EEG PATTERN RECOGNITION THROUGH MULTI-STREAM EVIDENCE COMBINATION

Morris, A.C.[1], Obermaier, B.[2] & Pfurtscheller, G.[2]

1. Dalle Molle Institute for Perceptual Artificial Intelligence, Martigny, Switzerland

2. Institute for Biomedical Engineering, DPMI, Graz, Austria

morris@idiap.ch, obermai@dpmi.tu-graz.ac.at, pfu@dpmi.tu-graz.ac.at

Abstract

EEG recordings provide an important means of brain-computer communication, but their classification accuracy is limited by unforeseeable variations in the signal due to artefacts or recogniser-subject feedback. A number of techniques were recently developed to address a related problem of recogniser robustness to uncontrollable signal variation which also occurs in automatic speech recognition (ASR). In this article we consider how some of the proved advantages of the "multi-stream combination" and "tandem" approaches in HMM/ANN hybrid based ASR can possibly be applied to improve the performance of EEG recognition.

1. Introduction

EEG signals are weak voltages resulting from the spatial summation of electrical potentials in the brain cortex, which can easily be detected by suitably placed electrodes on the scalp surface. They result from the superposition of three main types of brain potentials: oscillatory, event-related, and slow potential shifts. Different components of the EEG signal have been widely demonstrated to have measurable correlates with the brain activity involved in specific mental tasks.

1.1 Brain-computer interfaces

Some mental activities are more clearly correlated with detectable EEE patterns than others. Brain-computer interfaces (BCI) based on EEG recognition, with applications including brain-computer communication for people with severe motor disabilities, are being investigated by a number of groups [7, and references in 8]. The Graz BCI [11] exploits a particularly marked correlation which exists between the imagination of certain body movements and ERD (event-related desynchronisation) over specifically involved sensorimotor areas [12,13]. These ERD events cause changes in the in the alpha and beta bands (8-13 & 14-30 Hz) of the EEG signal which can be detected with good and improving reliability. In this article we consider whether the "multi-stream combination" and "tandem" approaches recently developed for robust speech recognition can also be successfully applied to the recognition of EEG signals for the Graz BCI. Any improvement in EEG classification can significantly increase ease of communication.

1.2 The EEG recognition task addressed

The recognition task here is the off-line classification of EEG Alpha and beta spectral power values from one electrode on each side of the head into two classes “left” and “right”. Classifiers are trained and tested on data from the same subject, there being no requirement for subject independent recognition. The subject is instructed to imagine moving their left or right arms to signify a “left” and “right” choice respectively.

2. Methods

The system mostly widely used in speech recognition is the Gaussian-mixture based HMM (Hidden Markov Model) [14,15]. The success of this model is due mainly to its ability to capture time structure in the input data without need for hand labelling. However, HMM performance degrades dramatically when noise conditions in operation do not match those used in training. For slowly varying noise this problem can be reduced using techniques such as noise robust features or noise estimation and subtraction. The following two multi-stream HMM/ANN hybrid HMM variations were recently developed to improve robustness to impulsive and rapidly varying noise.

2.1 Models as used in speech recognition

The multi-stream HMM/ANN (HMM/artificial neural network) system [5,6] was developed from the HMM in two stages. First the GMM (Gaussian mixture model) normally used with HMMs was replaced by an ANN, usually a one hidden layer MLP (multi-layer perceptron). ANNs are more discriminative than the GMMs normally used for HMM state probability estimation, and are better able to capture dynamic patterns from an extended sample window. Next, the single ANN was replaced by multiple ANNs which are trained separately on (preferably all possible) combinations of different input data streams. During recognition the outputs from these multiple “experts” are combined, taking into account the relative reliability estimated for each expert. This allows for combination of class evidence from different sources (such as audio and visual, or different audio time scales), and also permits sustained performance when one or more data streams are corrupted by rapidly changing noise.

