Audio Classification Using Dominant Spatial Patterns in Time-Frequency Space

Md. Khademul Islam Molla¹,² and Keikichi Hirose¹

¹ Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
² Department of Computer Science and Eng., The University of Rajshahi, Rajshahi, Bangladesh

{molla, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

This paper presents a novel audio discrimination algorithm using spatial features in time-frequency (TF) space. Three types of audio signals – speech, music without vocal and music with background vocal are taken into consideration for classification. The audio segment is transformed into TF domain yielding the spatial illustration of energy. Non-negative matrix factorization (NMF) is applied to TF space to extract a set of vectors which represents the dominant subspace of spatial energy distribution. The inverse Fourier transform is applied to individual dominant vectors to derive the features for audio discrimination. The classification is performed by using multiclass linear discriminant analysis (mcLDA). The experimental results show that the proposed algorithm is more noise robust and performs better than the recently reported methods.

Index Terms: Linear discriminant analysis, non-negative matrix factorization, speech/music discrimination, time-frequency representation.

1. Introduction

The efficient speech/music discrimination (SMD) has many potential applications including automatic speech recognition (ASR), multimedia indexing, low bit-rate audio coding etc. The SMD refers to segmenting audio stream and labeling each segment as speech or music. A number of SMD algorithms are developed using different acoustic features associated with various classification engines. In [1], the entropy of normalized spectral energy is used as the feature to discriminate radio recording into speech and music by employing dynamic programming and Bayesian network. The root mean square amplitude and zero crossing rates are used as the potential features in real time implementation of SMD [2]. Scheirer and Slaney [3] employed thirteen different acoustic features to train various types of multidimensional classifiers to solve the audio discrimination problem.

The classification accuracy of SMD algorithm and its robustness against noise directly affect the performance of ASR in adverse acoustic environment. It is necessary to select appropriate acoustic features and their effective combination to obtain the enhanced performance of SMD system. The time-frequency space of audio signal represents informative patterns in visual domain. Instead of using a number of acoustic features in a trivial way, the time-frequency based (spectrogram) features are considered in audio discrimination methods implemented by image recognition approach [5, 6, 7]. The texture property of spectrogram is adopted in [5]. The statistical characteristics of the energy distribution of individual block of spectrogram are used as the features in [6]. A multiple kernel learning technique is used to select potential features in our previous work [7]. In all cases, the full dimension of spectrogram is used in training and classification yielding the increase of computational cost, whereas, the use of dominant feature can improve the classification performance.

In this paper, we demonstrate an audio classification technique using the dominant spatial features derived from TF (spectrogram) domain. The audio segment is transformed into grayscale image using short-time Fourier transform (STFT) [6]. Distinctive spatial patterns for speech and music are noticed in the image of TF representation. Instead of using the overall image, a set of vectors representing the dominant patterns are extracted by employing nonnegative matrix factorization (NMF) [8]. Thus obtained vector space is used to derive the features for classification. The application of NMF on positive valued image like spectrogram is very much significant. In addition to the speech and music, discrimination of music with vocal is an important issue what is considered here. An efficient classification scheme of these three types of audio segments is introduced here. A multi-class linear discriminant analysis [9] is implemented to achieve the proposed audio discrimination.

2. Audio Classification Algorithm

The difference between speech and music is that in music the modulation is more synchronous over all bands or a wider frequency range, while it differs more in speech [10]. The narrow band energy of music is also more stable and long lasting along the temporal scale than that of the speech signal. The energy of speech has a high correlation, especially between adjacent bands, while music has a rather low correlation between all bands. Another potential characteristic of music is that it has common harmonic structure which is defined as the inverse Fourier transform of linear power spectrum with log-scaled frequency [11]. Unlike music, the speech signal does not consist of such harmonic structure, and this is the key decision factor of the proposed algorithm. The TF distribution (spectrogram) illustrates the time varying spectra with complete information of the signal in both time and frequency domain. The features representing the discriminant characteristics of speech and music signals can be derived from the spectrogram. The principal steps of the proposed algorithm can be expressed as:

i) The audio segment is transformed to TF domain (spectrogram using STFT) – image like spatial representation of energy density.

ii) The NMF is applied to derive a set of vectors representing the dominant spatial patterns of energy density in the spectrogram. The normalized summation of inverse Fourier transform of these vectors is used as features in the proposed audio discrimination.

iii) Three types of audio signals are taken into consideration for classification. The training and classification are performed by using linear discriminant analysis for multiple classes.

