# SPECTROGRAM BASED FEATURES SELECTION USING MULTIPLE KERNEL LEARNING FOR SPEECH/MUSIC DISCRIMINATION

*Sharmin Nilufar, Nilanjan Ray, M. K. Islam Molla*

University of Alberta, Edmonton, AB, Canada
Email:{sharmin, nray1, mdkhadem}@ualberta.ca

*Keikichi Hirose*

University of Tokyo, Tokyo, Japan
Email: hirose@gavo.t.u-tokyo.ac.jp

## ABSTRACT

This paper presents a multiple kernel learning (MKL) approach to speech/music discrimination (SMD). The time-frequency representation (spectrogram) implemented by short-time Fourier transform (STFT) of audio segment is decomposed by wavelet packet transform into different subband levels. The subbands, which contain rich texture information, are used as features for this discrimination problem. MKL technique is used to select the optimal subbands to discriminate the audio signals. The proposed MKL based algorithm is applied for SMD of a standard dataset. The experimental results show that the proposed technique yields noticeable improvements in classification accuracy and tolerance toward different noise types compared to the existing methods.

***Index Terms—*** Multiple kernel learning, spectrogram, speech/music discrimination, wavelet packet transform

## 1. INTRODUCTION

The speech/music discrimination (SMD) task is an important part of automatic speech recognition (ASR) systems, where it is used to disable the ASR when music or other classes of audio are present in automatic transcription of speech. Scheirer and Slaney [1] and Williams and Ellis [2] developed and evaluated different SMD systems for ASR of audio sound tracks. Another application that can benefit from SMD is low bit-rate audio coding. The existing SMD algorithms investigated different types of features for classification. The signal amplitude measured in root mean square (RMS) and zero-crossing (ZC) are used in real time implementation of SMD [3]. The dynamic programming and Bayesian networks with the entropy of the normalized spectral energy are applied for SMD of radio recording in [4]. In [1], thirteen different audio features were used to train different types of multidimensional classifiers, including a Gaussian maximum, a posteriori (MAP) estimator and a nearest neighbor classifier. The cepstral coefficients with Gaussian mixture and support vector machine (SVM) are used for classifying or segmenting speech and music in [5].

Although the acoustic features represent the characteristics of audio signals effectively, it is nontrivial to select and combine different types of acoustic features to obtain superior classification performance. The time-frequency space of the audio signal illustrates interesting patterns in the visual domain. Little attentions have been given to audio classification in the visual domain. The time-frequency based (spectrogram) features can be used in audio discrimination in the analogous way of image classification [6][7].

A novel idea of generating potential features from spectrogram is adopted here. The main idea is that, the segments of the analyzing audio stream is represented as spectrogram using STFT. We have proposed to use wavelet decomposed sunband images of the spectrogram as features to classify audio stream into two groups- speech and non-speech (music). Such type of feature illustrates both of spectral and temporal characteristics of the audio segment.

Wavelet transform provides an interesting multiscale analysis of images. Discrete wavelet packet transform (DWPT), which is the generalization of the discrete wavelet transform (DWT), provides richer subband analysis without the constraint of a dyadic decomposition. The decomposed subband images contain rich texture information which are suitable for further classification tasks. However huge number of features generated in wavelet packet transforms results in an extremely high dimensional feature space, most of which are redundant for accurate and efficient classification process. Selection of effective features from this high dimensional space is daunting. To determine subbands suited well for classification, it is important to identify an appropriate selection criterion and a search strategy to optimize this criterion.

In this paper, multiple kernel learning (MKL) [8] technique is investigated to select the optimal subbands to discriminate the two classes. It gives the flexibility of applying different kernel function based on our selected wavelet basis function. Moreover, MKL provides a generic method for selecting groups of subbands retaining computational feasibility [9].

