Content/Context-Adaptive Feature Selection for Environmental Sound Recognition

EnShuo Tsau, Sachin Chachada and C.-C. Jay Kuo
University of Southern California, Los Angeles, USA
E-mails: tsau@usc.edu, chachada@usc.edu and cckuo@sipi.usc.edu

***Abstract*—Environmental sound recognition (ESR) is a challenging problem that has gained a lot of attention in the recent years. A large number of audio features has been adopted for solving the ESR problem. In this work, we focus on the problem of automatic feature selection. Specifically, we propose two methods, called the content-adaptive and the context-adaptive feature selection schemes to achieve this goal. Finally, the superior performance of the proposed feature selection methods is demonstrated when they are applied to a medium-sized environmental database with a simple Bayesian network classifier.**

I. Introduction

The environmental sound recognition (ESR) problem arises in many interesting applications such as audio scene analysis, navigation, assisting robotics, and mobile device-based services. By audio scene analysis, we refer to the classification of a location (such as a restaurant, a playground or a rural area) based on its different acoustic characteristics. Audio data are available in challenging conditions such as lack of light and/or with visual obstructions. Besides, as compared with video, audio is relatively easy to store and process. The ESR technique can also be used to enhance the performance of speaker identification and language recognition with environmental sounds in the background.

Research on unstructured audio recognition, such as environmental sounds, has received less attention than that for structured audio such as speech or music. Only a few studies have been reported, and most of them were conducted with raw environment audio. To give a couple of examples, sound-based scene analysis was investigated in [1]–[3]. Because of randomness, high variance and other difficulties associated with environmental sounds, their recognition rates are poorer than those for structured audio. This is especially true when the number of sound classes increase. To overcome the insufficiency of MFCCs and other commonly-used features, Chu *et al.* [4] proposed a set of features based on the Matching Pursuit (MP) technique. Although the MP-based features provide good performance, their computational complexity is too high in real-time applications. Hence, the low-complexity CELP-based features are used in this work.

It is well known that the performance of ESR algorithms dramatically decreases when the number of sound classes increases (even with good features). It will need more good features for performance improvement. On the other hand, a larger number of features not only results in higher complexity but also demands more samples while there is no guarantee on performance improvement. That is because some features may help classify some classes but hamper the classification results for the others. Besides, it is not easy to train a classifier for better discriminant power in a higher dimension feature space. As a result, feature selection and reduction is an important task. In this chapter, we propose a novel content/context-based feature selection method to achieve this goal.

The rest of this paper is organized as follows. Various audio features are reviewed in Sec. II. The content-adaptive and the context-based feature selection methods are presented in Sec. III. Experimental results are shown in Sec. IV to demonstrate the superior performance of the proposed feature selection methods. Finally, concluding remarks and future research directions are given in Sec. V.

II. Review of Previous Work

Traditional feature selection attempts to find a feature subset that maximizes the utility function or a certain pre-defined performance metric. It assumes that the metric is a good approximation of the classification rate. Then, the problem can be re-formulated as the search of a feature subset to maximize a pre-defined metric. However, the feature set reduction problem is examined with all available sounds in the database, and the reduced feature subset is independent of a particular query sound.

Feature selection algorithms typically fall into the following two categories.

- Feature Ranking
  It ranks the features by a metric, and eliminates all features that do not achieve an adequate score.
- Subset Selection
  It searches the set of possible features for the optimal subset. It can be further classified into 3 types.
  - Wrappers
    Wrappers use a search algorithm to search through the space of possible features and evaluate each subset by running a model on the subset. Wrappers can be computationally expensive and have a risk of over fitting to the model.
  - Filters
    Filters are similar to Wrappers in the search approach. However, instead of being evaluated against a model, a simpler filter is used in the evaluation. Two popular filter metrics for classification are correlation and mutual information, although they are

not true distance measures in the mathematical sense, since they fail to obey the triangle inequality. They should rather be regarded as scores. These scores are computed between a candidate feature (or a set of candidate features) and the desired output category.

– Embedded Techniques
Embedded techniques are embedded in and specific to a model, for example, the decision tree. There are a large number of data analysis softwares available for feature selection in the public domain.

There are two main feature selection tasks: 1) selecting the evaluation criterion and 2) selecting the search algorithm. For the evaluation criterion selection, we have several choices, as discussed below.

