BIO Web of Conferences **1**, 00043 (2011)
DOI: 10.1051/bioconf/20110100043

Machine Learning Analysis of Binaural Rowing Sounds

Leonard Johard * Emanuele Ruffaldi * Pablo Hoffmann † Alessandro Filippeschi *

(*) PERCRO Lab, Scuola Superiore S.Anna, Pisa, Italy
(†) Department of Electronic Systems, Aalborg University, Denmark
E-mail: l.johard@sssup.it, e.ruffaldi@sssup.it, pfh@es.aau.dk

Abstract

Techniques for machine hearing are increasing their potentiality due to new application domains. In this work we are addressing the analysis of rowing sounds in natural context for the purpose of supporting a training system based on virtual environments. This paper presents the acquisition methodology and the evaluation of different machine learning techniques for classifying rowing-sound data. We see that a combination of principal component analysis and shallow networks perform equally well as deep architectures, while being much faster to train.

1 Introduction

Assessing and measuring expertise in sport is an important aspect of traditional training, but it assumes a special relevance when training is supported by robotics and virtual reality technologies. The execution of expert performance can be assessed by several means, motion and task performance being the most common. In some sports the sound of task execution is already recognized as being a signature of correct motion. Qualitative descriptions of what elite rowers use as indicative of optimal rowing technique include the sounds produced by blades and boat motion [6].

In the context of rowing training by means of virtual environments [9], it is important to evaluate expert performance in real rowing and we think that sound can be a source of information in addition to boat instrumentation. The aim of this work is indeed to evaluate the potential of rowing sound as a source of information for the automatic analysis and evaluation of performance. The specific goal is to explore the possibilities for developing a machine-hearing system that can locate the phase boundaries in real-time.

This work compares several supervised algorithms on an audio dataset recorded at the ears of expert rowers using a binaural audio-capturing system. Recordings were made on-boat during rowing on a real setting. This type of audio data is particularly challenging due to the presence of multiple sources of noise.

2 State of the Art

Machine hearing techniques have been used in the classification of several activities like sport for the purpose of extraction of events [10]. The first step of the classification is typically the identification of the relevant audio features, the mel-frequency cepstral coefficient (MFCC) and MPEG-7 descriptor being the most adopted ones. For example, the work reported in [11] performs a comparison of these features. The second step of the processing is the classification itself that can be distinguished between sequence recognition based on Hidden Markov Models (HMM) [11] or static recognition like one based on Support Vector Machines (SVM) [7] or Artificial Neural Networks (ANN) [8].

Research in classification tasks has in recent years seen rapid development in the field of Deep Belief Networks. These are systems that utilize many layers of neurons and greedy learning algorithms, producing effective feature extraction in the lower layers. They have proved successful in several classification tasks, including speech and audio [5].

3 Binaural audio capturing

Rowing sound was captured using a wearable binaural recording system capable of capturing high-quality audio without disrupting the hearing and doing of the user. The system consisted of two miniature microphones (Knowles Acoustics FG23629), a custom-built microphone amplifier, power supply, and a commercial portable digital recorder (Edirol R-09). All recording were in 48 kHz and stored in WAV format with 16-bit resolution. Three rowers participated in this audio-capturing. Two rowers had 5 years of experience (R1

and R2), and the last rower had 11 years of experience (R3). For each rower, recordings were made for several velocities spanning a range between 18 and 40 strokes per minute (SPM).

4 Segmentation

Binaural recordings were manually segmented according to two classes defined by the combination of two consecutive stroke phases. One class was defined by the sequence entry-drive (ED), and the other class was defined by the sequence finish-recovery (FR).

For the selection of relevant acoustic features a correlation analysis was performed on two of the four basic spectral descriptors defined in the MPEG-7 audio content description: audio spectrum envelope and audio spectrum centroid.

5 Methods

In the classification of audio data there are several possible approaches. In speech recognition HMMs have been the standard approach for a long time. However, this is based on the well-studied fact that speech consists of a continuous series of syllables, which is very close to the HMM's basic assumptions. We can not assume rowing sounds to have the same structure. For the general recognition problem we chose from a set of machine learning approaches with weaker assumptions. Some of the top performers are Support Vector Machines (SVMs), Nearest Neighbor-based methods and Artificial Neural Networks (ANNs). For real-time application from large amounts of data nearest neighbor-based methods quickly turn impractical, while ANNs and SVMs have been proven to be mathematically identical if regularization and objective functions are matching [1].

In classification there has recently been a great interest in deep network architectures, starting from a successful application by Hinton and Salakhutdinov [3]. Deep networks were previously found to be underperforming, which is often attributed to the overly greedy objective function resulting in poor generalization. Hinton and Salakhutdinov[3] introduced the idea of pretraining on unsupervised data to improve generalization. Although originally based on stochastic Restricted Boltzman Machines (RBM) units, the concept has been generalized to various forms of greedy layer-wise training of non-parametric machine learning algorithms.

