AUDIO-BASED AUTOMATIC DETECTION OF OBJECTIONABLE CONTENTS IN NOISY CONDITIONS USING NORMALIZED SEGMENTAL TWO-DIMESIONAL MFCC

Bong-Wan Kim[1], *Dae-Lim Choi*[1], *JaeDeok Lim*[2], *SeungWan Han*[2], *Yong-Ju Lee*[3]

[1]SiTEC, Wonkwang Univ., Korea
[2]Knowledge-based Information Security Research Division, ETRI, Korea
[3]Dept. of Computer Eng., Wonkwang Univ., Korea

ABSTRACT

The segmental two-dimensional Mel-frequency cepstral coefficient (STDMFCC) feature has been successfully used in recent studies to detect objectionable sounds, which implicitly represent both static and dynamic characteristics of signal. This study now proposes a new normalized STDMFCC to improve the content recognition performance in diverse noisy environments. Two tests were conducted to verify the performance of the proposed feature: First, an objectionable sound recognition test was conducted with 10-second clips to which white noises with diverse signal-to-noise ratios (SNRs) were added. The proposed feature in the test had an average error reduction rate (ERR) of 24.69% with respect to the STDMFCC. Second, a test was conducted based on the soundtrack that contained diverse channel environments and noises. The equal error rate (EER) of the proposed feature was 4.00% compared with 10.33% of STDMFCC, and the ERR was 61.29%.

Index Terms— Objectionable sound recognition, objectionable content detection, normalized STDMFCC, NSTDMFCC.

1. INTRODUCTION

With the development of Internet and multimedia technology, objectionable contents (especially pornographic contents) spread quickly, and they can be easily accessed by anybody. These contents can be harmful to young people, while causing social problems in some countries. Accordingly, studies have been conducted to automatically identify objectionable contents which were mostly based on the image-processing technology [1][2]. Recently, audio-based approach has been started as an alternative or a supplemental means to image-processing technology [3][4][5].

As with other classification or recognition problems, it is very important to choose proper features when detecting objectionable contents. Objectionable sounds are clearly characterized by repeated moans, heavy breathing, and skin contact sounds. The use of features that represent such static and dynamic characteristics of objectionable sounds will be helpful in detecting objectionable contents. The representative short time static feature is the Mel-frequency cepstral coefficients (MFCC) which are most successfully used in speech and speaker recognition area. Modulation data can be used as a dynamic feature for longer sections. In human speech perception, it is widely known that 16-Hz or less modulation frequency plays an important role for the intelligibility of speech [6].

Lim et al. suggested the two-dimensional Mel-frequency cepstral coefficients (TDMFCC) as features that represent short time MFCCs and their modulation information in matrix forms [7]. The TDMFCC matrix can be obtained by applying 2-D discrete cosine transform (DCT) to the successive logarithmic energies generated from Mel-scale bandpass filters. They used TDMFCC to classify bird species using bird sounds, and varied the analyzed segment lengths according to the syllable length of bird sounds. Kim et al. modified TDMFCC to detect objectionable sounds [4]. Because it is difficult to automatically detect the syllables under noisy environments, they proposed the segmental TDMFCC (STDMFCC), which is based on fixed-length of segments rather than on syllables. The results of tests showed that STDMFCC performs better than low-level features (short time energy, sub-band energy, etc.), perceptual features (spectral flux, spectral flatness, etc.), and MFCC-based features (MFCC, MFCC with time derivatives) [4] [5].

The soundtracks, as multimedia contents, may have diverse channel characteristics. There are compensation techniques that are widely known and successfully introduced: cepstral mean subtraction (CMS) [8] and cepstral mean and variance normalization (CMVN) [9]. These techniques are robust against the channel distortion and noise. In CMS and CMVN, the normalization of mean or variance is helpful to alleviate the disaccord between probability density functions (PDFs) of MFCC parameters for training and test data. These methods were not applied to the content detection studies that used STDMFCC features [4][5][7].

This study examines how the application of CMS and CMVN influence STDMFCC features and proposes a new normalization method. The proposed new normalization method applies the traditional CMVN to the static parts of STDMFCC and normalizes the variance of the entire local segment for the dynamic part to maintain the relative magnitude information. Via two tests, the performance of the proposed normalized STDMFCC (NSTDMFCC) was compared with the conventional STDMFCC and the STDMFCCs to which CMS and CMVN are applied. First, the recognition performance deterioration of each feature was comparatively examined according to the changes in SNR via the clip-unit, white-noise-added objectionable sound recognition test. Second, the performances of the features were compared with one another via the detection test on the soundtrack that contains diverse channel environments and noises. Both tests use Gaussian mixture model (GMM) to model the objectionable and normal sounds.