The “tandem” HMM/ANN approach [3] also makes use of the ANN’s ability to capture discriminative and dynamic features which the GMM is not able to model. But whereas in the multi-stream HMM/ANN mentioned above the ANN outputs are first combined and then interpreted by the system as “scaled state likelihoods” and used to substitute the state likelihoods usually provided by the GMM, in this case the ANN outputs (prior to squashing by the sigmoid activation function so as to represent probabilities) are not combined but are simply used to substitute the original input features. This is a non-linear generalisation of the commonly used LDA (linear discriminant analysis) technique which is already widely used for data preprocessing prior to classification. The tandem technique has recently proved very effective in noise robust speech recognition, beating all other methods on the Aurora benchmark (A.D. 2000).

2.2 Models used here for EEG classification

HMMs not used in current experiments. In [8] it was shown that HMMs can be used to some advantage for EEG recognition. However, the time structure of the EEG signal above that which is necessary for spectral analysis is very limited (or at least very little understood) in comparison to that for speech, for which the multi-stream models described above show a significant advantage over the baseline HMM. In this study we therefore consider only the problem of classifying EEG data (joint spectral data from two electrodes) at the level of individual time frames.

Some initial experiments were made in EEG decoding with the HTK HMM system, in which the entire EEG recording was treated as a single "whole word". The multi state HMMs used were unable to successfully identify any time structure in the sequence of EEG spectral frames. Visual inspection of the time course of spectral parameters for a random sample of different recordings from the same subject (Figure 2) show that temporally invariant structure is indeed very limited.

In speech the alternative to whole word modelling is phoneme based modelling. In the absence of an established lower level of recognition units for EEG signals, this approach is not possible. For this reason all of the experiments reported here are concerned only with left/right classification at the single frame level.

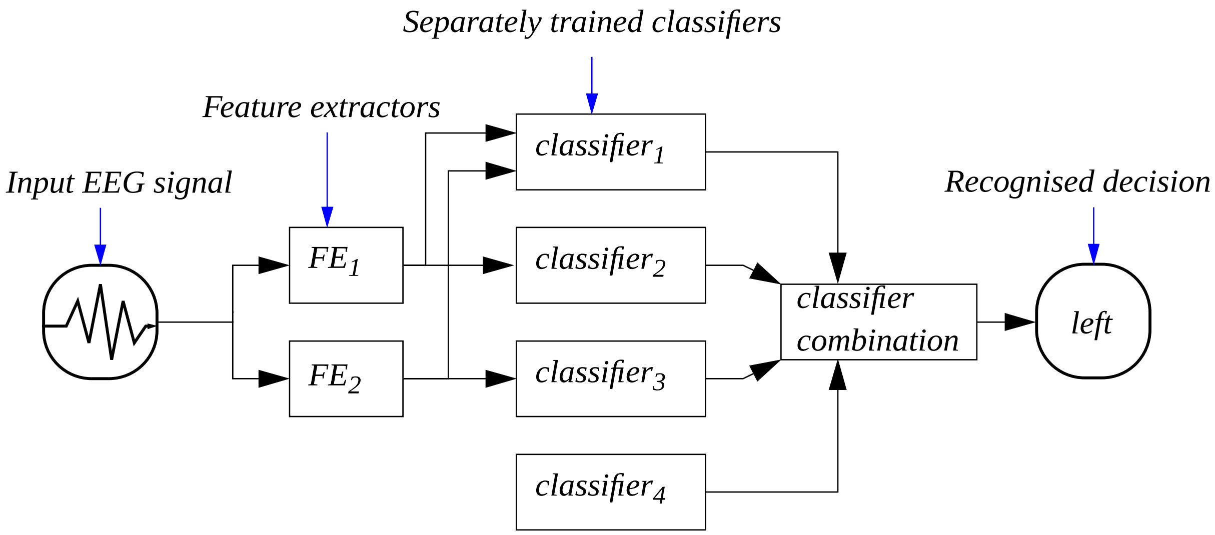

Figure 1. ***Multi-stream single frame EEG decoder.*** *Two or more types of feature vector are extracted from a short term EEG sample. In training a separate classifier (for example, an MLP) is trained on every combination of these feature vectors. In recognition the class probabilities from each classifier are combined by some rule (such as MAP) taking into account the relative reliability of each classifier..*

EEG adapted multi-stream recogniser. The multi-stream single frame EEG recognition system tested here is shown in Figure 1. This is like the multi-stream system used in ASR [6] except that the state-probabilities lattice decoder between the classifier combination unit and the recognised state sequence has been removed. Multiple streams here consisted only of subdivisions of the one feature stream available. A number of different classifiers were tested.