## 2. PROPOSED SMD ALGORITHM

The proposed SMD algorithm is designed to perform accurate discrimination of speech and music from an offline sys-

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

tem structure (recorded digital audio). The proposed algorithm consists of three main parts: (i) represent the audio signal as an image by time-frequency representation, i.e. generate the gray scale spectrogram, (ii) extract features by applying wavelet packet transform on spectrogram, and (iii) apply MKL for subband selection and discrimination. Each part is described in the following subsections.

### 2.1. Time-frequency Representation

The time-frequency representation (TFR) contains the complete information of an audio signal in both spectral and temporal domain [7]. The digital audio stream is divided into segments of fixed length for spectrogram generation using TFR. Each segment is transformed into TFR by applying STFT as follows:

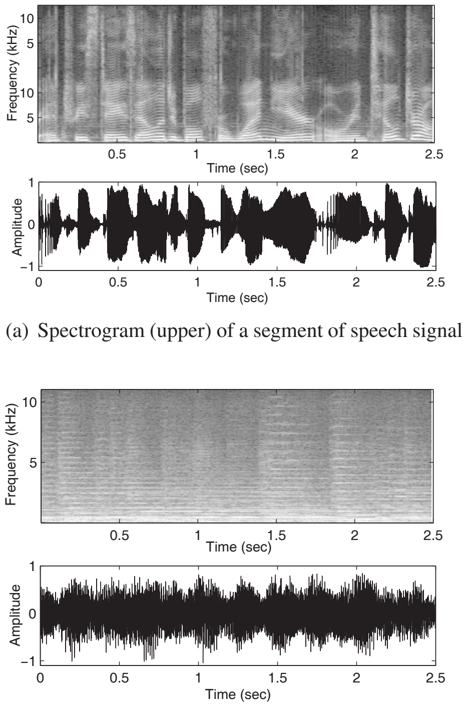
$$\psi\left\{x(n)\right\} = X(\tau, k) = \sum_{n=0}^{N-1} x(n)w(n-\tau)e^{-jnk} \qquad (1)$$

where $\psi$ is the STFT operator, $x(n)$ is the signal to be transformed, $w(n)$ is Hamming window function of length $N$. Then the linear spectrogram is the squared magnitude of STFT and given as: $S_{Linear}(\tau, k) = |X(\tau, k)|^2$. To implement the STFT, 20ms framing with 10ms overlap is used to reduce artifacts at the boundary. The human perception of sound is logarithmic and hence the log-spectrogram defined as: $S(\tau, k) = log\left\{S_{Linear}(\tau, k)\right\}$

The time-frequency matrix is normalized into grayscale normalized image, within the range $[0, 1]$:

$$H(\tau, k) = \frac{S(\tau, k) - S_{min}}{S_{max} - S_{min}} \qquad (2)$$

The spectrograms $H(\tau, k)$ and the corresponding speech and music signals are shown in Fig 1. It is critical to extract features that captures major temporal-spectral characteristics of signal to achieve a high accuracy in audio classification. The texture-like time-frequency representation usually contains distinctive patterns that capture different characteristics of speech and music signals. The spectral analysis shows that music signal is more harmonious than speech, since pure speech contains a sequence of tonal (voiced) and noise-like (unvoiced) sounds [6]. With the voiced part of speech signal, the energy is mostly concentrated within the lower frequency bands and at higher frequency bands for the unvoiced parts. It is observed from Fig. 1 that the music spectra change more slowly than speech and music can be regarded as a succession of segments of relatively stable notes, whereas speech is a series of alternate noisy and nearly steady parts. Hence, energy distribution of the speech signal in time-frequency domain is not continuous in both spectral (vertical) and temporal (horizontal) axes. The detail energy distribution is captured as rich texture information in the spectrogram. Such type of texture diversity suggests that wavelet analysis on spectrogram will generate highly discriminative features for the proposed audio classification task.



