Subband Spectral Histogram Feature for Improved Sound Recognition in Low SNR Conditions

Roneel V. Sharan and Tom J. Moir
School of Engineering
Auckland University of Technology
Private Bag 92006, Auckland 1142, New Zealand
Email: roneel.sharan@aut.ac.nz, tom.moir@aut.ac.nz

Abstract—**In this work, we use the subband intensity histogram values extracted from the spectrogram image of sound signals to form the feature vector for sound classification in an audio surveillance application. We propose two features based on this approach. Firstly, we extract the histogram features from the short time Fourier transform spectrogram image of sound signals, which we refer as the spectral histogram feature (SHF). Secondly, we apply the mel-filter to the spectrogram image before histogram feature extraction which we refer as the mel-spectral histogram feature (MSHF). When compared to baseline features from similar work, the SHF was shown to give significantly improved results in low SNR conditions with a higher overall classification performance. In addition, the MSHF produced even better results than the SHF with the added advantage of a lower feature dimension.**

Keywords—audio surveillance; mel-filter; spectral histogram feature; sound recognition; support vector machine

I. Introduction

A sound signal produces a certain texture which can be visualized using a spectrogram image, the intensity values of which represent the dominant frequency components against time. Capturing this information during feature extraction can improve the recognition rate of sounds in the presence of additive noise, provided the noise spectrum does not contain strong spectral peaks. This was demonstrated by the use of spectral subband centroids (SSCs) as supplementary features for improved robustness in speech recognition in [1]. A similar approach was taken for sound event recognition in [2], where the spectrogram image is divided into multiple blocks and second and third central moments are computed in each block which form the feature vector, referred as the spectrogram image feature (SIF). The SIF with reduced feature dimensions (RSIF) can be found in [3].

The intensity histogram of a spectrogram image also captures the distribution of spectral energy and is often used in digital image processing in a technique referred as histogram equalization (HEQ). A grayscale or a color image often has most of its intensity values concentrated within a particular range. HEQ tends to enhance the contrast of an input image by transforming its cumulative density function (CDF) so that the intensities are equally distributed. In some applications, it may be desired to match the CDF of an input image to a predetermined CDF or the CDF of a reference image. This primarily digital image processing technique has also been applied to feature vector components to compensate for nonlinear distortions caused by noise to the speech representation as in [4].

In this work, however, we use this technique as a feature extraction tool for classification of sounds in an audio surveillance application and propose two features. Firstly, we determine the histogram of intensity values in each frequency bin of the short time Fourier transform (STFT) spectrogram image and concatenate the histogram values to form the feature vector. We refer this as the spectral histogram feature (SHF). Secondly, we apply the mel-filter to the spectral values and use the histogram of intensity values in each channel to form the feature vector, which we refer as the mel-spectral histogram feature (MSHF).

Similar approaches have been taken in image texture classification [5] and for robust speech classification [6]. We extend this technique for texture classification of sound signal spectrogram images and the robustness of the proposed features is tested in the presence of different noise environments at different signal-to-noise ratios (SNRs). The performance is measured against mel-frequency cepstral coefficients (MFCCs), the SIF, and the RSIF.

The rest of this paper is organized as follows. Section II gives an overview of feature extraction. Section III is on the experimentations we carried out and the corresponding results while conclusions are given in Section IV.

II. Feature Extraction

The procedure for feature extraction is given in the following subsections with reference to Fig. 1.

A. MFCCs

In computation of MFCCs, firstly, the discrete Fourier transform (DFT) is applied to the windowed signal as

$$X(k,t)=\sum_{n=0}^{N-1} x(n)w(n)e^{\frac{-2\pi ikn}{N}}, \qquad k=0,...,N-1 \quad (1)$$

where N is the window length, $x(n)$ is the time-domain signal, $X(k,t)$ is the k^{th} harmonic corresponding to the frequency $f(k)=kF_s/N$ for the t^{th} frame, F_s is the sampling frequency, and $w(n)$ is the window function.

978-1-4799-8058-1/15/$31.00 ©2015 IEEE

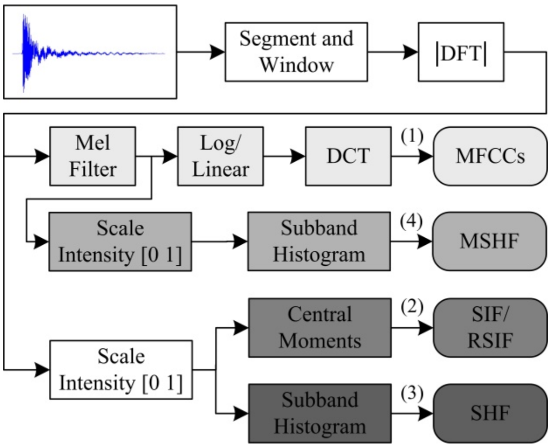


Figure 1. Steps in feature extraction: (1) path for MFCCs, (2) SIF and RSIF, (3) SHF, and (4) MSHF.