1) Distance-based measure
Distance measures are also known as separability, divergence or discrimination. One can maximize the inter-class distance using linear or non-linear metrics such as the Minkowski, the Euclidean or the Chebychev distance. Many probabilistic distances (e.g., Mahalanobis, Bhattacharyya, Divergence, Patrick-Fischer) can be simplified in the two-class case when the distribution of each class has a parametric functional form.
2) Margin-based measure
One can maximize the margin of a hyper-plane that separates two classes.
3) Information-based measure
It determines the information gain from a feature or mutual information between the feature and a class label or entropy. The maximal-relevance-minimal-redundancy (mRMR) criterion is shown to be equivalent to the maximal statistical dependency of the target class on the data distribution, but it is more efficient.
4) Dependence-based measure
The dependence-based measure is also called the correlation measure or the similarity measure. The correlation coefficient can be used to find the correlation between a feature and a class.
5) Consistency-based measure
The consistency-based measure attempts to minimize the number of features that separate classes inconsistently, where inconstancy is defined as two instances having the same value but different class labels.

The relation between these five measures and the classification error probabilities in terms of the performance bound has been widely studied. However, it is still an open question.

Given a criterion, the next question is how to find the optimal set of features. Exhaustive search is not practical. All optimal methods are based on Branch and Bound. However, they can only be applied to problems of lower dimensionality with a monotonic criterion (e.g. distance measure). Most often, people use sub-optimal solutions of polynomial complexity for problems of higher dimensionality or with a non-monotonic criterion. Examples of sub-optimal searches include: sequential selection, floating search, oscillating search, dynamic oscillating search, random space, evolutionary algorithms, memetic algorithms, relief algorithm, simulated annealing, tabu search, randomized oscillating search, etc. Other feature selection issues that need consideration are the determination of the feature size, feature acquisition cost, over-fitting and instability.

At the end of this section, we would like to review the Fisher Discriminant Analysis (FDA), since it will be used in our work. The Fisher linear discriminant provides an efficient tool for dimensionality reduction in statistical pattern recognition. Its main idea can be briefly stated as follows: Suppose that there are two kinds of sample points in a d-dimension data space. We wish to find a line in the feature space such that the projections of the sample points of two classes on this line can be separated as much as possible. To achieve this goal, one can define the Fisher discriminant ratio as

$$J(w) = (\tilde{m}_1 - \tilde{m}_2)^2/(\tilde{S}_1^2 + \tilde{S}_2^2),$$

where w is the direction vector of the separating line, $\tilde{m}_i$ and $\tilde{S}_i$ are the mean and the within-class scatter of the i-th class, respectively, and the tilde denotes the result after projection, for example, $\tilde{m} = w^T m$, where w is the linear discriminant function. The Fisher Discriminant Analysis (FDA) aims to find out a linear projection w that maximizes the Fisher ratio, $J(w)$.

III. Proposed Feature Selection Methods

To address the problem of a large number of classes, we consider a new methodology in this work. First, we study the content and the context of each audio sound, and identify the most useful content/context adaptive feature set. Then, the reduced feature set has a higher discriminant power for a particular query sound with respect to other sounds in the database. Specifically, we propose two methods; namely, the context-based method and the content-based method, to achieve this goal.

A. Context-Adaptive Method

The context-adaptive method divides the classification task into two stages: 1) context identification and 2) target classification as shown in Fig. 1 [5]. The main motivation is to divide all the classes into several contexts, where each context is a group of classes which share similar features or belong to the similar category. Then, we can use different context adaptive features to further classify the target within a context.

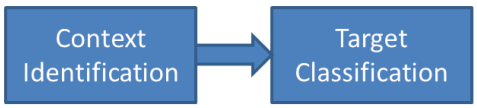

Fig. 1. The conceptual diagram of content/context-adaptive feature selection methods.

Context identification serves as a preprocessing unit before classification. We have the following two types of contexts.

- Norminal
Norminal contexts are a group of classes that belong to a

similar category with a physical meaning. For example, some similar sounds or similar situations in the environmental sound such as rain and water. Sometimes, it may not necessarily be a group of classes, it can be some other taxonomy or other factor which may affect the features, such as the weather condition in [5]. No matter what kind of context is used, additional information about the data is necessary.