This paper is organized as follows. Section 2 describes the proposed NSTDMFCC. Section 3 describes the GMM for the sound modeling. Section 4 contains test and analysis results, and Section 5 provides the conclusions of this study.

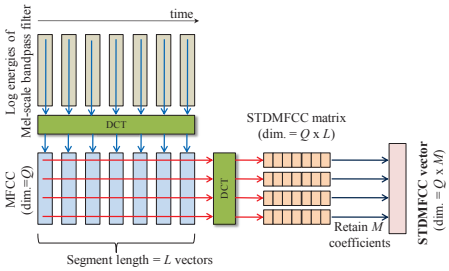

Fig. 1. STDMFCC matrix calculation and final feature vector construction.

2. NSTDMFCC

The STDMFCC matrix can be obtained by applying 2-D DCT to the successive logarithmic energies generated from Mel-scale bandpass filters. Since 2-D DCT can be divided into two 1-D DCTs, STDMFCC matrix, $C_t(q,n)$ at the tth frame is expressed as follows:

$$C_t(q,n) = \sum_{l=0}^{L-1} c_{t+l}(q) cos\left[\frac{\pi}{L}(l+0.5)n\right] \quad (1)$$
$$0 \le q < Q, 0 \le n < L,$$

wherein $c_t(q)$ is the qth MFCC coefficient, n is the modulation frequency index, Q is the dimension of MFCC, and L is the segment length to calculate STDMFCC. The equation above is obtained by slightly modifying the original equation to explicitly express the frame index and dimension of MFCC. STDMFCC implicitly expresses the static and dynamic characteristics of the audio signals within the analyzed segment. Since important information for sound recognition are concentrated on the lower-order coefficients, only the lower-order modulation coefficients are used as final feature vectors [4][5][7]. Figure 1 shows the aforementioned procedure.

In this study, STDMFCC feature vectors were extracted as follows: 13 dimensional MFCCs (12 MFCC plus c0) are calculated using 25ms Hamming window and 10ms frame shift. STDMFCC matrix was obtained by applying 1-D DCT to L successive MFCCs. L was determined as 48, which coincides with approximately 500ms. From the calculated matrix, eight lower-order modulation coefficients were selected per row to create the final feature vector. The resulting dimension of the feature vector was 104. Feature vectors were extracted from the segment for STDMFCC calculation at 250ms intervals.

The soundtracks in the multimedia contents can have diverse channel characteristics according to the devices and channels that are used to create the contents. In addition, according to the genre of contents, conversations, pieces of background music, sound effects, and noises that exist in the contents can worsen the detection performance. Therefore, a proper compensation technique is required to detect objectionable contents regardless of channel characteristics and noises. CMS and CMVN are widely known compensation techniques, which are being successfully used. CMS normalize the first moment of MFCCs as follows:

$$\hat{c}_t(q) = c_t(q) - \mu(q) \ , \quad (2)$$

and CMVN normalizes the first and second moments of MFCCs to produce unit variance:

$$\hat{\hat{c}}_t(q) = \frac{\hat{c}_t(q)}{\sigma(q)} \ , \quad (3)$$

wherein $\mu(q)$ is the mean of the qth MFCC components and $\sigma(q)$ is the standard deviation obtained from the audio data.

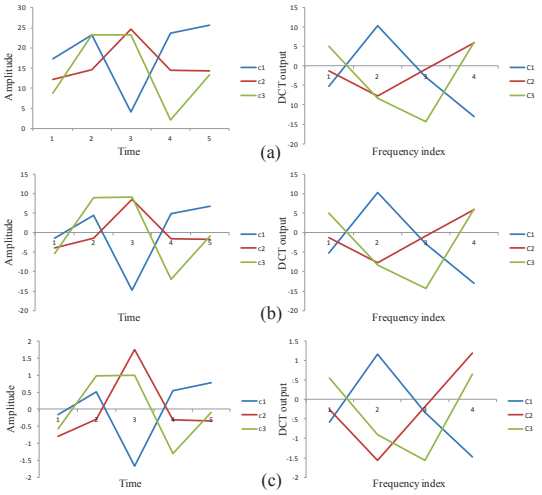

Fig. 2. Effects of CMS and CMVN on the dynamic information: (a) randomly generated 3 components and their DCT output, (b) CMS applied components and their DCT output, and (c) CMVN applied components and their DCT output.

