Cochleagram Image Feature for Improved Robustness in Sound Recognition

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 paper, we use the cochleagram image of sound signals for time-frequency analysis and feature extraction, instead of the conventional spectrogram image, in an audio surveillance application. The signal is firstly passed through a gammatone filter which models the auditory filters in the human cochlea. The filtered signal is then divided into small windows and the energy in each window is added and normalized which gives the intensity values of the cochleagram image. We then divide the cochleagram image into blocks and extract central moments as features. Using two feature vector representation methods, the results show significant improvement in overall classification accuracy when compared to results from literature employing similar feature extraction and representation techniques but using spectrogram images. The most improved results were at low signal-to-noise ratios.**

Keywords—audio surveillance; central moments; cochleagram; sound recognition; support vector machine

I. INTRODUCTION

Features extracted from the spectrogram image of sound signals have shown to be more noise robust in sound recognition than conventional features such as mel-frequency cepstral coefficients (MFCCs). In [1], after dividing the spectrogram image into blocks, central moments are extracted as features for robust sound event recognition, referred as the spectrogram image feature (SIF). In [2], we reduced the dimension of the SIF by using the mean and standard deviation of the central moment values along the row and column of the blocks, without compromising the classification accuracy, which we referred as the reduced spectrogram image feature (RSIF).

The spectrogram image is probably the most commonly used tool in time-frequency analysis of signals in both speech and sound recognition applications. The spectrogram image is generally formed by dividing the signal into smaller sections, referred as frames, and then applying discrete Fourier transform (DFT) to the windowed frames. The horizontal and vertical axis give time and frequency information, respectively. The frequency components are equally spaced along the vertical with constant bandwidth. However, most sound signals have greater frequency components in the lower frequency range and the information in these frequency components get compressed in this time-frequency representation.

A cochleagram [3] is similar to a spectrogram but a cochleagram uses the human auditory model for determining the center frequencies and bandwidth. A gammatone filter is often used for this purpose which is a linear filter modeling the frequency selectivity property of the human cochlea. It has more frequency components in the lower frequency range with smaller bandwidth and fewer frequency components in the higher frequency range with higher bandwidth. The most commonly used cochlea model is that proposed by Patterson et. al. [4]. It is a series of bandpass filters where the bandwidth is given by equivalent rectangular bandwidth (ERB). An efficient implementation of the gammatone filter bank has been provided in [5] which has been used for extracting gammatone cepstral coefficients (GTCCs) in both speech [6] and non-speech [7] recognition applications.

Time-frequency analysis and feature extraction using cochleagram images have a number of applications in areas of signal processing and pattern recognition. For example, features were extracted from cochleagram images in [8] in trying to improve the robustness in speech recognition. In [9], cochleagram features outperform a combination of common acoustic features in voice activity detection. Similar approach is also taken in [10] for audio separation purposes.

In this paper, we explore the applicability of cochleagram-based time-frequency analysis of sound signals for classification of sounds in an audio surveillance application. We test the effectiveness of the proposed approach with central moments as features and using two feature vector representation techniques from spectrogram analysis, SIF and RSIF, which, for the cochleagram, we refer as cochleagram image feature (CIF) and reduced cochleagram image feature (RCIF), respectively. We compare the results for the proposed features against MFCCs and the conventional spectrogram-derived features, SIF and RSIF, in different noise environments and at different signal-to-noise ratios (SNRs).

The rest of this paper is organized as follows. Section II gives an overview of time-frequency image formation and feature extraction. Section III is on experiments, results, and discussions while the conclusions are given in Section IV.

II. FEATURE EXTRACTION

In this work, we consider linear time-frequency images only since it showed greater robustness to logarithmic time-frequency images in [1, 2]. We first present the steps in generating the spectrogram and cochleagram images and then outline the procedure for feature extraction.

978-1-4799-8058-1/15/$31.00

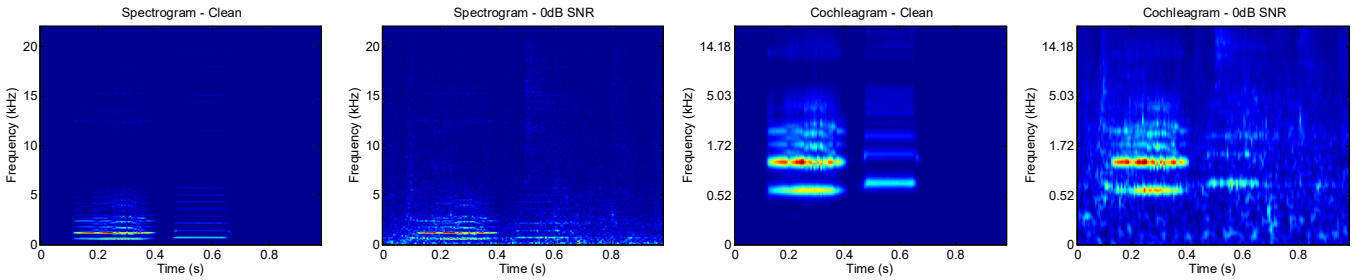

Figure 1. Spectrogram and cochleagram images for a sample sound signal. (a) Linear spectrogram image under clean conditions, (b) linear spectrogram image at 0dB SNR with factory noise, (c) linear cochleagram image under clean conditions, and (d) linear cochleagram image at 0dB SNR with factory noise.

A. Spectrogram

In generating the spectrogram image, firstly, the DFT is applied to the windowed signal as

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

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.

The linear values are obtained as

$$S(k,t)=\left|X(k,t)\right|. \quad (2)$$

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

$$I(k,t)=\frac{S(k,t)-\min(S)}{\max(S)-\min(S)}. \quad (3)$$

Illustration of spectrogram images, color representations are shown for the grayscale values for better visualization, under clean conditions and with the addition of noise at 0dB SNR can be found in Fig. 1(a) and (b), respectively.