- Artificial
  For some specific data, there is no clue or prior knowledge about the data and its features. As a result, the artificial taxonomy is adopted, *i.e.*, clustering. We cluster the data into one context with similar or nearby features. Intuitively speaking, those samples or classes should be under the same situation or under the same context although we have no idea what the context's physical meaning could be. In this situation, we artificially cluster all input sounds into a pre-defined number of contexts

We perform an extra stage of processing, *i.e.* context identification, because by doing so, we would be able to decrease the loading of the target classifier in terms of complexity and performance. This would result in fewer classes for the target classifier after context identification. Generally speaking, algorithm will demand fewer features since the number of sound classes within a context becomes smaller. This stage can also be helpful in avoiding ambiguous features which are only good for some classes while become notorious for other classes. We summarize the context-adaptive method below.

Context-Based Method: Nominal Context

Training Phase
1. Group the original classes into different contexts by nominal categories.
2. Pick up dominant feature set $F1$ by Fisher Ratio for context identification stage. 3. For each context, extract dominant feature set $F2$ using Fisher Ratio.
4. Use sequential forward search to decide $F1$ and $F2$ until the Fisher Ratio can not be increased further by adding more features. Testing Phase
1. Context Identification: Use feature set $F1$ to identify the context.
2. Depending on the result of context identification, one may use feature set $F2$ to identify the target.

Context-Based Method: Artificial Context

The algorithm is basically the same; the only difference is in the formation of contexts, as explained below.

1) Normalize the features.
   To have the clustering more meaningful, we have to normalize features. We do not want to have clustering for different scales. For example, the reflection coefficients, that are extracted from the LPC coefficients, are always less than 1 while pitch is usually in the range of hundreds.
2) Use PCA to reduce the feature to three linearly combined features.
   We use PCA for dimensionality reduction because high dimensional clustering may not be precise. Besides, if we need to use a probability-density-based classifier to estimate the density, a high dimensional feature space will be very challenging. Furthermore, PCA projects to the principal component regardless of class. Here, we are concerned with clustering rather than classifying; so it is fine to use the PCA.
3) Use K-means algorithm and the reduced features to cluster the context.
   For the reason that clustering in a high dimensional space is not practical, we use PCA to project the features into 3 dimensional space. K-means algorithm is then adopted to cluster the contexts. PCA is the easiest way to reduce the dimension here because the label is not important here. What we care is the main principal of the features. The best number of cluster is obtained when the Fisher Ratio can not be increased by more clusters.

For multiple classes, we need to identify within context classes. Here we need user-defined parameter to filter out non-relevant classes in a context. If a class has less than $c\%$ members in one context, we do not assign it to that context. We can set up an optimal cluster number by the following method:
1. We start with 2 clusters and get the Fisher Ratio R
2. We further try more clusters and calculate R.
3. We stop at K clusters until R is not improving.

The whole system diagram is shown in Fig. 2. It can be seen that in this multi-stage framework, a given test sample, based on its context, is evaluated by fewer classifiers as against the conventional single-stage systems. Also, at each stage, a small set of features are used, again depending on the context of the test sample. Hence, this multi-stage processing for multi-class systems can dramatically speed up the classification and reduce the number of features. By using different feature sets for different contexts, we can also improve the classification rate.

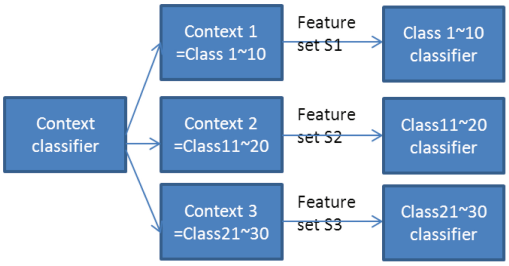

Fig. 2. Illustration of the context-based classifier design.

B. Content-Adaptive Method

The context-based method offers satisfactory results as shown in Fig. 3. However, it has some drawbacks. First,

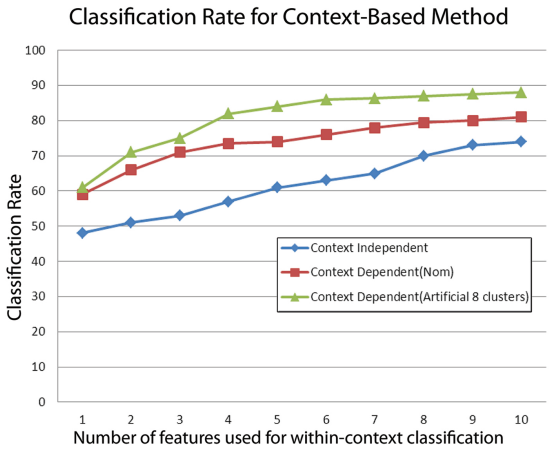

Fig. 3. Comparison of correct classification rates of context-based and context-independent feature selection methods.

its performance depends on an accurate context identification result. If the context is misclassified, there is no way to correct it. Second, there is a trade-off between the number of selected contexts and the number of within-context classes. On one hand, we want fewer within-context classes to simplify the feature selection task in each class. On the other hand, fewer within-context classes means a larger number of contexts, which makes the context identification problem more challenging. To alleviate this dilemma, we consider a content-based (or sample-based) method for feature selection.

