Pattern Recognition Letters xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

A learning approach to hierarchical feature selection and aggregation for audio classification

Paul Ruvolo [a,*], Ian Fasel [b], Javier R. Movellan [a]

[a] Machine Perception Laboratory, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA
[b] University of Arizona, Department of Computer Science, P.O. Box 210077, Tucson, AZ 85721-0077, USA

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Temporal modeling
Feature aggregation
Audio classification
Feature selection

ABSTRACT

Audio classification typically involves feeding a fixed set of low-level features to a machine learning method, then performing feature aggregation before or after learning. Instead, we jointly learn a selection and hierarchical temporal aggregation of features, achieving significant performance gains.

1. Introduction

Recognition of audio categories has become an active area of research in both the machine perception and robotics communities. Example problems of interest include recognition of emotion in the user's voice (Petrushin, 1999), music genre classification (Tzanetakis et al., 2001), language identification, ambient environment identification (Chu et al., 2006), and person identification.

The standard approach to auditory category recognition involves extracting acoustic features at short time scales, and then classifying longer intervals using summary statistics of the feature outputs across the longer intervals. This approach requires making two choices: (1) the set of low-level features and (2) the summary statistics. These two choices are related because the optimal statistic for describing the temporal distribution of a feature may depend on the characteristics of the feature itself. The best choices may be highly dependent on the particular problem and data being analyzed.

In this paper, we present a machine learning approach for making these two choices in a data-driven fashion, i.e. to learn an optimal set of low-level features *and* methods for aggregating these features across multiple time scales, *jointly*. In the proposed approach, learning is performed on a class of features that differ in their short time scale, medium time scale, and long time scale characteristics. By employing such a flexible set of features the learning process is more free (in comparison to typical approaches, see Section 1.1) to adapt to the characteristics of the task at any particular time scale, rather than being burdened by a suboptimal choice by the system designer. Short time scale spectral features are extracted over windows in the time scale of tens of milliseconds, relative contrast information from nearby spectral bands are combined at time scales of tenths of seconds, and long-term statistics of modulations in spectral contrast are combined at time scales of several seconds. The proposed approach is general purpose and can be applied to a wide range of audio classification tasks. Once training data has been collected, little additional effort is needed to generate a classifier. We test the approach on a variety of tasks and show that the proposed method achieves results comparable or superior to the state-of-the-art approaches that have been previously developed for each of the specific tasks.

1.1. Background and related work

Audio category recognition typically starts with the extraction of short time scale acoustic features using windows in the tens to a few hundred milliseconds range, such as fast Fourier transform coefficients (FFTC), discrete wavelet transform coefficients (DWT) (Mallat, 1999), Mel-frequency cepstral coefficients (MFCC) (Junqua and Haton, 1996), real cepstral coefficients (RECC) (Gold and Morgan, 2000), or MPEG7 low-level descriptors (e.g. spectral flatness) (Ntalampiras et al., 2008). In addition to these general purpose features, more specialized features have also been proposed to capture key perceptual dimensions of the audio signal. The literature is quite vast and several reviews are available (see e.g. Aucouturier and Pachet (2003) for a review, and McKinney and Breebaart (2003) for an experimental comparison of features for music retrieval). Some examples include features based on models of the human auditory system (e.g. Gammatone filters, see Glasberg and Moore, 1990; Hartmann, 1997), psychoacoustic features such as

* Corresponding author. Tel.: +1 650 279 8868; fax: +1 858 822 5242.
E-mail address: pruvolo@cs.ucsd.edu (P. Ruvolo).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2009.12.036

roughness (Daniel and Weber, 1997), sharpness (von Bismarck, 1974), pitch, amplitude and brightness (Wold et al., 1996), and music specific features such as beat-tracking (Scheirer, 1998).

Once a set of low-level features has been extracted, there are many ways to combine features over time. One approach is to first classify or model the short time features given the class labels, then combine these in a "bag of features" manner. For instance, (Barrington et al., 2007) model the vector of features at each point in time as independently generated from a Gaussian mixture model (GMM). The class-conditional probability of a longer time series is then the product of the individual feature vectors' class-conditional probabilities. Another example of this type of approach is to train a discriminative classification model on the short time scale features e.g. a Support Vector Machine (SVM) (Vapnik, 1995), then individual feature vector classifications are combined with a vote.

Another common approach is to compute various summary statistics of the short time scale features over the duration of an audio clip and use these summary statistics as input to a classifier. The salient difference from the previous class of methods is that aggregation is performed before rather than after learning. For example, to perform emotion recognition, Grimm et al. (2007) computed the mean, standard deviation, 25% and 75% quartiles, difference between minimum and maximum, and difference of quartiles of estimates of pitch, speaking rate, intensity, and MFCCs over the entire speech segment. The vector of summary statistics was then mapped into a continuous, three dimensional emotion space using a fuzzy logic system, and finally K-nearest neighbors (KNN) was used to classify segments into seven basic emotion categories. A more unusual type of temporal aggregation, proposed in Deshpande et al. (2001), involved extracting a 20 s sample from a song, converting the audio to an MFCC time–frequency "image", and applying a set of recursive image-texture features (originally developed for image retrieval by Bonet and Viola (1997)) to extract a 15,625 element feature vector, which is then classified with KNN and SVMs.

