International Journal of Fuzzy Logic and Intelligent Systems, vol. 10, no. 1, March 2010, pp. 7−11

# A Novel Speech/Music Discrimination Using Feature Dimensionality Reduction

**Ji-Soo Keum[1*], Hyon-Soo Lee[2] and Masafumi Hagiwara[1]**

**[1] Faculty of Science and Technology, Keio University, Yokohama, Japan**
**[2] Dept. of Computer Engineering, Kyung Hee University, Yongin, Republic of Korea**
**{keum, hagiwara}@soft.ics.keio.ac.jp, leehs@khu.ac.kr**

**Abstract**

In this paper, we propose an improved speech/music discrimination method based on a feature combination and dimensionality reduction approach. To improve discrimination ability, we use a feature based on spectral duration analysis and employ the hierarchical dimensionality reduction (HDR) method to reduce the effect of correlated features. Through various kinds of experiments on speech and music, it is shown that the proposed method showed high discrimination results when compared with conventional methods.

**Key Words** : speech/music discrimination, spectral duration analysis, hierarchical dimensionality reduction

## 1. Introduction

Speech/music discrimination is an important preprocessing part of multimedia applications such as speech recognition, speaker indexing, audio-visual based information retrieval, and human-robot interaction. In order to achieve remarkable results, extensive research has been conducted [1-6].

Speech/music discrimination can be divided into two main research parts. One is feature extraction, the other is discrimination. Previous research showed that the importance of feature selection in speech/music discrimination [2]. Based on many research results, we consider the feature extraction and combination in this research. In addition, we employ the well-known feature dimensionality reduction method to obtain better results.

Many features have been used to discriminate an audio segment into speech or music. The zero crossing rate (ZCR), short time energy (STE), and spectral analysis based features are widely used [1-7]. In addition, the mel frequency cepstral coefficient (MFCC) is often used [8]. However, there are many variations in real world audio signals such as speaking style, various musical genres and so on. Therefore, it is difficult to discriminate an audio signal into speech or music by using single feature alone [7].

In this paper, we propose an improved speech/music discrimination method based on a feature combination and dimensionality reduction approach. In order to improve discrimination ability, we used a feature based on spectral duration analysis. When we employ features, we should consider combinations of features. If the selected features are only composed of correlated features, it is not easy to obtain high performance. To reduce this situation, we employ the hierarchical dimensionality reduction (HDR) method to reduce the effect of correlated features [9].

This paper consists of the following section. In section 2, we briefly introduce the selected features for speech/music discrimination. In section 3, we describe the proposed speech/music discrimination method using spectral duration analysis and dimensionality reduction approach. In section 4, the experimental results are shown and finally we present conclusion with future research direction in section 5.

Manuscript received Dec. 7, 2009; revised Jan. 15, 2010
This paper received best paper award in the 10th International Symposium on advanced Intelligent Systems (ISIS 2009).
* Corresponding author

## 2. Selected Features

### 2.1 High Zero Crossing Rate Ratio

High zero crossing rate ratio (HZCRR) is defined as the ratio of the number of frames whose ZCR are above 1.5-fold average ZCR in an 1-second segment [4].

$$ZCR(n) = \frac{1}{2M}\sum_{m=1}^{M}|\mathrm{sgn}(x(m)) - \mathrm{sgn}(x(m-1))| \quad (1)$$

$$HZCRR = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(ZCR(n) - 1.5avZCR) + 1] \quad (2)$$

$$\mathrm{sgn}(x) = \begin{cases} 1, & if\,(x \geq 0) \\ -1, & otherwise \end{cases} \quad (3)$$

### 2.2 Low Short Term Energy Ratio

Low short term energy ratio (LSTER) is defined as the ratio of the number of frames whose STE are less than 0.5 times of average STE in an 1-second segment [4].

$$STE(n) = \frac{1}{M}\sum_{m=0}^{M-1}[x(m)w(M-m)]^2 \quad (4)$$

$$LSTER = \frac{1}{2N}\sum_{n=0}^{N-1}[\mathrm{sgn}(0.5avSTE - STE(n)) + 1] \qquad (5)$$

In Eqs.(1)~(5), $m$ is the sample index and $M$ is the number of samples in the analysis window, and sgn(·) is the sign function. And, $n$ is the frame index and $N$ is the number of frames in the analysis segment, and $w(\cdot)$ is the window function.