The underlying assumption of the sample-based method is that we can collect enough statistical information from a query sample to decide its distinctive features using the Fisher Ratio. This provides an automatic feature selection process targeting at a specific input instance. To ensure sufficient statistics, the length of query samples plays an important role. In our experiments, the length of a query audio sequence is 3 seconds. The process can be summarized as follows.

1) Use the Fisher ratio to rank important features within each class.
2) Use multiple test samples from one audio frame to calculate the statistics.
3) Compare the statistics of the test sample with that of the training data using the KL divergence to eliminate unlikely candidate classes.

It is worthwhile to explain the last step in the above description. It is not proper to perform classification by directly matching the distinctive feature set of the test sample and those in the training database since it is too complicated to compare distinctive features of similar sounds in the database. However, if the distinctive features selected from a test sample are very different from those in a training class, we can claim that they are quite different.

The above filtering process helps decrease the load of the classifier in the next stage. In other words, this allows us to choose a simple classifier of lower complexity, which is a side benefit of our feature selection method.

IV. Experimental Result

A. Experimental Setup

In the experiments, we collected 30 classes of environmental sounds from the BBC audio data. These were sounds associated with the following:

- Transportation (7): airplane, car, motorcycle, train, helicopter, ship and elevator.
- Weather (3): rain, thunder and wind.
- Sports(3): table tennis, tennis and basketball.
- Rural Areas (3): bird, insect and stream.
- Animals(5): dog, chicken, sheep, horse and pig.
- Indoors(3): telephone, bell and clock.
- Human(3): crowd chatting, crowd applause and baby crying.
- Special(3): machine gun, tank and vacuum cleaner.

All collected data were re-sampled at 8KHz and normalized to a one-minute-long audio clip from a mono channel. Some pre-processing and filtering operations were used to filter out silence as well as irrelevant or noisy parts of environmental sounds. The CELP features were extracted by modifying the standard code of ITU-T G.723.1 [6]. There were total 19403 instances in the feature space with roughly an equal number in each class. Since we collect statistics for every 3 seconds, we take the average of the 3-second features as one sample point. The adopted feature set includes the following:

- CELP
- MFCC
- Amplitude Modulation
- Auto-Correlation
- Energy
- Envelope
- Envelope Shape Statistics
- Zero Crossing rate
- Perceptual Sharpness
- Spectral Flatness
- Spectral Shape Statistics

The result of using the Bayesian Network classifier is presented below.

B. Results and Discussion

Context-Based Feature Selection Method

In Fig. 3, the x-axis is the number of features adopted in the within-context classification (*i.e.*, the number of distinctive features) and the y-axis is the correct classification rate. We compare the performance of the context-based selection method and the context-independent features (*i.e.*, we use all the features and PCA result regardless of context). We can clearly see that the performance of the context-based method is better than that of the context-independent one for all classes. For the context-based method, the artificial context outperforms the nominal one since environmental sounds may have totally different characteristics even if they belong to a similar category.

TABLE I
THE CONFUSION MATRIX OBTAINED WITH THE CELP FEATURES AND THE BAYESIAN NETWORK CLASSIFIER.