An intermediate approach is to perform aggregation of time segments within a medium-scale time-window (of e.g. several seconds) and then perform classification at the window level (Tzanetakis et al., 2001). The classification result of multiple segments are then combined with a vote. A systematic comparison of the effect of window length and the feature type using the window-voting approach was performed by Bergstra et al. (2006). Similar to Grimm et al. (2007) and others, mean values for each feature in the time window were first computed and then fed into a classifier. The authors settled on time windows of about 2–5 s classified with AdaBoost. Although the learning method was restricted to combining features from one feature set over a window (i.e. learning was not involved in long time scale aggregation), this method was able to win first prize in genre classification and second prize in artist classification at the 2005 MIREX (Music Information Retrieval Evaluation eXchange) contest.

In this paper, we use a machine learning algorithm to simultaneously solve the problem of selecting the class of short time scale features and performing aggregation and classification over multiple time scales. We do so by defining a novel set of features, called Spectro-Temporal Box-Filters (STBFs), that include in their parameterization both the low-level feature space and the medium and long time scale aggregation. STBFs are capable of capturing ambient, transient, and periodic signals over medium and long time scales. The learning method we use, GentleBoost (Friedman et al., 2000), sequentially selects STBFs according to a classification criterion, thereby jointly optimizing the feature type and multiscale aggregation method for the specific problem at hand.

Our choice of representation at the medium and long time scales extends and parameterizes many of the features previously used for audio pattern recognition based on correlation/ derivation of local spectro-temporal patterns (see e.g. Abe and Nishiguchi, 2002). In (Casagrande et al., 2005a) a boosting technique was used to learn local spectral patterns similar to one of the features we define here, however, they did not employ learning to aggregate across time. By combining short time scale feature extraction and temporal aggregation into a joint parameterization, we implicitly define a family of several millions of spectral-temporal features. A similar idea for formalizing very large sets of audio features has been explored by Pachet and Roy (2007), who proposed a variety of low-level analytic operators and a genetic-algorithm method for learning arbitrary compositions of these features, thereby defining a set of billions of candidate features.

This paper extends our previous work on STBFs (Ruvolo et al., 2008; Ruvolo and Movellan, 2008), by proposing a hierarchical approach to combine features at multiple time scales. The results suggest that using machine learning framework to jointly select from a rich class of features and of aggregation methods can result in dramatic performance gains for a wide range of problems.

2. Spectro-temporal box filters

Fig. 2 shows a graphical representation of an STBF indicating the three time scales at which information is captured. Each STBF is parameterized by:

1. *Time scale: tens of milliseconds*. A set of low-level features that are extracted over short time scales (e.g. Mel Frequency Cepstral Coefficients or Short Time Fourier Transform Coefficients).
2. *Time scale: hundreds of milliseconds*. A box-filter that serves to summarize local responses of the low-level features in order to model the intermediate temporal dynamics.
3. *Time scale: seconds*. A periodic sampling scheme and summary statistic that aggregates the responses of the intermediate-level temporal time scale.

Fig. 1 describes the steps involved in learning and applying a classifier. First the auditory signal is preprocessed and the short time scale feature channels are extracted. Next, a bank of STBFs (learned using GentleBoost) are applied and the outputs combined to make a binary classification. For multiclass problems, a set of binary classifiers (one for each possible non-empty subset of classes versus the rest) are trained, and the output of these classifiers are combined into a single n-category classifier.

The specifics of each of the models at each of these time scales are given in the next three subsections. Each subsection fills in the details of one particular level of the overall system architecture presented in Fig. 6. The description begins from the bottom level of this architecture diagram and proceeds upwards.

2.1. Short time scale features

While a wealth of short time scale audio feature descriptors have been proposed in the literature (e.g. Gold and Morgan, 2000; Junqua and Haton, 1996; Wold et al., 1996 and others as described in Section 1.1), there is little consensus on what types of features are best for various different tasks. Rather than attempting to guess the best set beforehand, we allow the machine learning method to select from a large set of possible features. For the experiments in this document we use Mel Frequency Cepstral Coefficients (MFCCs), Sones (Fastl and Zwicker, 1990), and Linear Predictive Cepstral Coefficients (LPCCs), however, there is nothing in our algorithm to prevent additional low-level features from being added.

For each of these short time scale features, the duration of the time windows over which they are computed can impact perfor-

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036

Train-Time Algorithm

1. Compute low-level feature channels over short time scales (see Figure 3).
2. Use GentleBoost to choose a set of Spectro-Temporal Box Filters to solve multiple binary classification problems.
3. Combine the output of the binary classifiers using multinomial logistic regression to produce an n-category classifier.

Run-Time Algorithm

1. Compute low-level feature channels over short time scales (see Figure 3).
2. Apply bank of Spectro-Temporal Box Filters selected during the training process.
3. Combine output of the filters into binary classifiers.
4. Combine output of binary classifiers into an n-category classifier.

Fig. 1. General description of the approach at train-time and run-time.

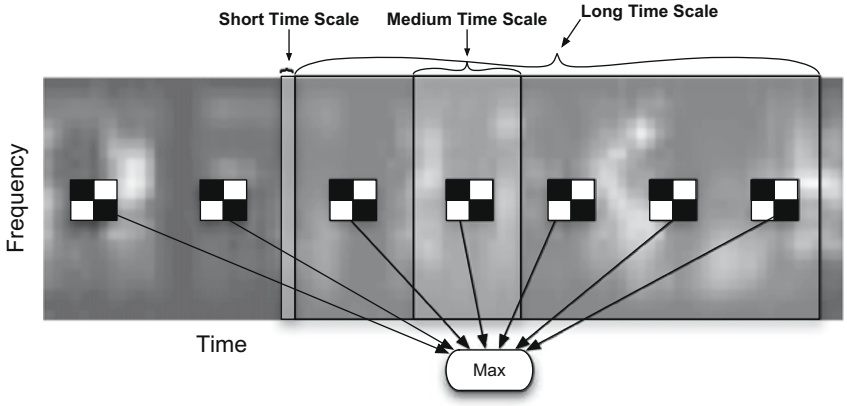

Fig. 2. A spectro-temporal box filter. An STBF combines information over three different time scales. The shor -time scale corresponds to the temporal window of the low-level features. The intermediate time scale corresponds to the box-like kernel that computes local changes in the low-level feature channels. The long time scale consists of a summary statistic of the intermediate time scale outputs over a longer window.