The individual steps of the proposed algorithm are described in the following subsections.
2.1. Grayscale spectrogram

The spectrogram obtained by STFT is an image like representation of energy density of any signal. The content of any signal is properly localized along both time and frequency scales in time-frequency space. The audio stream is divided into fixed length segments. The TF representation of each segment \( x(t) \) is derived by STFT as \( F(\tau, f) = \mathcal{A}[x(t)] \); where \( \mathcal{A}[\cdot] \) is the STFT operator. The 20ms Hamming window with 10ms overlap and 512 points FFT are used to implement the STFT. The linear spectrogram is the squared magnitude of \( F(\tau, f) \). The human perception of sound is in logarithmic scale and hence the log spectrogram is computed as:

\[
S(\tau, f) = \log|F(\tau, f)|^2.
\]

To derive the nonnegative TF space, the spectrogram \( S(\tau, f) \) is transformed to positive coordinate: \( S(\tau, f) = \sqrt{S_{\min}} + |S_{\max} - S_{\min}| \); where \(|\cdot|\) is the absolute value operator. Finally, the TF matrix is normalized into grayscale image, with the range scaled between \([0, 1]\):

\[
X(\tau, f) = \frac{\tilde{S}(\tau, f) - \tilde{S}_{\min}}{\tilde{S}_{\max} - \tilde{S}_{\min}}
\]

Thus obtained TF representation (spectrogram) is a nonnegative matrix to which NMF is applied for feature extraction. The spectrograms of three types of audio segments are illustrated in Fig. 1.

The time-frequency space illustrates detailed spectro-temporal energy distribution which can be used to derive potential features in audio classification. Some dominant patterns are noticed representing the discriminative characteristics of different audio signals: speech, music and with background vocal (speech). The spectrogram image shows that the music signal is more harmonious than speech. The TF space (spectrogram) of music with vocal has the mixed effects of both (speech and music). The energies are concentrated at the lower and relatively higher frequencies for the tonal and non-tonal parts of the speech respectively. The music spectra are repeated with (almost) regular patterns along the frequency and change more slowly over time than those of the speech signal. The energy distribution of the music signal can be treated as the succession of segment of relatively stable notes, whereas, the energy of speech is not continuous in both spectral (vertical) and temporal (horizontal) axes. The spatial energy distribution of music with vocal is appeared as the mixture of the energy patterns of music and speech (Fig. 1(c)).

2.2. Feature extraction

The spectrogram image contains a noticeable amount of redundant information which decreases the performance of audio classification. Only the dominant energy patterns are required to enhance the classification accuracy. The texture-like time-frequency representation usually contains distinctive patterns that capture different characteristics of the audio signals. It is critical to extract features that capture major spectro-temporal characteristics of signal to achieve a high accuracy in audio discrimination. The NMF is employed here to extract the dominant spatial patterns from the time-frequency space of the audio signals.

A useful data representation typically represents the latent structure in the data explicit, and often reduces the dimensionality of the data so that further computational methods can be applied. The NMF is a recent method for deriving such a representation [12]. In the TF matrix \( X_{\text{norm}} \) obtained by Eq. (1), each column is an \( n \)-dimensional nonnegative vector representing the spectrum of a short time frame. Using NMF, \( X \) can be approximated by two new matrices \( W_{ao} \) and \( H_{ao} \) as \( X \approx WH \) or:

\[
X_{ia} = (WH)_{ia} = \sum_{r=1}^{m} W_{ir}H_{ra}
\]

