Audio Classification Based on Sparse Coefficients

Syed Zubair [1], Wenwu Wang [2]

Centre for Vision, Speech and Signal Processing , University of Surrey
Guildford, GU2 7Xh, United Kingdom
[1] s.zubair@surrey.ac.edu, [2] w.wang@surrey.ac.uk

***Abstract*—Audio signal classification is usually done using conventional signal features such as mel-frequency cepstrum coefficients (MFCC), line spectral frequencies (LSF), and short time energy (STM). Learned dictionaries have been shown to have promising capability for creating sparse representation of a signal and hence have a potential to be used for the extraction of signal features. In this paper, we consider to use sparse features for audio classification from music and speech data. We use the K-SVD algorithm to learn separate dictionaries for the speech and music signals to represent their respective subspaces and use them to extract sparse features for each class of signals using Orthogonal Matching Pursuit (OMP). Based on these sparse features, Support Vector Machines (SVM) are used for speech and music classification. The same signals were also classified using SVM based on the conventional MFCC coefficients and the classification results were compared to those of sparse coefficients. It was found that at lower signal to noise ratio (SNR), sparse coefficients give far better signal classification results as compared to the MFCC based classification.**

I. Introduction

Audio signals acquired from an uncontrolled natural environment have different types of contents e.g. speech, music and environmental sounds. In content based retrieval system, different contents such as voiced, unvoiced speech and music are needed to be distinguished from each other. This has been done by extracting discerning features from audio signal and then using them for signal and content classification.

Signal classification is in general a two-step process. First signal features are extracted and then a classifier is used to discriminate the signals. A lot of research in this area has been conducted in last two decades with the methods proposed mainly differing in the types of features and classification techniques used [1], [2], [3].

Various time, frequency and time-frequency representations have been used in the literature for generating audio features. For example, zero crossing rate (ZCR) [4], [5] and short-time energy (STE) [6], [7], together with their variations are the low level time domain features that have been used extensively. The frequency domain features that have been used are line spectral frequencies (LSF) [8], 4 Hz modulation energy, spectral centroid, spectral flux [9] and mel-frequecny cepstral coefficients (MFCCs) [3], [7]. Some other features have also been used based upon psychoacoustics which measures perceptual loudness, roughness [10], etc.

The second stage in the classification involves the selection of the type of classifiers. A number of different classifiers have been used for audio signal discrimination. Gaussian mixture models (GMM) [4], [11], K nearest neighbour (KNN) [9], [8], [12], neural network (NN) [13], [14] and hidden Markov model (HMM) [6], [15] along with variations of each classifier have been used at different stages of an audio classification algorithm.

Nowadays there is an increasing interest in sparse signal representation for various applications. Many signals are either sparse in some specific domain or they can be made sparse by using some machine learning techniques [16], [17], [18]. This inherent or manufactured sparsity of audio signals has many benefits in terms of a lower computational complexity and less demand of the resources. Hence these sparse coefficients have high potential to be used as signal features. Sparse representations or coefficients have successfully been employed in some applications like denoising [17] and coding [18], however less attention has been given by researchers to their use in signal classification. In this paper, we propose an audio classification algorithm where sparse coefficients are used as features for the discrimination of speech and music with the application of the SVM classifier. We evaluate its performance as compared to the use of conventional features.

This paper has been divided in the following sections. Section II discusses the K-SVD algorithm for dictionary learning and the OMP for sparse coding. Section III describes our algorithm for speech and music classification. Section IV presents the experiments performed and their results and the conclusion is given in section V.