mance. For instance, work on speech recognition has demonstrated the benefit of including low-level feature descriptors that operate over multiple time scales (Tyagi and Bourlard, 2003). Again, rather than forcing a choice *a priori*, we allow the learning method to select which time scales are most appropriate for the task. For the experiments performed in this document we included Sone features extracted over three time scales corresponding to 32 ms, 64 ms, and 128 ms of audio per feature. The two other low-level feature channels, MFCCs and LPCCs, were only included at the 32 ms time scale.

Each of these low-level feature descriptors is represented as a two-dimensional map. The treatment of the individual feature channels over time as a two-dimensional map is appropriate given that the low-level channels for a particular feature type have a logical ordering (for instance MFCCs that act on neighboring frequency bands) and thus it is natural to represent the extracted features over the duration of the audio segment as a map where one dimension is time and the other dimension is a particular feature channel. Fig. 3 shows a schematic of the low-level feature extraction process. In this case the raw PCM signal is processed into five two-dimensional feature maps.

2.2. Medium time scale features

STBFs attempt to characterize medium time scale auditory structure by computing local temporal statistics of the short time scale features. These medium time scale models are represented by box-like kernels that compute both temporal and feature channel derivatives. Box-filters (McDonnell, 1981; Shen and Castan, 1985; Heckbert, 1986) are characterized by rectangular, box-like kernels, a property that makes their implementation in digital computers very efficient. Their main advantage over other filtering approaches, such as those involving Fourier Transforms, is apparent when non-shift variant filtering operations are required (Heck-

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036

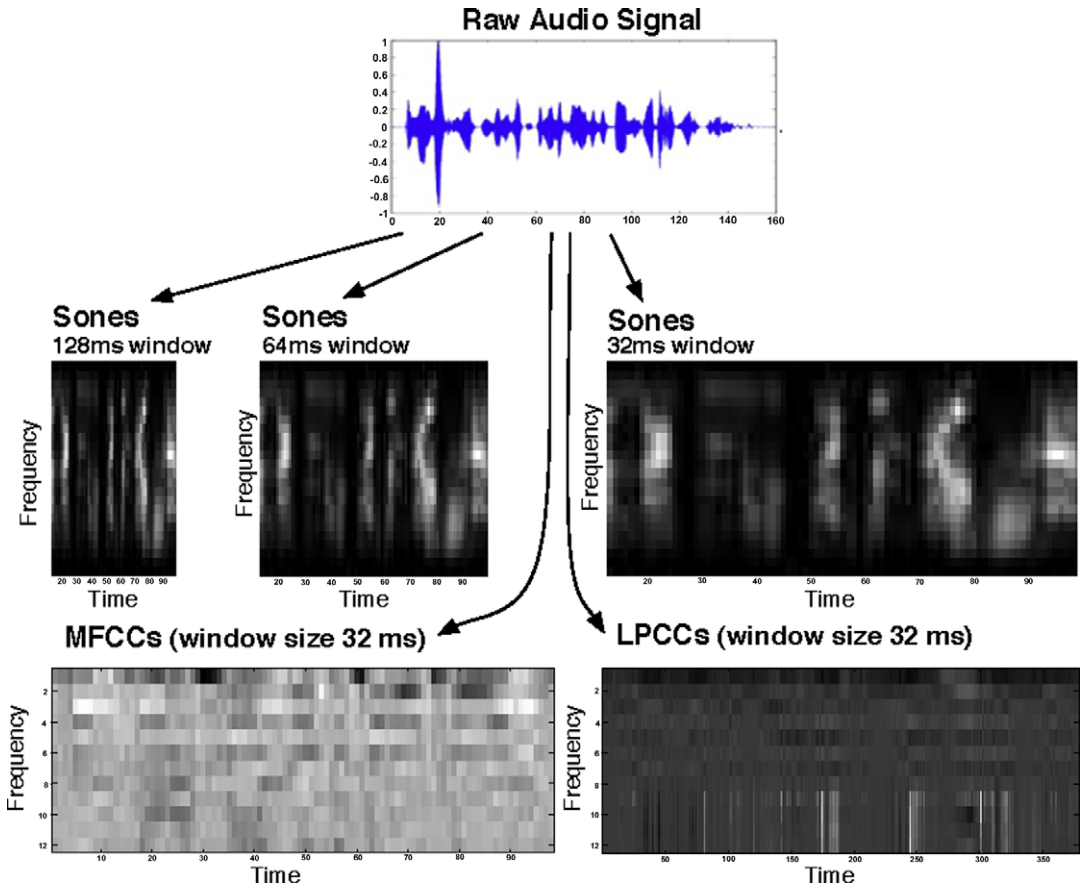

Fig. 3. Top: the original 1-d temporal audio signal. Middle: the Sone feature extracted over 3 different window lengths (32 ms, 64 ms, and 128 ms). Bottom left: MFCC features extracted over 32 ms windows. Bottom tight: LPCCs extracted over 32 ms windows.

bert, 1986). Box-filters became popular in the computer graphics community (McDonnell, 1981; Shen and Castan, 1985; Heckbert, 1986) and have recently become one of the most popular features used in machine learning approaches to computer vision (Viola and Jones, 2004). They also have been proposed previously as a method for capturing medium time scale structure in audio (Casagrande et al., 2005b).