where \( X_{ia} \) is a single element of \( X \) defined in the transformed domain, \( r \) is chosen such that \((n+m)r < nm \). Each column of \( W \) contains a basis vector while each column of \( H \) contains the weights needed to approximate corresponding column in \( X \). In order to estimate the factorized matrices, the objective function is defined by

\[
h = \sum_{i=1}^{n} \sum_{a=1}^{m} \{X_{ia} \log WH\}_{ia} - (WH)_{ia}\}
\]

subject to the non-negativity constraints of the factorization. The log spectrogram is normalized within the range \([0, 1]\) representing a grayscale image and hence it does not affect the derivation of the objective function. The objective function can be related to the likelihood of generating subspace in \( X \) from the bases \( W \) and \( H \). The iterative rules to reach the local maximum of the objective function are given by [13]:

\[
W_{ia} \leftarrow W_{ia} \frac{X_{ia}H_{ia}}{(WH)_{ia}}
\]

\[
H_{ia} \leftarrow H_{ia} \frac{W_{ir}X_{ia}}{(WH)_{ia}}
\]

The initialization is performed using positive random initial condition for both \( W \) and \( H \). The fidelity of the approximation enters the updates through the quotient \( X_{ia}/(WH)_{ia} \) [12]. The update rules preserve the non-negativity of \( W \) and \( H \) and also constrain the columns of \( W \) to sum to unity. This sum constraint is a convenient way of eliminating the degeneracy.
associated with the invariance of $WH$ under the transformation $W \rightarrow W\Lambda H \rightarrow \Lambda^2 H$, where $\Lambda$ is a diagonal matrix.

The basis vectors in $W$ represent the latent harmonic structure of audio signals varying along frequency bands, whereas, the column vectors of $H$ illustrate the time varying energy distribution. It does not provide noticeable distinctive characteristic of the audio signals and hence not used in classification. The first $r$ columns of $W$ are taken representing the dominant spatial energy patterns to form the characteristic vector as: $E = [w_1^T, w_2^T, \ldots, w_r^T]$ ; where $w_i^T$ represents the transpose of $w_i$. The first component ($w_1$) of speech, music and music with vocal are shown in Fig. 2.

![Figure 2](image)

Figure 2. The first component ($w_1$) of characteristic vector for speech (top), music (middle) and music with vocal (bottom panel). The first dominant vector of speech signal is similar to rippling on a trend; music exhibits a regular variation of energy; the music with vocal illustrates regular energy variation superimposed on a high energy trend.

The higher energy of audio signal is concentrated at the lower frequency region due to the pitch period. The energy up to the upper pitch boundary (500Hz) is discarded from each vector $w_i$ ($i=1,2,\ldots,r$) to minimize the dynamic range. The zero padding is performed at the end of the modified vectors to retain its original dimensions.

![Figure 3](image)

Figure 3. The feature vectors (obtained by Eq. (5)) of different audio signals: speech (top), music (middle) and music with vocal (bottom). The quefrency means inverse of frequency.

The inverse Fourier transformation is applied to each component of vector $E$ and summed together to obtain the feature vector. Unlike cepstrum which is the inverse Fourier transform of the log-scaled power spectrum, the inverse Fourier transform of the linear-scale power spectra as well as frequency is used to derive the features. The feature vector $y$ corresponding to the characteristic vector $E$ is computed as:

$$y = N_i \sum_{i=1}^{n} \mathfrak{z}^{-1}(w_i)$$  \hspace{1cm} (5)$$

where $\mathfrak{z}^{-1}()$ is inverse Fourier transform operator and $N_i()$ is an amplitude normalization operator. The relatively higher energy peaks are appeared in the feature space due to the regular energy repetition (as shown in Fig. 2) of music signal. The feature vectors of different audio signals illustrated in Fig. 3 indicate that unlike speech, the music signal contain the repetition of regular energy patterns along the spectral axis.

