature, based on which an SVM (non-linear in this work) is constructed and is used as the classifier for distraction detection.

We finally note that it is only for a binary classification task that the SDA model coincides with the elastic net regression proposed by [27]. For multiple classification tasks, the SDA model is a general framework to derive sparse coefficients $\{\beta_k\}$. In this case, the non-zero dimensions of different β_k are usually different, so the words corresponding to all these non-zero dimensions of all the coefficients $\{\beta_k\}$ have to be selected as keywords.

2.2.2 Sparse SVM

A shortcoming of the SDA-based approach resides in the discrepancy between the objective functions used in the feature selection and the mental status classification: the former is based on the minimum square error, and the latter is based on the maximum margin, if SVM is used. A better approach is to use the same objective

function/model to classify mental status. The sparse SVM model is a good candidate because it is a sparse version of the SVM and both are based on the maximum margin.

We follow the formulation in [32]. First, we define x_n as a training sample and $t_n \in \{+1, -1\}$ as its label. The linear SVM holds a classification boundary $w^T x + b = 0$ where w and b are model parameters, and it predicts the target for x_n (i.e., the category assignment y_n) as follows:

$$y_n = w^T x_n + b. \tag{4}$$

The model training involves optimizing the following regularized hinge function with respect to w and b:

$$C \sum_{n=1}^{N} \xi_n + \frac{1}{2}||w||_2^2 \quad s.t. \quad t_n y_n > 1 - \xi_n, \tag{5}$$

where N is the number of training samples, and ξ_n is a slack variable that represents the cost term of x_n: $\xi_n = 0$ if x_n is inside or on the correct margin boundary, otherwise $\xi_n = |t_n - y_n|$. In addition, $||w||_2^2$ is the regularization term, and C is a tunable hyperparameter to trade off the cost and regularization. From the constraint of Eq. (5), one can show that the distance from the margin to the decision boundary remains to be 1, and so any data x_n is misclassified if $\xi_n > 1$.

As pointed by [30], the l_2 norm $||w||_2^2$ leads to a dense vector of the optimal w. In order to obtain a sparse w, an l_1 norm can be used to substitute for or append to the l_2 norm, leading to the following cost function:

$$\sum_{n=1}^{N} \xi_n + \gamma||w||_2^2 + \lambda||w||_1, \quad s.t. \quad t_n y_n > 1 - \xi_n, \tag{6}$$

where γ and λ are two model hyperparameters for trading off the hinge cost and the regularization. A larger λ drives more dimensions of w to zeros, which in turn vanishes contributions of more features when conducting model inference, according to Eq. (4). Therefore, a sparse SVM leads to a natural way for feature selection. As in SDA, the words corresponding to the non-zero coefficients in w are selected as significant dimensions, and the selected dimensions comprise the low dimensional features to build a non-linear SVM model for mental status classification.

In this work, we employ the template first-order conic solver (TFOCS) to optimize the sparse SVM. TFOCS is a general framework for solving a variety of convex cone problems, including the problem of Eq. (6) [33].

We notice that using a linear sparse SVM to conduct feature selection has been studied in some publications. For example, [30] proposed a quite similar approach to ours, where a linear sparse SVM is used to choose significant dimensions and a non-linear SVM conducts classification. The difference is that [30] worked on

v-Support Vector Regression (SVR) and did not involve the l_2 term in Eq. (6). [31] provided another form of sparse SVM, where the maximum number of non-zero dimensions was treated as a constraint, and a convex relaxation approach was employed to optimize the model. To the authors' best knowledge, this paper is the first application of the sparse SVM model to distraction detection (mental status classification).

Comparing the SDA-based and sparse SVM-based feature selection (Eq. (3) and Eq. (6)), we notice that both are based on sparse constraints in the form of an elastic net regularization. The only difference resides in the objective function when optimizing the model coefficients β (in SDA) or w (in sparse SVM): the former is the regularized square error while the latter is the regularized hinge cost. Since the classifier in the detection component is an SVM in this work, the sparse SVM-based approach tends to be more consistent to the classification component.

2.3 Mel Fbank extraction

Feature extraction from time series such as EEG data involves various signal processing steps. The most popular processing scheme is perceptual-oriented filtering. We borrow the Mel filter bank (Fbank) pipeline from speech processing research to extract features for distraction detection. The main steps of Fbank extraction is presented in Fig. 1. The signal is first split into equal-length segments, or frames, and then passes a time-domain window to mitigate the boundary effect. Each windowed frame is then converted to the frequency domain by fast Fourier transform (FFT), and then a set of Mel filter banks are applied in the frequency domain to extract the energy of each frequency band. These energies of subbands are then compressed by logarithm, resulting in the Fbank features. Note that a property of Fbank features is that the frequency resolution is higher in the low frequency area than in the high frequency area. In speech processing, this matches the character of the human auditory system as the former is more sensitive in low frequency. This character is also attractive for processing EGG signals: it has been demonstrated that information related to mental status is mostly within the low-frequency components of EEG signals [14, 16, 17, 18, 19, 20].

