35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Discriminative Tandem Features for HMM-based EEG Classification

Chee-Ming Ting*, *Member, IEEE*, Simon King, *Senior Member, IEEE*, Sh-Hussain Salleh and A. K. Ariff

***Abstract*—We investigate the use of discriminative feature extractors in tandem configuration with generative EEG classification system. Existing studies on dynamic EEG classification typically use hidden Markov models (HMMs) which lack discriminative capability. In this paper, a linear and a non-linear classifier are discriminatively trained to produce complementary input features to the conventional HMM system. Two sets of tandem features are derived from linear discriminant analysis (LDA) projection output and multilayer perceptron (MLP) class-posterior probability, before appended to the standard autoregressive (AR) features. Evaluation on a two-class motor-imagery classification task shows that both the proposed tandem features yield consistent gains over the AR baseline, resulting in significant relative improvement of 6.2% and 11.2% for the LDA and MLP features respectively. We also explore portability of these features across different subjects.**

***Index Terms*- Artificial neural network-hidden Markov models, EEG classification, brain-computer-interface (BCI).**

I. Introduction

A brain-computer interface (BCI) is a communication system which translates specific brain activity, typically measured by scalp-recorded electroencephalogram (EEG) signals into output commands. In standard EEG-based BCI system, the EEG pattern represented by compact feature vector is identified automatically using a pattern classifier to the associated class of mental state. Various classification methods have been used in BCI research. Common approaches are discriminative classifiers which use discriminant decision hyperplanes directly estimated to maximize the separability between classes. Discrimination of EEG using simple linear discriminant analysis (LDA) [1] and nonlinear classifier such as artificial neural networks (ANNs) [2]-[3] and support vector machine (SVM) [4] show good classification results. However, these classifiers are prone to over-fitting and have poor generalization. Improving generalization capabilities by incorporating regularization, however, introduces additional parameters to be selected empirically.

Instead of using hard decision boundary, the alternative generative classifiers choose the class model, typically probabilistic, that most likely generates the sample, which gives better generalization to the test-set. Hidden Markov model (HMM), a dynamic generative model, has been applied in BCI research [5] to better describe the temporal changes of EEG features which static classifiers such as NNs cannot naturally model, e.g. spectral pattern of event-related (de)synchronization (ERD, ERS). However, the generative models trained by maximizing the likelihood of in-class data, does not guarantee discrimination against out-of-class data. The classification performance by these methods, however, is still far from satisfactory when fewer channels are used [6]. This may be due to the highly non-stationarity of the EEG signals typically in poor signal-to-noise (SNR) condition, which renders finding optimal decision hyperplanes difficult if not impossible.

To improve the discrimination of HMM classifier, [7] developed a hybrid ANN-HMM framework using discriminatively trained ANNs to replace the conventional Gaussian mixture models (GMMs) to generate the posterior probabilities of HMM states, while maintaining the underlying Markov structure for temporal modeling. Extension using pre-trained deep neural networks [8] gives significantly better performance than HMM-GMM-based system in large-vocabulary speech recognition task. Other solution involves discriminative training of HMMs [9] by modifying the estimation objective function to be discriminative. However, both approaches are computationally expensive and require substantial modifications on the well-established HMM-GMM framework for which many effective techniques such adaptation have been developed.

Recent work uses the so-called *tandem* approach firstly introduced by [10], in which the class-posterior probability outputs of an ANN classifier are used as additional input features to the conventional HMM-GMM recognizer. These features usually undergo further transformation (such as log-transform and principle component analysis (PCA)) to match the Gaussian modeling assumptions before being augmented to the standard feature set. The main advantage of this approach is that the ANNs can provide discriminative features as complementary to the generative classifiers, and consume less training effort than the discriminative training of HMMs. Besides, ANN can take multiple frames of feature vectors which enable wider context modeling. This approach has been widely used for speech recognition based on the phone-posterior features derived from multilayer perceptrons (MLPs), and shows impressive error reduction, consistently in wide variety of tasks [11]-[12]. Besides, such features are shown to be portable across different domain and language, based on assumption that they share certain similar phonetic

This work was supported by Universiti Teknologi Malaysia (UTM) and Ministry of Higher Education (MOHE) Malaysia under Research University Grant (GUP) –TIER 1 Q.J130000.2545.04H21.

C.-M. Ting is with the Center for Biomedical Engineering (CBE), UTM, 81310 Skudai, Johor, Malaysia (e-mail: cmting1818@yahoo.com).