In is observed that the energy of feature vector of speech is decreasing exponentially (no energy repetition event is present); for music, there are some higher valued energy peaks indicating the presence of harmonic structure and for the music with vocal consists of both the effects. The obtained features demonstrate the distinguishing characteristics of the corresponding audio signals.

### 2.3. Classification by multi-class LDA (mcLDA)

The basic idea of linear discriminant analysis (LDA) is to find a linear transformation that best discriminate among the classes and classification is then performed in the transformed space based on some metric. A multi-class LDA is adopted here to classify the audio stream into three different classes – speech, music and music with vocal (speech).

Consider that we have $k$ $p$-dimensional samples $y_1, y_2, \ldots, y_k$ obtained by Eq. (5); where $y_{\ell} = (y_{\ell1}, y_{\ell2}, \ldots, y_{\ell p})^T$ belonging to any of $c$ different classes. In multi-class LDA, it is required to maximize the ratio of intra-class scatter to the inter-class scatter among all the competing classes [9]. The intra-class matrix can be calculated as:

$$\psi = \sum_{i=1}^{c} \sum_{y_{\ell} \in C_i} (y_{\ell} - \overline{y}_i)(y_{\ell} - \overline{y}_i)^T$$  \hspace{1cm} (6)$$

The inter-class scatter matrix is derived by

$$\xi = \sum_{i=1}^{c} k_i (\overline{y}_i - \overline{y})(\overline{y}_i - \overline{y})^T$$  \hspace{1cm} (7)$$

where $k_i$ is the number of training samples and $\overline{y}_i$ is the mean of $i^{th}$ class. The total mean vector $\overline{y}$ is given by

$$\overline{y} = \frac{1}{k} \sum_{i=1}^{c} k_i \overline{y}_i$$  \hspace{1cm} (8)$$

After obtaining $\psi$ and $\xi$ the linear transformation $\phi$ can be derived by solving the generalized eigenvalue problem: $\xi \phi = \psi \phi$. The upper bound of the rank of $\psi$ and $\xi$ are $k-c$ and $c-1$ respectively. Once the transformation $\phi$ is given, the classification is then performed in the transformed space on the distance metric (Euclidean distance). Then upon the arrival of a new instance $z$ (the feature vector of an audio segment from testing set), it is classified to the $g^{th}$ class as:

$$\text{arg min}_{g} \phi^T (z) \phi (\overline{y}_g)$$  \hspace{1cm} (9)$$

where $\phi^T (z)$ is the function to measure the distance and $\overline{y}_g$ is the centroid of $g^{th}$ class. Multiple discriminant analysis provides an elegant way for classification using discriminant features.

### 3. Results and Discussion

The efficiency of the proposed audio discrimination algorithm is evaluated using “music-speech” database. The datasets developed by Scheirer and Slaney are considered as
benchmark and used in multi-features speech/music disermination (MFD) [3]. This corpus contains speech, music without vocal (mWTv) and music with background vocal (mWv) each containing 80 samples of 15 seconds long. The audio samples are collected by digitally sampling (22 kHz sampling frequency) an FM tuner with different stations, contents and styles. The experimental results are presented for all samples which are jackknifed into four cuts with ¼ cut used to train the classifier and ¼ cut for testing [4]. The results are given as the true classification (in %) of the audio segment.

The first experiment is conducted to classify speech and music without vocal (mWTv) for all of 160 samples. The spectrogram is generated (using 512 point FFT of 20ms Hamming with 10ms shifting) for the audio segment of length 2.5sec [3, 4, 7]. Each sample of length 15sec consists of 7 segments. A set of vectors representing spectral energy patterns are extracted from the spectrogram using NMF as described in subsection 2.2. Only six dominant vectors selected which are used to derive proposed features. The classification is performed with the extracted features using mcLDA. The performance of the proposed algorithm is compared with MFD [3], posterior probability based features (PPF) [4] and multiple kernel learning (MKL) [7] using the mentioned “music-speech” datasets (in same experimental conditions) as shown in Table 1. The proposed multi-class LDA (mcLDA) with NMF based features performs better than the others. Although the classification accuracy of the proposed method is very close to MKL [7], the required computational cost is lower than that of MKL.

