Mixture of experts for audio classification: an application to male female classification and musical genre recognition

Hadi Harb, Liming Chen, Jean-Yves Auloge

Mixture of experts for audio classification: an application to male female classification and musical genre recognition

Hadi Harb, Liming Chen, Jean-Yves Auloge
LIRIS CNRS FRE 2672 Lab.
Dept. Mathematiques Informatique, Ecole Centrale de Lyon
{hadi.harb, liming.chen, Jean-Yves.Auloge}@ec-lyon.fr

Abstract

In this paper we report the experimental results obtained when applying a mixture of experts to the problem of audio classification for multimedia applications. The mixture of experts is based on neural networks as individual experts and the Piecewise Gaussian Modeling was used for audio signal representation. Experimental results on two audio classification problems, male/female classification and musical genre recognition, show a clear improvement of using a mixture of experts in comparison to one individual expert.

1. Introduction

Several audio classification problems necessitate a large amount of training data in order to obtain acceptable performance by an audio classifier. The need for a large amount of learning data may complicate the learning process of the classifier, especially if the neural networks with error back propagation training algorithm are used. One way to overcome this complexity is to divide the classification problem into a set of simpler sub-problems and to train a set of experts on the obtained sub-problems. The combination of these experts constitutes the mixture of experts. In this paper we show that the use of a mixture of audio experts may simplify the training process while improving the classification results. Two audio classification problems are studied, male/female classification and musical genre recognition.

2. PGM-MLP audio experts

An expert in this paper is an audio classifier that estimates the *a posteriori* probabilities of an audio class given an acoustic observation. The audio classifier used in this paper is a Piecewise Gaussian Model – Multi Layer Perceptron (PGM-MLP) classifier [2]. A short overview of the PGM-MLP audio classifier will be presented in this section.

Audio classification into concepts, or semantic classes, is a problem where humans can easily surpass today's machines. It can be argued that modelling some aspects of the human perception of audio semantic classes when developing audio classifiers may lead to an improvement over a solution to the problem from an engineering point of view alone. The PGM-MLP classifier is based on perceptually motivated features in an attempt to include our conclusions from psychoacoustic experiments.

In order to model some effects observed from psychoacoustic experiments on the ability of human subjects to classify stimuli into predefined semantic concepts, we proposed that the classification of a stimulus is based on the current acoustic observation and its relation to the context (past acoustic observations). An auditory memory model is then proposed where we suppose that the classification of a stimulus is based on the refreshed memory and not on the actual acoustic observation with no relation to the past. One special case of the auditory memory model is called the Piecewise Gaussian Modelling (PGM).

Let $S(i)=\{X_1,...X_N\}$ be the set of acoustic vectors, Mel Frequency Spectral Coefficients (MFSC) for instance, for class C_i. The index of the acoustic vectors is the time index. That is, X_j and X_{j+1} are consecutive vectors in time. The PGM consists of modelling each set of T consecutive acoustic vectors contained in a time window called the Integration Time Window (ITW) by one Gaussian Model. $X_j,...X_{j+T} \rightarrow G(\mu, v)$. The Gaussian model is expressed by its mean vector and diagonal covariance matrix. Therefore $S(i)$ is modelled by a set of Gaussians. The duration of ITW is typically 100, corresponding to 1 second, and an overlap between consecutive windows is allowed.

The normalized concatenation of the mean and variance vectors for each Gaussian constitutes the

Midlevel feature vector characterising the respective *ITW* window.

The midlevel features, called PGM-features, are used as the input of a Multi Layer Perceptron (MLP). The MLP is trained on the PGM features to describe the *ITW* windows. The number of input neurons of the neural network is equal to the elements of the PGM-feature vector and the number of output neurons is equal to the number of classes. Therefore, the PGM features are the basic features used for training and for testing. And the audio classifier is called PGM-MLP classifier.

3. Mixture of PGM-MLP experts

The use of a mixture of experts consists of splitting a classification problem into several simpler sub-problems where individual experts are trained and the combination of the trained individual experts constitutes the mixture of experts, Figure 1. The use of mixture of experts has been shown advantageous for several classification problems, especially in the image analysis literature [1].

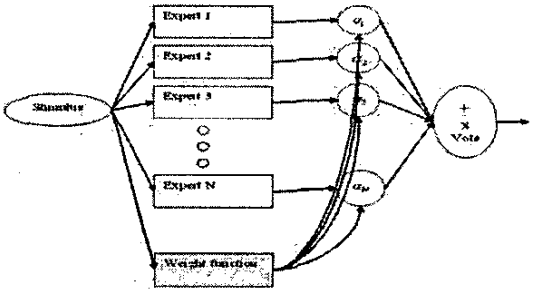

Figure 1 A general architecture of a mixture of experts

When designing a mixture of experts, several parameters that seriously affect the performance need to be set. The first parameter to be set is normally the number of experts. Second, the function used to combine the individual experts' outputs need to be suitably chosen for the task in hand. We consider in this work that a mixture of experts is accurate if its components, namely the individual experts, are accurate. Therefore, the choice of the number of experts is based on two constraints: 1- the accuracy of an individual expert has to be relatively acceptable for the classification problem, and 2- the expert has to be trained on as little data as possible. This clearly requires a validation dataset and a training dataset. The amount of training data for each individual expert is increased continuously and the performance of the trained expert is evaluated on the validation dataset. A compromise between the training time and the accuracy of the trained experts permits an optimal choice of the number of experts, or indirectly the amount of training data for each individual expert.

