DISCRIMINATING TWO TYPES OF NOISE SOURCES USING CORTICAL REPRESENTATION AND DIMENSION REDUCTION TECHNIQUE

Shiva Sundaram and Shrikanth Narayanan
Speech Analysis and Interpretation Laboratory (SAIL),
Dept. of Electrical Engineering- Systems, University of Southern California,
3740, McClintock Ave, EEB 400, Los Angeles, CA 90089. USA
email: ssundara@usc.edu, shri@sipi.usc.edu

ABSTRACT

Content-based audio classification techniques have focused on classifying events that are both semantically and perceptually distinct (such as speech, music, environmental sounds etc.). However, it is both useful and challenging to develop systems that can also discern sources that are semantically and perceptually close. In this paper we present results of our experiments on discriminating two types of noise sources. Particularly, we focus on machine-generated versus natural noise sources. A bio-inspired tensor representation of audio that models the processing at the primary auditory cortex is used for feature extraction. To handle large tensor feature sets, we use a generalized discriminant analysis method to reduce the dimension. We also present a novel technique of partitioning data into smaller subsets and combining the results of individual analysis before training pattern classifiers. The results of the classification experiments indicate that cortical representation performs 25% better than the common perceptual feature set used in audio classification systems (MFCCs).

Index Terms— Noise classification, audio classification, discriminant analysis for tensor representation, cortical representation, auditory scene analysis.

1. INTRODUCTION

Content-based audio systems rely on clustering, segmentation and classification of distinct acoustic source types through their within-class signal similarities. These systems rely on direct mapping between signal level feature vectors and their classes to achieve the end result. They group the vast possibilities of general audio classes into a handful of application specific classes such as *speech, environmental sounds, music* etc [1]. To generalize, it will be necessary to explicitly increase the number of recognizable groups resulting in increased complexity of the system (for example, more heuristic rules in [2] would be required). In [3] the authors developed a generalizable, mid-level representation scheme, where each instance (of frame based analysis) of an audio signal is classified into perceptual *speech-like*, *harmonic* and *noise-like* categories. This representation was successfully used to segment vocal sections in popular songs using a *maximum a posteriori* scheme. To tackle more complex scenes, futher categorization of sounds would be necessary.

In this paper, we build upon ideas of attribute based audio representation presented in [3]. Particularly, we aim at connecting signal representations to audio attributes, and focus on further analyzing the *noise-like* class. We present a classification system to discern between two noise sources: machine generated and other natural noise sources. *Machine-generated* noise are audio from sources such as computer printers, telex machines, vehicle engines, air plane propellers etc. Examples of *other noises* are sounds of wind, waves on a seashore, rainfall, leaves rustling. etc. The discrimination of the two noise categories is challenging because they are both semantically and acoustically similar and they are usually categorized without any distinction as non-speech or environmental sounds in systems such as [1, 2]. Other noise classification systems follow the content-based approach of trying to explicitly classify individual noise classes such as car, plane, train etc., using elaborate Hidden Markov Models ([4] lists a comprehensive list of such systems). As mentioned earlier, for a generalizable mid-level representation, it is desirable to classify noises into categories based on signal attributes (such as suggested here) rather than classes based on canonical names. Classifying noise categories has applications in context recognition [4], scene change detection and indexing [5], context-aware listening for robots [6] and also in background/foreground audio tracking [7].

The contributions of this work are as follows. First, as features for the noise classification task, we use a bio-inspired approach involving a model of processing at the primary auditory cortex. This has been applied to the speech non-speech discrimination (SNS) problem successfully [8]. As indicated by our experimental results for noise classification, the cortical representation (CR) exceeds the performance of the commonly used Mel-frequency cepstral co-efficients (MFCCs). Since CR is a multi-dimensional tensor, training pattern classifiers using this data becomes prohibitive due to large vector dimensions ($\sim 10^3$). For dimension reduction, we use a generalization of the Fisher discriminant analysis (FDA). However, this also is not feasible on large training data for pattern classifiers. This issue is exacerbated by the large dimension of the data. To address this, we propose a new technique of data partitioning, and combining the results of analysis on the individual subsets. We show that discriminant analysis of a large data set can be made practical by breaking the problem into smaller sets and using the information from each of this subset. In the next section, details of this representation