II. Dictionary Learning and Sparse Coding

Dictionary is a transformation matrix that is used to represent a signal in a specific domain, e.g. the frequency domain. Such a dictionary is usually obtained by a predefined function such as discrete cosine transform (DCT). These dictionaries can also be adapted from signals with some specific structure, e.g. sparsity, hence they give a succinct representation of the signal given some constraint on the learning process. In our discussion, the specific constraint is sparsity. The objective function for a dictionary learning of an input signal $\mathbf{Y}$ with sparsity constraint is given as

$$\| \mathbf{Y} - \mathbf{AX} \|_F^2 \;\; s.t. \; \forall i \; \| \mathbf{x}_i \|_0 \leq T_0 \tag{1}$$

where $\mathbf{A}$ is a dictionary matrix, $\mathbf{X}$ is a coefficient matrix, T_0 is a small positive value measuring the sparsity of vector $\mathbf{x}_i$ and

$\| \cdot \|_0$ is l_0 norm representing the number of non-zero values in vector $\mathbf{x}_i$. $\| \mathbf{A} \|_F$ is called Frobenius norm and is defined as $\| \mathbf{A} \|_F = \sqrt{\sum_{ij} \mathbf{A}_{ij}^2}$.

Dictionary learning is a two-step process. In the first step, given input signal $\mathbf{Y}$ and dictionary matrix $\mathbf{A}$, sparse coefficients $\mathbf{x}_i$ are calculated. In the second step, with the given signal and coefficients matrix $\mathbf{X}$ calculated in the previous step, dictionary vectors called atoms are updated. In this paper, we use the K-SVD algorithm for dictionary learning where OMP is used for sparse coding of the signal of interest.

A. K-SVD Dictionary Learning Algorithm

For two-step process of dictionary learning, the K-SVD algorithm [17] uses OMP [19] for dictionary coefficients calculation and singular value decomposition (SVD) for calculating and updating dictionary atoms. In the dictionary learning step, $\mathbf{AX}$ is decomposed into N rank-1 matrices by selecting a dictionary element $\mathbf{a}_k$ and its corresponding coefficient vector $\mathbf{x}_T^k$ which is the kth row in matrix $\mathbf{X}$, where subscript T in $\mathbf{x}_T$ represents the sprasity level of $\mathbf{x}$.

$$\begin{aligned} \| \mathbf{Y} - \mathbf{AX} \|_F^2 &= \| \mathbf{Y} - \sum_{j=1}^{N} \mathbf{a}_j \mathbf{x}_T^j \|^2 \\ &= \| (\mathbf{Y} - \sum_{j \neq k} \mathbf{a}_j \mathbf{x}_T^j) - \mathbf{a}_k \mathbf{x}_T^k \|^2 \\ &= \| \mathbf{E}_k - \mathbf{a}_k \mathbf{x}_T^k \|^2 \end{aligned} \tag{2}$$

where $\mathbf{E}_k$ is the error term formulated by excluding an arbitrarily selected dictionary element from $\mathbf{A}$. Now SVD is used to find the closest rank-1 matrix that effectively minimizes the error. After removing columns from $\mathbf{E}_k$ that do not use $\mathbf{a}_k$, the SVD of $\mathbf{E}_k$ yields $\mathbf{U\Delta V}^T$, where the first column of $\mathbf{U}$ gives updated dictionary atom $\mathbf{a}_k$ and the first column of $\mathbf{V}$ multiplied by $\Delta(1,1)$ gives the coefficient vector $\mathbf{x}_T^k$ corresponding to the dictionary atom. Iterating through the two steps of dictionary learning, the K-SVD produces a dictionary that approximates the given signal $\mathbf{Y}$ sparsely.

B. Orthogonal Matching Pursuit Algorithm

To calculate sparse coefficients of an input signal with a given dictionary, the OMP algorithm [19] projects the input signal on the subspace spanned by the dictionary atoms. The atom which strongly correlates with the signal or its residual is selected and used for calculation of the coefficients. The whole algorithm works as follows:

- Initialize the residual $\mathbf{r}_0$ to be the input signal vector $\mathbf{y}_i$ and coefficient vector $\mathbf{x}_0$ to zero.
- At step k, a new atom is selected according to the following optimization problem

$$\lambda_k \in \arg\max_{\omega \in \Omega} | \langle \mathbf{r}_{k-1}, \mathbf{a}_\omega \rangle | \tag{3}$$

  where Ω is the index set of all the atoms in the dictionary and λ_k is the index of the atom.