We will compare the efficiency of some of these approaches to the result of the more classic approach using shallow artificial neural networks and principal component analysis (PCA). Our aim is to find a suitable supervised algorithm for use on audio data, as well as contribute to the understanding and benchmarking of deep network architectures.

An identical series of 20 consecutive time windows of data representing a total of 850 ms were used as input to all networks. In each time window 11 features have been selected, 10 for the spectral envelope and 1 for the centroid, summing up to 220 scalars per classification point. All the computations have been performed under Mathworks MATLAB R2010b running under Windows 7 64-bit on a Intel Core i7 940 at 3.06GHz using 8 logical cores and with 6GB of memory.

We present below the different approaches investigated.

5.1 Shallow networks

The shallow networks had a single hidden layer and performed descent on the objective function using the scaled conjugate gradient. Performance in early trials converged as the amount of neurons increased. We chose 20 neurons, after which addition of neurons resulted in no increase of performance. We used a tangent sigmoid activation function, as trials with the radial basis function performed slightly worse.

Trials replacing early stopping with the more sophisticated Bayesian regularization gave no change in performance on our data. Since this regularization is impractical for the larger deep network architectures, early stopping was chosen to ease comparison.

PCA was chosen as an additional data pretreatment for comparison with the lower layers of the deep architectures. While the ANN will be a comparison for the classifying abilities, PCA will be a comparison for the feature extraction aspect of especially the lower layers in a deep network.

5.2 Deep Belief Networks

A deep belief network here refers to a network consisting of restricted Boltzmann machines. They form directed networks similar to neural networks, but use stochastic outputs similar to Boltzmann machines.

We used units with linear energy and trained them by contrastive divergence [3]. Initially the hidden layers were determined by greedy layerwise training. The resulting weights were used as initialization and followed by supervised training with backpropagation of the whole network.

Equal layer sizes of 1000 units were used. Remaining parameters were kept as in Hinton and Salakhutdinov [3].

5.3 Autoassociator Networks

The other deep network is known as a stacked autoassociator network [4].

Two types of autoassociator networks were tested: first non-linear greedy layer-wise autoencoders, each consisting of a layer of neural network. Each layer of a greedy layer-wise autoencoders is trained much like a regular shallow ANN. Each hidden layer is trained by adding an additional output layer and training it to reconstruct the input data. After training is complete, the output layer is discarded and only the encoder is kept. The output of the encoder is used as input data for training the next layer. Once all layers are trained, an output layer is added to the top and trained to mimic the final target. For the final training we allow errors to backpropagate through the earlier layers.

Because of deterministic and continuous output we can use smaller layer sizes than in the case of deep belief nets. Layers consisting of 40 neurons each were determined to be sufficiently large in this case. Early trials of layer sizes of up to 1000 neurons did not result in any significant increase of performance.

The second type of deterministic deep network is trained in a greedy backpropagating fashion. With this approach layers are added and trained one-by-one, with the target always being the training data itself. Lower layers are fed backpropagating signals and are not kept constant during training of higher layers.

After adding the final layer the whole network was trained to match the final target. Like in the greedy layer-wise approach 40 nodes were sufficient in early tests.

6 Results

All networks were trained on data sets from two rowers and evaluated on a test data set from a third rower, thus measuring the intrasubject generalization ability. Each data set included strokes from all the different velocities performed by the respective rowers. Manual classification was used as training class and for validation of the third rower.

The test set performances are represented in Table 1. Error is expressed as the fraction of misclassified samples and timing is the running time on the test computer. Fig 1 shows the training classes and classification plotted on the principal components and the same data mapped by the final nodes of an autoassociator network, where the final layer size is restricted to two nodes.

7 Conclusions

On our data none of the deep architectures outperformed the simple PCA/ANN approach. This approach is both fast and efficient in the recognition task, which makes it ideal for phase recognition. The next step will be to implement it within the reinforcement learning framework of boundary placement in real-time. Two possible approaches will have to be considered: heuristic methods and reinforcement learning algorithms. While the former approach is faster and easier to implement, the second benefits from reaching

Table 1: *Performance of different machine learning approaches.*

Method	Train error	Test error	Train time
ANN, shallow	0.39	0.40	6 min
PCA+ANN, 1	0.43	0.40	4 min
PCA+ANN, 2	0.41	0.40	4 min
PCA+ANN, 10	0.30	0.23	5 min
PCA+ANN, 210	0.28	0.21	7 min
AA, prog, 1	0.43	0.44	21 min
AA, prog, 2	0.40	0.43	23 min
AA, prog, 210	0.26	0.35	27 min
AA, greedy, 210	0.36	0.43	23 min
DBN	0.17	0.29	180 min

Number in method column is the number of principal component used for PCA or the size of the output layer for autoassociator networks. Error measurement is the fraction of misclassified test data.