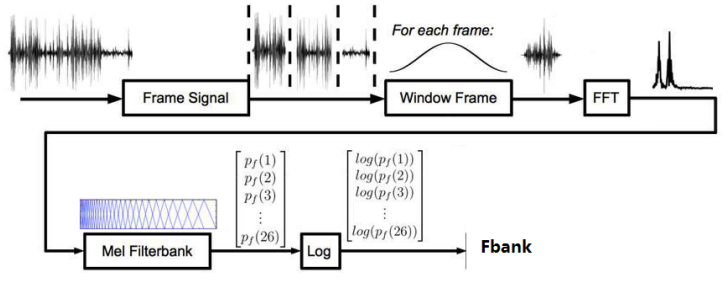

Figure 1 Processing steps for Mel Fbanks

3 Methods

3.1 Participants

Twelve healthy participants (6 males, 6 females) took part in the laboratory session, which involved a driving simulator. Their average age was 32.3 years (range from 25 to 39, SD=4.52). The average number of years since they obtained their first driver licenses was 7.12 (SD=2.55), and the average estimated annual mileage was 16152.63 kilometers (SD=6638.14). All participants had normal or corrected-to-normal vision, a valid driver's license, and reported being free of psychiatric or neurological disorders. 2.2. Materials A STISIM® driving simulator (STISIM-DRIVE M100K) was used in the experimental study. The STISIM simulator was installed on a Dell Workstation (Precision 490, Dual Core Intel Xeon Processor 5130 2GHz) with a 256MB PCIe×16 nVidia graphic card, Sound Blaster® X-Fi^{TM} system, and Dell A225 Stereo System. It includes a Logitech Momo® steering wheel with force feedback, a gas and a brake pedal. The driving scenario was presented on a 27-inch LCD with 1920×1200 pixels resolution. A Neuroscan system including one Quik-Cap, Nuamps Express, and SCAN software, was utilized to record EEG activity and perform the EEG frequency analysis. Nuamps Express is a 40-channel digital EEG recording system, and SCAN provides a full research-grade-data processing tool to remove noise and artifacts or decompose complex signals. Electrodes were also placed on each earlobe for use as reference points. In addition to the driving simulator and EEG recording system, one more system (the stimuli generation computer) was used to synchronize driving behavioral and EEG data. The initialization of each driving task and the timing of each keystroke were recorded by the driving simulator, and read by the stimuli generation computer. These timings triggered signals that were coded accordingly and sent to the EEG recording computer as markers simultaneously through LabJack® interface (LabJack Corporation, Colorado, USA). These markers were used as bases of extracting frequency features from the continuous EEG data and behavioral indices from continuous driving signals.

3.2 Experimental task

3.2.1 Driving task

Participants drove along a straight suburban street with one lane in each direction. The subject vehicle (SV; vehicle driven by the participants) was equipped with a simulated cruise control system that engaged automatically at 70 km/h and disengaged when drivers pressed the brake pedal. The participants were instructed to follow the vehicle in front of them (lead vehicle, LV) and to use the cruise control as much as possible. The participants performed three driving tasks during each of the six drives. The first task was to follow the LV and respond to six LV braking events during each drive. The timing of each braking event was determined by the status of the IVIS task. During the events, the LV braked at a rate of 0.2 g until it reached a minimum speed of no more than 30 km/h and the participant had braked at least once. Following a brief, random delay (0 to 5 s), the LV accelerated at a rate of 0.25 g until it reached a speed of 40 km/h. The second task was to keep the SV from drifting toward the lane boundaries and to drive in the center of the lane as much as possible. The final task was to detect the appearance of a pedestrian on

the right side of the road in the driving scene by pressing a button on the steering wheel. The pedestrian appeared about three times per minute and was visible, on average, for approximately 2.8 s.

3.2.2 Auditory IVIS Task

During four of the six drives, participants interacted with the IVIS: an auditory stock ticker. The auditory, purely cognitively loading, task used in HASTE, was based on the visual Continuous Memory Task, described in Veltman and Gaillard (1998). The task (henceforth referred to as the ACMT—Auditory Continuous Memory Task) was to keep an updated count of two target sounds, presented in sequence among two non-target sounds. Each target sound was counted separately, so the subject had to keep track of two counts in parallel. The current sum was read out aloud after each drive. The experimenter annotated their responses manually. The subjects were trained before the experimental trials to distinguish between the target and the non-target sounds.

3.3 Procedure

Upon arrival, participants completed an informed consent form, a questionnaire inquiring their demographic information and driving experience along with a vision test. Drivers with at least four years of driving experience and normal or corrected-to-normal vision ability were allowed to participate in this experiment. Participants were seated in a comfortable chair and wore the filtered Quik-Cap sensor cap. The reference electrodes were placed on the left and right mastoids and the ground electrode was placed mid-forehead. The horizontal and vertical EOGs were recorded with electrodes placed 10 mm away from the outer canthi of both eyes and below and above the left eye. Electrical impedances at each electrode site were reduced to less than 5 kOhms. After the cap was set up, the participants were instructed to sat quietly and close their eyes. EEG and EOG signals were sampled at 1000 Hz and continuously recorded for up to 15 minutes (i.e., baseline condition). After the baseline recording, participants were provided with a brief description of the experimental task. Participants then went through the 15-minute practice drive session which allowed them to become familiar with the driving simulator controls including steering wheel, speedometer, throttle, and brake pedal. They were free to ask questions during the practice drive session. The formal experimental session began after participants fully understood the task and felt comfortable operating the driving simulator. It consisted of six test blocks and each lasted for 15-20 minutes. Participants were given a 5-minute break between two test blocks. Continuous EEG, EOG, and driving signals were recorded and synchronized. EEG and EOG data and sampled at 1000 Hz and driving signals were sampled at 10 Hz. The whole experiment was completed within 2 hours. All participants were paid at a rate of $20.00 per hour.

4 Results

We cast the distraction detection task to a mental status classification task, i.e., train a classifier that discriminates normal and distraction status. The performance is evaluated in terms of frame classification error rate (FER). The database involves

EEG data from 8 subjects, and each subject recorded 6 session, where the first four sessions are positive (distracted status), and the rest two sessions are negative (normal status).

4.1 Psychology derived features

4.1.1 SVM approach on all channels

The first experiment employs psychology derived features. The most representative 3 features are derived from each channel, and there are 34 channels in total. The features are derived from the alpha, beta and theta bands which have been shown related to mental status. The sampling rate is 10 Hz. Positive samples (distraction) and negative samples (normal) are balanced by random sampling some negative data. The SVM model is selected as the classifier, for which the LibSVM package is used for model training and inference[1].