(a) Spectrogram (upper) of a segment of speech signal (lower)

(b) Spectrogram (upper) of a segment of music signal (lower)

**Fig. 1**. Time domain signals and corresponding log-spectrograms of (a) speech and (b) music

### 2.2. Wavelet Packet Feature Extraction

Multi-scale analysis of discrete wavelet transform has proved to be an effective approach to analyze texture images [10]. In this step, wavelet packet decomposition is applied to the spectrogram to generate a number of subbands at different levels. At the first level of decomposition the spectrogram image is decomposed into one approximation and three (i.e.horizontal, vertical and diagonal) detail images [10]. The approximation and the detail images are then decomposed into a second-level approximation and detail images, and the process is repeated. The subband decomposition of two-dimensional DWPT analyzes a spectrogram simultaneously at different resolution levels and orientations. Each subband has important discriminatory properties for images and as such can be used as a feature for audio classification.

### 2.3. Multiple Kernel Learning (MKL)

The overcomplete structure of the wavelet packet transform encourages the selection of the optimal set of wavelet features for classification. In the case of wavelet packet feature selection method, an image is only needed to be decomposed into the wavelet subbands selected by any selection method during the training stage.

Here we map the subband selection problem to multiple kernel learning problem designing suitable kernel functions on the wavelet decomposed subband images. MKL is an interesting approach for designing and integrating kernels to address several, challenging real world applications in computer vision involving several different and heterogeneous data sources [8]. MKL simplifies feature / feature-group selection to kernel selection in kernel-based methods. If the basis kernel functions are defined on individual subbands then their corresponding weights determine the importance of that subband in the classification process.

MKL framework proposed an optimization over the coefficients in a convex combination of kernel functions. The weight of each kernel is optimized during training. Given a set of $K$ kernels, a combined kernel function may be defined as the weighted sum of the individual kernels.

$$\boldsymbol{k}(\boldsymbol{x}_i^l, \boldsymbol{x}_j^l) = \sum_{k=1}^{K} \beta_k \boldsymbol{k}_k(\boldsymbol{x}_i^l, \boldsymbol{x}_j^l) \tag{3}$$

where $\beta_k \geq 0$ and $\sum_k \beta_k = 1$. Here the sub kernel $\boldsymbol{k}_k$ is computed within the subband images $x_i^l$ and $x_j^i$ where $l$ is the subband index. Thus the optimized combination coefficients can be used to understand which subband features of the examples are of importance for discrimination. The MKL framework find sparse weighting $\beta_k$ which can quite easily interpret the resulting decision function. There are several techniques available for combining multiple kernels. One promising technique to combine multiple kernel functions is discussed in [9]. Here the kernel weights are incorporated into the standard SVM objective function. For a set of N training instances $X_1, \cdots, X_N$ with label $y_i \in \{-1, 1\}$, the optimal set of weights are those that maximize the margin. These can be obtained by minimizing the following objective function:

$$\begin{array}{rl} min & \frac{1}{2}\sum_{k=1}^{K} \frac{1}{\beta_k} \|\boldsymbol{w}_k\|_2^2 + C\sum_{i=1}^{N} \xi \\ w.r.t. & \beta, \boldsymbol{w}_k, \boldsymbol{b}, \boldsymbol{\xi} \\ s.t. & y_i \left( \sum_{k=1}^{K} \boldsymbol{w}_k^T \phi_k(\boldsymbol{X}_i, \lambda + b) \geq 1 - \xi_i \right) \\ & \xi_i \geq 0 \ \forall i \ \ \beta_k \geq 0 \ \ \forall k \ \ \sum_{k=1}^{K} \beta_k = 1 \end{array} \tag{4}$$

where $\boldsymbol{w}_k$ is the SVM weight and $b$, $\xi$ and $C$ are standard SVM bias, slack variable and regularization term.