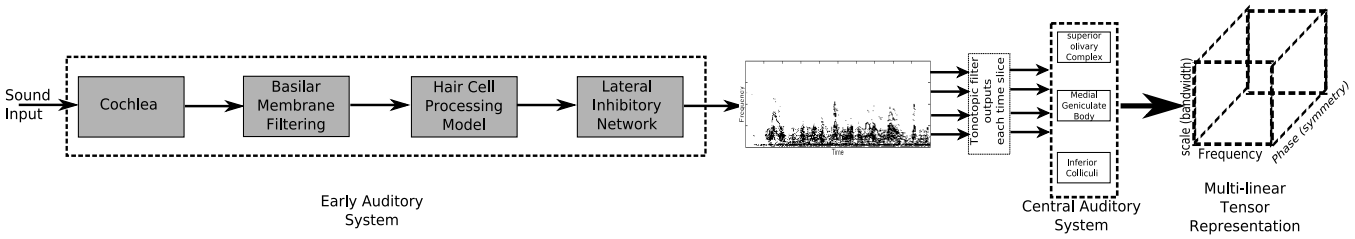

Fig. 1. Summary of processing from input sound to the multi-dimensional tensor representation of cortical processing [9, 10, 8].

followed by the proposed dimension reduction method is presented. Finally, results of pattern classification algorithms are presented. We also compare performance of the CR with the performance of the MFCCs.

2. AUDITORY CORTICAL REPRESENTATION (CR)

While features such as the popular MFCCs are based on the processing at the early auditory system, the CR is based on processing of sound at the central auditory system [10]. This is modelled as a re-analysis of the input spectra from the early auditory processing stage along the logarithmic frequency axis, a scale axis (local frequency bandwidth) and a phase axis which is a measure of the local symmetry of the spectrum. Since each time-frequency slice of an input audio signal is measured with respect to these three axes, the analysis results in a tensor (n-mode) representation.

Figure 1 illustrates the processing stages which finally result in a tensor representation. The output of the early auditory system is a time-frequency representation of the input signal. Here the input sound signal is filtered by the basillar membrane at different centre frequencies along the tonotopic frequency axis, followed by an differentiator stage, a non-linearity , low-pass filtering and finally, a lateral inhibitory network [9]. This is the input to the central auditory system, which is analogous to the early auditory system, except all the transformations are along the tonotopic frequency axis. The processing is modelled as a double affine wavelet transform of the frequency axis at different scales and phase. The *mother function* wavelet is a negative second derivative of the normal Gaussian function.

The result of this analysis is a 3-mode tensor $A(f : f_c, \phi, \lambda) \in \mathcal{R}^{D_1 \times D_2 \times D_3}$. Here, f is the tonotopic frequency at different centre frequencies f_c, ϕ is the symmetry (or phase) and λ is the scale factor or the dilation factor of the wavelet function. In our experiments, the analysis was performed using 64 bank filters ($D_1 = 64$) at 12 phase (or symmetry) values ($D_2 = 12$), 5 scale values ($D_3 = 5$). Therefore $A(\cdot, \cdot, \cdot)$ is a $64 \times 12 \times 5$ tensor.

3. DIMENSION REDUCTION

The classification experiments presented in this paper follow a data-driven approach. A total of 1.38 hours of data (from 217 clips) was collected from the BBC sound effects library (`http://www.soundideas.com`). A preprocessing stage first converted all the 2-channel 44.1 kHz uncompressed audio files into 1-channel 16kHz channels. Then, using an audio editor (`http://audacity.sourceforge.net/`) each clip was manually segmented to remove silence sections, extraneous impulsive sounds and other non-noise segments. In the process, to facilitate data analysis, the data was also grouped into *machine-noise* and *other-noise*. For analysis, audio frames of 40 millisecond duration, were extracted every 10 milliseconds after it was multiplied with a Hamming window. After the processing stages in the early and the central auditory system, a tensor of dimension $64 \times 12 \times 5$ is obtained for each frame. Basically, as a vectorized 1-mode representation, a vector of length $64 \times 12 \times 5 = 3840$ was extracted for each frame (effectively 5×10^5 vectors). For comparison of performance, MFCCs (39 dimensions: 13 order $+\Delta + \Delta\Delta$) were also extracted.