Two experiments are conducted: the first 'Single-subject' experiment trains a single SVM for each subject[2], and the second 'Multi-subject' experiment trains a single SVM with the data from all the subjects, but test is conducted on each subject[3].

The results are shown in Table 1. It can be seen that with either approach, the performance on the training set is almost 100% correct, while the performance on the test set is pretty low. This suggests severe over-fitting. It is understandable as the training data is very limited (about 900 positive samples and 300 negative sample per subject), and the dimensionality is pretty high (102). The over-fitting problem exists with linear models as well.

Table 1 Results of SVM, based on psychological features of all channels.

Training Approach	Training subject	Test subject	Training FER	Test FER
Multi	1-8	1	-	48.55
Multi	1-8	2	-	49.00
Multi	1-8	3	-	49.47
Multi	1-8	4	-	49.70
Multi	1-8	5	-	49.08
Multi	1-8	6	-	50.30
Multi	1-8	7	-	49.70
Multi	1-8	8	-	48.93
Multi	1-8	1-8	0.02	48.93
Single	1	1	0.00	48.54
Single	2	2	0.00	48.85
Single	3	3	0.00	38.16
Single	4	4	0.00	49.70
Single	5	5	0.00	41.54
Single	6	6	0.00	49.40
Single	7	7	0.00	49.70
Single	8	8	0.00	47.43

4.1.2 SVM approach on selected channels

To alleviate the sever over-fitting problem, the feature dimensionality needs to be reduced. From previous studies, it has been shown that some channels are more related to mental status change. By this knowledge, we choose the most prominent 5 channels, leading to 15 features.

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvm/
[2] /nfs/disk/work/wangd/research/psychology/test/t2.2
[3] /nfs/disk/work/wangd/research/psychology/test/t2.3

Again, a 'Single-subject' experiment that trains a single SVM for each subject[4], and a 'Multi-subject' experiment that trains a single SVM for all subjects[5] are conducted.

The results are shown in Table 2. It can be seen that the over-fitting is alleviated a little bit, but the performance is still unacceptable. This indicates that the psychology-based feature dimension is not ideal.

Table 2 Results of SVM, based on psychological features of selected channels.

Training Approach	Training subject	Test subject	Training FER	Test FER
Multi	1-8	1	-	42.42
Multi	1-8	2	-	51.77
Multi	1-8	3	-	46.46
Multi	1-8	4	-	47.45
Multi	1-8	5	-	48.92
Multi	1-8	6	-	53.45
Multi	1-8	7	-	54.29
Multi	1-8	8	-	48.64
Multi	1-8	1-8	9.70	49.29
Single	1	1	0.00	43.80
Single	2	2	0.15	49.92
Single	3	3	0.40	47.36
Single	4	4	0.00	45.35
Single	5	5	0.05	49.23
Single	6	6	6.90	57.06
Single	7	7	0.05	50.30
Single	8	8	0.15	49.24

4.1.3 SDA approach: Multi-subject model

In this experiment, we use the same psychologically derived data (all channels), but employ SDA to select the most representative features (no-zero dimensions). Once the features are selected, a simple linear model (logistic regression) is applied to conduct the classification. The data from all the 8 subjects are used to train the SDA and the classifier, and then test on each subject as well as the entire test data. The results on the entire training set and test set are reported in Fig. 2 and Fig. 3 respectively [6], where the sparsity of SDA is set in different values so that the dimensionality of the selected features changes from 1 to 102. It can be seen that on the training set, more dimensions lead to better performance, while on the test set, there is an optimal dimensionality that leads to the best test performance. This confirms the over-fitting problem, and indicates that SDA can help select most prominent features so that the over-fitting problem can be alleviated. We can see that the best performance with SDA is much better than with the simple SVM.

The results on each subject are presented in Fig. 4, where each subject is represented by a curve. We observe that there is a large variance among subjects: some subjects can obtain very good performance, while others exhibit rather poor.

4.1.4 SDA approach: Single-subject model

Motivated by the great variance among subjects, we train a specific SDA for each subject in this experiment. The results on the training sets and test sets are presented in Fig 5 and Fig. 6 respectively.[7]

[4] /nfs/disk/work/wangd/research/psychology/test/t2.2
[5] /nfs/disk/work/wangd/research/psychology/test/t2.3
[6] /nfs/disk/work/wangd/research/psychology/test/t4.0/model
[7] /nfs/disk/work/wangd/research/psychology/test/t4.0/model.single-subject

From the results on the training data, it can be observed that the single models can learn each subject very precisely, therefore obtains rather good performance for each subject, compared to the multi-subject model shown in Fig. 2. The results on test sets do not shown much advantage compared to those obtained with the multi-subject model shown in Fig. 4; however, very bad subjects as Fig. 4 disappear. This again suggests that subject variability is an important factor, and single-subject modeling is necessary. However, the subject-specific model suffers from more data sparsity, leading to more serious over-fitting. This is why the highest performance obtained by the single-subject models is even worse than the one obtained with the multi-subject model.

4.2 Fbank features

The psychological features involve human-discovered information, however it is also possible that the man-made features overlook some important information that helps discriminating mental status. This experiment employs Fbanks for norm-distraction classification. Fbanks are developed by signal processing society and more concerns with auditory perception. We investigate whether the general features can deliver reasonable performance with EEG data.

The window length is empirically set to 512, FFT length to 256, the number of filer banks to 25. The original 36-channel EEG data are used, for which the sampling rate is 1000 Hz. We experiment with two scenarios: in the first scenario, models are constructed for each channel (single-channel model), and in the second scenario, data from multiple channels are used (multi-channel model).