The triangular filters are equally spaced on the mel-scale [7] and the adjacent filters overlap such that the lower and upper end of a filter are located at the center frequency of the previous and next filter, respectively, while the peak of the filter is at its center frequency. The output of the m^{th} filter can then be determined as

$$E(m,t)=\sum_{k=0}^{\frac{N}{2}-1} V(m,k)\left|X(k,t)\right|, \qquad m=1,2,\ldots,M \tag{2}$$

where $E(m,t)$ represents the filter bank energies, M is the total number of mel-filters, and $V(m,k)$ is the normalized filter response.

The MFCCs are then obtained as the discrete cosine transform (DCT) of the log compressed filter bank energies given as

$$c(l,t)=\sqrt{\frac{2}{M}}\sum_{m=1}^{M}\log\left(E(m,t)\right)\cos\left(\frac{\pi l}{M}(m-0.5)\right) \tag{3}$$

which is evaluated from $l=1,2,\ldots,L$ where L is the order of the cepstrum.

We also report results using linear-MFCCs where no compression is applied to the filter bank energies before computing the cepstral coefficients and which was shown to be more noise robust in [3].

B. Spectrogram-Derived Features

We first give an overview of generating the spectrogram and mel-filtered spectrogram images. Feature extraction for SIF and RSIF is given next. Finally, feature extraction for SHF and MSHF is presented.

1) Generating Grayscale Spectrogram

The linear values are firstly obtained from the DFT values as

$$S(k,t)=\left|X(k,t)\right|. \tag{4}$$

These values are then normalized in the range $[0,1]$ which gives the grayscale spectrogram image intensity values. The normalization is given as

$$I(k,t)=\frac{S(k,t)-\min(S)}{\max(S)-\min(S)}. \tag{5}$$

The computation of the mel-filtered spectrogram values takes a similar approach, however, instead of normalizing $S(k,t)$, the filter bank output values, as computed in (2), are utilized.

Illustration of the two spectrogram images under clean conditions and with the addition of noise at 0dB SNR are given in Fig. 2. HSV color representations are shown for the grayscale values for better visualization.

2) SIF and RSIF

The central moments are extracted as features from the spectrogram images which form the SIF and the RSIF. For computing the central moments, the time-frequency image is divided into blocks and the v^{th} central moment for any given block of image is then determined as

$$\mu_v=\frac{1}{K}\sum_{i=1}^{K}\left(I_i-\mu\right)^v \tag{6}$$

where K is the sample size or the number of pixels in the block, I_i is the intensity value of the i^{th} sample in the block, and μ is the mean intensity value of the block.

3) SHF and MSHF

Given a grayscale spectrogram image $I(k,t)$ with G grayscale intensity levels, the intensity histogram can be determined as

$$h(k,g)=\sum_{t=1}^{N_t}\begin{cases}1, & \text{if } I(k,t)=g\\ 0, & \text{otherwise}\end{cases} \tag{7}$$

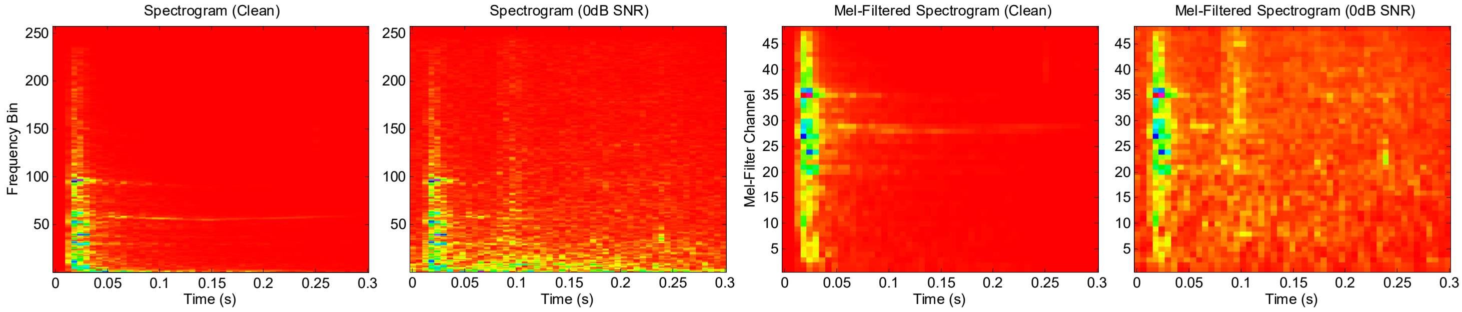
where $h(k,g)$ is a matrix with the count of grayscale intensity value g in the frequency bin k, $g=1,2,\ldots,G$ and $k=0,1,\ldots,N/2-1$, and N_t is the number of frames or pixels along the horizontal. The concatenation of the histogram values from each frequency bin forms the SHF.