Certainly, due to the multi-scale analysis of the spectral profile, the extracted data contain large amounts of redundancy. Also, it is not feasible to train pattern classifiers on this raw, large dimensional data set. To make it practical, the dimension of the data needs to be reduced before training the classifier. In [8] the authors reduce the dimension using a generalization of the principal component analysis (PCA) for this 3-mode tensor, and successfully apply it to robust speech/non-speech discrimination. Although PCA is a powerful dimension reduction technique, it focuses on finding the best representation of the data with fewer principal components. For discriminatory pattern classification, however, discriminant analysis is more appropriate.

For the work presented here, a generalization of the matrix discriminant analysis for the tensor case was implemented. This was originally proposed in [11] and successfully used for face recognition tasks. Although the discriminant analysis of tensor representation (DATER) algorithm is a suboptimal solution and works iteratively and by unfolding the tensor in each dimension, it is still not directly practical for the data set extracted here. This is made feasible here by introducing a further sub-optimality by partitioning the data into smaller sets and performing localized discriminant analysis. Since, the partitioning is not restricted to a given discriminant analysis method, we also apply it to perform FDA. Our experimental results indicate that splitting the data does not deteriorate performance, and as shown for the MFCC case, it improves the classification result. Next the DATER algorithm is discussed, followed by the proposed data partitioning modification.

3.1. DATER Algorithm

Like FDA, the DATER algorithm seeks to find matrices $\mathbf{U}= \{U_1 \in \mathcal{R}^{D_1 \times m_1}, U_2 \in \mathcal{R}^{D_2 \times m_2}, \ldots, U_N \in \mathcal{R}^{D_N \times m_N}\}$ where, $m_k < D_k \; \forall \; k$ such that, (for data set tensor $\mathbf{X} \in \mathcal{R}^{D_1 \times D_2 \ldots \times D_N \times N_S}$, N_S is the total number of sample tensors),

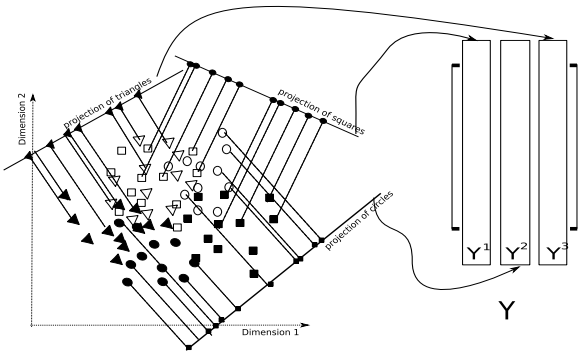

Fig. 2. Illustration of data partitioning. Data points belong to 2 classes (hollow and solid). The whole data is partitioned into 3 subsets (circles, squares and triangles). The whole data mapped onto each projection (resulting from discriminant analysis of each subset) are augmented column-wise to form the resulting data matrix Y

$$(U_k|_{k=1}^{k=N})=\underset{U_k|_{k=1}^{k=N}}{\operatorname{argmax}} \frac{\sum_c n_c \|\bar{\mathbf{X}}_{\mathbf{c}}\times_1 U_1 \ldots \times_N U_N - \bar{\mathbf{X}}\times_1 U_1 \ldots \times_N U_N\|}{\sum_i \|\mathbf{X_i}\times_1 U_1 \ldots \times_N U_N - \bar{\mathbf{X}}_{\mathbf{c_i}}\times_1 U_1 \ldots \times_N U_N\|}$$

Here $\bar{\mathbf{X}}_{\mathbf{c}}$ is the mean of tensors belonging to class c . $\bar{\mathbf{X}}$ is the overall mean. n_c is the number of samples in its respective class. $\mathbf{X}_i$ is the i^{th} sample tensor of class c_i. $\times_k U_k$ represents the unfolding of a tensor along the k^{th} dimension (into a matrix) and multiplication with U_k [12]. This is equivalent to maximizing the between class scatter and minimizing the within class scatter in FDA.