4.2.1 Single-channel SDA: Multi-subjects model

We first experiment with the scenario of single-channel (i.e., each channel is modelled independently) and multi-subject (i.e., training data are from all subjects) models. The performance is evaluate the entire training and test data. The dimensionality of the selected features by SDA is changed from 1 to 25 (the number of Fbanks), and the results are shown on each channel[8].

The results are shown in Fig. 7, where each curve represents a single channel, and the x-axis shows the feature dimensionality from 1 to 25. We first that observe different channels perform very differently. This is expected, as the sensors placed on different positions on the head tend to receive very different signals. On the other hand, it can be seen that sparsity impacts the performance significantly. When only a small number of features are available, performance on different channels are much more different; with more features selected, the variance among channels tends to be small. This is perhaps because the training data involves multiple subjects, by which the variety on subjects may alleviate the variety on channels.

For a more clear comparison on different channels, we choose the best sparsity on each channel, and presents the corresponding performance in Fig. 8. It can be observed that channel 2 and 25 are the most discriminative. The best result is 25.39%, which is better than the best result (38.68%) obtained with the psychological features as shown in Fig 3. This is a highly promising result and indicates that psychological knowledge is useful, but maybe not the best. Raw features like

[8] /nfs/disk/work/wangd/research/psychology/test/t4.1/model

Fbanks derived from signal processing might be sufficiently good to deliver even better results than man-made psychological feaures.

4.2.2 Single-channel SDA: Single-subject model

In the next experiment, we build SDA models for each subject and each channel. This is motivated by the observation in the previous sections that there is great variety among subjects. We present the results on each subject[9]. The results are shown in Fig. 9 to Fig. 16. Each figure presents each subject, and in each picture, each curve represents a single channel.

From these results, we can observe that for most subjects, the performance on the best channel with the best sparsity can reach very high, almost close to zero. This on the one hand indicates that the Fbank feature is very powerful in discriminating mental status, with appropriate sparse constraints applied. On the other hand, it also shows great variability among channels and subjects. The best channels are significantly different from one subject to another, which implies that the spatial patterns on the EEG data for mental distraction are very complex and highly depends on individuals. It is difficult to design a unified model to achieve good detection for all people, and we must train subject-dependent models.

To further analyze the variety of subjects, the best channel and the corresponding PER of each subject are presented in Table 3. It can be seen that for every subject, the best channel is very unique, and the best performance that can be achieved is also quite different.

Table 3 The best channel and FER on test data on each subject. The model is single subject and single channel.

Subject	Best Channel	Test FER
1	11	2.46
2	23	1.25
3	17	0.00
4	33	17.08
5	28	23.36
6	16	0.00
7	2	0.00
8	26	0.00

4.2.3 Multi-channel SDA: Single-subject model

In this experiment, we try to use data from multiple channels. Again, Fbanks of 25 dimensions are derived from each channel, and then the Fbanks from all the 36 channels are concatenated together. The SDA model is trained for each subject, and a logistic regression model is trained as the classifier[10]. We test several sparsity settings (the number of non-zero dimensions in SDA): 10, 20, 30, 40, 50. The results for the 8 subjects on training and testing data are shown in Table 4. It can be seen that for all the subjects, the performance on the training sets are almost 100%, while on test sets, the performance is pretty bad for most subjects. This suggests severe over-fitting. More robust approaches such as group sparsity are under investigation.

[9] /nfs/disk/work/wangd/research/psychology/test/t4.1/model.single
[10] /nfs/disk/work/wangd/research/psychology/test/t4.2/model.t2

Table 4 PER on each subject with SDA based on Fanks. The model is single subject and multiple channel.

Subject	FER%				
1	9.90	0	0	0	0
2	0	0	0	0	0
3	0	0	0	0	0
4	1.04	0.04	0	0	
5	0	0	0	0	0
6	0	0	0	0	0
7	0	0	0	0	0
8	0	0	0	0	0
1	100	56.95	55.24	55.47	50.09
2	12.38	26.05	27.27	27.47	27.46
3	14.40	7.59	7.61	9.71	7.44
4	63.18	51.61	35.84	31.60	32.47
5	49.70	49.83	49.80	49.80	49.75
6	48.28	46.34	29.38	29.59	29.06
7	0	0	0	0	0
8	41.99	87.58	84.35	73.45	70.00

5 Conclusions

We investigated the problem of distraction detection in car driving, using EEG data. Our experiments showed that simply applying regular machine learning methods such as SVM suffers from very severe over-fitting problem and can not be practically used. We therefore employ SDA to select the most promising features in a group fashion. Experiments demonstrated that the SDA-based feature selection is highly effective. Additionally, we found general Fbanks features can achieve much better performance compared to the psychologically derived features, demonstrating the capability of signal processing methods in psychological analysis. Finally, we observed a large variety among subjects: the most discriminative channels are different for different subjects, and the optimal sparsity levels are also different. This suggests that in order to get an effective distraction detection, the system must be carefully tuned to suite each individual.

Acknowledgement

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011.