S. King is with the Centre for Speech Technology Research (CSTR), University of Edinburgh, UK. (e-mail: Simon.King@ed.ac.uk).

Sh-Hussain Salleh is with the Center for Biomedical Engineering, UTM, 81310 Skudai, Johor, Malaysia (e-mail: hussain@fke.utm.my).

A. K. Ariff is with the Center for Biomedical Engineering, UTM, 81310 Skudai, Johor, Malaysia (e-mail: armouris@gmail.com).

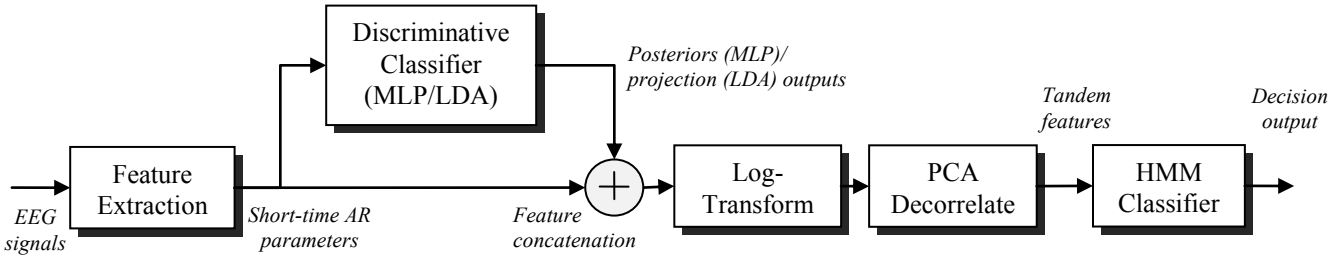

Figure 1. Block diagram of the proposed tandem feature extraction for EEG classification.

contexts [12]. Specifically, the MLPs can be trained from data from other domain to generate tandem features for in-domain data. The out-of-domain (OOD) MLP features can give comparable to better results than the features trained on in-domain data. There are no or limited studies using tandem features for biomedical signal classification other than speech.

This paper proposes the use of discriminative features in tandem to improve the HMM-based classification of single-trial EEG. The effectiveness of the MLP features for speech is due to the model's complexity to learn non-linear separability, and the availability of large dataset to train it reliably. However, for BCI, this advantage of using complex model is limited by small amount of EEG data available. LDA is simple, has less parameters to train and robust to overfitting [13]. Besides, the commonly used LDA is only slightly outperformed by the nonlinear methods on EEG classification [4]. We further propose LDA-based features on EEG for BCI. Thus, two sets of tandem features are extracted from the linear and non-linear discriminative classifiers respectively: projection output features of LDA and posterior features derived from MLP. We use short-time autoregressive (AR) parameters as baseline features to train these discriminative tandem feature extractors at frame basis, which are then appended with the generated tandem features. The back-end classifier uses HMM with GMM observation density trained using Viterbi algorithm. We compare the LDA- and MLP-based features on two-class motor-imagery classification using dataset IIIa from BCI Competition III. We investigate subject-specific and average tandem features derived from the MLP trained from data of one subjects and all subjects respectively. We also address the portability of tandem features across subjects.

II. Tandem Features for EEG Classification

Fig. 1 shows the proposed tandem framework for EEG classification, where two discriminative classifiers are trained to extract additional input features to the conventional HMM classifier. We focus on two-class classification in this study. The EEG signal undergoes standard feature extraction to generate sequence of AR feature vectors estimated from short-time segments. We derive two kinds of features from the outputs of a typical linear and a non-linear static EEG classifier trained on the AR features at frame level.

A) LDA projection features: These features are derived as the projection output of a two-class LDA decision function given an input feature vector $\mathbf{x}$ [14]

$$D(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (1)$$

$$b = -0.5\mathbf{w}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \quad (2)$$

$$\mathbf{w} = (\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) \quad (3)$$

where $\mathbf{w}$ is projection weight vector and b is the bias value. $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ are the mean vector and covariance matrix for class i. This output is an un-calibrated projected value that is not probabilistic, but can be used for discrimination. The output can be mapped to class-posterior probability using a fitted sigmoid function, as performed on SVM outputs [15]. The input to the LDA can be context window of successive frames of feature vectors.