In our work we use six types of box filters. The particular types of box-filters (see Fig. 4) are taken directly form the computer vision literature (Viola and Jones, 2004). This is an extension over the previous work of (Casagrande et al., 2005a) in which only two types of box-filters were utilized. The filter response of a box filter to a feature map is given by the sum of the feature channel values in the white regions minus the sum of the feature channel values in the black regions. A motivation for this particular choice of box filters in the domain of audio is that they unify many previously proposed mid-level audio descriptors (such as computing temporal derivatives of spectral energy) while providing a large number of new intermediate time scale features. For instance, while temporal energy derivatives are quite ubiquitous in the audio classification literature, the center surround filter in Fig. 4 computes a statistic that is quite novel in the field of audio classification (Fig. 5).

2.3. Long time scale features

Past work (Bergstra et al., 2006) on aggregating features over long time scales has shown that using simple summary statistics (such as mean and standard deviation) over long time windows can increase performance over directly classifying the short time scale features. In our work we use a similar approach, but instead of summarizing the low-level feature responses using a collection of statistics, we summarize the outputs of the mid-time scale box-filters (see Section 2.2). Also, since we provide a much richer class of long time scale models that can be selected from during learning, we do not have to commit ourselves to a particular feature summarization method, but can let the learning algorithm adaptively choose the summary statistics (which may be different for each different feature) that work best for the audio category in question.

Fig. 4. The six box-filter kernels used for medium time scale temporal modeling. These box-filters compute local frequency and temporal derivatives when applied to low-level feature maps (see Section 2.1)

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036

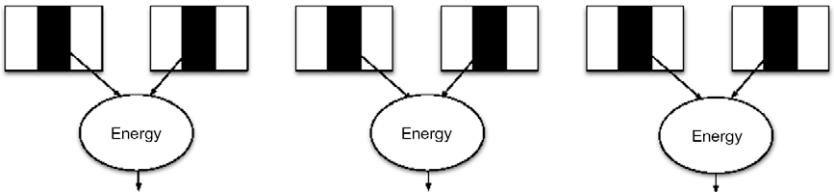

Fig. 5. An "Energy" STBF. Each pair of subsequent box-filter outputs, x and y is combined $\left(\sqrt{x^2+y^2}\right)$ to produce a feature output that is then fed into the summary statistic.

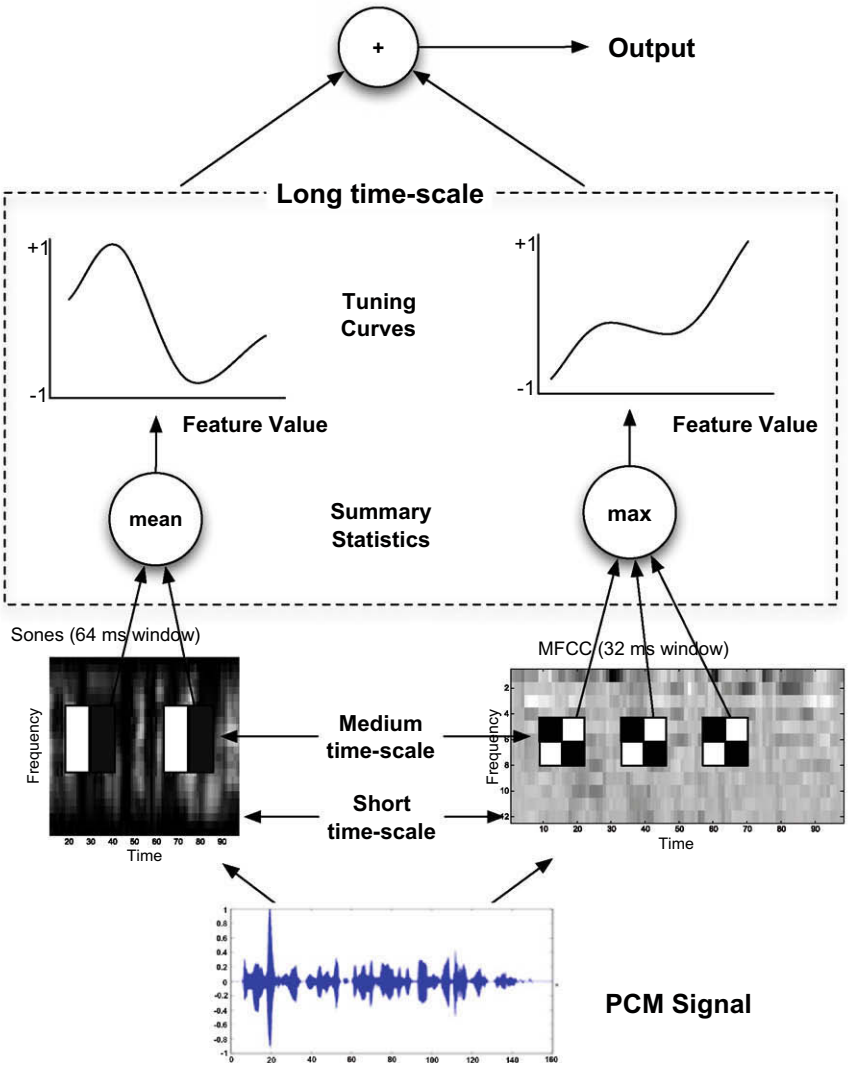

Fig. 6. A schematic of an example two-feature STBF classifier. The classifier converts the raw pcm data to two low-level feature representations, applies box-filters to extract intermediate time scale dynamics, applies a summary statistic to each sequence of feature outputs, passes the resulting values through non-linear tuning curves, and finally combines the results over all features additively. Note that learning is involved in every stage of this process.