References

1. M. A. Recarte and L. M. Nunes, *Driver Distractions*, CRC Press, 2008.
2. Jane C Stutts, Donald W Reinfurt, Loren Staplin, and Eric A Rodgman, "The role of driver distraction in traffic crashes," 2001.
3. Jonathan Mosedale, Andrew Purdy, and Eddie Clarkson, "Contributory factors to road accidents," *London: Department of Transportation*, 2004.
4. Kristie Young, Michael Regan, and M Hammer, "Driver distraction: A review of the literature," *Distracted driving*, pp. 379–405, 2007.
5. Lawrence T Lam, "Distractions and the risk of car crash injury: The effect of drivers' age," *Journal of Safety Research*, vol. 33, no. 3, pp. 411–419, 2002.
6. Sheila G Klauer, Thomas A Dingus, Vicki L Neale, Jeremy D Sudweeks, David J Ramsey, et al., "The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data," 2006.
7. Ralph H Craft and Brian Preslopsky, "Driver distraction and inattention in the usa large truck and national motor vehicle crash causation studies," in *1st International Conference on Driver Distraction and Inattention (DDI 2009)*, 2009.
8. Kristie Young, John D Lee, and Michael A Regan, *Driver distraction: Theory, effects, and mitigation*, CRC Press, 2008.
9. Michael A Regan, Charlene Hallett, and Craig P Gordon, "Driver distraction and driver inattention: Definition, relationship and taxonomy," *Accident Analysis & Prevention*, vol. 43, no. 5, pp. 1771–1781, 2011.
10. Samuel G Charlton and Nicola J Starkey, "Driving without awareness: The effects of practice and automaticity on attention and driving," *Transportation research part F: traffic psychology and behaviour*, vol. 14, no. 6, pp. 456–471, 2011.
11. Yanchao Dong, Zhencheng Hu, Keiichi Uchimura, and Nobuki Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE transactions on intelligent transportation systems*, vol. 12, no. 2, pp. 596–614, 2011.
12. Michal Gruberger, Eti Ben Simon, Yechiel Levkovitz, Abraham Zangen, and Talma Hendler, "Towards a neuroscience of mind-wandering," *Frontiers in human neuroscience*, vol. 5, pp. 56, 2011.
13. Jibo He, Ensar Becic, Yi-Ching Lee, and Jason S McCarley, "Mind wandering behind the wheel performance and oculomotor correlates," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 53, no. 1, pp. 13–21, 2011.
14. Jungang Qin, Christopher Perdoni, and Bin He, "Dissociation of subjectively reported and behaviorally indexed mind wandering by eeg rhythmic activity," *PloS one*, vol. 6, no. 9, pp. e23124, 2011.
15. Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven, "Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, space, and environmental medicine*, vol. 78, no. Supplement 1, pp. B231–B244, 2007.
16. Claire Braboszcz and Arnaud Delorme, "Lost in thoughts: neural markers of low alertness during mind wandering," *Neuroimage*, vol. 54, no. 4, pp. 3040–3047, 2011.
17. Stephen Cunningham, Mark W Scerbo, and Frederick G Freeman, "The electrocortical correlates of daydreaming during vigilance tasks.," *Journal of Mental Imagery*, 2000.
18. Alan T Pope, Edward H Bogart, and Debbie S Bartolome, "Biocybernetic system evaluates indices of operator engagement in automated task," *Biological psychology*, vol. 40, no. 1, pp. 187–195, 1995.
19. T Jane Pritzl, "The effect of experimentally-enhanced daydreaming on an electroencephalographic measure of sleepiness," 2003.
20. Chin-Teng Lin, Shi-An Chen, Li-Wei Ko, and Yu-Kai Wang, "Eeg-based brain dynamics of driving distraction," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*. IEEE, 2011, pp. 1497–1500.
21. Tulga Ersal, Helen JA Fuller, Omer Tsimhoni, Jeffrey L Stein, and Hosam K Fathy, "Model-based analysis and classification of driver distraction under secondary tasks," *IEEE transactions on intelligent transportation systems*, vol. 11, no. 3, pp. 692–701, 2010.
22. Martin Wollmer, Christoph Blaschke, Thomas Schindl, Björn Schuller, Berthold Farber, Stefan Mayer, and Benjamin Trefflich, "Online driver distraction detection using long short-term memory," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 574–582, 2011.
23. Fabio Tango and Marco Botta, "Real-time detection system of driver distraction using machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 2, pp. 894–905, 2013.
24. Yulan Liang, John Lee, and Michelle Reyes, "Nonintrusive detection of driver cognitive distraction in real time using bayesian networks," *Transportation Research Record: Journal of the Transportation Research Board*, , no. 2018, pp. 1–8, 2007.
25. Yulan Liang, Michelle L Reyes, and John D Lee, "Real-time detection of driver cognitive distraction using support vector machines," *IEEE transactions on intelligent transportation systems*, vol. 8, no. 2, pp. 340–350, 2007.
26. Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B*, pp. 267–288, 1996.
27. Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

28. Line Clemmensen, Trevor Hastie, Daniela Witten, and Bjarne Ersbøll, "Sparse discriminant analysis," *Technometrics*, vol. 53, no. 4, pp. 406–413, 2011.
29. Daniela M Witten and Robert Tibshirani, "Penalized classification using fisher's linear discriminant," *Journal of the Royal Statistical Society: Series B*, vol. 73, no. 5, pp. 753–772, 2011.
30. Jinbo Bi, Kristin P. Bennett, Mark Embrechts, Curt Breneman, and Minghu Song, "Dimensionality reduction via sparse support vector machines," *The Journal of Machine Learning Research*, vol. 3, pp. 1229–1243, 2003.
31. Mingkui Tan, Li Wang, and Ivor W Tsang, "Learning sparse SVM for feature selection on very high dimensional datasets," in *Proceedings of the International Conference on Machine Learning*, 2010, pp. 1047–1054.
32. Christopher M Bishop, *Pattern recognition and machine learning*, vol. 1, Springer, New York, 2006.
33. Stephen R Becker, Emmanuel J Candès, and Michael C Grant, "Templates for convex cone problems with applications to sparse signal recovery," *Mathematical Programming Computation*, vol. 3, no. 3, pp. 165–218, 2011.