B) MLP posterior features: These features are derived from the class-posteriors estimated by MLP; here, a standard feed-forward network with single hidden layer is used. The network output is a vector of posterior probabilities, each element associated with each class to be identified [10] (in our case, 2 for two-class classification). However, to match the 1-dimensionanl LDA features, only one output is used, with the training target set to 1 for one class and 0 for the other. The network is trained using the variable learning-rate gradient descent algorithm.

These discriminative features are then concatenated with the standard AR features. However, both the augmented feature vectors are highly correlated and non-Gaussian, and hence not suitable for modeling by GMMs with diagonal covariance. Thus, the augmented features are Gaussianized by taking their logarithms, and decorrelated using single PCA transform with further dimensionality reduction to retain only the significant components [10]-[12]. Applying transform on the augmented features including the baseline AR parameters rather than on the tandem features alone as in [10]-[12], is due to that the AR features used here have not been Gaussianized and decorrelated implicitly as the Mel-frequency cepstral coefficients (MFCCs), the baseline features for speech do in [10]-[12]. These transformed features are then fed to the HMM-GMM classification system trained with standard maximum-likelihood (ML) estimation.

III. Experimental Results

We investigate the performance of the tandem features on single-trial EEG-based motor imagery classification using dataset IIIa, a subset of BCI Competition III dataset [16]. The task is classification of four-class cued motor imagery EEG (left hand, right hand, foot or tongue movements). The

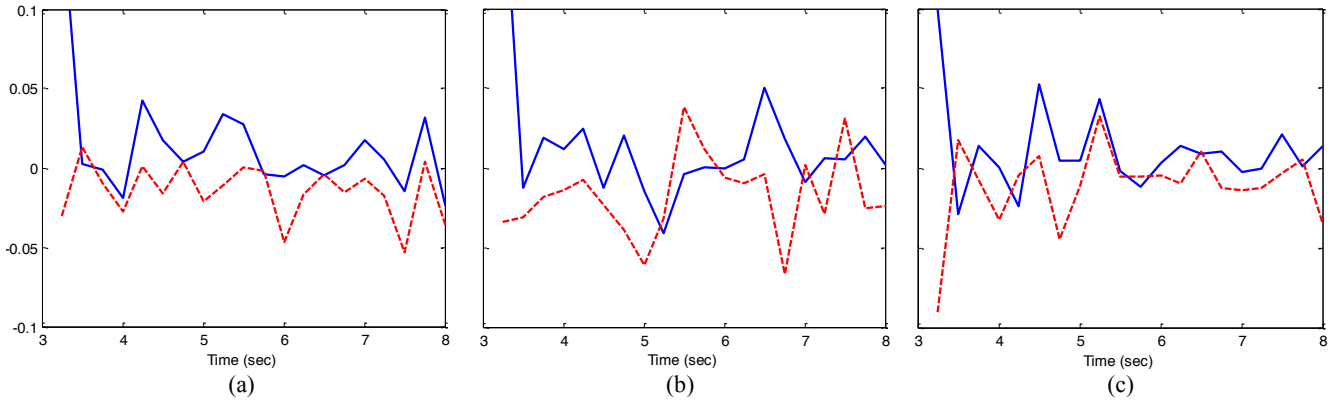

Figure 2. Averaged time courses of LDA tandem features extracted from the left-hand (——) and right-hand (·····) motor imagery EEG data from the training set for subject k3 (a), k6 (b) and l1 (c). The cue onset is presented at 3 s. The plots are obtained by ensemble averaging over trials.

TABLE I. NUMBER OF TRAINING AND TEST TRIALS FOR EACH SUBJECT.

Subject	# of training trials		# of test trials	
	Left hand	Right hand	Left hand	Right hand
k3	36	37	38	38
k6	21	26	22	16
l1	20	20	23	19

TABLE II. CLASSIFICATION RESULTS IN ACCURACY (%) USING AR BASELINE, LDA AND MLP TANDEM FEATURES.

Feature set	Subject			Mean
	k3	k6	l1	
AR baseline	59.21	61.08	52.86	57.72
SS LDA tandem	61.84	68.47	53.66	61.32
SD LDA tandem + PCA	56.58	66.19	48.86	57.21
SS MLP tandem	64.47	66.48	61.56	64.17
SS MLP tandem + PCA	57.89	61.65	56.41	58.65
AS MLP tandem	59.21	65.63	61.21	62.02
AS MLP tandem + PCA	60.53	63.92	53.78	59.41

database consists of three subjects each recorded 60-channel EEG data with sampling frequency 250 Hz, with 60 trials per class. We focus on two-class classification of left and right hand movement, using only two unipolar channels i.e. the C3 and C4. The numbers of training and test trials for each subject are shown in Table I. Use of very few channels along with small amount of training data per subject make the classification task very challenging.