1. LDA
2. linear MLP (n inputs -> 1 sigmoid output)
3. one hidden-layer MLP [1] (n inputs -> 10 hidden units -> 1 sigmoid output)
4. SVM (support vector machine) [2]
5. GMC (Gaussian mixture classifier) (as HMM with no hidden states or transition probs).

For two-class LDA, if X is the data matrix, with data vectors as columns, μ_1 and μ_2 are the class means, and μ_{12} is the joint data mean, then the single discriminant projection vector is given by $u = (XX')^{-1}(\mu_2 - \mu_1)$. If $a = u' \mu_{12}$, then LDA places x in class 1 if $u'x < a$, else in class 2.

MLPs were trained using the cross-entropy objective [1], with a fixed "momentum" of 0.5, adaptable "learning rate", with one weight update per full cycle through the training set. Training stopped after 4000 iterations, or (most usually) when the cross validation error started to increase.

The GMC was initialised with k-means clustering, followed by EM iterative estimation of maximum likelihood parameters (the only classifier here not to use discriminative training).

Multiple classifier combination was tested using each of the following rules:

1. MAP rule (Maximum A-Posteriori): select the class which has the maximum probability among all classes and all classifiers (max posterior probability => max probability correct)
2. Majority rule: select the class which is most often selected when MAP selection is used by each individual classifier.
3. Average rule: Form the average of corresponding class probability outputs from each expert, then select the class which has the maximum average probability [1].

EEG adapted tandem based recogniser. When the HMM is removed from the "tandem" HMM/ANN system, all that remains is the multi classifier based non-linear discriminative preprocessor, so that the recognition process is left incomplete. The decision not to use HMMs in this study was not taken until after the results were obtained for the above mentioned tests with a baseline HMM. Tandem preprocessing should be tested by anyone using HMMs for EEG recognition. We could test multiple classifier combination by some of the above classifiers in future.

2.3 Test data preparation

60 "left" and 60 "right" 8.0 s recordings were made from 2 electrodes, on opposite sides of the head, for each of 4 subjects, denoted k4, k1, j4, i6. Features extracted were two log band power coefficients, for bands 8-12 Hz (alpha) and 16-24 Hz (beta), obtained using 5th order Butterworth filters, with a frame shift of 1/128 s, giving 1024 frames per recording. The signal to "start thinking" appeared after 3.0 s into the recording, and only samples from 4.25 to 8.0 s (frames 545-1024) were used. In Figure 2 it can be seen that each spectral coefficient varies over time and between recordings in (what appears to be) a very unpredictable way. Figure 3 shows the correlation matrix and principal factor for each subject.