## 3. RESULTS AND DISCUSSION

The performance of the proposed discrimination method is evaluated using two labeled datasets. First dataset constructed by Scheirer and Slaney called ”music-speech” corpus are considered as benchmark and used in multi-features speech/music discrimination (MFD) [1]. Another dataset contains the speech of BBC radio news and different types of music including jazz, rock, pop, folk, instrumental and classical music collected from internet. The ”music-speech” corpus contains speech and music datasets each containing 80 audio samples of 15 second long. The samples were collected by digitally sampling an FM tuner using a variety of stations, content styles, and noise levels by Scheirer and Slaney. Detail description of the dataset can be found in [1]. The results are reported for all 160 samples, jacknifed into four cuts, with $3/4$ cut used to tune several parameters of the classifier and $1/4$ cut used for test as in [2]. Results are given in accuracy which represent proportion of true classification results in the dataset.

In the first phase of experiment, $2.5$ second segment [1][2] of audio signal are used to generate spectrogram. Each spectrogram is then decomposed up to level 3 using wavelet packets. Thus there are 84 decomposed subbands generate from each spectrogram. MKL is applied for sparse selection and weighting of these subbands. Different types of kernel are applied to see the performance. However shift invariant circular convolution kernel proposed in [11] perform the best among all other kernel functions considered here. The performance of our classification system are compared with MFD [1] and PPF [2] using the same dataset and similar experimental conditions as shown in Table 1. The proposed MKL based approach with circular convolution kernel performs better than others.

The effect of noise is studied here to evaluate the robustness of our method. The training is only performed using the original data of “music-speech” dataset and the testing is performed after adding different types of real world noises with the test set. The performance of different noise level of various noise types are illustrated in Table 2.

**Table 1**. Classification accuracy on “music-speech” corpus

| Methods | | Speech(%) | Music(%) | Overall(%) |
|---|---|---|---|---|
| MKL | Circ-conv Kernel | 98.96 | 99.17 | 99.06 |
| | Gaussian Kernel | 98.96 | 97.92 | 98.44 |
| | Polynomial Kernel | 98.33 | 98.33 | 98.33 |
| PPF [2] | | 99.17 | 98.33 | 98.75 |
| MFD [1] | | | | 98.60 |

The noisy data is used in training before performing the testing with noisy audio segments in [6]. Here, the noisy data is not included in the training set. However, the performance of the classification in the presence of noise on the test dataset is much better than that of [6]. This proves the robustness of the proposed features against noise. The main distinguishing feature is the energy continuity of spectrogram which is usually more continuous in case of music. When more noise is added, the energy continuity in music spectrogram is increased while decreasing the discreteness of energy in speech spectrogram. Hence, some speech segmented is recognized as music that increases the error rate to discriminate speech from music as illustrated in Table 2.

**Table 2**. Classificaion accuracy on "music-speech" corpus at the presence of different kind of noises.

| Noise type | Noise | Speech (%) | Music (%) | Overall(%) |
|---|---|---|---|---|
| | Clean | 98.96 | 99.17 | 99.06 |
| Babble | 20dB | 98.33 | 99.17 | 98.75 |
| | 10dB | 98.33 | 99.58 | 98.96 |
| | 0dB | 95.83 | 100.00 | 97.92 |
| Destroyer | 20dB | 97.50 | 99.38 | 98.44 |
| | 10dB | 94.38 | 99.79 | 97.08 |
| | 0dB | 63.54 | 100.00 | 81.77 |
| Jet | 20dB | 97.50 | 99.58 | 98.54 |
| | 10dB | 95.00 | 100.00 | 97.50 |
| | 0dB | 70.83 | 100.00 | 85.42 |
| HFChannel | 20dB | 96.88 | 99.38 | 98.13 |
| | 10dB | 95.00 | 99.79 | 97.40 |
| | 0dB | 78.33 | 100.00 | 89.17 |

Additional experiments are also carried out with different size of segment lengths on the first dataset. The proposed algorithm is also robust with respect to the segment lengths. The lower segment length is more effective in multimedia retrieval and classification applications. The variation of performance (using MKL based algorithm) with the change of segment length (1.0s to 3.5s) is not so much as shown in Fig. 2. The performance of the proposed method is also tested on the second dataset to discriminate the speech of BBC radio news from different types of music signals as shown in Fig. 3.