When we apply a DCT on a data, only 0[th] coefficient of DCT output is influenced by the DC component. Therefore, if STDMFCC is calculated using CMS-applied MFCCs, the normalization is applied only to the 0[th] modulation coefficients that express channel characteristic and not to dynamic coefficients. Therefore, the channel distortion compensation effect is still effective to STDMFCC. When CMVN is applied, however, it influences the dynamic coefficients of STDMFCC. The dynamic coefficients, which can be important queues for recognizing objectionable sounds, have different relative magnitude due to the cepstral variance normalization.

Figure 2 illustrates the time trajectories of three randomly generated components (c1, c2 and c3), their DCT output (C1, C2 and C3), and the results of applying CMS and CMVN. As shown in Figure 2(a) and 2(b), the application of CMS did not change the dynamic information in the DCT output, but the application of CMVN changed it (Figure 2(c)). Especially for the frequency indices 2 and 4, the relative magnitudes and orders of C2 and C3 changed.

Therefore, a new NSTDMFCC was proposed to overcome this problem. NSTDMFCC matrix $NC_t(q,n)$ can be obtained as follows:

$$NC_t(q,n) = \begin{cases} \dfrac{C_t(q,n) - L \cdot \mu(q)}{\sigma(q)} & n = 0 \\ \dfrac{C_t(q,n)}{\sigma_{t,s}} & 0 < n < L \end{cases} \quad (4)$$

wherein $\sigma_{t,s}$ is the standard deviation of all MFCC coefficients in the segment which can be calculated as follows:

$$\sigma_{t,s} = \sqrt{\frac{1}{Q \cdot L - 1} \sum_{q=0}^{Q-1} \sum_{l=0}^{L-1} \left[c_{t+l}(q) - \mu_{t,s}\right]^2} \ , \quad (5)$$

and $\mu_{t,s}$ is the segmental mean which can be calculated as follows:

$$\mu_{t,s} = \frac{1}{Q \cdot L} \sum_{q=0}^{Q-1} \sum_{l=0}^{L-1} c_{t+l}(q) \ . \quad (6)$$

The proposed NSTDMFCC applies the conventional CMVN to the 0th modulation coefficients. On the other hand, simple variance normalization is applied to the dynamic coefficients using a single value to maintain relative magnitude information. In this study, c0 was excluded from the calculation of $\mu_{t,s}$ and $\sigma_{t,s}$ to avoid bias because its dynamic range differed from those of other components, while it was included in normalization.

3. GMM FOR SOUNDS MODELING

GMMs have been successfully used in speaker recognition and sound classification fields. A GMM is a parametric probability function represented as a weighted sum of Gaussian mixture densities, as shown in the following equation.

$$p(x|\lambda) = \sum_{i=0}^{M-1} w_i g(x|\mu_i, \Sigma_i) \quad (7)$$

where x is feature vector; w_i is the weight for ith mixture; M is the number of mixture components in model λ; and $g(x|\mu_i, \Sigma_i)$ is the probability density function with mean vector μ_i and covariance matrix Σ_i.

We made two GMM models: positive model (λ_{obj}) for objectionable sounds and negative general model (λ_{gen}) for non-objectionable sounds. For classification, the test sounds are classified by length-normalized likelihood ratio as follows:

$$\Lambda(X) = \frac{\sqrt[T]{\prod_{t=0}^{T-1} p(x_t|\lambda_{obj})}}{\sqrt[T]{\prod_{t=0}^{T-1} p(x_t|\lambda_{obj})} + \sqrt[T]{\prod_{t=0}^{T-1} p(x_t|\lambda_{gen})}} \geq \theta \quad (8)$$

where X is the observed feature vectors, T is the number of feature vectors and θ is the decision threshold, which is empirically determined.

4. EXPERIMENTAL RESULTS

For the objectionable sound recognition test, the volume of the sound clip DB that was used in [4] and [5] was extended by about 1.5 times. For objectionable sounds, 211 source soundtracks were collected from the Internet. The collected soundtracks include hidden camera, self-produced video, adult video, and adult broadcast. For training, 2,040 clips were extracted from 110 soundtracks; for testing, 1,012 clips were extracted from 101 soundtracks. Objectionable sound clips were collected so that a wide range of objectionable sound strength and performer's gender could be evenly included. For non-objectionable sounds, 487 source soundtracks were collected from user created contents (UCCs), TV programs, DVDs, music videos, audio, and music CDs. The genres of the collected soundtracks included culture and current events, children's program, drama, entertainment, news, instrumental music, pop music, and sports. For training, 2,089 clips were extracted from the 247 soundtracks. For testing, 1,128 clips were extracted from the 240 soundtracks.