%	Airplane	Bird	Insect	Motor	Rain	Rest.	Stream	Thunder	Train	Wind
Airplane	88.4	–	–	–	–	1.9	–	0.2	5.1	4.4
Bird	–	96.8	–	0.1	–	1.6	0.3	0.2	1.1	–
Insect	–	–	99.6	–	–	0.4	–	–	–	–
Motor	0.1	–	–	90.4	–	5.7	–	0.3	3.5	–
Rain	–	–	–	–	99.0	0.3	0.4	0.1	–	–
Rest.	1.0	2.2	–	8.1	0.1	77.9	1.4	2.6	6.8	0.1
Stream	–	0.2	–	–	0.3	1.0	97.7	0.2	0.5	–
Thunder	1.9	0.6	0.1	3.0	0.3	7.5	3.8	78.8	3.4	0.7
Train	5.1	0.7	–	5.0	0.1	7.1	0.1	0.7	81.3	–
Wind	–	–	–	–	–	–	–	1.3	–	98.7

Content-Based Feature Selection Method

We compare the performance of the MP features proposed by Chu *et al.* in [4] and the PCA-based features with the content-dependent features in Fig. 4, where the x-axis is the number of features adopted in the within-context classification (*i.e.*, the number of distinctive features) and the y-axis is the correct classification rate. It is clear that the proposed content-based feature selection method outperforms the MP and the PCA-based features.

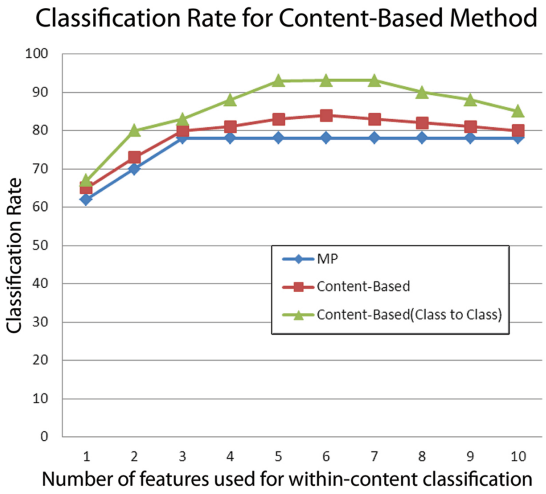

Fig. 4. Comparison of correct classification rates of the MP features, the PCA-based features and the proposed content-dependent features.

Length of samples

We show the correct classification rate as a function of the number of samples from a query instance in Fig. 5. As shown in the figure, the performance becomes saturated when the length is longer than 3 seconds. Then, the performance degrades slightly when the input length is longer than 20 seconds. This is caused by a larger variation of query environmental sounds over a longer time interval.

Confusion matrix

Table I shows the confusion matrix for 10 out of the 30 classes. We see that the performance of most classes is quite good. The result is also consistent with our intuition. For instance, the train station sound and the restaurant sound can be confused more easily due to their similar environments.

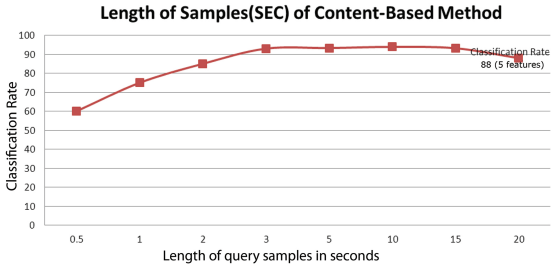

Fig. 5. The correct classification rate as a function of the sample length (in the unit of seconds).

V. CONCLUSION AND FUTURE WORK

Two novel feature selection methods (*i.e.* context-based and content-based) were proposed to solve the ESR problem in this work. The methods were applied to a medium-sized ESR database that contains 30 environmental sound classes. The content-based method offers the best classification result with a correct classification rate of 95.2% using the Bayesian network classifier. The proposed solution is scalable to a larger size problem by nature. Furthermore, the context-based and the content-based feature selection processes can be used in cascade to yield an even better result. These are good future research topics.

REFERENCES

[1] B. Clarkson, N. Sawhney, and A. Pentland, "Auditory context awareness via wearable computing," in *In Proceedings of The 1998 Workshop on Perceptual User Interfaces (PUI98)*, 1998.

[2] V. Peltonen, "Computational auditory scene recognition," Master's thesis, Tampere University of Tech., Finland, 2001.

[3] A. Eronen, V. Peltonen, J. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 1, pp. 321 – 329, jan. 2006.

[4] S. Chu, S. Narayanan, and C.-C. Kuo, "Environmental sound recognition with time-frequency audio features," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 6, pp. 1142 –1158, aug. 2009.

[5] C. Ratto, P. Torrione, and L. Collins, "Exploiting ground-penetrating radar phenomenology in a context-dependent framework for landmine detection and discrimination," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 49, no. 5, pp. 1689 –1700, may 2011.

[6] ITU-T, *ITU-T Recommendation G.723.1 : Transmission Systems and Media, Digital Systems and Networks*, 2006.