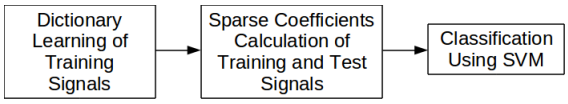

Fig. 1. Block diagram of the proposed audio signal classification system.

- Let $\Lambda_k = \{\lambda_1, \ldots, \lambda_k\}$ list the atoms that have been chosen at step k, then the kth approximant (coefficient) is calculated as

$$\mathbf{x}_k = \arg\min_{\mathbf{x}} \| \mathbf{y}_i - \mathbf{x} \| \quad s.t. \ \ \mathbf{x} \in \text{span}\{\mathbf{a}_\lambda : \lambda \in \Lambda_k\} \tag{4}$$

  This minimization can be performed incrementally by standard least-square techniques. The residual is then calculated as $\mathbf{r}_k = \mathbf{r}_{k-1} - \langle \mathbf{r}_{k-1}, \mathbf{a}_{\lambda_k} \rangle \mathbf{a}_{\lambda_k}$.

III. Speech Music Classification Algorithm

In this section, we describe an algorithm based on dictionary learning and sparse coding for the task of speech and music classification. In this algorithm, sparse coefficients of speech and music signals are used as discriminating features and the SVM is used as a classifier [20]. A block diagram of the audio classification system is shown in Figure 1. Dictionaries for the training speech and music signals are learned and used to find sparse coefficients of the training and the testing signals. Those sparse coefficients are used for audio signal classification.

A. Dictionary Learning of Training Signals By K-SVD

Dictionaries learning process for the set of training signals is shown in Figure 2. Two sets of training signals $y_\mathbf{s}$ and $y_\mathbf{m}$, one for speech and other for music respectively, are used. Each signal is passed through the K-SVD algorithm to get its dictionary. Before applying K-SVD, each one-dimensional signal y_{s_i} and y_{m_i} are firstly converted to two dimensional matrix $\mathbf{Y}_{s_i}$ and $\mathbf{Y}_{m_i}$, respectively, where i represents the index of audio signal in the set of training signals. These two matrices are fed to K-SVD to get their dictionary and coefficient matrices, $\mathbf{A}_{s_i}, \mathbf{X}_{s_i}$ and $\mathbf{A}_{m_i}, \mathbf{X}_{m_i}$, respectively. All speech signal dictionaries obtained are combined together to form a single large dictionary $\mathbf{A_s}$ such that

$$\mathbf{A_s} = [\mathbf{A}_{s_1}, \mathbf{A}_{s_2}, \ldots, \mathbf{A}_{s_k}] \tag{5}$$

where $\mathbf{A}_{s_i}$ is the dictionary of each speech signal y_{s_i} with a dimension $R^{n \times p_i}$ and $1 \leq i \leq K$, where p_i is the number of atoms in the dictionary $\mathbf{A}_{s_i}$, K is the total number of speech signals in the training set and $\mathbf{A_s}$ is the whole dictionary of

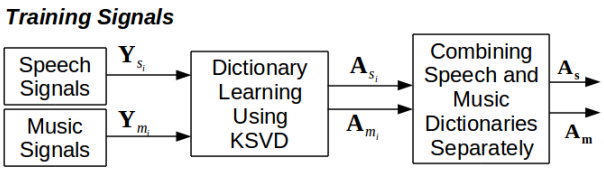

Fig. 2. Dictionaries learning for speech and music training signals.

all training speech signals with dimension $\mathbf{A_s} \in R^{n \times p}$ where $p = \sum_{i=1}^{K} p_i$.

In the same way, an overall large dictionary $\mathbf{A_m}$ for all music training signals is obtained by combining the dictionaries from individual music training signals i.e. $\mathbf{A_m} = [\mathbf{A}_{m_1}, \mathbf{A}_{m_2}, \ldots, \mathbf{A}_{m_k}]$ where $\mathbf{A_m} \in R^{n \times q}$, $\mathbf{A}_{m_i} \in R^{n \times q_i}$ and $1 \leq i \leq K$, where q_i is the number of atoms in the dictionary $\mathbf{A}_{m_i}$, K is total number of music signals in the training set and $q = \sum_{i=1}^{K} q_i$. These two dictionaries $\mathbf{A_s}$ and $\mathbf{A_m}$ are used to obtain the sparse coefficients of the training and testing signals, both from speech and music classes, in the sparse coding stage.