All sound clips were 10s long, and were digitized in 16 bits per sample with 11 kHz sampling rate in a mono-channel. White noises were applied to the test DB at SNRs of 15 dB, 10 dB, and 5 dB to compare the feature performances for different SNRs. The training of GMMs was performed using the training DB without white noises.

Features were extracted in the manner described in Section 2. For GMMS, we evaluated the feature set with 128 mixtures of GMM for each sound class. In order to evaluate the effectiveness of proposed new feature, we evaluated the performance of conventional STDMFCC (TD), STDMFCC from CMS applied MFCCs (CMS.TD), and STDMFCC from CMVN applied MFCCs (CMVN.TD) and proposed NSTDMFCC (NTD). The performance was evaluated in terms of equal error rate (EER) where false positive rate and false negative rate are equal. Table 1 shows the performances of proposed NTD and other features.

Table 1. Objectionable sound recognition performance of the proposed feature (NTD) and other features in the white noise environment for 128 Mixtures of GMM.

Feature	EER (%) for noise conditions			
	Original	15dB	10dB	5dB
TD	3.75	3.47	4.46	7.98
CMS.TD	**2.57**	3.72	5.35	9.22
CMVN.TD	3.56	4.81	5.75	6.89
NTD	2.96	**3.17**	**3.77**	**4.91**

First of all, all normalized features outperformed the conventional TD in the original condition which white noises are not added. CMS.TD had a good performance in the original environment, but its performance deteriorated with the decrease in SNR. This seems to have been caused by the decrease in SNR which affected the mean value for CMS. CMVN.TD had a worse performance than CMS.TD, but its performance decrease slope with respect to the noise level increase was gentler than that of CMS.TD. This seems to have been caused by the automatic gain control (AGC) effect according to the variance normalization of CMVN.

NTD, which was proposed in this study, had a performance decrease of absolute 0.39% compared with CMS.TD in the original environment; it performed best in all the other environments. Its performance decrease was also most gentle with respect to the noise level increase. This seems to have been caused by efficiently addressing the channel distortion via CMVN in the static part of the feature. Moreover, in the dynamic part, each coefficient maintained it's a relative magnitude via local segment variance normalization. NTD had an average EER of 3.70% compared with 4.92% of the conventional TD, which made an error reduction rate (ERR) of 24.69%.

This test is limited in that the actual noises in the multimedia contents can significantly differ from white noises, and that few actual contents have such strong and constant noises as those in this test. Therefore, the entire soundtrack that contained diverse channel environments and noises was used to detect objectionable contents, rather than proceeding to another artificial tests with other noises.

To examine the performance of the proposed feature in real noise environment, a soundtrack DB for testing was established. For the non-objectionable class, a total of 600 soundtracks were collected (60 news, 60 documentaries, 49 music videos, 62 sports, 60 entertainments, 192 dramas, and 117 movies). The total length was 546h. A total of 600 soundtracks were collected for objectionable class, with a total length of 577h. Objectionable soundtracks contained performers' conversation, music, and others, as well as diverse noises. All the collected soundtracks were not the same as those in the DB for the preceding test, and they were digitalized in the same audio format.

Special attention was given to the feature extraction. Most soundtracks may have long mute sections when they start and end. Mute sections can also exist when scenes change. These mute sections must be excluded before the calculation of normalization

coefficients or feature vector extraction. Because c0 of MFCC shows the characteristics similar to those of short time logarithmic energy, it was used for muted region detection. The MFCC vector that had a c0 value smaller than the threshold was excluded from the normalization coefficient calculation. The feature vector was also not extracted from the segment that contained such an MFCC vector. The features were extracted in three conditions with threshold 0 (no mute section excluded), 30, and 40. The training clip data that were used for the previous test was used for the GMM model training.

Table 2. Objectionable soundtrack detection performance of NTD and other features for 128 mixtures of GMM.

Feature	EER (%) for c0 thresholds			Min EER (%)	ERR (%)
	0	30	40		
TD	11.33	10.33	10.33	10.33	-
CMS.TD	6.83	4.83	5.17	4.83	53.23
CMVN.TD	10.33	7.33	7.00	7.00	32.26
NTD	**4.33**	**4.00**	**4.17**	**4.00**	61.29

Table 2 shows the performances of NTD and other features. The test results showed that the use of c0 threshold for excluding the mute section significantly influenced the performance. For CMS.TD and CMVN.TD, there was an absolute EER of 2% – 3% between threshold 0 and 30, which revealed that the exclusion of mute section is important in the soundtrack-based detection of objectionable content. The soundtrack-based performances of the features had a lower EER than the clip-based performance; however, it was similar to the order in the original environment.