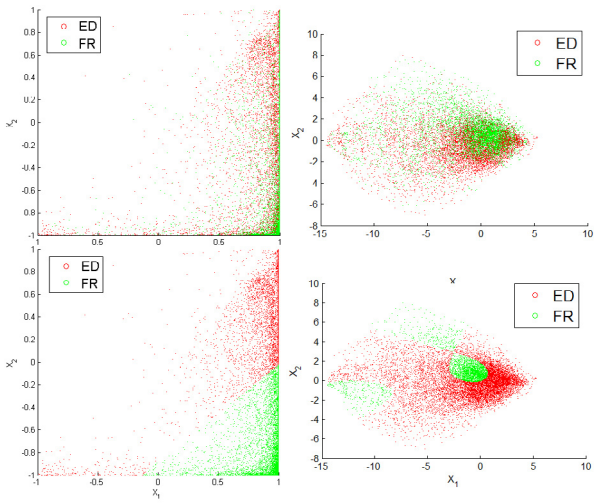

Figure 1: *Top right: True classes placed on the two dimensional output layer of a progressive autoencoder. Bottom right: Class recognition of the autoencoder in the output later. Top left: True classes placed along the two largest principal components. Bottom left: Class recognition using the two largest principal components.*

the locally optimal policy. However, the local optima of policy ascent methods can be rather poor, so the relative efficiency of this approach is still an open question.

Perhaps surprisingly, no network was able to outperform the combination of PCA with 220 components and ANN. Since the raw data vector has 220 components, the benefits are derived from the decorrelation of data rather than data reduction. It seems that the decorrelation has a positive effect on the regularization, which makes sense from the perspective that the normalized inputs are generally assumed to be uncorrelated for weight decay to be compatible with the generalized Tikhonov regularization. It seems reasonable to assume a similar relation using early stopping.

Similar results were achieved by Ballan et al. [2] on noisy sound data, where deep belief networks also performed considerably worse than shallow networks.

Future direction of this work will be to apply the same method on the larger raw Fourier series and extending the number of participating subjects.

8 Acknowledgements

This work was supported by the SKILLS Integrated Project (IST-FP6 #035005, http://www.skills-ip.eu) funded by the European Commission.

References

[1] P. Andras. The Equivalence of Support Vector Machine and Regularization Neural Networks. In *Neural Processing Letters*, 65, pages 97–104, 2002.

[2] L. Ballan, A. Bazzica, M. Bertini, A. Del Bimbo, and G. Serra. Deep networks for audio event classification in soccer videos. In *Proceedings of the 2009 IEEE international conference on Multimedia and Expo*, ICME'09, pages 474–477, Piscataway, NJ, USA, 2009. IEEE Press.

[3] G. E. Hinton and R. R. Salakhutdinov. Reducing the Dimensionality of Data with Neural Networks. *Science*, 313(5786):504–507, July 2006.

[4] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, pages 473–480, New York, NY, USA, 2007. ACM.

[5] H. Lee, Y. Largman, P. Pham, and A. Y. Ng. Unsupervised feature learning for audio classification using convolutional deep belief networks. *Science*, 22:1–9, 2009.

[6] V. Lippens. Inside the rower's mind. *Rowing faster. Human Kinetics, Inc*, pages 185–194, 2005.

[7] L. Lu, H. Zhang, and S. Li. Content-based audio classification and segmentation by using support vector machines. *Multimedia Systems*, 8(6):482–492, 2003.

[8] S. Rein, M. Reisslein, and T. Sikora. Audio content description with wavelets and neural nets. In *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*, volume 4, pages iv–341. IEEE.

[9] E. Ruffaldi, A. Filippeschi, C. A. Avizzano, and M. Bergamasco. Skill modeling and feedback design for training rowing with virtual environments. In D. Kaber and G. Boy, editors, *Proceedings of 3rd Conference on Human Factors and Ergonomics*, number 978-1-4398-3491-6, pages 832–841. CRC Press / Taylor & Francis, Ltd., 2010.

[10] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang. Audio events detection based highlights extraction from baseball, golf and soccer games in a unified framework. In *icme*, pages 401–404. IEEE, 2003.

[11] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. Huang. Comparing MFCC and MPEG-7 audio features for feature extraction, maximum likelihood HMM and entropic prior HMM for sports audio classification. In *icme*, pages 397–400. IEEE, 2003.