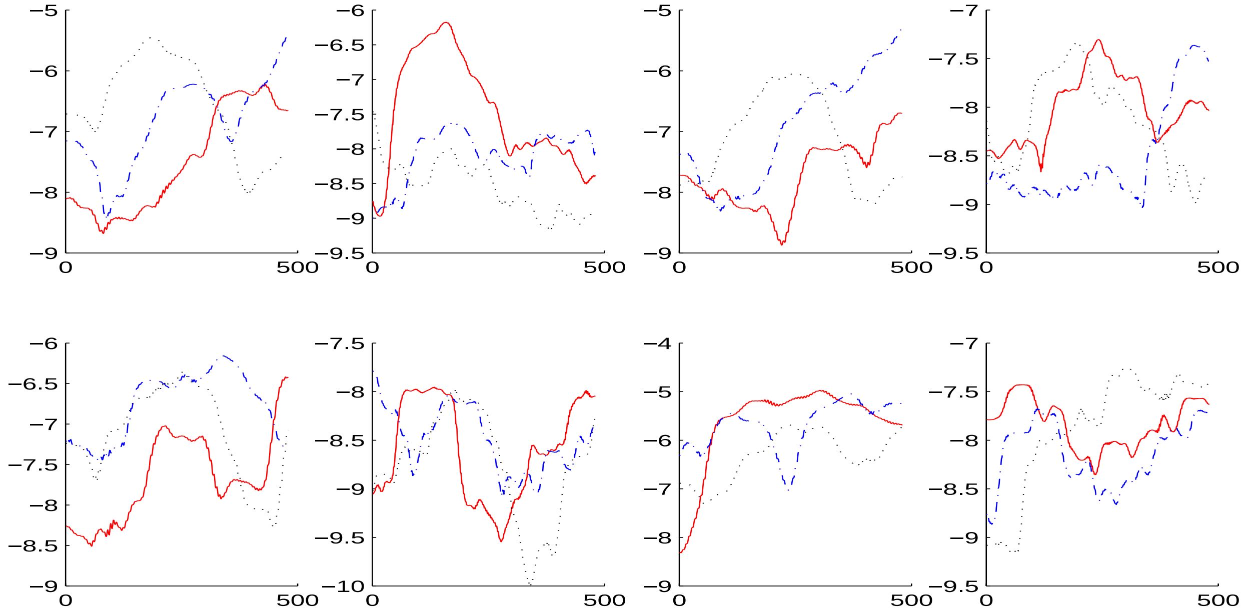

Figure 2. Top row shows the time course for log spectral power coefficient 1 (left plot) to 4 (right plot) from three separate complete "left" recordings (frames 545-1024) from subject k4 (the easiest to recognise!). Bottom row shows same for three complete "right" recordings.

```
k4 correlation                                      j4 correlation
    1.0000    0.4705    0.4436    0.2294                1.0000    0.1211    0.3169    0.0788
    0.4705    1.0000    0.1235    0.2034                0.1211    1.0000   -0.0169    0.2165
    0.4436    0.1235    1.0000    0.5795                0.3169   -0.0169    1.0000    0.2190
    0.2294    0.2034    0.5795    1.0000                0.0788    0.2165    0.2190    1.0000
   principal factor                                     principal factor
   -0.4544   -0.5467    0.5762    0.4034               -0.0167   -0.8774    0.2411    0.4145

i6 correlation                                      k1 correlation
    1.0000    0.5887    0.3810    0.2076                1.0000    0.3432    0.4362    0.0033
    0.5887    1.0000    0.2979    0.3687                0.3432    1.0000    0.1225    0.2822
    0.3810    0.2979    1.0000    0.3996                0.4362    0.1225    1.0000    0.1717
    0.2076    0.3687    0.3996    1.0000                0.0033    0.2822    0.1717    1.0000
   principal factor                                     principal factor
   -0.7051   -0.2254    0.5242    0.4210               -0.7907    0.2585    0.5543   -0.0268
```

Figure 3. Correlation matrix and corresponding LDA projection unit vector for each subject

3. Results

Results comparing recognition accuracy for each of seven multi-classifier combination schemes with its single classifier baseline are shown in Figure 5. Each result is an average over 10 jackknifed selections of 3 parts for training, one for cross validation, and one for testing, from 5 subdivisions of the data set for each subject. A one standard deviation error bar is shown with each result. 480 data frames were available with each recording, but only one frame in 10 was used. Results were also obtained using one frame in 5 instead of 10, and with 5, 10 and 15 concatenated frames, but no improvement was obtained and these results are not shown.