### 2.3 Pitch Ratio

Pitch ratio (PR) is defined as the ratio of frames with human pitch to the total number of frames in an audio segment [10].

$$PR = NP / NF \qquad (6)$$

where $NP$ is the total number of frames that have a pitch and $NF$ is the total number of the frames in a segment. The PR was proposed to classify an audio segment into speech or nonspeech in security monitoring. It also shows a high discrimination ability on speech/music discrimination [7].

### 2.4 Spectral Duration

Spectral duration (SD) was recently proposed in our previous research for speech/music discrimination [5-6]. In section 3, we introduce the maximum spectral duration in detail to combine with the selected features.

## 3. Proposed Speech/Music Discrimination

The proposed speech/music discrimination method consists of two parts. Figure 1 shows the proposed speech/music discrimination method. Firstly, we extract features and reduce the feature dimensionality. Secondly, we classify an audio segment using the well-known k-nearest neighbor (k-NN) classifier.

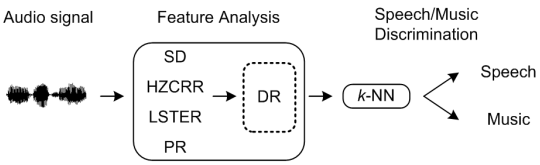


Fig. 1 Proposed speech/music discrimination method

### 3.1 Spectral Duration Analysis

Through the analysis of spectral information, we can find the difference between speech and music in spectral duration. Also, we can analyze the harmonic structure of the audio segment. The spectral peak tracks exist in a specific frequency bin and have a duration time. The speech segment has a spectral peak track that exists in the lower frequency band. It has a short duration characteristic compared with a spectral peak track of music [3], [6-7].

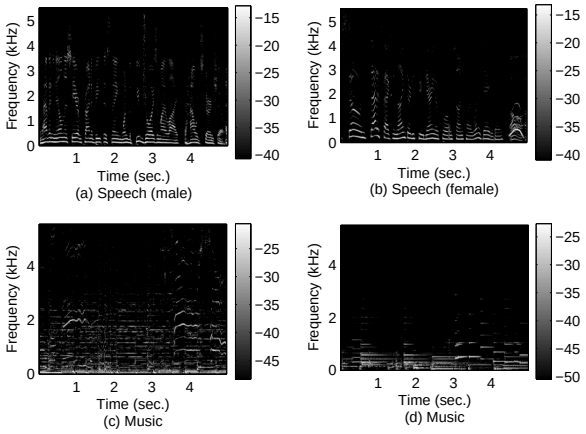


Fig. 2 Examples of spectrogram for speech and music

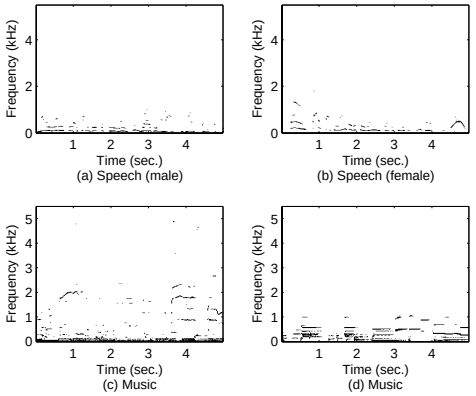


Fig. 3 Examples of spectral peak tracks for speech and music

The procedure of the extraction of maximum spectral duration consists of three parts. First, we compute the fast Fourier transform (FFT) spectrum for a 1-second of audio segment. And, we apply threshold ($\theta$) to consider important spectral peaks and neglect small amplitudes.

$$CX_n(k) = \begin{cases} 1, & if\,(X_n(k) \ge \theta) \\ 0, & otherwise \end{cases} \qquad (7)$$

where $X_n(k)$ is the FFT spectrum and $CX_n(k)$ is the clipped spectrum at the $n$-th frame by threshold $\theta$, and $k$ is the frequency bin. The $\theta$ is determined relative to the maximum magnitude of the segment. The $\theta$ is calculated by $\max(X(k))\times\eta$, and $\eta$ is the scaling factor. Figure 2 and Fig.3 show examples of the spectrogram and the extracted spectral peak track for speech and music segments, respectively.