Our temporal models at long time scales are defined by a number of parameters. The following is a summary of the individual dimensions of variation:

1. *Phase*: at what position in time the first mid-level feature response is sampled.
2. *Sampling interval:* how often to sample the mid-level feature response in time. This periodic sampling is designed to capture properties such as beat, rhythm, and cadence.
3. *Moment*: either use the raw outputs of the mid-level features or use the squared deviation from the mean feature response of the mid-level feature over the entire segment.
4. *Energy*: filters can be characterized by an "Energy" filter or the raw box-filter output. In our work we compute an "Energy" value for our intermediate features that is inspired by the computation of the power spectrum in Fourier Analysis (Bloomfield, 2000). An "Energy" value is produced by aggregating across the output of two box-filters that are separated by half a sampling interval. The energy of a filter is given by $\sqrt{a^2+b^2}$ where a and b are the filter outputs of a medium time scale feature and the same feature shifted by half a sampling interval.
5. *Summary statistic:* the summary statistics are applied to the sequence of mid-level feature outputs, producing the final output of the feature. The summary statistics considered in this

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036

work are all possible quantile values (in the interval [0,1]) as well as the mean. Since it would be impossible to exhaustively search this infinite set of quantiles, specific quantile values are sampled uniformly at random during learning. Each of these summary statistics can be seen as a method of converting local evidence from the mid-level features to an estimate over a longer time scale.

In concert, these individual dimensions of the long time scale temporal model can capture a wide range of acoustic phenomena. For instance, it may be the case that a particular intermediate-level time feature captures a salient characteristic of an auditory category (for instance the beat of a bass drum). By using temporal models with various sampling intervals in combination with this particular intermediate-level feature a classifier could distinguish between music that contains sporadic base drum beats or a sustained beat throughout the composition.

2.4. Feature selection and learning

We use GentleBoost (Friedman et al., 2000) to construct a classifier that combines a subset of all possible STBFs. Where each STBF includes both a particular short time scale, medium time scale, and long time scale model. GentleBoost is a popular method for sequential maximum likelihood estimation and feature selection. At each round of boosting, a transfer function, or "tuning curve", is constructed for each STBF which maps feature response to a real number in $[-1, 1]$. Each tuning curve is computed using non-parametric regression methods to be the optimal tuning curve for the corresponding STBF at this round of boosting (see Fasel et al., 2005 for details). The feature plus tuning curve that yields the best improvement in the GentleBoost loss function is then added into the ensemble, and the process repeats until performance no longer improves on a holdout set. In this way, GentleBoost simultaneously builds a classifier and selects a subset of good STBFs.

At each round of boosting, an optimal tuning curve (see Fig. 6) is constructed and training loss is computed for each feature under consideration for being added to the ensemble. To speed up search for the best feature to add (since brute-force search through all possible features would be very expensive) we employ a search procedure known as Tabu Search (Glover and Laguna, 1997). Tabu search is a method of stochastic local search that is very similar to a genetic algorithm. First, a random set of n filters are selected and evaluated on the training set, and are used to initialize the "tabu list" of filters already evaluated in this round. The top $k \leqslant n$ of these filters are then used as the starting points for a series of local searches. From each starting filter, a set of new candidate filters are generated by replicating the filter and slightly modifying its parameters (sampling interval, phase, etc.). If any of these features are not already in the tabu list, they are evaluated and then added to the list. If the best feature from this set improves the loss, it is retained and the local search is repeated until a local optimum is reached. After the local search has been completed for each of the initial k best features, the feature and tuning curve which achieved the greatest reduction in the loss function is added into the ensemble classifier.

With this method, the amount of time needed to train a classifier scales linearly with the number of examples. On a computer with a 2.66 GHz Dual-Core Intel Xeon processor it takes approximately 1 h to train a classifier on a dataset of audio that is roughly 40 min in length.

3. Evaluation

We performed experiments on two standard datasets and on a new dataset collected from an early childhood education center. To assess whether or not the hierarchical temporal modeling presented in this document gains us anything over the more simplistic schemes, we compared the approach proposed here with two other popular approaches. The first aggregates low-level features over longer time scales by computing means and standard deviations from individual feature channels, as in Tzanetakis et al. (2001) and others, and then feeding the resulting aggregated features into a Support Vector Machine. In each experiment a series of timbral features were computed (MFCCs, LPCCs, zero crossing rate, spectral centroid, spectral rolloff). We refer to this approach as "Simple Summary". The second approach, due to Casagrande et al. (2005b), is similar to ours in that it uses similar features and learning algorithms (box-filters applied to spectrograms, and AdaBoost (Freund and Schapire, 1996)). However, Casagrande's approach lacks both the integration across multiple time scales and the diversity of intermediate time-scale features, which are both key aspects of our method. We refer to this approach as "Intermediate Aggregation".

3.1. Recognition of emotion from speech: Berlin dataset

The Berlin Emotional Database (Burkhardt et al., 2005) consists of acted emotion from five female and five male German actors. Each utterance in the database was classified by human labelers into seven emotional categories: anger, boredom, disgust, fear, joy, neutral, and sadness. Five long utterances and five short utterances are given by each speaker for each of seven emotional categories. Speech samples that are correctly classified by at least 80% of the human labelers and classified by 60% of labelers as being natural were selected for training and testing.