The algorithm finds $\mathbf{U}$ iteratively by re-projecting the tensor $\mathbf{X}$ along $U_k \ \forall \ k$ at the end of each iteration. Similar to FDA, a generalized eigenvector problem is solved, to determine a mapping that maximizes the inter-class scatter and minimizes the within-class scatter (using the unfolded tensor). When the stopping criterion is met, the algorithm outputs the matrices $\mathbf{U}=\{U_1 \in \mathcal{R}^{D_1 \times m_1}, U_2 \in \mathcal{R}^{D_2 \times m_2}, \ldots, U_N \in \mathcal{R}^{D_N \times m_N}\}$. After projecting $\mathbf{X}$ along the matrices $\mathbf{U}$, we obtain a smaller dimension tensor $\mathbf{X}' \in \mathcal{R}^{m_1 \times m_2 .. \times m_N \times N_S}$.

3.2. Partitioning data

As it will become clearer later, since the DATER algorithm involves unfolding of the tensors along each dimension, it is impractical to use it for large training sets. However, by partitioning the data into smaller sets, it is possible to use the algorithm for each subset. This is the partitioning modification proposed in this work. This method is explained below:

1. Randomly partition the whole data $\mathbf{X}$ into P sets $\mathbf{L_j}$ such that $\mathbf{X} = \{\mathbf{L_1}|\mathbf{L_2}|\ldots|\mathbf{L_P}\}$. *i.e,* $\mathbf{L_j} \in \mathcal{R}^{D_1 \times D_2 \ldots \times D_N \times N_j}$ and $\sum_{j=1}^{j=P} N_j = N_S$ (total number of sample tensors)
2. FOR $j = 1, 2, \ldots, P$
   - Execute DATER algorithm on $\mathbf{L_j}$ and obtain $\mathbf{U}^j$.
   - Project tensor samples $\mathbf{X}$ along $\mathbf{U}^j$ and obtain $\mathbf{X^{j'}} \in \mathcal{R}^{m_1^j \times m_2^j .. \times m_N^j \times N_S}$
   - let $Y^j = matricize(\mathbf{X^{j'}})$. i.e, By vectorizing each tensor, obtain $Y^j \in \mathcal{R}^{N_S \times (m_1^j \cdot m_2^j \ldots m_N^j)}$
3. END
4. Let $Y = \{Y^1|Y^2|\cdots|Y^P\} \in \mathcal{R}^{N_S \times \sum_{j=1}^{j=P}(m_1^j \cdot m_2^j \ldots m_N^j)}$ (column-wise augmentation)

This partitioning technique, illustrated in 2 dimensions is shown in figure 2. As a more concrete example, the data used in the current work can be considered. From the 1.38 hours of audio data, $N_S \approx 5 \times 10^5$ (each $\mathcal{R}^{D_1=64 \times D_2=12 \ \times D_3=5}$) tensors, belonging to $c = 2$ classes are extracted. For the DATER algorithm, unfolding along the first dimension would result in a $64 \times 3 \cdot 10^7$ matrix, a size that is prohibitive from a computational standpoint. This is also a problem when the tensor data set is unfolded along the other dimensions. But, if the data is partitioned into $P = 50$ smaller sets, ($N_j = 10^4 \forall \ j$), unfolding would result in (50 times) smaller matrices. This would also result in $P = 50$ $\mathbf{U^j}$ projection matrix sets. Since it is a 2 class problem, $m_k^j = 1 \forall \ \ j, k$. Therefore each $\mathbf{U^{j'}}$ results in a mapping from $64 \times 12 \times 5$ space to a 1-dimensional line. By column-wise augmentation, this results in $\sum_{j=1}^{j=P}(m_1^j \cdot m_2^j \ldots m_N^j) = 50$, and $Y \in \mathcal{R}^{N_S=(5 \cdot 10^5) \times 50}$ which is the reduced dimension data set available for training (instead of the initial $5 \cdot 10^5 \times 3840$ set). Each column of the resulting matrix Y is the projection of *all* the data points on projections obtained from individual partitions. While figure 2 is shown in 2 dimensions, the data points are actually in a very high dimensional space. It however, also shows that this partitioning procedure is not specific to tensor representation. As illustrated, it can also be used for FDA of 2-mode data. This partitioning is also used for FDA to compare CR with MFCCs. Next, the results of classification experiments are presented.