Second, we find a connected spectral peak track from the clipped spectrum and generate a histogram. Most of the spectral peak track exists within 1 kHz band, therefore we generate the histogram below 1 kHz frequency band. When we generate a histogram, we detect 92ms connected spectral peak track by considering the duration of vowel length. The duration of the shortest vowel is about 70 ms and the longest is about 110 ms in Korean [6]. Equation (8) is a normalized histogram

by the number of frames in a segment. Where the analyzed frame size is 46 ms and overlap is 23 ms.

$$SD(k) = \frac{1}{N-1}\sum_{n=1}^{N-1}[CX_{n-1}(k)\cdot CX_n(k)\cdot CX_{n+1}(k)] \qquad (8)$$

Finally, we get the maximum spectral duration to discriminate an audio segment. We called the SD as maximum spectral duration feature (MSDF) or maximum spectral duration (MSD) in the previous work [6-7].

$$SD = \max(SD(k)). \qquad (9)$$

Figure 4 shows the probability distribution of speech and music for selected features.

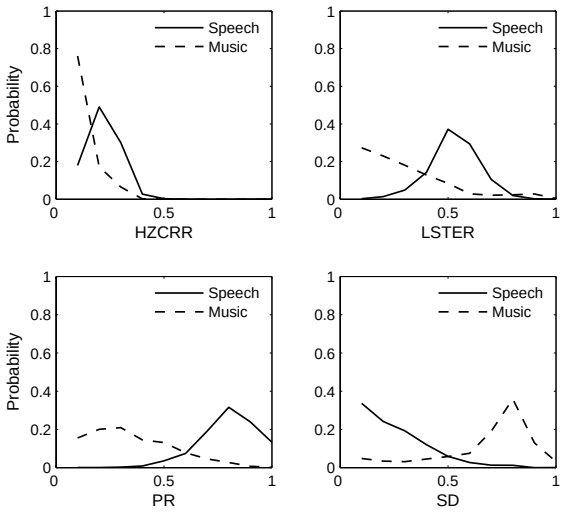


Fig. 4 Probability distribution of speech and music

### 3.2 Feature Dimensionality Reduction

In order to reduce an effect of dominated or correlated features in combinations, we employ the HDR as preprocessing. The HDR combines correlated features in every iteration, and reduces the dimensionality of the feature space. In Eq. (10), the $\rho_{ij}$ is correlation coefficient.

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}. \qquad (10)$$

where $0 \le \rho^2_{ij} \le 1$, with $\rho^2_{ij} = 0$ for uncorrelated features and $\rho^2_{ij} = 1$ for completely correlated features [9]. The $\rho^2_{ij}$ works as a similarity function between features. When two features have large $\rho^2_{ij}$, they are merged into one feature. In this research, we simply averaged two features based on an assumption that each feature has equal weight.

When we apply the HDR to the combination of 'HZCRR, LSTER, PR, SD', it was merged as sequence ((HZCRR, (LSTER, PR)), SD), and the 'HZCRR, PR, SD' was merged as ((HZCRR, PR), SD), respectively, for example. Where the (·) means merged feautre with the highest correlation, and the innermost (·) is merged first, followed by the next outer (·) in sequence. In the HDR, the merging condition is that when the $\rho^2_{ij}$ is larger than defined threshold 0.75, we merged two features.

## 4. Experimental Results

We performed experiments using the speech/music data in [2]. 44 files were selected for speech data that were not mixed with other speakers. For music data, 41 files were selected that contain both vocal and nonvocal. The half of the speech and music data was used for training and the remainder was used for testing.

We constructed all combinations of the selected features for performance evaluation. The tested segment length is 0.5, 1.0, and 1.5 seconds. The analyzed frame size of a segment is 46 ms and the overlap is 23 ms in feature extraction. We performed various experiments changing the number of nearest neighbors (1, 3, 5, 7, 9, 11, 15, 25), and the scaling factor on spectral duration analysis. In addition, the feature dimensionality approach was compared with the principal component analysis (PCA).

Figure 5 shows the receiver operation characteristic (ROC) curve for the selected features. The best result of the proposed method is shown by a circle. The combination of 'HZCRR, PR, SD' outputs the best performance.

Table 1 shows the discrimination result without pre-processing on different segment length and combination. Table 2 shows the experimental results for the feature combination of 'HZCRR, PR, SD' on 1-second of segments. The accuracy was calculated using different numbers of nearest neighbors. The result given is the average accuracy and standard deviation.