To ensure speaker independence, we performed 10-fold leave one speaker out cross validation. That is we trained our system 10 times each time leaving one speaker out of the training set and testing performance on the speaker left out. Each classifier consisted of 15 STBFs selected by the GentleBoost algorithm. In order to make a multi-class decision, we trained all possible non-empty subsets of emotions versus the rest. For a seven-way classification experiment this makes a total of 63 binary classifiers. To make the final classification decision, multinomial ridge logistic regression (Movellan, 2006) was applied to the continuous outputs

Table 1
A confusion matrix for our method on the Berlin EMO database. The cell in the ith row and jth column represents the fraction of samples with of emotion i classified as emotion j. The recognition rate using 10-fold leave one speaker out cross validation is 78.7%.

	Anger	Boredom	Disgust	Fear	Joy	Neutral	Sadness
Anger	**0.9291**	0	0	0.0236	0.0472	0	0
Boredom	0	**0.7468**	0.0506	0.0253	0	0.0886	0.0886
Disgust	0.1316	0.0263	**0.6842**	0	0.0263	0.0263	0.0886
Fear	0.1091	0	0	**0.7091**	0.0545	0.0727	0.0545
Joy	0.3906	0	0.0156	0.0469	**0.5**	0.0469	0
Neutral	0	0.0897	0	0.0128	0.0128	**0.859**	0.0256
Sadness	0	0.0377	0	0.0189	0	0.0377	**0.9057**

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036

Table 2
A summary of the classification accuracies obtained from applying each of the three methods to the three datasets. An entry of N/P indicates not performed. The values listed in the table represent percentage of correct classifications for a particular method on a particular dataset.

	Berlin	Music versus speech	Cry detection
Our approach	**78.7%**	**98.4%**	**95%**
Simple Summary	61.7%[a]/ 50.9%[b]	95.1%	88.5%
Intermediate Aggregation	59.1%[c]/ 65.7%[d]	93%[e]	N/P

[a] It is the result of training "Simple Summary" using a multiclass support vector machine.
[b] Binary detectors are combined using multinomial ridge logistic regression to make the final classification decision.
[c] Only the box-filters are used in the original work.
[d] Additional box-filters are used.
[e] Intermediate Aggregation" was obtained from Casagrande et al. (2005b).

of each of the 63 binary detectors. The confusion matrix of the final system on the hold out set is presented in Table 1. The overall recognition rate on this seven-way classification task was 78.7%. The "Simple Summary" approach fared worse. In Table 2 we report the results of "Simple Summary" using both a multiclass Support Vector Machine as well as multinomial ridge logistic regression to combine the outputs of 63 binary Support Vector Machines. The better of these two accuracies (obtained using a multiclass SVM) is 61.7%. This discrepancy in accuracy is evidence that the ability to jointly learn a classifier and select aggregate features can result in large gains in performance.

In order to gain more insight into how our method was achieving gains we tried the "Intermediate Aggregation" method (Casagrande et al., 2005a) on the Berlin dataset. However, in order to minimize the number of variables we used the Sonogram as the input to this method as opposed to the raw power spectrum (which was done in the original work). Since in (Casagrande et al., 2005a) smoothing is used to boost performance, we average the outputs of each classifier (which each give an output every 50 ms) over the entire audio clip. As in our approach, multinomial ridge logistic regression is used to combine the outputs of these 63 binary classifiers to make the final classification decision. The result for the system of Casagrande et al. is 59.1% with the original features (Casagrande et al., 2005a) and 65.7% when including an additional box-filter type that computes contrasts in spectral energy across frequency bands (see the left-most box-filter in Fig. 4). The substantial increase in performance when considering the additional box-filter type hints at its importance in emotion recognition. However, we have not analyzed in detail any additional causes for the performance boost that our method enjoys. In the future we will attempt to isolate the various factors that contribute to advantages of our system over the "Intermediate Aggregation" approach.

3.2. Detection of crying in a preschool environment

The original motivation for the work we are presenting here was to develop audio-based perceptual primitives for social robots that need to interact with children. A key problem we found was the need to recognize whether children are crying at any given moment. In this section we evaluate the performance of the proposed method on the problem of detecting whether or not a short audio segment (a few seconds) does or does not contain infant crying. The dataset was collected in the typically noisy atmosphere of the preschool and thus is more challenging and realistic that many auditory category recognition databases that are collected in pristine laboratory conditions. This actually highlights why it is useful to have a method which automatically selects the appropriate feature representations and temporal aggregation, as other systems which are fine tuned for speech or music genre recognition may not be appropriate for the arbitrary classification categories we need for our robots such as this one.

To train a cry detector, we collected audio from one full working day at Classroom 1 of the Early Childhood Education Center (ECEC) at UCSD in San Diego. We then had two coders label each two second segment for the presence or absence of children crying. The inter-labeler agreement on this dataset was 94%. The proportion of segments containing children crying is 24%. The database is publicly available at http://mplab.ucsd.edu.

Classification experiments were conducted using various lengths of audio context. The label of the clip was then obtained by using a majority vote of all the labels given over the shorter two second windows. The particular method of evaluation was 25-fold cross validation. The segment boundaries were selected to include one or more salient events (e.g. a crying session or a particular song). Each fold leaves out one particular continuous segment of audio collected from the preschool, rather than leaving out a particular crier (or speaker as is done in the Berlin experiments). The segments were all recorded in the same room of the preschool in order to minimize the risk of allowing a system to overfit to the idiosyncrasies of the background noise or acoustic characteristics of a particular room. Table 2 displays a comparison of the classification accuracy between the method we are proposing here and the previously proposed "Simple Aggregation" method on the task of detecting cry in an 8 s segment of audio. Our approach (using 15 STBFs selected using GentleBoost) outperforms the "Simple Aggregation" approach by a margin of 95% to 88.5% on the measure of classification accuracy.