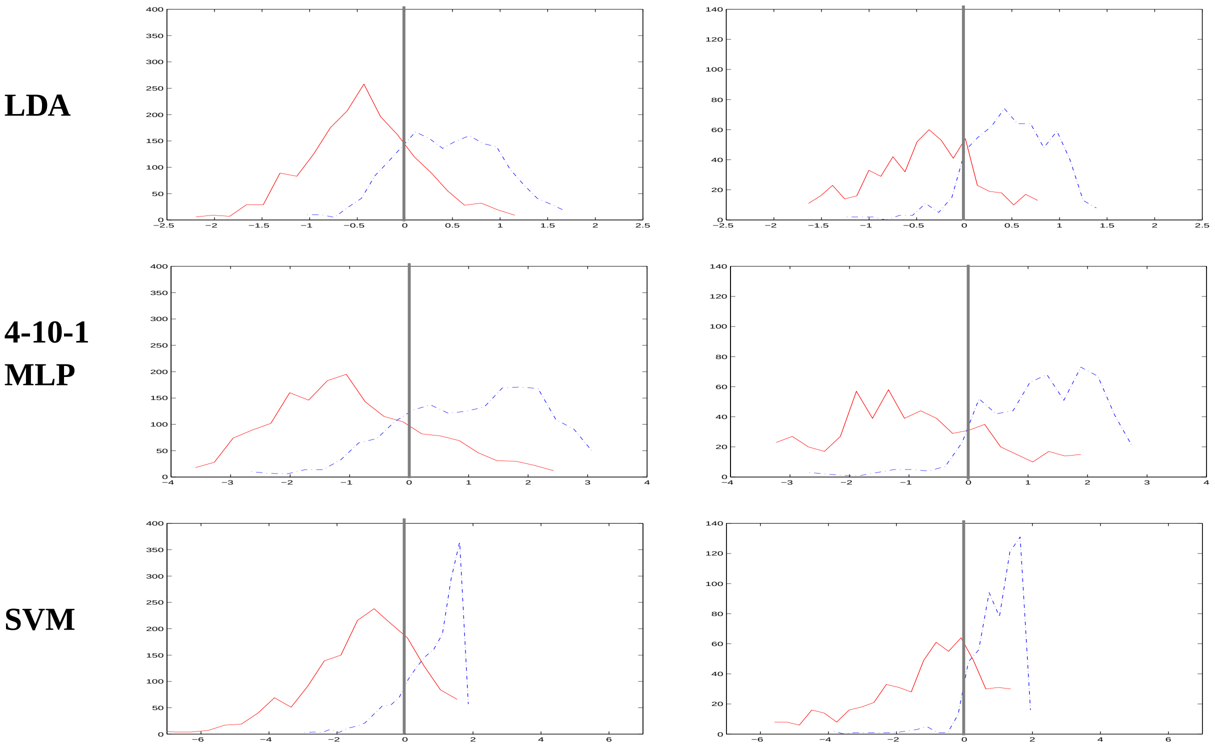

Figure 4. Histograms of projected training data (left) and corresponding test data (right) using LDA (top), MLP (centre) and SVM (bottom) for one training/test subset for one subject (k4). "Left" data = solid, "right" data = dashed. Training set is parts 1,2,3, test set is part 5. Respective test data %correct scores for this example are 83.9, 83.0 and 82.9 respectively (MLP values are pre-squashed MLP class probability outputs).

Conclusion

Results show that whatever advantage the multi-classifier methods tested may have in speech recognition, this does not transfer to EEG recognition under any of the frame level classification schemes tested here. The SVM performs marginally better than other classifiers, but none of the multi-classifier systems show any advantage over the corresponding single classifier baseline, or over the simple LDA classifier (which is also far quicker to train). Full covariance GMC did not improve over diagonal covariance. There was no significant % difference between the MAP, Maj and Avg combination rules. "VSE" in schemes 6 & 7 refers to an experiment with "vector space expansion" intended to correspond to the concatenation of spectral time difference features in ASR. Here all combinations of log feature sums were used in order to capture spectral correlations. This gave 1% improvement for LDA, but a significant degradation for SVM. Note that examination of the principal factors in Fig. 3 shows that in 3 cases out of 4 the LDA decision is just saying "choose left if the log power sum is greater in electrode 1, else choose right".