A subject-dependent HMM-GMM system is built for each subject using Viterbi training algorithm. Single-trial HMMs of 2 and 3 states with Gaussian mixture components per state varied from 1, 2, 4, and 8, are trained and the optimal model is selected based on the best classification result. The EEG segments during motor-imagery from 3 to 8s are used for analysis. The baseline features used to represent the spectral changes of ERD during motor imagery, are 12-dimensional short-time AR features (6 coefficients from each channel) computed over each window of 250ms without overlapping. Single frame is used as input as we found that use of window of frames does not give any performance improvement, which may due to that the 250ms window is sufficient to capture the wide contexts of ERD changes.

We extract the LDA features using the linear discriminant function (1) estimated by Equations (2) and (3). Fig. 2 shows the averaged time courses of LDA features (plotted as function of frames) extracted from the left-hand and right-hand motor imagery EEG data of the train-set for each subject. Generally, subject k3 shows the clearest hemispheric differentiation, followed by k6 and l1. The best frame discrimination is observed at earlier period for subject k3 and l1 (4 to 5s) and later for k6 (6 to 7s), where prominent ERDs are expected to occur. These best discriminating regions are consistent with where the best frame classification results using LDA are achieved on the same dataset [6]. The good separability of the left- and right-hand features obtained, is expected to provide additional discriminative information for improving the subsequent classification. The MLP features are extracted by a network with a hidden layer of 15 units and a single non-linear output unit (a structure of $12\times15\times1$), trained with variable learning rates with maximum 500 training iterations as stopping criterion. The single discriminative feature is appended to the AR features, giving augmented vector of 13 dimensions, further reduced to 10 after PCA transform. Additional experiments show that the performance decreases when the log-transform is applied. For classification, the increase of computational time to calculate these tandem features was insignificant.

Table II shows the classification results for the baseline AR features, and after concatenation with LDA and MLP tandem features with and without PCA transform. For MLP features, we also compare networks trained on a subject-specific (SS) basis and over all subjects (AS). It can be seen that both tandem features without PCA, significantly outperform the baseline AR features for all subjects, giving relative averaged accuracy improvement of 6.2% and 11.2% for LDA and MLP features respectively. This also indicates that the more complex non-linear model can generate better discriminative features. Applying PCA transform actually degrades the performance of tandem features but still improve over the baseline generally. Possible reason is that the discriminative power of the single tandem feature might loss after averaged projection along with the more dominant high-dimensional baseline AR features. All subjects are hypothesized to exhibit

TABLE III. CROSS-SUBJECT CLASSIFICATION RESULTS USING OUT-OF-SUBJECT MLP FEATURES TRAINED ON OTHER SUBJECTS.

Feature set	Subject			Mean
	k3	k6	l1	
AR baseline	59.21	61.08	52.86	57.72
In-subject MLP tandem	64.47	66.48	61.56	64.17
MLP tandem trained on k3	-	62.78	51.14	
MLP tandem trained on k6	55.26	-	53.66	
MLP tandem trained on l1	56.58	58.24	-	
Mean	55.92	60.51	52.40	56.28
In-subject MLP tandem + PCA	57.89	61.65	56.41	58.65
MLP tandem trained on k3 + PCA	-	67.05	55.95	
MLP tandem trained on k6 + PCA	56.58	-	55.95	
MLP tandem trained on l1 + PCA	51.32	65.06	-	
Mean	53.95	66.06	55.95	58.65

some similar subject-independent pattern of ERD despite great inter-subject variability. The average MLP tandem despite being trained on all subjects outperform the baseline with slightly poorer performance over the SS model, suggesting its use for subject-independent classification in BCI research.

To further evaluate the subject-independency of the MLP and hence the portability of the generated tandem features across subject, we perform cross-subject classification, where MLPs trained on one subject for tandem feature extraction for other subjects. The results are shown in Table III. It is clearly seen that the posterior features trained on out-of-subject data are able to match the performance of in-subject trained features. While the all-subject trained MLP can offer gains over the baseline, the posterior features trained solely on single other subject are slightly underperformed. This indicates that these cross-subject MLPs fail to generalize as the average models, due to the large inter-subject variability. Use of PCA transform performs slightly better than the baseline, possibly because of the imposed feature selection and decorrelation.