B. Sparse Coding of Training and Testing Signals

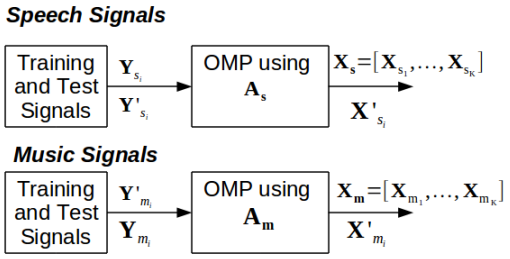

Fig. 3. Sparse coefficients extraction using OMP.

Sparse coefficients matrices of training and testing signals $\mathbf{X_s}$, $\mathbf{X}'_{s_i}$ and $\mathbf{X_m}$, $\mathbf{X}'_{m_i}$, from speech and music class respectively, are obtained by using the OMP algorithm [19]. OMP finds sparse coefficients by projecting input signals on the subspace spanned by dictionary atoms. Here we find sparse coefficients of input signals -speech and music training and testing signals- by projecting them onto the subspace spanned by the dictionary atoms of each class. Speech testing and training signals are encoded using the dictionary $\mathbf{A_s}$ and music testing and training signals are encoded using the dictionary $\mathbf{A_m}$. Sparse coefficient matrix $\mathbf{X_s}$ for speech training signals is obtained by combining all the coefficients matrices of each individual signal in speech training set. The sparse coefficient matrix $\mathbf{X_m}$ for music training signals, is also obtained in the same way. This process is depicted in Figure 3. The sparse coefficients of these training signals are used to train the SVM model for signal classification.

The sparse coefficients of speech and music test signals $\mathbf{X}'_{s_i}$ and $\mathbf{X}'_{m_i}$ are also obtained by the OMP algorithm. These sparse coefficients are to be classified as speech or music signals in the classification stage.

C. Signal Classification by SVM

Our motive for finding sparse coefficients is to use them for speech and music discrimination in the signal classification stage where the SVM is used as a classifier.

The diagram of the proposed audio signal classification system is shown in Figure 4. We use SVM for our binary classification problem where one class belonging to music signals has class label $z_i = +1$ and the other class belonging to

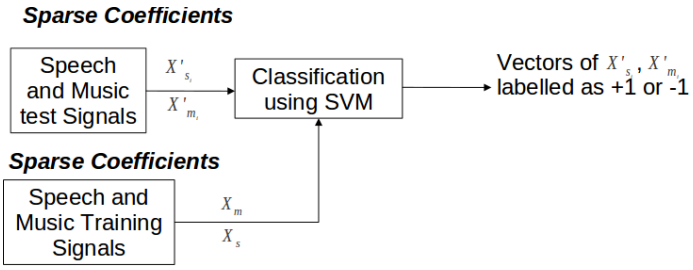

Fig. 4. Sparse coefficients based speech and music classification using SVM.

speech signal has class label $z_i = -1$. The training data points used to define the feature space (planes in terms of SVM) are vectors $\mathbf{x}_{s_i} \in R^p$ and $\mathbf{x}_{m_i} \in R^p$ from sparse coefficients matrices, $\mathbf{X_s} \in R^{p \times j}$ and $\mathbf{X_m} \in R^{p \times l}$, each for speech and music class, respectively.

A signal to be classified, in the form of sparse coefficient vector $\mathbf{x}'$ obtained in the sparse coding stage, is passed through SVM and its label z' is evaluated. If z' for the majority of sparse coefficient vectors $\mathbf{x}'$ is positive, it is considered as belonging to the music class otherwise to the speech class.

IV. Experiment and Results

For SVM classification, sparse coefficients of the training data are first used to train the SVM model. The class of test signals is then determined by passing the test signals in the form of sparse coefficients through SVM.

