SPEECH/MUSIC CLASSIFICATION USING BLOCK BASED MFCC FEATURES

Vikaskumar Ghodasara ghodasara.vikas@gmail.com

Daimi Syed Naser sndaimi123@iitkgp.ac.in

Shefali Waldekar shefali.waldekar@gmail.com

Goutam Saha gsaha@ece.iitkgp.ernet.in

Electronics & Electrical Communication Engineering Department,
Indian Institute of Technology Kharagpur, India

ABSTRACT

Classifying an audio stream as either speech or music is receiving wide spread attention due to its varied applications. In this paper, we propose a novel block based mel frequency cepstral coefficient (MFCC) feature extraction method for music and speech classification. We found that the proposed features give better classification accuracy as compared to conventional MFCC features and zero crossing rate (ZCR) features. Here, we use support vector machine (SVM) classifier with 3-fold cross validation scheme. Evaluation is done on GTZAN music/speech dataset. Further, we investigate the effect of number of blocks, size of each block and number of filter banks on the classification performance.

1. INTRODUCTION

The problem of discrimination between speech and music signal has assumed great importance for bandwidth efficient transmission. This approach has been used by Spina et al [1] for automatic speech recognition of general audio data by making the speech/non-speech decision after classifying the audio into seven different classes. In the past few years, different feature extraction techniques have been proposed for speech and music discrimination. In [2] pioneered this area by presenting a zero-crossing technique and thereafter different features for classification of speech and music were proposed in recent years.

However, better results were apparent through the use of frequency domain features such as cepstral, pitch and zero crossing rate (ZCR) along with Gaussian mixture models (GMMs) [3]. Mel frequency cepstral coefficients (MFCCs) gained popularity because the best results obtained were from the cepstral domain.

A system proposed in [4], based on cepstral and spectral features using GMMs with fusion of outputs from four GMMs, reported an accuracy of 90.25% from the evaluation performed on 12 hours of French Radio broadcast.

Lack of standard and freely available speech and music database makes comparison of audio indexing techniques a complicated matter. However, results from other databases can still give a fair idea of the performance.

This paper presents an analysis of speech/music classification system using MFCC based features.

2. ANALYSIS OF SPEECH AND MUSIC SIGNALS

2.1 MFCC

The mel frequency cepstrum has proven to be highly effective in recognizing structure of music signals and in modeling the subjective pitch and frequency content of audio signals. Psychophysical studies have found the phenomena of the mel pitch scale and critical band, and the frequency scale-warping to the mel scale has led to the cepstrum domain representation. The mel scale is defined as

$$\boldsymbol{F_{mel} = 2595 \times log_{10}\left(1 + \frac{f}{700}\right)} \tag{1}$$

where $\boldsymbol{F_{mel}}$ is the logarithmic scale of $\boldsymbol{f}$ normal frequency scale. The steps used to compute MFCC features are as follows:

Step-1: Pre emphasis (by factor 0.97)

Step-2: Framing (20ms and 50% overlap)

Step-3: Windowing

Step-4: Discrete Fourier Transform

Step-5: Mel scaled filter bank

Step-6: Energy computation

Step-7: Log

Step-8: Discrete Cosine Transform

Step-9: MFCC

Pre-emphasis, which is a high pass filtering process, is used to boost the spectrum of voiced sounds which suffer a steep roll-off of 6 dB/octave towards high frequency region. Pre-emphasis is followed by framing and windowing of the filtered signal. Speech is quasi-stationary signal due to slow varying nature of vocal tract. Therefore, to estimate the spectral characteristics, signal should be analyzed over shorter duration frames (20-30ms). Adjacent frames are overlapped to preserve the boundary information. This is followed by estimation of power spectrum, multiplication with filter banks, logarithm of the result and discrete cosine transform (DCT) to generate de-correlated feature vectors.

2.2 Block Based MFCC

In MFCC computation, as seen in above section, DCT is applied on all the log energy coefficients at a time. To fully de-correlate the features DCT is applied in blocks, similar to the approach used in image processing. In [5], block DCT based MFCC was shown to outperform full-band MFCC in speaker recognition system. In [5] DCT is applied on two blocks. They have observed that non-overlapping blocks with first block covering approximately the log energies of frequency bands equal to span of first formant and second block on remaining two formants, gives best performance.

2.3 Proposed Feature Extraction Method

Spectral features like MFCC are widely used to discriminate speech and music. The frequency band range for speech and music are 20 Hz – 4 kHz and 20 Hz – 20 kHz respectively. Figure 1 shows log power spectrum of speech and music.