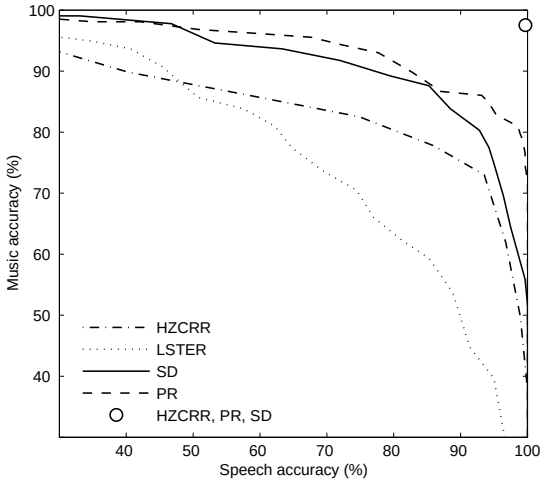


Fig. 5 ROC curve for the selected features and proposed method

Table 1. Speech/Music discrimination result by feature combination (%)

| Segment Duration | 0.5 sec. | | | 1.0 sec. | | | 1.5 sec. | | |
|---|---|---|---|---|---|---|---|---|---|
| Feature Combination | Speech | Music | Avg. | Speech | Music | Avg. | Speech | Music | Avg. |
| HZCRR, LSTER | 93.10 | 65.73 | 79.41 | 93.84 | 75.29 | 84.57 | 90.77 | 75.25 | 83.01 |
| PR, SD | 98.16 | 90.56 | 94.36 | 98.62 | 95.00 | 96.81 | 98.79 | 96.13 | 97.46 |
| HZCRR, LSTER, PR | 97.35 | 91.54 | 94.44 | 98.39 | 94.75 | 96.57 | 98.65 | 98.00 | 98.33 |
| HZCRR, PR, SD | 97.82 | 94.85 | 96.34 | **99.45** | **97.04** | **98.25** | **99.79** | **98.06** | **98.92** |
| HZCRR, LSTER, PR, SD | **98.07** | **95.48** | **96.77** | 98.58 | 96.46 | 97.52 | 98.86 | 97.81 | 98.34 |

Table 2. Speech/Music discrimination result for the 'HZCRR, PR, SD' combination (%)

| Feature Combination | Preprocessing | Speech | Music | Avg. |
|---|---|---|---|---|
| HZCRR, PR, SD | - | 99.45 ± 0.68 | 97.04 ± 0.55 | 98.25 ± 0.46 |
| | HDR | **99.68 ± 0.54** | **97.54 ± 0.83** | **98.61 ± 0.56** |
| | PCA | 98.02 ± 0.48 | 96.21 ± 0.25 | 97.12 ± 0.34 |

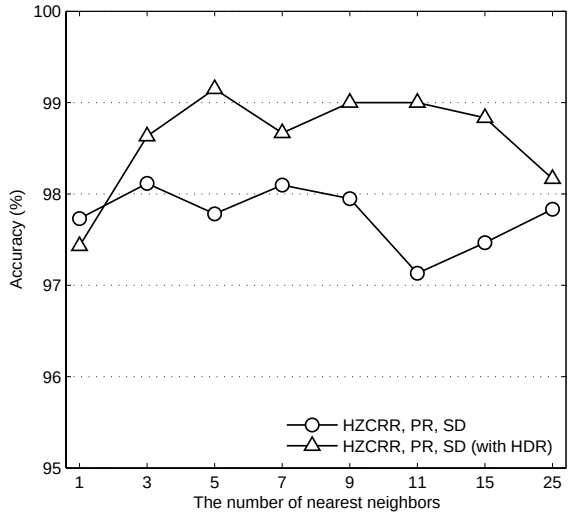


Fig. 6 Speech/music discrimination result by changing of the number of nearest neighbors (when the scaling factor $\eta$ is 1.7)

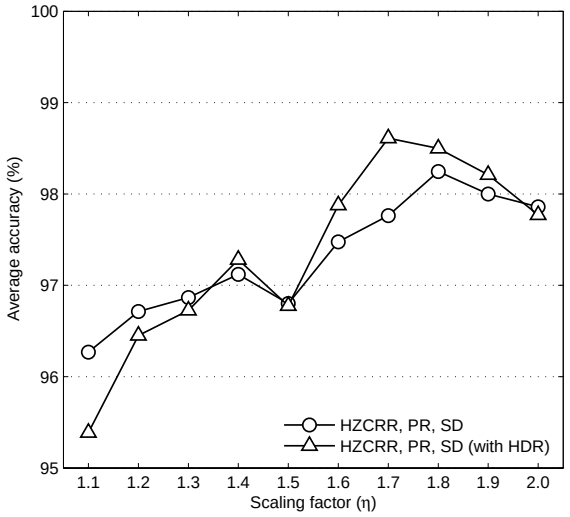


Fig. 7 Speech/music discrimination result by changing of the scaling factor in spectral duration analysis

Fig. 6 and Fig.7 show the result of speech and music discrimination by changing of the number of nearest neighbors and the scaling factor in spectral duration analysis, respectively. The proposed method was not sensitive to the selection of the number of neighbors and the scaling factor.