1) Dictionary Learning for Training Data: 35 different speech signals from TIMIT [21] database are used which include male and female speakers speaking different sentences with different style and accent. Each signal has a different duration ranging from 1.5 seconds to 5 seconds. However the total duration for 35 speech signals is 128 seconds sampled at 16 kHz. Other training data belongs to the music class which is composed of 16 piano signals with different notes, taken from the University of Iowa Musical Instruments Database [22]. These piano signals are sampled at 44.1 kHz having a total duration of 90 seconds.

To get the dictionaries representing the speech and music subspace, we first divide the whole 128 seconds long speech signal into four equal segments each of 32 seconds long. Dictionary for each speech segment is then learned using K-SVD. In the K-SVD algorithm, the signal formed as a matrix is given as an input to obtain the dictionary matrix and sparse coefficient matrix. Hence the 32 seconds speech signal is first converted to a matrix of size 128×3907 which is then passed through the K-SVD. The dictionary thus obtained is of size 128×150. Four such dictionaries are obtained for each of the four segments of speech. All these dictionaries are combined together to form a large dictionary $\mathbf{A_s}$ of size 128×600 that represents the speech signal subspace.

In the same way, to get the dictionary $\mathbf{A_m}$ of the music data, 90 seconds long music signal is also divided into 4 music segments each of length about 22 seconds. Before applying K-SVD, each music segment is converted to a matrix of size

128×7500, which after passed through K-SVD gives four music dictionaries each of size 128×150. Again all these four music dictionaries are combined together to form a large dictionary $\mathbf{A_m}$ of size 128×600 representing the music signal space.

2) Sparse Coefficients Calculation: Two dictionaries $\mathbf{A_s}$ and $\mathbf{A_m}$ for speech and music respectively, are used to get sparse coefficients by using the OMP greedy algorithm. For this purpose, the whole 128 seconds long speech signal and 90 seconds long music signal are converted into frames with each having 128 samples. These frames are used to calculate the sparse coefficient vectors of the training signal on a frame-by-frame basis. Each sparse coefficient vector $\mathbf{x}_{s_i}$ and $\mathbf{x}_{m_i}$ of dimension 600 has maximum 13 non-zero values. We also compare the sparse coding based features with other features such as MFCC. For this purpose, these frames are also used to get the MFCC coefficients of the training signals with each having a dimensionality 13. MFCC and sparse coefficients of the test signals are also computed and fed as inputs to the SVM algorithm.

3) Classification Results: In this binary classification system, where classification is based on sparse vectors, a positive label is assigned to music sparse coefficient vector and a negative label to speech sparse coefficient vector. Each test signal converted to sparse coefficient vectors of size 600 is passed through SVM to get class labelled vectors as the output. If the number of vectors with positive labels is greater than that of the negatively labelled vectors, then the signal is classified as music signal, otherwise speech signal.

For comparison, we also perform classification experiments based on the MFCC coefficients. 10 speech and 10 music signals are used to in the test stage to evaluate the classification performance using sparse coefficients and MFCC. Different levels of white Gaussian noise has been added to the testing signals to evaluate the robustness of the classification method. The classification percentage with different values of signal to noise ratio (SNR) is given in Table I.

TABLE I
SPEECH AND MUSIC SIGNALS CLASSIFICATION RESULTS

SNR (dB)	% classification based on MFCC		% classification based on Sparse Coefficients	
	Speech	Music	Speech	Music
0	100	100	100	100
-5	100	100	100	100
-10	100	100	100	100
-15	90	100	100	100
-17	70	100	90	100
-20	20	100	60	100

From Table I we can observe that for lower SNRs, the sparse coefficients give better performance as compared to the MFCC. Specifically, for SNR at -17 and -20 dB using sparse coefficients, 9 and 6 speech signals out of 10 testing signals were correctly classified as compared to 7 and 2 using MFCC, respectively.