NTD outperformed other features in all cases, and were less affected by the mute section exclusion than other features. This seems to have been caused when NTD used the variance information in the local segment. NTD had an ERR of 61.29% with respect to the TD.

It was not enough to compare the performances with a single value; the performance of each feature was examined with diverse thresholds. The c0 threshold that had the best EER was chosen by feature, and the receiver operating characteristic (ROC) curves were drawn in the Figure 3. The true positive rate of the y-axis must generally be 0 – 1 in the ROC curve, but as the performances of all features were good with an area under curve (AUC) of 0.95, only the range of 0.8 – 1 was represented for easy comparison. In the ROC curve, the proposed NTD had better performance than other features in almost all sections.

5. CONCLUSION

In this study, a new method was proposed to recognize the objectionable sounds in diverse noise environments by normalizing the conventional STDMFCC feature. The proposed NSTDMFCC allows the relative magnitude information of dynamic coefficients to be maintained by applying CMVN to the 0^{th} modulation part and normalizing the variance of the entire local segment in the dynamic part. Using two tests, the performance of the proposed NSTDMFCC was compared with the conventional STDMFCC and the STDMFCCs to which CMS and CMVN were applied. First, an objectionable sound recognition test was conducted with sound clips to which white noises were added. The proposed feature in the test was more stable at diverse noise levels and had an average ERR of 24.69% with respect to the STDMFCC. In the soundtrack-based detection test, the proposed feature had an ERR of 61.29% with respect to the conventional STDMFCC. Therefore, it seems that the proposed feature is more robust than STDMFCC against diverse noises and levels.

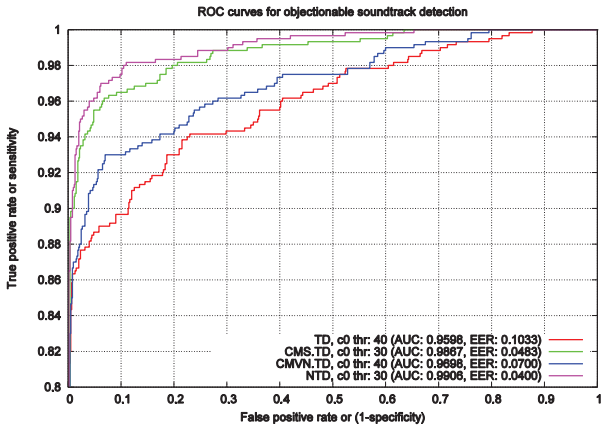

Fig. 3. ROC curves for objectionable soundtrack detection.

6. ACKNOWLEDMENTS

This research was supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2011-09914-06003).

7. REFERENCES

[1] M. Hammami, Y. Chahir, and L. Chen, "WebGuard: a Web filtering engine combining textual, structural, and visual content-based analysis," *IEEE Trans. Knowledge and Data Engineering*, Vol. 18(2), pp. 272-284, 2006

[2] J.-F. Yang, and X.-J. Shen, "Research on Key Technologies of Content-Based Erotic Image Filtering and It's Application," Proc. ICMLC 2006, pp. 3786-3792, 2006

[3] Z.Q. Shi, B.Y. Gao, J.Q. Han and Z. Wu, "Study of Objectionable Sound Recognition based on Histogram Features and SVM," Proc. CISP '09, pp. 1-4, 2009

[4] M. J. Kim, Y. Kim, J.D. Lim, H. Kim, "Automatic Detection of Malicious Sound Using Segmental Two-Dimensional Mel-Frequency Cepstral Coefficients and Histogram of Oriented Gradients," Proc. ACM Multimedia2010, pp. 887-890, Oct. 2010.

[5] J.D. Lim, S.W. Han, B.C. Choi, B.H. Chung, and C.H. Lee, "Classifying of objectionable contents with various audio signal features," Proc. ITCS 2010, pp.1-5, Aug. 2010

[6] H. Dudley, "Remarking Speech," *J. ACOUST. SOC. AM.*, Vol. 11(2), pp. 169-177, 1939

[7] C.-H. Lee, C.-C. Han, and C.-C. Chuang, "Automatic Classification of Bird Species from Their Sounds Using Two-Dimensional Cepstral Coefficients," *IEEE Trans. Audio, Speech, Lang. Process.*, Vol. 16, No. 8, pp. 1541-1550, Nov. 2008

[8] B. S. Atal, "Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification," *J. ACOUST. SOC. AM.*, Vol. 55(6), pp. 1304-1312, 1974

[9] O. Viikki, K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, Vol. 25, pp. 133-147, Aug. 1998