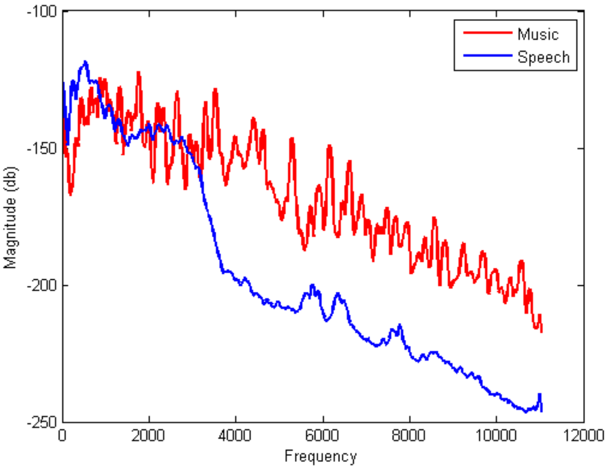

Figure 1. Log power spectrum of speech and music

In [5], it is observed that first formant frequency range is around 1 kHz and remaining two formants between 1 kHz – 4 kHz. Here, we compute MFCC over three blocks; the third block range is from 4 kHz to half of the sampling frequency, because in case of music, the sampling frequency required is greater than 8 kHz (the usual sampling frequency for speech). The Cepstral coefficients $\{X_i\}$, using non-overlapping three blocks, with first and second block size q and r respectively, can be expressed as:

$$\{X_i\}_{i=1}^{q-1} = \sqrt{\frac{2}{q}}\sum_{j=0}^{q-1}\Psi(j+1)\times cos\left(\frac{\pi i(2j+1)}{2q}\right) \qquad (2)$$

$$\{X_i\}_{i=q}^{q+r-2} = \sqrt{\frac{2}{r}}\sum_{j=0}^{r-1}\Psi(j+q+1)\times cos\left(\frac{\pi(i-q+1)(2j+1)}{2r}\right) \qquad (3)$$

$$\{X_i\}_{i=\alpha-1}^{p-2} = \sqrt{\frac{2}{p-\alpha}}\sum_{j=0}^{p-(\alpha+1)}\Psi(j+\alpha+1)\times cos\left(\frac{\pi(i-(\alpha-2)(2j+1)}{2(p-\alpha)}\right) \qquad (4)$$

Where $\alpha = q + r$, and Ψ is vector of p filter banks log energies. One approach is to use Gaussian Mixture Model (GMM) as classifier to classify MFCC features. Another approach is to use GMM to generate super vectors of MFCC features and then use SVM to classify them. Both these approaches require huge computation. Here, instead of using GMM we used statistical parameters like mean and standard deviation over all frames as feature vectors.

3. RESULTS

For evaluation we used GTZAN [6] music/speech dataset which has total 128 audio files; 64 are speech files and 64 are music files. Each file is around 30s long, with sampling frequency 22.05 kHz, 16-bit wav format. We divided both speech and music (64 files) into three set (21, 21, 22 files). Using two sets we trained SVM classifier and remaining one set was used for testing. The results of three-fold cross validation are tabulated. It is found that 40 filters and three blocks give better performance than any other combination. In two-block MFCC, first block is of 0-1 kHz and second block is of 1 kHz - $(f_s/2)$ kHz. In three block case, the frequency range is 0-1 kHz, 1-4 kHz and 4 - $(f_s/2)$ kHz for first, second and third block, respectively, where f_s is sampling frequency.

No. Of filter banks	MFCC	MFCC (2 blocks)	MFCC (3 blocks)
20	95.70	94.53	92.18
40	97.65	96.09	**98.43**
60	96.48	96.87	97.65

Table 1. Results for Standard deviation as statistical property

No. Of filter banks	MFCC	MFCC (2 blocks)	MFCC (3 blocks)
20	78.51	76.56	76.56
40	78.51	76.17	77.34
60	75.39	73.43	73.04

Table 2. Results for Mean as statistical property

4. CONCLUSION

The performance of block-based MFCC as a feature and SVM-based speech/music classification system for varying number of filter banks, number of blocks, and mean, standard deviation as a statistical parameter is investigated. The results indicate that the best choice is of 40 filter banks with three blocks and standard deviation as statistical parameter, when compared to other combinations. For the database employed in this work, this system can attain a music/speech classification accuracy of up to 98.43%.

5. REFERENCES

[1] Spina, Michelle S., and Victor W. Zue. "Automatic transcription of general audio data: Preliminary analyses." In Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, vol. 2, pp. 594-597. IEEE, 1996.

[2] Saunders, John. "Real-time discrimination of broadcast speech/music." In icassp, pp. 993-996. IEEE, 1996.

[3] Carey, Michael J., Eluned S. Parris, and Harvey Lloyd-Thomas. "A comparison of features for speech, music discrimination." In Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on, vol. 1, pp. 149-152. IEEE, 1999.

[4] Sénac, Christine, and Eliathamby Ambikairajah. "Audio indexing using feature warping and fusion techniques." In Multimedia Signal Processing, 2004 IEEE 6th Workshop on, pp. 359-362. IEEE, 2004.

[5] Sahidullah, Md, and Goutam Saha. "Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition." *Speech Communication* 54, no. 4 (2012): 543-565.

[6] GTZAN Music/Speech Dataset: http://marsyasweb.appspot.com/download/data_sets/