Learning Sparse Dictionaries for Music and Speech Classification

M. Srinivas, Debaditya Roy and C. Krishna Mohan
Department of Computer Science and Engineering
Indian Institute of Technology Hyderabad - 502205, India
Email: {cs10p002, cs13p1001, ckm}@iith.ac.in

***Abstract*—The field of music and speech classification is quite mature with researchers having settled on the approximate best discriminative representation. In this regard, Zubair et al. showed the use of sparse coefficients alongwith SVM to classify audio signals as music or speech to get a near-perfect classification. In the proposed method, we go one step further, instead of using the sparse coefficients with another classifier they are directly used in a dictionary which is learned using on-line dictionary learning for music-speech classification. This approach removes the redundancy of using a separate classifier but also produces complete discrimination of music and speech on the GTZAN music/speech dataset. Moreover, instead of the high-dimensional feature vector space which inherently leads to high computation time and complicated decision boundary calculation on the part of SVM, the restricted dictionary size with limited computation serves the same purpose.**

Keywords. *Music and Speech Classification, Sparse Representation, Dictionary Learning*

I. Introduction

Discriminating speech from other types of audio signals for e.g. music has become increasingly important due to its application in various domains like automatic transcription of news from correspondents in non-ideal recording environments where the speech signal is frequently interspersed with background noise or music. Also, it is highly applicable in defense applications where tremendous background noise overpowers the speech signal when communicating to headquarters from the battlefield.

The problem of of music-speech classification has been viewed from various perspectives and one of them is representing the audio signals with some known models to reveal the difference in structure of the two types of signals, music and speech. Ruiz-Reyes et al. [1] attempted to identify speech signals from a extracted from news and television programs based on the features derived from the fundamental frequency (F0) estimation of the audio signal. F0, also known as pitch portrays the periodicity of signal patterns in time domain. The entire audio signal is divided into frames with categoric distinction between voiced and unvoiced frames. The relative periodicity of these two types of frames is different in speech and music signals. These frequency patterns were then applied to a genetic-fuzzy classifier in conjunction with k-NN to achieve discrimination. Shirazi et al. [2] addressed the classification problem by modelling a sinusoidal representation of the audio signals to recognize the variance of birth frequencies and duration of the longest frequency track. The resulting features from this model namely, high zero crossing rate, low short time energy ratio, mean of spectral roll-off frequency, spectral centroid and mel-frequency cepstral coeffcients (MFCC) to both gaussian mixture model (GMM) and support vector machine (SVM) classifiers on their own music-speech corpus. Mubarak et al. [3] introduced modulation features for classification where amplitude modulation (AM) and frequency modulation (FM) features were extracted from a gammatone filter bank applied to the XM2VTS database. In addition cepstral coeffcients (CC) were also presented as input features to a GMM classifier where the output class was determined by comparing the outputs of the GMMs for both the music and speech signals.

Lim et al. [4] presented a real-time SVM based speech-music classifier for the selectable mode vocoder (SMV) codec to efficiently encode the input audio streams. The method of skipping some frames based on inter-frame correlation was introduced to reduce computation. This method required the classfication of signals by SVM based on six different features, since, the bit length required for coding speech signals is much higher and more number of frames can be skipped from music than speech signals. Ajmera et al. [5] gave an Hidden Markov Model (HMM) based classification framework. The entire audio signal was represented as an ergodic 2-state HMM model with the states signifying speech and non-speech (music) parts of the signal. The entropy in non-speech segments was observed to be higher and this was used to obtain posterior probablities of the signal segment. These probabilities were then used as input to the HMM model parameters, mainly transition and observation sequence probabilities and Viterbi algorithm was applied to align the input signal to the HMM states leading to classification.

The remainder of this paper is arranged as follows. Section II explains the proposed sparsity based classification method in detail with an insight into ODL. Section III deals with the experiments performed on the GTZAN music-speech dataset and relevant discussions. Section IV presents the conclusions and possible directions for future exploration.

II. Proposed Method

This section provides details on the process of learning sparse dictionaries for music and speech classification. At

first, feature extraction is covered in detail. In the next subsection, a gentle introduction to online dictionary learning is given. Finally, the entire procedure of construction of sparse dicitonaries for each of the classes is discussed in detail.

A. Feature Extraction

Tzanetakis et al. [**?**] introduced the GTZAN music-speech classification. The feature set proposed in this paper is collectively known as Marsyas [7] which includes feature groups like timbral texture, rhythmic and pitch content based features alongwith MFCCs. For the proposed method, 45 features were chosen, some form the Marsyas feature set others from various other experiments on audio signal classification. The chosen features consist of 39 MFCCs which are widely regarded as the feature coefficients that bestow the most discriminative characteristics of an audio signal. Alongwith MFCC, six other features were also chosen to further divulge the distinct characteristics of speech and music. These were energy entropy block, short time energy, zero-crossing rate, spectral roll-off, spectral centroid and spectral flux features giving a total of 45 features per frame of the . While the MFCCs are fine on their own the addition of these features helped in the design of an overcomplete dictionary for both music and speech which is discussed in the next subsection.