That EEG recognition did not benefit from any of the potential advantage of these multi-stream techniques suggests two main problems with the systems tested here. One is the need to combine data streams which do not repeat but *add complementary information*. In future one could try combining spectral power with AAR and Hjorth AMC features. Another is the need for improved

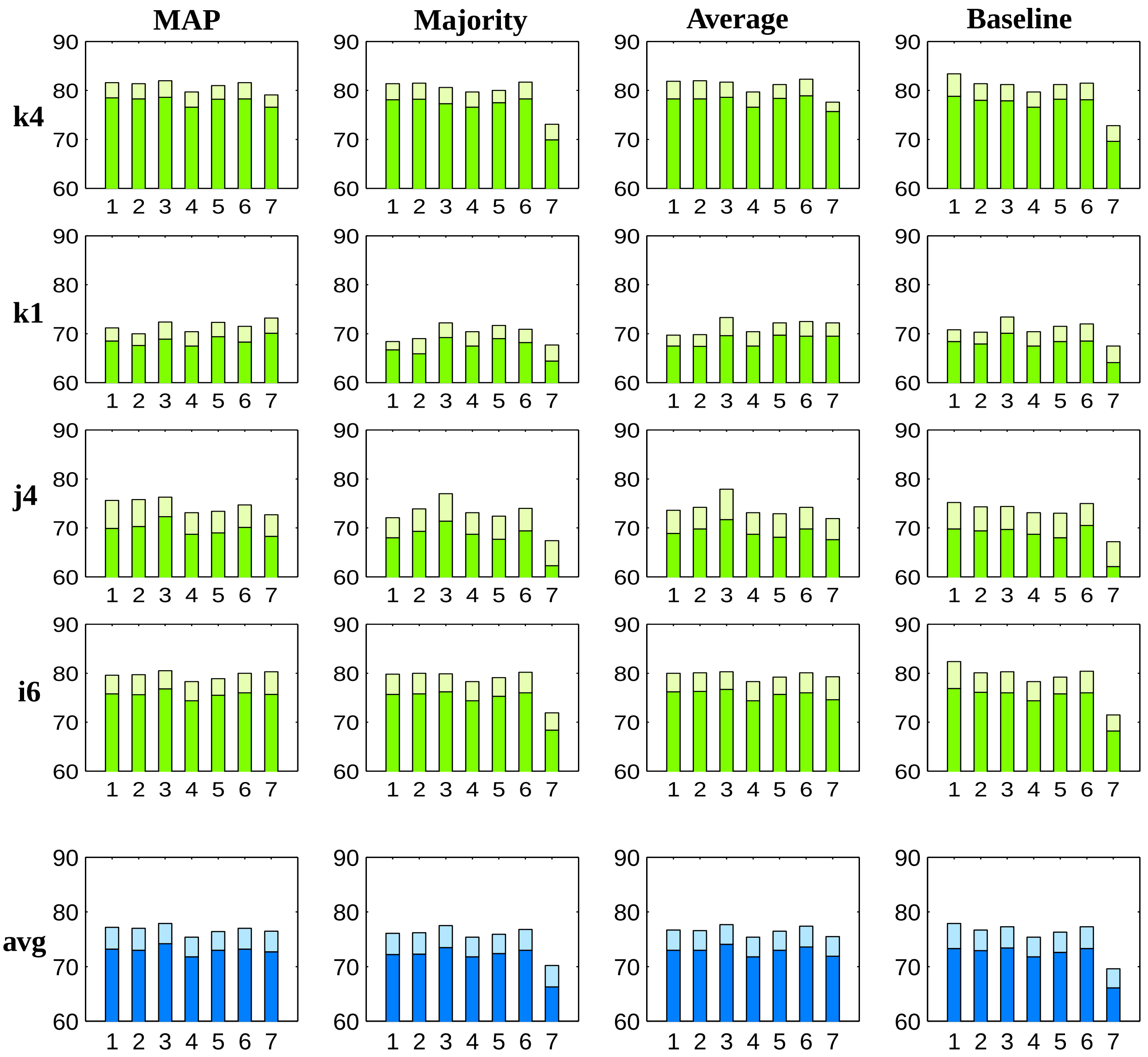

Figure 5. Comparison of multi-classifier combination %accuracy (columns 1-3) with single classifier baseline. Percentage accuracy results are shown for seven different classifier types (1:LDA, 2:MLP, 3:SVM, 4:GMC, 1-mix, 5:GMC:, 3-mix, 6:MLP with VSE, 7:SVM with VSE). Each system first performs independent classification with data from each possible selection of 1 or more coefficients from the 4 dimension the data vector. Separate classifiers are then combined using either the MAP, "majority" or "average" rule. Rows 1 to 4 give results for each subject. Row 5 shows average accuracy results over all subjects.

modelling of the time structure of EEG signals, so as to tap the modelling power of HMMs.

Acknowledgements

This work was supported by the EC/OFES (European Community / Swiss Federal Office for Education and Science) RESPITE project (REcognition of Speech by Partial Information TEchniques).

References

[1] Bishop, C. (1995) **Neural Networks for Pattern Recognition**, Clarendon Press, Oxford.

[2] Collobert, R. & Bengio, S. (2001), “SVMTorch: support vector machines for large-scale regression problems”, Journal of Machine Learning Research, Vol.1, pp. 143-160.

[3] Hermansky, H., Ellis, D. & Sharma, S. (2000), “Tandem connectionist feature stream extraction for conventional HMM systems”, ICASSP-2000, http://www.dcs.shef.ac.uk/research/groups/spandh/projects/respite/publications/ellis_2_icassp_00.pdf

[4] Hermansky, H., Tibrewela, S. & Pavel, M. (1996) “Towards ASR on partially corrupted speech”, Proc ICSLP’96, pp. 462-465.