4. RESULTS

The performance of a 5 nearest neighbour (5NN) classifier and decision stump classifier with AdaBoost, as a function of the number of projections (or columns) used in the matrix Y is shown in figure 3. Using only 1 or 2 projections, ($Y \in \mathcal{R}^{N_S \times 2}$) the average accuracy (and the true positives rate) of the classifiers using the CR is about 95%. This is about 18-25% better than the performance of MFCCs for the same number of projections.

For MFCCs, as the number of projections (columns of Y) increases, the classifier accuracy increases. This trend can be observed in both classifiers. However, for the CR, there is no significant increase with more projections (F-measure=98.4% for 5NN and 94.1% for AdaBoost for 10 projections). Even with 10 projections, the performance of MFCC features (F-measure=93.5% for 5NN and 71.2% for AdaBoost with 10 projections) is less than the performance of using just 1 or 2 projections of the CRs. However, for MFCCs, by partitioning the data and using multiple projections, this performance is better than the average baseline accuracy that is obtained by FDA on the whole data set (also shown in figure 3).

5. DISCUSSION AND CONCLUSION

In this work, discrimination of two-types of noise sources: machine generated (such as vehicle noise, engine noise, printers, fax and telex machines etc.), versus other natural noise sources (rainfall, waves on a seashore, blowing wind etc.)

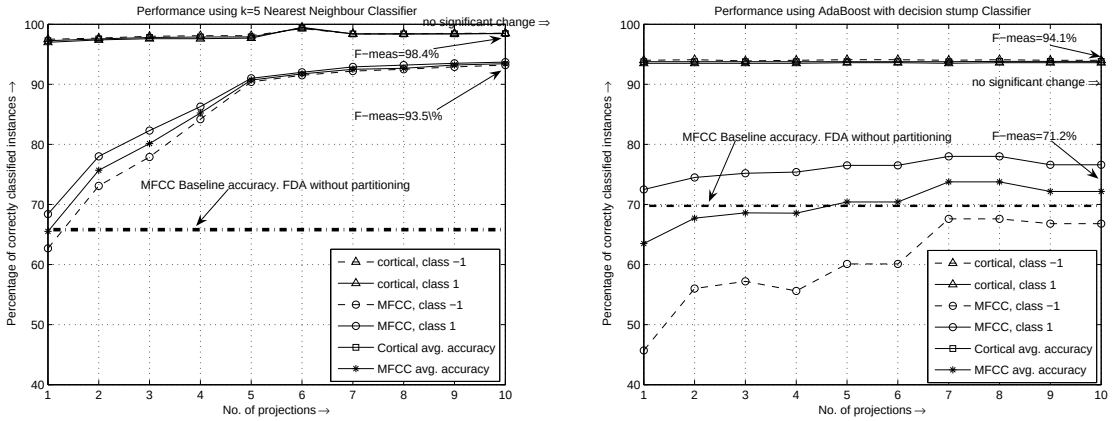

Fig. 3. 5 nearest neighbour (left), AdaBoost (right) classifier results (average accuracy and true positives rate) of the cortical representation (CR) versus MFCCs as a function of number of projections included. (90/10% train/test split,1 -machine generated, -1 -other noise)

was presented. Performance of the pattern classifiers was better using the cortical representation (CR) as opposed to using MFCCs. To reduce the dimension of the extracted data, a generalized version of the discriminant analysis for multidimensional tensor representation was used. Discriminant analysis of this large data set was made tractable by a new data partitioning technique. Intuitively, discriminant analysis on a subset of the whole data gives rise to a projection that is optimal for each smaller subset. This gives rise to many projections for the whole data set. When the information obtained from multiple projections is used for classification, it results in higher classification accuracy (as opposed to single projection obtained by discriminant analysis on the whole data). This partitioning technique makes training classifiers of large dimensional data sets feasible. It can also be extended to real-time processing problems.