When combining the outputs of individual experts, the decision of each individual expert is weighted. Several combination functions can be used, such as the sum, the multiplication or a majority vote. From our experiments we found that the sum and the multiplication functions perform equally while surpassing the majority vote. Consequently the combination function used is the sum function. However, the choice of the weighting scheme is important in this case.

Let $A = \{a_1, a_2, ..., a_M\}$ be the training data set with $A = A_1 \cup A_2 \cup A_3 \ldots A_N$, where N corresponds to the number of experts and A_i is the training data set for the i^{th} expert, E_i. The output of the mixture of experts is obtained by $P(X/Cj) = \sum_{i=1}^{N} \alpha_i P_{Ei}(X/Cj)$ with α_i is the weight of the i^{th} expert, E_i. Cj corresponds to the label of the class

3.1. Choosing weights

The choice of the weights for the outputs of the experts can be done in a simple manner such as giving equal weights, or by using more complex methods. In this work we compared several strategies for the choice of the weights, using a gate network, using the novelty between the experts' experience, and using an estimation of the error rate of the experts.

3.1.1. Gate network. When a gate network is used, the weights α_i correspond to the outputs of an expert trained to classify an acoustic observation into A_i, the training data sets for the individual experts. The output of the mixture of experts becomes $P(X/Cj) = \sum_{i=1}^{N} P(X/A_i) P_{Ei}(X/Cj)$.

3.1.2. Novelty of experts. An expert E_i with outputs correlated with the outputs of other experts of the mixture will probably provide less novel experience to the mixture than an expert E_i which outputs are less correlated with those of the other experts. That is, we propose to give less weight to the experts that do not bring a novelty to the mixture of experts than to those providing novel experience. To measure the novelty brought by expert E_i given expert E_j, we train E_i on the

training data of E_j. The training time, or the number of epochs, is an indicator on the novelty.

Let $T(E_i, E_j)$ be the training time needed to train expert E_i on the training data of the E_j expert. The weight given to the expert E_i is proportional to the sum of the novelties, training times, brought given the all other experts. $\alpha_i = \sum_{j=1} T(E_i, E_j)$

3.1.3. Using error estimation. It can be argued that an expert providing less error rate on a development dataset may be more efficient on the test data in a real world application than an expert providing a high error rate. The weight chosen for the expert E_i is inversely proportional to its error rate on a development data.

Recall that one important reason for using a mixture of experts is simplifying the learning process. We studied empirically the effect of using different weighting schemes of a mixture of experts on the overall training time as on the accuracy of the classifier. The male/female classification problem is used for this task. The database used for the choice of an optimal configuration of the mixture of experts contains radio recordings in French and British English. The signal was compressed at a low compression ratio, namely mp3 8Kbps. The training data corresponds to 600 seconds of male speech and 600 seconds of female speech while the test data contains 1000 seconds of male speech and 1000 of female speech. The training data was clustered in a sequential manner; hence no unsupervised clustering algorithm was used. The use of an unsupervised clustering algorithm is necessary when the training and/or the testing data contains recordings from different sources. Each source will probably constitute a cluster in the feature space. The data used in this paper is relatively homogenous; therefore the use of unsupervised clustering algorithms will not probably improve the performance. The number of experts is set to 3.

Table 1 The effect of the weighting scheme on the accuracy and the training complexity of a mixture of experts

	Equal	Novelty	Error	Gate
Training	Fast	Slow	Slow	V. Slow
Accuracy %	81.3	83.5	83.6	78.7

As it is shown in Table 1, choosing the weights of the experts in accordance to their novelty or to an estimation of their performance improves the accuracy of a mixture of experts in comparison to the use of equal weights. However, in our application where we can compromise some decrease in the overall performance if the learning process can be simplified, the choice of the simplest solution of equal weights is motivated. Notice that using equal weights permits an incremental learning where a new expert is trained on newly available training data and then plugged into the mixture of experts. This property is particularly important in practice. Surprisingly, the use of a Gate network shows relatively low accuracy rate. It is possible that these results are due to the inexistence of clear homogenous clusters for the training data in the feature space. Therefore, the gate network's role is not well defined. This role would be defined better if the training/testing data contains recordings from different sources and hence presenting homogenous clusters in the feature space.