[5] Morris, A.C., Hagen, H. & Bourlard, H. (2001) “MAP Combination of Multi-Stream HMM or HMM/ANN Experts”, Proc. Eurospeech’01 (in press), & IDIAP-RR 01-14, ftp://ftp.idiap.ch/pub/reports/2001/rr01-14.ps.gz

[6] Morris, A.C., Hagen, A., Glotin, H. & Bourlard, H. (2001), “Multi-stream adaptive evidence combination for noise robust ASR”, Speech Communication, vol.34, pp.25-40.

[7] Millán, J., Mourinño, Babiloni, F., Cincotti, F., Varsta, M. & Heikkonen, J. (2000), “Local neural classifier for EEG-based recognition of mental tasks”, Proc. IJCNN 2000.

[8] Obermaier, B. (2001), “Design and implementation of an EEG based ‘virtual keyboard’ using Hidden Markov models”, PhD thesis, Technische Universität Graz, Institut für Elektro- und Biomedizische Technik, Abteilung für Medizinische Informatik, Graz, Austria

[9] Obermaier, B., Munteanu, C., Rosa, A. & Pfurtscheller, G. (2000) “Asymmetric hemisphere modeling in an off-line brain-computer interface”, in Revision IEEE Systems, Man and Cybernetics.

[10] Pearce, D. & Hirsch, H.-G. (2000) “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, Proc. ICSLP’00, Vol.4, pp.29-32.

[11] Pfurtscheller, G. **Brain-Computer Interface**, http://www.dpmi.tu-graz.ac.at/~guger/bci/BCI_team.htm

[12] Pfurtscheller, G., Neuper, C., Schlögl, A., & Lugger, K. (1998) “Separability of EEG signals recorded during right and left motor imagery using adaptive autoregressive parameters”, IEEE Trans. Rehab. Engng. 6(3), pp.316-325.

[13] Pfurtscheller, G. & Neuper, C. (1992) “Simultaneous EEG 10 Hz desynchronisation during finger movements”, NeuroReport 3, pp.1057-1060.

[14] Rabiner, L. (1989), “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, 77.

[15] Rabiner, L. & Juang, B.-H. (1993) **Fundamentals of Speech Recognition**, Prentice Hall Signal Processing Series. Englewood Cliffs, NJ: Prentice Hall.