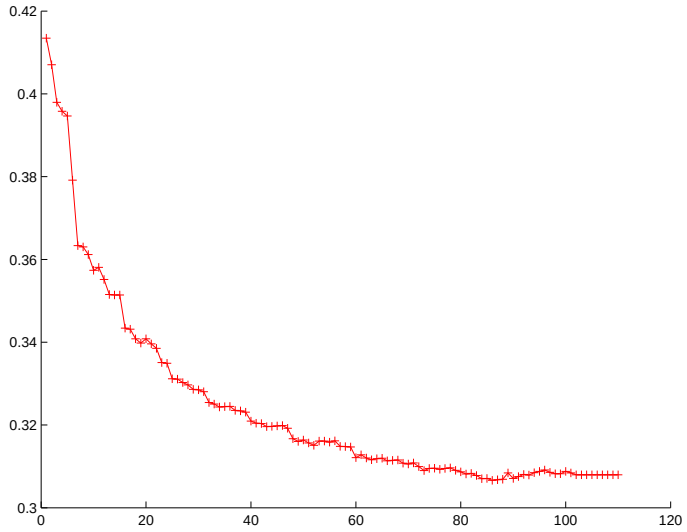

Figure 2 PER on the entire train set, with SDA based on psychological features of all channels. Models are trained on data of all subjects.

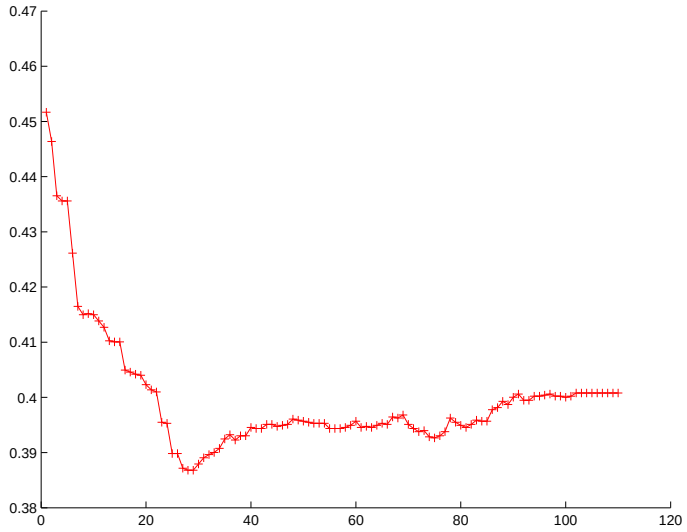

Figure 3 PER on the entire test set, with SDA based on psychological features of all channels. Models are trained on data of all subjects.

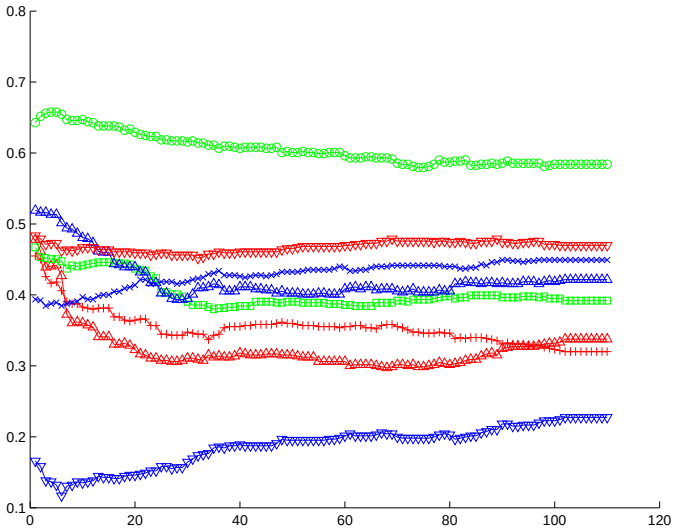

Figure 4 PER on each subject, with SDA based on psychological features of all channels. Models are trained on data of all subjects.

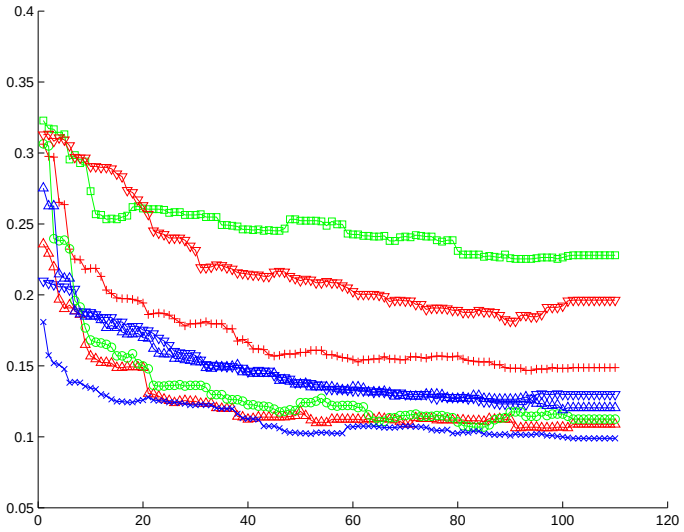

Figure 5 PER on training data of each subject, with SDA based on psychological features of all channels. Models are trained on data of each subject.

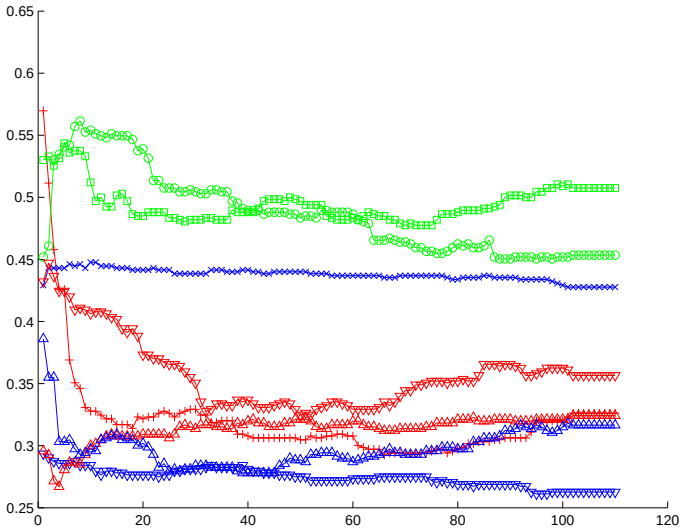

Figure 6 PER on test data of each subject, with SDA based on psychological features of all channels. Models are trained on data of each subject.