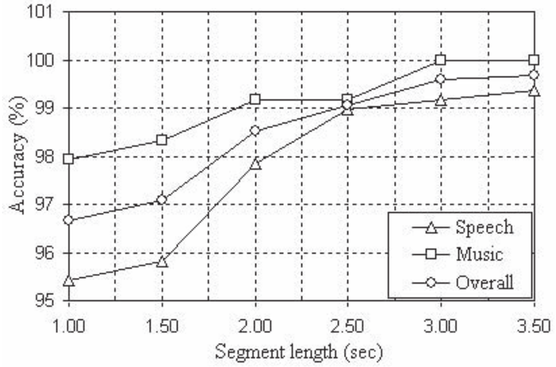


**Fig. 2**. Performance with the variation of segment length

## 4. CONCLUSIONS

A novel algorithm for speech/music discrimination of digital audio is proposed. The visual perception of time-frequency space of audio signal is used to derive potential textual features using the combination of wavelet packet transform and MKL. The important characteristic of this algorithm is its robustness toward noise. We also note that training can proceed on noise-free audio signals, while the classification can be done on noisy signals. Such type of features would be useful in real world applications. The proposed MKL based approach is versatile to the datasets, i.e. by learning with limited dataset, it can be used to discriminate a wide classes of speech-music data. Also it is less sensitive to the segment length.

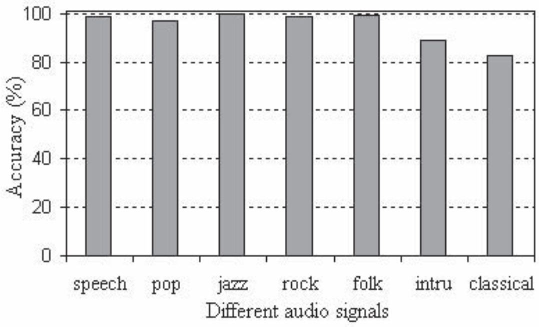


**Fig. 3**. Classification accuracy of speech and different types of music

## 5. REFERENCES

[1] Eric Scheirer and Malcolm Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *ICASSP'97*, 1997, pp. 1331–1334.

[2] Gethin Williams and Daniel P. W. Ellis, "Speech/music discrimination based on posterior probability features," in *EUROSPEECH*, 1999.

[3] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155 – 166, feb. 2005.

[4] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "A speech/music discriminator of radio recordings based on dynamic programming and bayesian networks," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 846 –857, aug. 2008.

[5] Pedro J. Moreno and Ryan Rifkin, "Using the fisher kernel method for web audio classification," in *ICASSP'00*, 2000, pp. 2417–2420.

[6] Jonathan Dennis, T Dat, and H Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 130–133, 2011.

[7] Guoshen Yu and Jean-Jacques Slotine, "Audio classification from time-frequency texture," in *ICASSP'09*, 2009, pp. 1677–1680.

[8] L. Gert, C. Nello, B. Peter, and El G. Laurent, "Learning the kernel matrix with semi-definite programming," *Journal of Machine Learning Research*, vol. 5, pp. 2004, 2002.

[9] Sören Sonnenburg, Gunnar Rätsch, Christin Schäfer, and Bernhard Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[10] S. Arivazhagan and L. Ganesan, "Texture classification using wavelet transform," *Pattern Recogn. Lett.*, vol. 24, pp. 1513–1521, June 2003.

[11] S. Nilufar, N. Ray, and H. Zhang, "Optimum kernel function design from scale space features for object detection," *IEEE ICIP*, 2009.