For the MSHF, (7) can be modified as

$$h(m,g)=\sum_{t=1}^{N_t}\begin{cases}1, & \text{if } I(m,t)=g\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

where $m=1,2,\ldots,M$.

It should be noted that in the event $G=2$, the grayscale image is essentially converted to a binary image.



(a) (b) (c) (d)

Figure 2. Spectrogram images for a sound signal from *construction* sound class. (a) Spectrogram image under clean conditions, (b) spectrogram image at 0dB SNR with factory noise, (c) mel-filtered spectrogram image under clean conditions, and (d) mel-filtered spectrogram image at 0dB SNR with factory noise.

III. Experimental Evaluation

A description of the sound database used in this work is given first followed by an overview of the noise conditions and the experimental setup. We then present results using the baseline features which includes MFCCs and the spectrogram image features, SIF and RSIF. Finally, results using the proposed SHF and MSHF are presented.

A. *Sound Database*

The sound database has a total of 1143 files belonging to 10 classes: *alarms*, *children voices*, *construction*, *dog barking*, *footsteps*, *glass breaking*, *gunshots*, *horn*, *machines*, and *phone rings*. The sound files are largely obtained from the Real World Computing Partnership (RWCP) Sound Scene database in Real Acoustic Environment [8] and the BBC Sound Effects library [9]. All signals in the database have 16-bit resolution and a sampling frequency of 44100 Hz. More details about the sound database and its comparison with that used in other similar work can be found in [3].

B. *Noise Conditions*

The performance of the different features is investigated under three different noise environments taken from the NOISEX-92 database [10]: *speech babble*, *factory floor 1*, and *destroyer control room*. The signals are resampled at 44100 Hz and the overall performance is measured in clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs.

C. *Experimental Setup*

For all experiments, signal processing is carried out using a Hamming window of 512 points (11.61 ms) with 50% overlap. Support vector machine (SVM) is used for classification where the classification accuracy is given in percentage as *number of correctly classified test samples* divided by the *total number of test samples*. Being a binary classifier, we use the one-against-all (OAA) method [11] for multiclass classification where the classifier that has the highest output function assigns the class. In [3], the OAA method was shown to give the best overall performance when compared against three other multiclass classification methods and against the k-nearest neighbor (kNN) classifier.

All results are reported using a nonlinear SVM with a Gaussian radial basis function kernel as it was found to give the best results. The classifier parameters, refer to [3], were tuned using cross validation where, instead of maximizing the classification accuracy under each noise condition, samples from all noise conditions were used at once to get the best overall classification accuracy. For all experimentations, the classifier is trained with two-third of the clean samples with the remaining one-third data used for testing under clean and noisy conditions.

TABLE I. CLASSIFICATION ACCURACY USING BASELINE FEATURES

Feature	Clean	20dB	10dB	5dB	0dB	Average
Log-MFCC	98.43	92.83	73.14	57.57	43.31	73.05
Linear-MFCC	99.21	93.53	86.09	70.87	47.16	79.37
SIF	91.60	91.34	88.80	67.19	40.51	75.89
RSIF	92.13	92.04	89.33	78.57	53.37	81.08

D. *Results and Discussions*

1) *Baseline Features*

The first baseline method uses MFCCs as features. The feature vector for each frame is 39-dimensional: 13 MFCCs using a 20-filter bank system, plus deltas, and accelerations. The overall size of the feature vector for a signal is $39 \times N_t$. We present results using log-MFCCs and linear-MFCCs. For log-MFCCs, we apply logarithmic compression to the filter bank energies before computing the cepstral coefficients while no compression is applied in the case of linear-MFCCs. After data normalization, a 78-dimensional final feature vector is formed by concatenating the mean and standard deviation for each dimension.

The second baseline method uses the SIF and the RSIF. For the SIF, the spectrogram image is divided into 9×9 blocks and second and third central moments are computed in each block. These values are then concatenated into a column vector which forms a 162-dimensional feature vector. For the RSIF, the mean and standard deviation of the central moment values along the row and column of the blocks are concatenated to form a 72-dimensional final feature vector.