IV. CONCLUSION

In this paper, we have presented a tandem approach for single-trial EEG classification by using two discriminative classifiers as additional feature extractors for the conventional generative HMM-GMM classification system. Our results on two-class motor-imagery EEG classification show that both the proposed LDA and MLP tandem features provide substantial gains when augmented to the baseline AR features consistently for each subject, with the best relative accuracy improvement of 11.2% by the non-linear MLPs. This suggests that the tandem features can bring discriminative information complementary to the baseline features. The posterior features from MLP trained on all subjects, an average model which are shared across subject, are able to improve the baseline, suggesting some degree of subject-independency of the features. However, MLP features trained solely on a single subject fail to generalize to other subjects. Future work will investigate cross-task portability of the tandem features trained on multiple subjects, for BCI research. Besides, the performance evaluation will be extended using more complex four-class classification task or other databases with more subjects.

ACKNOWLEDGEMENT

The authors would like to thank the reviewers for valuable comments.

REFERENCES

[1] G. Pfurtscheller, C. Neuper, A. Schlogl and K. Lugger, "Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters," *IEEE Trans. Rehabil. Eng*., vol. 6, no. 3, pp. 316–325, 1998.

[2] C. Neuper, A. Schlogl and G. Pfurtscheller, "Enhancement of left-right sensorimotor EEG differences during feedback-regulated motor imagery," *J. Clin. Neurophysiol.*, vol. 16, no. 4, pp. 373–382, 1999.

[3] E. Haselsteiner and G. Pfurtscheller, "Using time-dependent neural networks for EEG classification," *IEEE Trans. Rehabil. Eng*., vol. 8, no. 4, pp. 457–463, 2000.

[4] D. Garrett, D. A. Peterson, C. W. Anderson, and M. H. Thaut, "Comparison of linear, nonlinear, and feature selection methods for EEG signal classification," *IEEE Trans. Rehabil. Eng*., vol. 11, no. 2, pp. 141–144, 2003.

[5] B. Obermaierb, C. Gugera, C. Neuperb, and G. Pfurtschellera, "Hidden Markov models for online classification of single trial EEG data," *Pattern Recognition Letters*, vol. 22, no. 12, pp. 1299–1309, 2001.

[6] A. Schlögl, F. Lee, H. Bischof, G. Pfurtscheller, "Characterization of four-class motor imagery EEG data for the BCI-competition 2005," *J. Neural Eng*., , vol. 2, no. 4, pp. 14–22, 2005.

[7] N. Morgan and H. Bourlard, "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach." *Signal Processing Magazine*, 25-42, May 1995.

[8] G. E. Dahl, D. Yu, L. Deng and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Trans. On Audio, Speech and Language Process*., vol. 20, no. 1, pp. 30-42, 2012.

[9] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition—A unifying review for optimization-oriented speech recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, Sep. 2008.

[10] H. Hermansky,D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems", *in Proc. ICASSP*, pp. 1635–1638, 2000.

[11] D. P. W. Ellis, R. Singh, and S. Sivadas, "Tandem acoustic modeling in large-vocabulary recognition," in *Proc. ICASSP*, pp. 517–520, 2001.

[12] A. Stolcke, F. Grezl, M.-Y. Hwang, L. Xin, N, Morgan and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, vol 1, pp. 321–324, 2006.

[13] K. R. Müller, C. W. Anderson, and G. E. Birch, "Linear and nonlinear methods for brain-computer interfaces," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 11, no. 2, pp. 165–169, Jun. 2003.

[14] C. Vidaurre, A. Schlogl, R. Cabeza, R. Scherer and G. Pfurtscheller, "Study of on-line adaptive discriminant analysis for EEG-based brain computer interfaces," *IEEE Trans. Biomed. Eng.,* vol. 54, no. 3, pp. 550–556, Mac. 2007.

[15] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA: MIT Press, 1999.

[16] B. Blankertz, K. R M¨uller, G. Curio, T. M. Vaughan, G. Schalk, J. R. Wolpaw, A. Schl¨ogl, C. Neuper, G. Pfurtscheller, T. Hinterberger, M. Schr¨oder, and N. Birbaumer, "The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 1044–1051, Jun. 2004.