In the second phase, the classification is performed on three types of audio signals – speech, mWTv and mWv using mcLDA with NMF based proposed algorithm. Various real world noises are added in different noise level to simulate the performance in noisy environment. The training is performed with the original data and the testing is made with the noisy data. The performances with noise are illustrated in Table 2. In contrast to [6], the method does not require training with noisy data and hence provides its robustness against noise. It is noted that the classification performance of speech signal is decreased with increasing the noise contamination. The noisy spectrogram enhances music-related characteristics yielding some speech segments are erroneously classified as music.

The length of segment is an important issue in audio processing. Although performance is a main issue, the smaller segment length is a better choice in most of the audio applications. The effect of segment length (1.0s to 3.5s) on the performance of proposed method is studied as shown in Fig. 4. It is observed that the performance variation is not so much rigorous after the segment length of 2sec. Also the effect of the number of dominant vectors obtained by NMF is studied as illustrated in Fig. 5. The best performance is obtained when 6-7 vectors are considered. The large number of vectors includes more detail of energy distribution in addition to the dominant patterns yielding the decrease of performance.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Speech</th>
<th>mWTv</th>
<th>mWv</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposed</td>
<td>99.16</td>
<td>99.40</td>
<td>99.28</td>
<td></td>
</tr>
<tr>
<td>MKL[7]</td>
<td>98.96</td>
<td>99.17</td>
<td>99.02</td>
<td></td>
</tr>
<tr>
<td>PPF[4]</td>
<td>99.17</td>
<td>98.33</td>
<td>98.75</td>
<td></td>
</tr>
<tr>
<td>MFD[3]</td>
<td></td>
<td></td>
<td>98.60</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. The comparison of different algorithms in term of classification accuracy (%) on ‘music-speech’ corpus

<table>
<thead>
<tr>
<th>Noise</th>
<th>Speech</th>
<th>mWTv</th>
<th>mWv</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>98.22</td>
<td>99.30</td>
<td>98.15</td>
<td>98.56</td>
</tr>
<tr>
<td>Babble</td>
<td>97.24</td>
<td>99.35</td>
<td>98.10</td>
<td>98.23</td>
</tr>
<tr>
<td></td>
<td>96.15</td>
<td>99.64</td>
<td>96.72</td>
<td>97.50</td>
</tr>
<tr>
<td></td>
<td>94.34</td>
<td>100.00</td>
<td>90.55</td>
<td>94.96</td>
</tr>
<tr>
<td>Jet</td>
<td>97.23</td>
<td>99.38</td>
<td>98.13</td>
<td>98.25</td>
</tr>
<tr>
<td></td>
<td>94.50</td>
<td>100.00</td>
<td>94.12</td>
<td>96.21</td>
</tr>
<tr>
<td></td>
<td>71.38</td>
<td>100.00</td>
<td>89.86</td>
<td>87.08</td>
</tr>
<tr>
<td>White</td>
<td>94.60</td>
<td>99.68</td>
<td>96.35</td>
<td>96.88</td>
</tr>
<tr>
<td></td>
<td>89.27</td>
<td>100.00</td>
<td>91.75</td>
<td>93.67</td>
</tr>
<tr>
<td></td>
<td>68.24</td>
<td>100.00</td>
<td>84.15</td>
<td>84.13</td>
</tr>
</tbody>
</table>

Table 2. The classification accuracy (%) of speech, mWTv, mWv using the proposed method on “music-speech” corpus at presence of different types of noises

4. Conclusions

A noise robust algorithm is proposed here to discriminate speech, music without vocal and music with background vocal. The dominant spatial patterns obtained by applying NMF on time-frequency space (spectrogram) are used to derive discriminative features. The superiority of this algorithm is that it can classify the mentioned three different types audio signals rather than simple speech/music discrimination reported by existing methods. It is less sensitive to different types of noise and the length of audio segment. These characteristics are very much useful in practical applications.
5. References