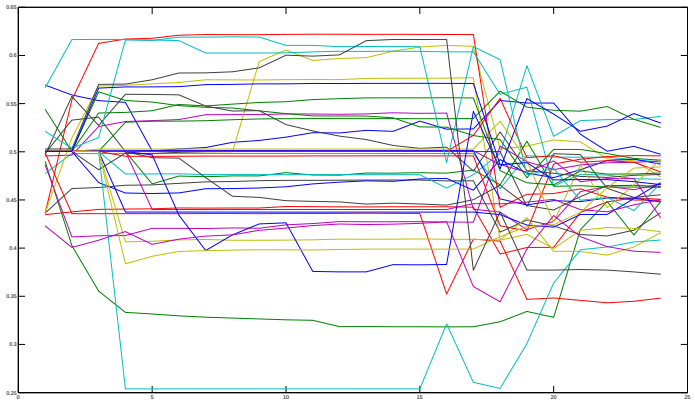

Figure 7 PER on entire test data set, with SDA based. The model is single-channel and multi-subject. Each curve represents a channel.

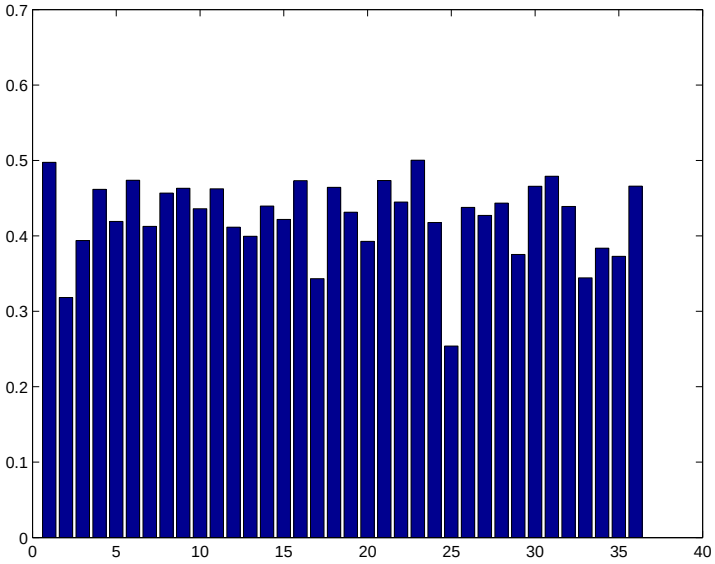

Figure 8 Best PER on entire test data set with optimal sparsity, with SDA based on Fbank. The model is single-channel and multi-subject.

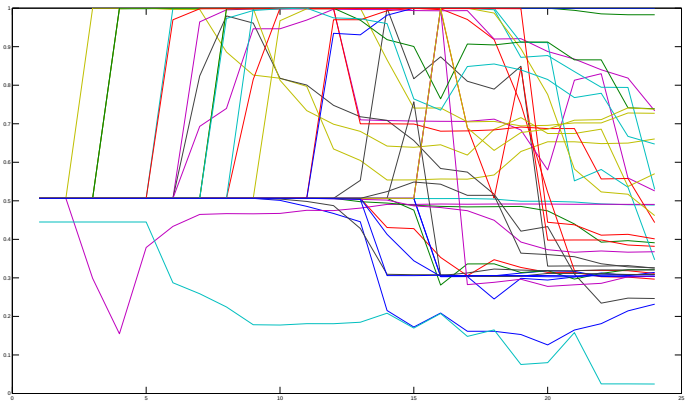

Figure 9 PER on test data for subject 1, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

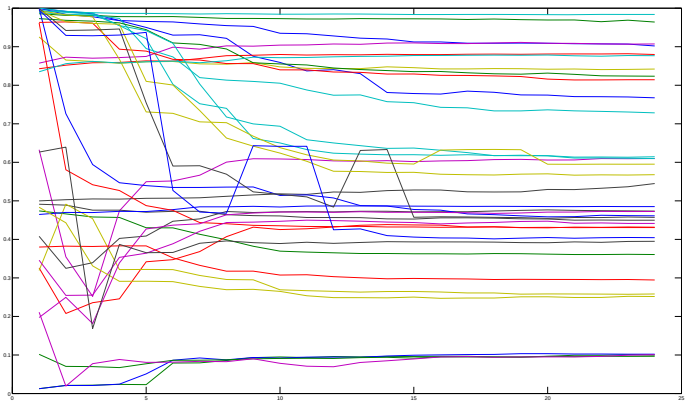

Figure 10 PER on test data for subject 2, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

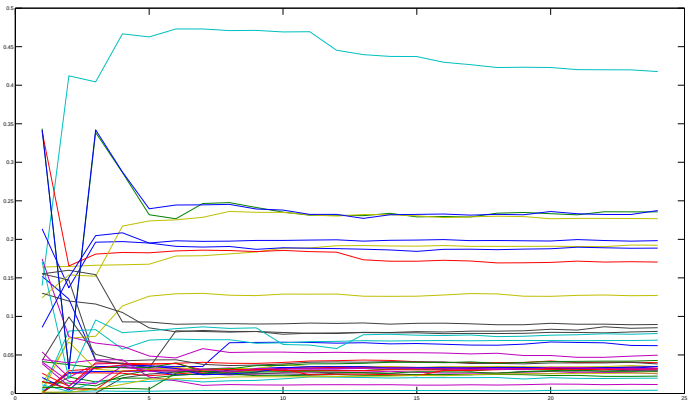

Figure 11 PER on test data for subject 3, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