According to the result of our experiments on speech and music discrimination, we can verify the importance of feature selection. Also, the employed HDR contributes to obtaining higher performance compared with PCA, and the result of the proposed method with HDR yields stable speech/music discrimination result for the selection of neighbors.

## 5. Conclusion

In this paper, we have proposed an improved speech/music discrimination method based on feature combination. To discriminate an audio segment into speech or music, we combined recently proposed features and introduced a feature derived from spectral duration analysis. In addition, we have employed feature dimensionality reduction method to reduce the effect of correlated feature combination. The proposed method is simple and efficient while giving high performance.

Our future work will be focused on a weighting scheme of each feature in dimensionality reduction phase.

## References


[1] J. Saunders, "Real-time discrimination of broadcast speech/ music," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 993-996, 1996.

[2] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1331-1334, 1997.

[3] T. Zhang and J. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 441-457, 2001.

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

[5] J.-S. Keum and H.-S. Lee, "Speech/music discrimination using spectral peak feature for speaker indexing," *Proc. IEEE Int. Sym. on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 323-326, 2006.

[6] J.-S. Keum, S.-K. Lim, and H.-S. Lee, "Speech/music discrimination using spectrum analysis and neural network," *The Journal of the Acoustical Society of Korea*, vol. 26, no. 5, pp. 207-213, 2007.

[7] J.-S. Keum, H.-S. Lee, and M. Hagiwara, "An improved speech/nonspeech classification based on feature combination for audio indexing," *IEICE Trans. on Fundamentals*, vol. E93-A, no. 4, 2010.

[8] J.R. Deller, J.G. Proakis, and J.H.L. Hansen, *Discrete Time Processing of Speech Signals*, Prentice Hall, 1987.

[9] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley & Sons, 2001.

[10] AR Abu-El-Quran and RA Goubran, "Security-monitoring using microphone arrays and audio classification," *IEEE Trans. on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1025-1032, 2006.



**Ji-Soo Keum**

He received his M.E. and Ph.D. degrees in computer engineering from Kyung Hee University, Republic of Korea, in 2000 and 2008, respectively. Since 2008, he has been with Keio University, where he is now a visiting researcher. His research interests are speech processing, pattern recognition, and kansei engineering.



**Hyon-Soo Lee**

He received his B.S. degree in electronic engineering from Kyung Hee University, Republic of Korea, in 1979, and M.E. and Ph.D. degrees in electrical engineering from Keio University, Japan, in 1982 and 1985, respectively. Since 1985, he has been with Kyung Hee University, where he is now a Professor. From 1999 to 2000, he was a visiting scholar at Oregon State University and University of California, Irvine. From 2005 to 2008, he was Dean of the College of Electronics & Information, and the Graduate School of Information and Communication. His research interests are computer architecture, parallel processing, neural networks, and pattern recognition.



**Masafumi Hagiwara**

He received his B.E., M.E. and Ph.D. degrees in electrical engineering from Keio University, Yokohama, Japan, in 1982, 1984 and 1987, respectively. Since 1987, he has been with Keio University, where he is now a Professor. From 1991 to 1993, he was a visiting scholar at Stanford University. He received the Niwa Memorial Award, Shinohara Memorial Young Engineer Award, IEEE Consumer Electronics Society Chester Sall Award, Ando Memorial Award, Author Award from the Japan Society of Fuzzy Theory and Systems (SOFT), Technical Award and Paper Award from Japan Society of Kansei Engineering in 1986, 1987, 1990, 1994, 1996, 2003, and 2004, respectively. His research interests include neural networks, fuzzy systems, evolutionary computation and kansei engineering. Dr. Hagiwara is a member of IEICE, IPSJ, SOFT, IEE of Japan, Japan Society of Kansei Engineering, JNNS and IEEE (Senior member).