4. Experimental Results

The use of a mixture of PGM-MLP experts was compared experimentally to the use of one individual PGM-MLP expert on two classification problems, male/female classification and musical genre recognition. The chosen problems generally require a large amount of training data that may be heterogeneous. For example, the training data of a male/female classifier will easily include thousands of seconds of speech from different languages containing a variety of capture conditions.

In both experiments the sum function and equal weights were used for the combination of the individual experts' outputs, and hence simplifying the training process. Sequential segmentation of the audio material was performed since in this case the audio data that are close in time constitute a homogeneous cluster in the feature space. It is supposed that an unsupervised clustering algorithm such as a K-means algorithm would provide a near-sequential segmentation of the audio data.

4.1. Male/female classification

The data used in this experiment was obtained from the switchboard database. The choice of the switchboard database which corresponds to recordings from the telephone network in the US English was mainly made in order to make a fair comparison to existing techniques developed for gender identification. It is also important to notice that the classification accuracies presented in the section 3 were estimated on a highly noisy speech material obtained at a low compression ratio (8 Kbps).

The training data consists of 1000 seconds of male speech and 1000 seconds of female speech obtained from 9 male speakers and 9 female speakers respectively. The test data consists of 1000 seconds of male speech and 1000 seconds of female speech obtained from 10 male and 10 female speakers.

We show in Table 2 the classification accuracies of a mixture of 3 experts for two time precisions, 1 and 5 seconds. In the first case the decision is made on the ITW window and hence for 1 second windows, while in the second case the decision is averaged for every 5 consecutive ITW windows in order to smooth the results over durations of 5 seconds. It is necessary to estimate the accuracy of the proposed classifier ruming on 5 seconds chunks of the data in order to make a fair comparison to existing systems in the literature. The classification accuracy of one single expert trained on the totality of the training data is also shown in the table. As we can see on the table, the use of a mixture of experts slightly improves the performance in comparison to the use of one single expert while simplifying considerably the training process. Another conclusion from the table is that with low time precisions, the accuracy is considerably improved. The classification accuracy obtained by a mixture of PGM-MLP experts As compared to the best accuracies reported in the literature, namely 97 % for 5 seconds segments in [3], is clearly acceptable.

Table 2 Accuracies in % of a mixture of experts and one expert trained on the totality of the training data for 1 s and 5 s time precisions

	Male	Female	Average
MoE(3)1 s	96	90	93.0
MoE(3)5 s	99	95	97.5
Expert 1 s	95	90	92.5
Expert 5 s	97	97	97.0

4.2. Musical genre classification

We collected a database from online radio stations containing 6 musical genres including: Metal, Hip Hop, New Metal, Smooth Jazz, Soft Pop Rock, and Disco. For each genre 2400 seconds are available, 800 seconds were selected for the training process and the rest of the data (1600 seconds) were used for testing. Two mixtures of experts were used in this experiment, one mixture containing 2 experts and another containing 3 experts. We show the classification accuracies of the mixtures of experts using 2 and 3 experts and that of one expert trained on the totality of the training data in Table 3. The improvement in this case is considerable.

Table 3 Accuracies of a mixture of experts and one expert trained on the totality of the training data

	Expert	MoE (2)	MoE (3)
Me	44.0	43.5	55.8
HH	24.8	71.8	59.5
NM	47.0	55.8	58.8
SJ	75.5	73.5	75.5
SPR	58.8	54.0	59.3
Di	59.0	52.8	60.5
Average	51.5	58.6	61.6

The improvement may be due to the great variability of the audio signal within each of the classes, musical genres for instance, making that the individual experts specialize well on their training data and hence minimizing the correlation between the errors of different individual experts. However, with a time precision of 4 seconds, the performance of the mixture of PGM-MLP experts is noticeably better than those reported in [4].

5. Conclusion

This paper presented the use of a mixture of PGM-MLP audio classifiers for the problems of audio classification necessitating a big amount of training data. It was empirically shown that for the cases where no clear homogenous clusters of the training data exist in the feature space, the use of a simple weighting scheme is advantageous. The mixture of PGM-MLP experts was used for male/female classification and musical genre recognition. For both problems, the mixture of experts improves the performance and simplifies the learning process in comparison to the use of one single expert trained on the totality of the training data.

6. References

[1]. Gutta S., Huang J. R. J., Jonathon P., Wechsler H., *Mixture of Experts for Classification of Gender, Ethnic Origin, and Pose of Human Faces*, IEEE Transactions on Neural Networks, VOL. 11., NO. 4, pp 948-960, July 2000

[2]. Hadi Harb, Liming Chen, *Robust Speech Music Classification Using Spectrum's First Order Statistics and Neural Networks*, Proceedings of the IEEE ISSPA2003, July 1-4, Paris – France, 2003

[3]. Parris E. S., Carey M. J., *Language Independent Gender Identification*, Proceedings of the IEEE, ICASSP 96, pp 685-688, 1996

[4]. Tzanetakis G., Cook P. *Musical genre classification of audio signals* IEEE Transactions on Speech and Audio Processing, vol. 10, no. 5, July 2002