The classification accuracy values using the baseline features are given in Table I. MFCCs give the highest classification accuracy under clean conditions and at 20dB

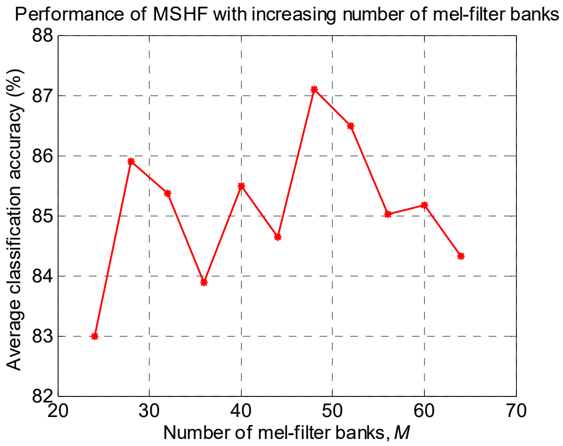


Figure 3. Average classification accuracy value for MSHF with increasing number of mel-filter banks

SNR with linear-MFCCs proving to be much more noise robust than log-MFCCs. However, the RSIF gives superior performance at 10dB, 5dB, and 0dB SNRs and a better overall classification accuracy than MFCCs and the SIF. A detailed analysis of the results can be found in [3].

2) *SHF and MSHF*

We now present results using the proposed histogram features. For the SHF, the histogram of the intensity values are computed in each frequency bin and concatenated to form the feature vector. In addition, we use only two histogram bins in each subband since it gave the best results. In general, it was observed that the average classification accuracy decreased as the number of histogram bins increased. With $N = 512$, the total number of frequency bins is 256 and with two histogram bins in each frequency bin, the final feature vector for the SHF has a dimension of 512. For the MSHF, the upper and lower cut-off frequencies are set as $^{F_s}/_2$ and $^{F_s}/_N$, respectively. We also consider various number of filter banks in the range 24 to 64 at intervals of 4 where the feature vector in each case is given as $2 \times M$. The average classification accuracy value for MSHF with increasing values of M is plotted in Fig. 3. The highest average classification accuracy was achieved at $M = 48$.

The results for SHF and MSHF ($M = 48$) are given in Table II. Looking at the results for SHF, there has been an increase in the average classification accuracy when compared to RSIF, the best performing baseline feature. The SHF gives lower classification accuracy than RSIF under clean conditions and at 20dB and 10dB SNRs, however, there is significant improvement in the classification accuracy at 5dB and 0dB SNRs.

Furthermore, at 87.10%, the average classification accuracy using MSHF is significantly better than all baseline features. In addition, it is also better than the results for SHF. Similar to the SHF, the MSHF outperforms the RSIF at 5dB and 0dB SNRs.

TABLE II. CLASSIFICATION ACCURACY USING SHF AND MSHF

Feature	Clean	20dB	10dB	5dB	0dB	Average
SHF	87.40	87.31	86.88	85.65	72.97	84.04
MSHF	92.39	92.39	91.69	86.26	72.79	87.10

However, unlike the SHF, the MSHF gives marginally better performance than RSIF at 10dB SNR and comparable results under clean conditions and at 20dB SNR. In addition, MSHF has the added advantage of a significantly lower feature dimension when compared to SHF. With $M = 48$, the MSHF feature dimension is 96, which is 2.67 times less than the SHF.

IV. CONCLUSION

The proposed histogram features, SHF and MSHF, were shown to outperform the baseline features in low SNR conditions. Out of the two features, the MSHF gave the best overall classification accuracy and the feature vector dimension is also much lower than the SHF. While the MSHF generally performs much better than all the baseline features in the presence of noise, it is not able to match the classification accuracy of MFCCs under clean conditions which is its only disadvantage.

REFERENCES

[1] K. K. Paliwal, "Spectral subband centroid features for speech recognition," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1998, pp. 617-620.

[2] J. Dennis, H. D. Tran, 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.

[3] R. V. Sharan and T. J. Moir, "Noise robust audio surveillance using reduced spectrogram image feature and one-against-all SVM," *Neurocomputing,* vol. 158, pp. 90-99, 2015.

[4] Á. De La Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing,* vol. 13, no. 3, pp. 355-366, 2005.

[5] X. Liu and D. Wang, "Texture classification using spectral histograms," *IEEE Transactions on Image Processing,* vol. 12, no. 6, pp. 661-670, 2003.

[6] B. Gajic and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 14, no. 2, pp. 600-608, 2006.

[7] D. O'Shaughnessy, *Speech communication: human and machine.* Addison-Wesley Pub. Co., 1987.

[8] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000, pp. 965–968.

[9] *BBC Sound Effects Library*. Available: http://www.leonardosoft.com

[10] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication,* vol. 12, no. 3, pp. 247-251, Jul. 1993.

[11] V. N. Vapnik, *Statistical learning theory.* New York: Wiley, 1998.