V. CONCLUSION

We have presented a method of using learned dictionaries to extract signal features for music and speech classification. We learned two different dictionaries with each representing speech or music. Using those dictionaries, we calculated the sparse coefficients of speech and music. We found that those sparse coefficients were very good representatives of signal features that can be used for speech and music discrimination. Our preliminary results show that the sparse coefficients outperform the MFCC features for the task of music and speech classification, particularly for noisy data.

ACKNOWLEDGEMENT

Syed Zubair is funded by International Islamic University, Islamabad, Pakistan under the faculty / research development program. The authors are grateful to Fei Yan and Muhammad Awais for helpful discussions on classification.

REFERENCES

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-based classification, search and retrieval of audio," *IEEE Transactions on Multimedia*, vol. 28, pp. 27–36, 1996.

[2] C. West and S. Cox, "Features and classifier for the automatic classification of musical audio signals," in *International Conference on Music Information Retrieval*, 2004, pp. 531–537.

[3] M. J. Carey, E. S. Parris, and H. Lloyd-Thomas, "A comparison of features for speech, music discrimination," in *International Conference on Acoustics, Speech and Signal Processing*, vol. 1, March 1999, pp. 149–152.

[4] J. Saunders, "Real-time discrimination of broadcast speech/music," in *International Conference on Acoustics, Speech and Signal Processing*.

[5] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[6] T. Zhang and C. J. Kuo, "Hierarical classification of audio data for archiving and retrieving," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 6. Phoenix, Ariz, USA.: IEEE, 1999, pp. 3001–3004.

[7] J. T. Foote, "Content-based retreival of music and audio," in *SPIE*, 1997, pp. 138–147.

[8] K. El-Maleh, M. Klein, G. Petrucci, and P. Kabal, "Speech / music discrimination for multimedia applications," in *IEEE International Conference on Acoustic, Speech and Signal Processing*, vol. 6. Istanbul, Turkey: IEEE, June 2000, pp. 2445–2446.

[9] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech / music discriminator," vol. 2, Munich, Germany, April 1997, pp. 1331–1334.

[10] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and models*, ser. Springer Series on Information Science, 1999.

[11] G. Williams and D. P. W. Ellis, "Speech/music discrimination based on posterior probability features," in *Proceedings of the 6th European Conference on Speech Comunication and Technology*, Budapest, Hungry, September 1999, pp. 687–690.

[12] L. Lu, H. J. Zhang, and H. Jiang, "Content analysis for audio classification and segmentation," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 7, pp. 504–516, 2002.

[13] Z. Liu, J. Huang, Y. Wang, and I. T. Chen, "Audio feature extraction and analysis for scene classificaion," in *Proceedings of the 1st IEEE Workshop on Multimedia Signal Processing*, Princeton, NJ, USA, June 1997, pp. 343–348.

[14] J. G. A. Barbedo and A. Lopes, "A robust and computationally efficient speech / music discriminator," *Journal of the Audio Engineering Society*, vol. 54, no. 7-8, pp. 571–588, 2006.

[15] J. Ajmera, I. McCowan, and H. Bourlard, "Speech/music segmentation using entropy and dynamism features in a hmm classification framework," *Speech Communication*, vol. 40, no. 3, pp. 351–363, 2003.

[16] K. Delgado, J. F. Murray, B. D. Rao, K. Engan, T. Lee, and T. J. Sejnowski, "Dictionary learning algorithms for sparse representations," *Neural Computation*, vol. 15, pp. 349–396, 2003.

[17] M. Aharon, M. Alad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.
[18] M. Yaghoobi, T. Blumensmith, and M. E. Davies, "Dictionary learning for sparse approximations with the majorization method," *IEEE Trans. on Signal Processing*, vol. 57, no. 6, pp. 2178–2190, 2009.
[19] J. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. on Inform. Theory*, vol. 50, no. 10, pp. 2231–2242, 2004.
[20] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and other Kernel Based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000.
[21] J. S. Garofolo and et al, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," *Linguistic Data Consortium, Philadelphia*, 1995.
[22] *The University Of Iowa Musical Instruments Samples Database*, Online: http//theremin.music.uiowa.edu.