In previous work (Ruvolo and Movellan, 2008) we compared our approach to SOLAR (Hoiem et al., 2005). SOLAR is a system designed to detect audio events in complex audio environments. SOLAR is a general purpose system engineered to detect auditory events in the presence of background noise, and consequently appears to provide a suitable comparison system on the task of detecting crying children in a noisy classroom environment. The ROC curve for our system and SOLAR are given in Fig. 7. The task for each system was to decide whether a given 4-s clip of audio contained children crying. Our system achieved much better per-

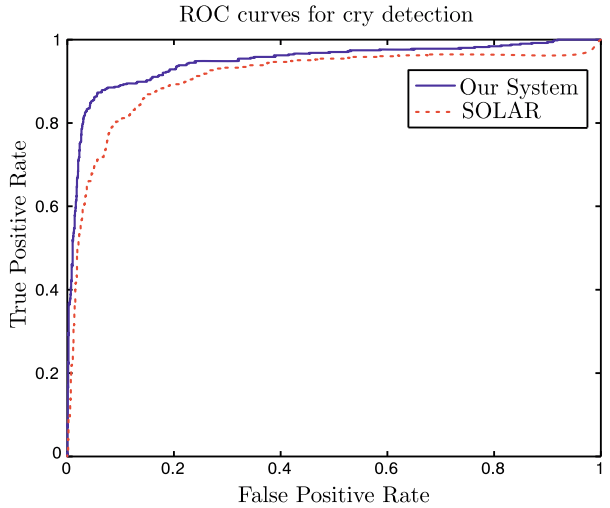

Fig. 7. Comparison of our method to that of (Hoiem et al., 2005) on the task of predicting whether a given 4-s window contained a child crying. Our system significantly ($p < 0.0001$) outperforms (A' of 0.9467) the method in (Hoiem et al., 2005) (A' of 0.9093).

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036

formance, 0.9467 area under the ROC, versus 0.9093 for SOLAR. The area under the ROC curve is a commonly used statistic to measure the performance of a classifier in a way that is not affected by the bias in the class label distribution.

3.3. Discrimination of speech versus music

In (Scheirer and Slaney, 1997), Scheirer and Slaney present a robust system for discriminating speech and music. A subset of the database used in this work has been made available publicly. As a point of comparison with the published result of (Casagrande et al., 2005b) we train each of the three methods on the task of discriminating 15 s clips into two groups: speech and music. The corpus contains 120 training and 61 testing segments. The results of this analysis are given in Table 2. The "Simple Aggregation" approach was second best (although the performance is sensitive to the window size used for feature aggregation). The approach proposed here performed best and yielded an accuracy of 98.4% on the testing data.

4. Conclusion

Auditory signals have rich temporal structure operating at multiple time scales. Finding methods to capture this multi-scale structure is a central issue in audio classification. Traditional approaches to speech recognition tackled the time scale problem using machine learning methods, e.g. low-level features get combined with HMMs that can be composed at the scale of phonemes, words and sentences. This HMMs are then trained using machine learning methods. For general purpose audio classification problems it is important to develop alternative approaches that can go beyond the limitations of traditional HMMs while maintaining the proven success of learning methods.

Here we proposed an approach (STBFs) that allowed the use of learning methods to select low-level auditory features and to aggregate them at multiple time scales. The proposed approach is general purpose and performed very well in a wide range of tasks, when compared to other popular approaches in the literature. One key issue for future research is to continue exploring new alternatives for capturing and aggregating information at multiple time scales. One possibility is to use HMMs which are ubiquitous in speech recognition but has yet to become a mainstay in the field of general audio category recognition as the long time scale feature model for STBFs. If such an approach is pursued care must be taken to maintain the fast learning performance of our current system.