From the high correct classification results, it can be concluded that the two noise types in question are effectively discernible using the cortical representation. The performance differential with respect to MFCC features is significant. This better performance can be attributed to the high resolution multi-scale spectral analysis of the cortical processing, which in effect, also naturally captures temporal properties of audio (due to time-frequency duality). Whereas, with MFCCs this had to be approximated with delta (Δ) and delta-delta ($\Delta-\Delta$) features.

As suggested in [3], as a part of future work, we would like to use this type of bio-inspired feature based discrimination ability to robustly represent various audio scenes in an attribute mid-level representation (*speech-like*, *harmonic*, *machine-noise like*, etc.) and use higher-level decision to classify a given audio scene.

6. REFERENCES

[1] T. Zhang and C.-C. J. Kuo, "Audio Content Analysis for Online Audiovisual Data Segmentation and Classification," in *IEEE Trans. on Speech and Audio Processing*, May 2001, vol. 9, pp. 441–457.

[2] H.J. Zhang L. Liu and H. Jiang, "Content Analysis for Audio Classification and Segmentation," in *IEEE Trans. on Speech and Audio Processing*, October 2002, vol. 10, pp. 504–516.

[3] S. Sundaram and S. Narayanan, "An attribute-based approach to Audio description Applied to Segmenting Vocal sections in Popular Music Songs," in *International Workshop on Multimedia Signal Processing (MMSP) 2006*, October 2006.

[4] D. Smith L. Ma and B. Miller, "Context Awareness using Environmental Noise Classification," in *In Proc. EUROSPEECH 2003*, 2003, pp. 2237–2240.

[5] H.-J. Zhang R. Cai, L. Lu and L.-H. Cai, ""Highlight Sound Effects Detection in Audio Stream"," in *In Proceedings of the International Conference on Multimedia and Expo (ICME) 2003.*, July 2003, vol. 3, pp. 37–40.

[6] S. Chu S. Narayanan, C.-C. Kuo and M. J. Mataric, "Where am I? Scene Recognition for Mobile Robots using Audio Features," in *In Proc. of the International Conference on Multimedia and Expo (ICME) 2006.*, July 2006.

[7] R. Radhakrishnan and A. Divakaran, "Generative Process Tracking for Audio Analysis," in *IEEE International Conf. on Acoustics, Speech and Audio Processing*, May 2006, pp. V1–V4.

[8] N. Mesgarani M. Slaney and S. A. Shamma, "Discrimination of Speech From Nonspeech Based on Multiscale Spectro-Temporal Modulations," in *IEEE Trans on Audio, Speech and Language Processing*, May 2006, vol. 14, pp. 920 – 930.

[9] K. Wang X. Yang and S. A. Shamma, "Auditory representations of acoustic signals," in *IEEE Trans. on Information Theory*, March 1992, vol. 38, pp. 824–839.

[10] K. Wang and S. A. Shamma, "Spectral shape analysis in the central auditory system," in *IEEE Trans. on Speech and Audio Processing*, September 1995, vol. 3, pp. 382–395.

[11] Q. Yang S. Yan, D. Xu, X. Tang L. Zhang, and H. J. Zhang, "Discriminant analysis with tensor representation," in *In Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR) 2005*, 2005, pp. 526–532.

[12] B. W. Bader and T. G. Kolda, "Algorithm 8xx: MATLAB Tensor Classes for Fast Algorithm Prototyping," in *ACM Transactions on Mathematical Software*, December 2006, vol. 32.