B. Sparse Dictionaries for classification

A signal $x \epsilon \mathbb{R}^m$ projects a sparse approximation over a dictionary D in $\mathbb{R}^{m\times k}$, with k columns each referred to as an "atom", when the linear combination of some atoms from D form a signal $\hat{x}$ that is "close" to x. The basic idea of classification is to represent the test data as a sparse linear combination of training data acquired from a dictionary.
For any class C_i, the examples belonging to this class are close to each other in a lower dimensional subspace. Let the p^{th} class have K_p training samples and the total number of training samples is denoted by $\{y_i^N\}$ where $i = 1, 2, \ldots, K_i$ and $K_1, K_2, \ldots, K_N$ are training samples corresponding to classes $C_1, C_2, \ldots, C_N$. Let b be a input vector belonging to the p^{th} class, then it is represented as a linear combination of the training samples belonging to class p.

$$b = D_p\Phi_p \tag{1}$$

where D_p is a $m \times K_p$ dictionary whose columns are the training samples in the p^{th} class and $\mathbf{\Phi}_p$ is a sparse vector for the same class. The two main steps involved in the method are :

1) *Training: Dictionary Construction* During training, dictionary for each class is formed from the training features using ODL [6]. Then $D = [D_1, \ldots, D_N]$ are calculated using the equation.

$$(\hat{\mathbf{D}}_\mathbf{i}, \hat{\mathbf{\Phi}}_\mathbf{i}) = \arg\min_{\mathbf{D}_i, \Phi_i} \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2}\|\mathbf{C}_i - \mathbf{D}_i\Phi_i\|_2^2 + \lambda\|\Phi_i\|_1 \tag{2}$$

where $\mathbf{C}_i = \hat{\mathbf{D}}_i\hat{\mathbf{\Phi}}_i, \quad i = 1, 2, \ldots, N$.

2) *Testing: Sparsity Calculation* During testing, the sparse vector Φ for given test feature is found in the test dataset $B = [b_1, \ldots, b_l]$. Using the dictionaries of training samples $D = [D_1, \ldots, D_N]$, the sparse representation Φ satisfying $D\Phi$=B is obtained by solving the following optimization problem:

$$\begin{aligned} \Phi_j &= \arg\min_\Phi \tfrac{1}{2}\|\mathbf{b_j} - \mathbf{D}\Phi_j\|_2^2 \\ &\text{subject to} \|\Phi_j\|_1 \le T_1, \\ \hat{i} &= \arg\min_i \|\mathbf{b_j} - \mathbf{D}\delta_i(\Phi_j)\|_2^2 \;\; j = 1, \cdots, t \end{aligned} \tag{3}$$

where δ_i is a characteristic function that selects the coefficients. Then b_j is assigned to C_i associated with the i^{th} dictionary. The sparsest dictionary for a given testing data is found using l_1 -lasso algorithm. The test example is assigned to the dictionary corresponding to class.

III. Experimental Results and Discussion

The chosen dataset for demonstrating our approach was the GTZAN music/speech dataset [7]. The dataset contains 64 audio clips each for the music and speech class. Each clip is 30 seconds long with variable number of frames. We adopted a 75-25 split for testing and training with 4-fold cross validation. The selection of training set was done randomly but it was made sure that every clip is tested for atleast once. So, for each fold 47 clips were chosen for training of the dictionary and the remaining 17 for testing.
The ensemble of features chosen gave rise to very large matrix which was about 175000×45 (variability due to uneven number of frames in each clip) for each class) in training and about 35000×45. This represented the classical case of an over-determined system for which an overcomplete dictionary was formed. A misclassification rate comparison of the ODL classifier with some of the other well-known classifiers is presented in Table I.

TABLE I
Comparison of error rates among classifiers on GTZAN music-speech dataset

Classifier	**Error Rate (in %)**	
	Music	Speech
Back Propagation NN	1.5	2.9
Bayesian	10.9	19.8
Kernel SVM	26.2	3.9
k-NN	5.8	22
Proposed(ODL)	0.0	0.0

Table I shows that all the classfiers except the back propagation NN perform well on either music or speech but not both. This can be attributed to the fact that the feature vector space for music and speech is not well discriminated for classification by these classifiers. However, the success of the proposed method validates the fact that even with overcomplete dictionaries which represent the data in considerably

lower dimension than the input feature vector space we can still attain total classification.

IV. Conclusion

A music-speech classification approach based on sparse representation of a set of feature descriptors using on-line dictionary learning on the GTZAN music-speech dataset is proposed in this paper. Perfect classification is achieved using the proposed approach which demonstrates that sparse representation is clearly a semantically rich description. Moreover, we have shown it outperforms most of the established classification systems making it a viable choice for representing audio signals with the only constraint on the representation being the features extracted form which the dictionary is designed. In future, this work can be extended to applications like voice artist recognition from a given audio clip.

References

[1] Nicolás Ruiz-Reyes, Pedro Vera-Candeas, JE Muñoz, S García-Galán, and FJ Cañadas. New speech/music discrimination approach based on fundamental frequency estimation. *Multimedia Tools and Applications*, 41(2):253–286, 2009.

[2] Jalil Shirazi and Shahrokh Ghaemmaghami. Improvement to speech-music discrimination using sinusoidal model based features. *Multimedia Tools and Applications*, 50(2):415–435, 2010.

[3] Omer Mohsin Mubarak, Eliathamby Ambikairajah, Julien Epps, and Teddy Surya Gunawan. Modulation features for speech and music classification. In *Communication systems, 2006. ICCS 2006. 10th IEEE Singapore International Conference on*, pages 1–5. IEEE, 2006.

[4] Chungsoo Lim and Joon-Hyuk Chang. Efficient implementation techniques of an svm-based speech/music classifier in smv. *Multimedia Tools and Applications*, pages 1–26, 2014.

[5] Jitendra Ajmera, Iain McCowan, and Hervé Bourlard. Speech/music segmentation using entropy and dynamism features in a hmm classification framework. *Speech Communication*, 40(3):351–363, 2003.

[6] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pages 689–696, New York, NY, USA, Jun. 2009. ACM.

[7] G. Tzanetakis. Marsyas(music analysis,retreival and synthesis for audio signals). http://marsyas.info/.