References

Abe, M., Nishiguchi, M., 2002. Self-optimized spectral correlation method for background music identification. In: Proc. IEEE ICME'02, Lausanne, pp. 333–336.
Aucouturier, J., Pachet, F., 2003. Representing musical genre: A state of the art. J. New Music Res. 32 (1), 1–12.
Barrington, L., Chan, A., Turnbull, D., Lanckriet, G., 2007. Audio information retrieval using semantic similarity. In: IEEE Internat. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Vol. 2, pp. 725–728.
Bergstra, J., Casagrande, N., Erhan, D., Eck, D., Kégl, B., 2006. Aggregate features and ADABOOST for music classification. Machine Learning 65 (2–3), 473–484.
Bloomfield, P., 2000. Fourier Analysis of Time Series: An Introduction (Wiley Series in Probability and Statistics). Wiley-Interscience.
Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeiser, W., Weiss, B., 2005. A database of german emotional speech. In: Interspeech Proceedings.
Casagrande, N., Eck, D., Kégl, B., 2005a. Frame-level audio feature extraction using ADABOOST. In: Proc. 6th Internat. Conf. on Music Information Retrieval. University of London, London, pp. 345–350.
Casagrande, N., Eck, D., Kégl, B., 2005b. Geometry in sound: A speech/music audio classifier inspired by an image classifier. In: ICMC 2005, Barcelona, Spain.
Chu, S., Narayanan, S., Kuo, C.-C.J., Matarić, M.J., 2006. Where am I? Scene recognition for mobile robots using audio features. In: IEEE Internat. Conf. on Multimedia & Expo (ICME).
Daniel, P., Weber, R., 1997. Psychoacoustical roughness: Implementation of an optimized model. Acustica 83, 113–123.
De Bonet, J., Viola, P., 1997. Structure driven image database retrieval. In: Advances in Neural Information Processing, Vol. 10.
Deshpande, H., Singh, R., Nam, U., 2001. Classification of music signals in the visual domain. In: Proc. COST G-6 Conf. on Digital Audio Effects (DAFX-01), Limerick, Ireland.
Fasel, I., Fortenberry, B., Movellan, J.R., 2005. A generative framework for real-time object detection and classification. Comput. Vision Image Understanding 98, 182–210.
Fastl, H., Zwicker, E., 1990. Psychoacoustics Facts and Models. Springer-Verlag, Berlin, Heidelberg, Germany.
Freund, Y., Schapire, R.E., 1996. Experiments with a new boosting algorithm. In: Proc. 13th Internat. Conf. on Machine Learning. Morgan Kaufmann, pp. 148–146.
Friedman, J., Hastie, T., Tibshirani, R., 2000. Additive logistic regression: A statistical view of boosting. Ann. Statist. 28 (2), 337–374.
Glasberg, B.R., Moore, B.C.J., 1990. Derivation of auditory filter shapes from notched-noise data. Hearing Res. 47, 103–138.
Glover, F.W., Laguna, M., 1997. Tabu Search. Kluwer Academic Publishers.
Gold, B., Morgan, N., 2000. Speech and Audio Signal Processing: Processing and Perception of Speech and Music. Wiley.
Grimm, M., Kroschel, K., Mower, E., Narayanan, S., 2007. Primitives-based evaluation and estimation of emotions in speech. Speech Comm. 49, 787–800.
Hartmann, W.M., 1997. Signals, Sound, and Sensation. American Institute of Physics Press, Woodbury, New York.
Heckbert, P.S., 1986. Filtering by repeated integration. In: Internat. Conf. on Computer Graphics and Interactive Techniques, pp. 315–321.
Hoiem, D., Ke, Y., Sukthankar, R., 2005. SOLAR: Sound object localization and retrieval in complex audio environments. In: IEEE Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 5, pp. 429–432.
Junqua, J., Haton, J., 1996. Robustness in Automatic Speech Recognition. Kluwer Academic, Boston.
Mallat, S.G., 1999. A Wavelet Tour of Signal Processing. Academic, New York.
McDonnell, M.J., 1981. Box-filtering techniques. Comput. Graph. Image Process. 17 (1).
McKinney, M.F., Breebaart, J., 2003. Features for audio and music classification. In: ISMIR 2003, 4th Internat. Conf. on Music Information Retrieval, Baltimore, MD, USA.
Movellan, J.R., 2006. Tutorial on multinomial logistic regression. MPLab Tutorials. http://mplab.ucsd.edu.
Ntalampiras, S., Potamitis, I., Fakotakis, N., 2008. Automatic recognition of urban soundscenes. In: Tsihrintzis, G.A., Virvou, M., Howlett, R.J., Jain, L.C. (Eds.), New Directions in Intelligent Interactive Multimedia, Studies in Computational Intelligence, vol. 142. Springer, pp. 147–153.
Pachet, F., Roy, P., 2007. Exploring billions of audio features. In: Eurasip (Ed.), Proceedings of CBMI 07.
Petrushin, V., 1999. Emotion in speech: Recognition and application to call centers. In: Proc. Conf. on Artificial Neural Networks in Engineering (ANNIE '99).
Ruvolo, P., Fasel, I.R., Movellan, J.R., 2008. Auditory mood detection for social and educational robots. In: ICRA, pp. 3551–3556.
Ruvolo, P., Movellan, J.R., 2008. Automatic cry detection in early childhood education settings. In: Proc. ICDL, pp. 204–208.
Scheirer, E., 1998. Tempo and beat analysis of acoustic musical signals. J. Acoust. Soc. Amer. 103 (1), 588–601.
Scheirer, E., Slaney, M., 1997. Construction and evaluation of a robust multifeature speech/music disciminator. In: Proc. ICASSP.
Shen, J., Castan, S., 1985. Fast approximate realization of linear filters by translating cascading sum-box technique. In: Proc. CVPR, pp. 678–680.
Tyagi, V., Bourlard, H., 2003. On multi-scale fourier transform analysis of speech signals. IDIAP Research Report 03-32.
Tzanetakis, G., Essl, G., Cook, P., 2001. Automatic musical genre classification of audio signals. In: Proc. Internat. Symp. on Music Information Retrieval (ISMIR), Bloomington, IN, USA, pp. 205–210.
Vapnik, V.N., 1995. The Nature of Statistical Learning Theory. Springer-Verlag, Heidelberg, DE.
Viola, P., Jones, M., 2004. Robust real-time object detection. Internat. J. Comput. Vision. 57 (2), 137–154.
von Bismarck, G., 1974. Sharpness as an attribute of the timbre of steady sounds. Acustica 30, 159–172.
Wold, E., Blum, T., Keislar, D., Wheaton, J., 1996. Content-based classification, search, and retrieval of audio. IEEE Multimedia 3 (2).

Please cite this article in press as: Ruvolo, P., et al. A learning approach to hierarchical feature selection and aggregation for audio classification. Pattern Recognition Lett. (2010), doi:10.1016/j.patrec.2009.12.036