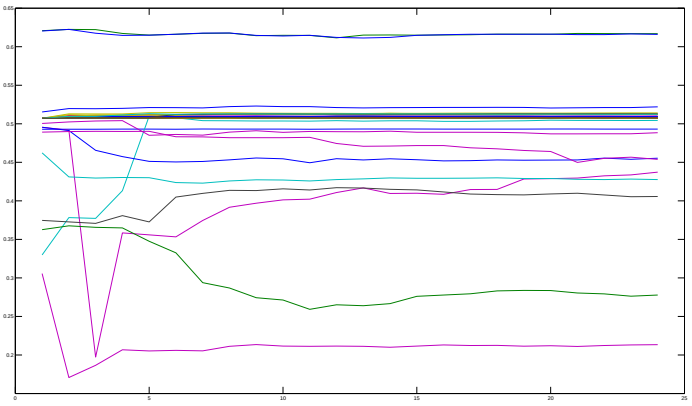

Figure 12 PER on test data for subject 4, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

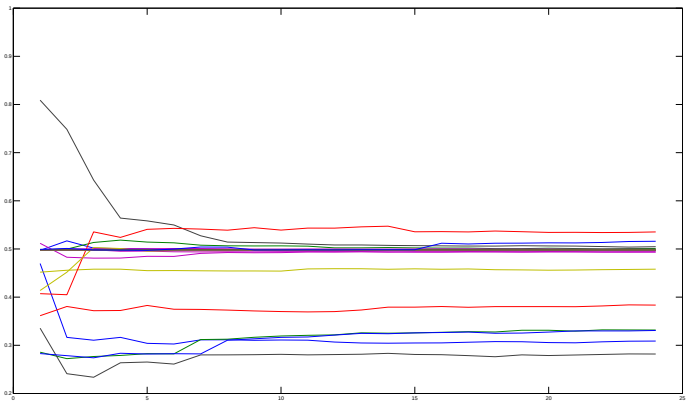

Figure 13 PER on test data for subject 5, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

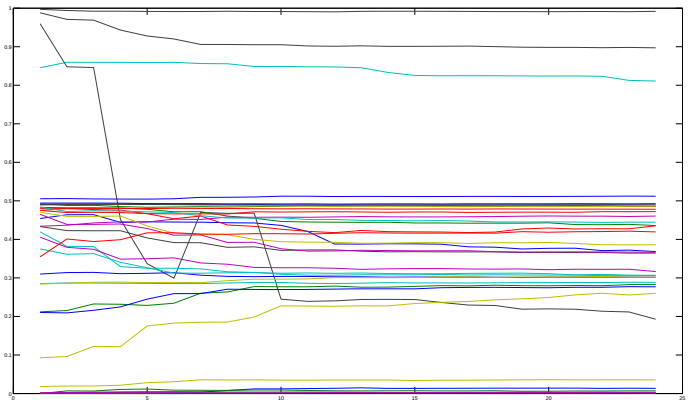

Figure 14 PER on test data for subject 6, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

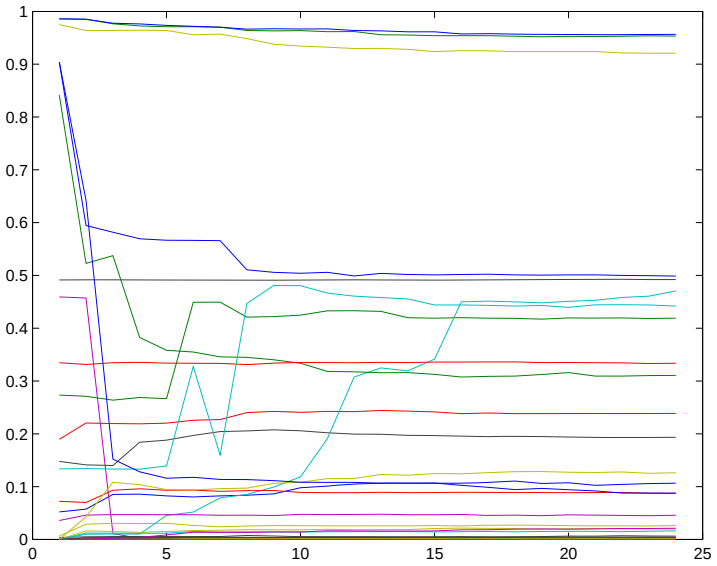

Figure 15 PER on test data for subject 7, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.

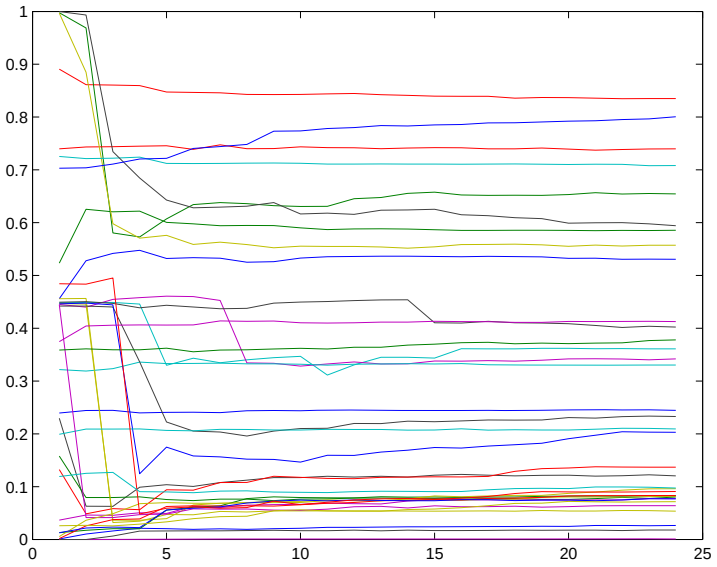

Figure 16 PER on test data for subject 8, with Fbank as features and model trained with the data of this subject. Each curve represents the model trained with the data of a particular channel.