B. Cochleagram

The formation of the cochleagram requires gammatone filter banks which are a series of bandpass filters the impulse response for which can be given as [4]

$$g(r)=Ar^{j-1}e^{-2\pi Br}\cos(2\pi f_c r+\phi) \quad (4)$$

where A is the amplitude, j is the order of the filter, B is the duration of the impulse response or filter bandwidth, f_c is the center frequency of the filter, ϕ is the phase, and r is the time.

The ERB is used to describe the bandwidth of each cochlea filter in [4]. ERB is a psychoacoustic measure of the auditory filter width at each point along the cochlea and can be given as

$$f_{c,ERB}=\left[\left(\frac{f_{c,Hz}}{Q_{ear}}\right)^p+(B_{min})^p\right]^{1/p} \quad (5)$$

where Q_{ear} is the asymptotic filter quality at high frequencies and B_{min} is the minimum bandwidth for low frequency channels. The bandwidth of a filter can then be approximated as $B=1.019\times f_{c,ERB}$. The three commonly used ERB filter models are given by Glasberg and Moore [11] ($Q_{ear}=9.26$, $B_{min}=24.7$, and $p=1$), Lyon's cochlea model as given in [12] ($Q_{ear}=8$, $B_{min}=125$, and $p=2$), and Greenwood [13] ($Q_{ear}=7.23$, $B_{min}=22.85$, and $p=1$).

The human cochlea has thousands of hair cells which resonate at their characteristic frequency and at a certain bandwidth. In [5], the mapping between center frequency and cochlea position is determined by integrating the reciprocal of (5) with a step factor parameter to indicate the overlap between filters. This can then be inverted to find the mapping between filter index and center frequency which can be given as

$$f_{cm}=-Q_{ear}B_{min}+(f_h+Q_{ear}B_{min})e^{-ms/Q_{ear}} \quad (6)$$

where $m=1,2,\ldots,M$, M is the number of gammatone filters, f_h is the maximum frequency in the filter bank, and s is the step factor given as

$$s=\frac{Q_{ear}}{M}\log\left(\frac{f_h+Q_{ear}B_{\min}}{f_l+Q_{ear}B_{\min}}\right) \quad (7)$$

where f_l is the minimum frequency in the filter bank.

We use a 4th order gammatone filter with four filter stages and each stage a 2nd order digital filter as given in [5]. The gammatone filter was implemented using the Auditory Toolbox for Matlab [14]. After filtering the signal with the gammatone filter, the energy in the windowed signal for each frequency component is added which can be given as

$$C(m,t)=\sum_{n=0}^{N-1}\left|\hat{x}(m,n)\right|w(n),\qquad m=1,\ldots,M \quad (8)$$

where $\hat{x}(m,n)$ is the gammatone filtered signal, $C(m,t)$ is the m^{th} harmonic corresponding to the center frequency f_{cm} for the t^{th} frame.

These values are then normalized using (3) to get the grayscale cochleagram image intensity values. Illustration of cochleagram images under clean conditions and with the addition of noise at 0dB SNR can be found in Fig. 1(c) and (d), respectively, using the same sound signal as the spectrogram images of Fig. 1(a) and (b).

C. Central Moments

The central moments are extracted as features from the time-frequency images. 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 determined as

$$\mu_v = \frac{1}{K}\sum_{i=1}^{K}(I_i - \mu)^v \tag{9}$$

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.

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, we present results using cochleagram image features, CIF and RCIF.

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 [15] and the BBC Sound Effects library [16]. 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 [2].

B. Noise Conditions

The performance of the different features is investigated under three different noise environments taken from the NOISEX-92 database [17]: *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 [18] for multiclass classification where the classifier that has the highest output function assigns the class. In [2], 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.

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

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 [2], 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.

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-filterbank system, plus deltas, and accelerations. The overall size of the feature vector for a signal is $39 \times N_t$, where N_t is the number of frames in the signal, which is different in each case depending on the length of the signal. We present two sets of results using this method, 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 features derived from the spectrogram image of the sound signal, namely 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 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

TABLE II. CLASSIFICATION ACCURACY USING CIF AND RCIF

ERB Filter Model	CIF						RCIF					
	Clean	20dB	10dB	5dB	0dB	Average	Clean	20dB	10dB	5dB	0dB	Average
Glasberg and Moore [11]	92.13	91.78	90.73	85.74	63.08	84.69	94.75	94.58	94.14	89.68	65.44	87.72
Lyon [12]	91.60	91.25	90.46	83.38	58.88	83.11	95.01	94.40	93.35	89.59	65.44	87.56
Greenwood [13]	93.18	93.09	92.21	89.06	63.95	**86.30**	94.75	94.75	94.58	91.69	69.38	**89.03**

analysis of the results can be found in [2].

2) Cochleagram Image Features

For obtaining the CIF and RCIF, we follow the same procedure as SIF and RSIF, respectively, but using a cochleagram image instead of spectrogram image. For best comparison, we create a similar experimental setup as the spectrogram image features. To get the same time-frequency image resolution, we use 256 gammatone filters ($M = 256$) and the same window size ($N = 512$). The classification accuracy values using CIF and RCIF for the three ERB filter models are given in Table II.

When compared to the SIF and the RSIF, the proposed CIF and RCIF generally show improvement in classification accuracy under all noise conditions, respectively. In both cases, there is significant improvement in the overall classification accuracy with the most improved results under noisy conditions, 10dB, 5dB, and 0dB SNRs, in particular. The Glasberg and Moore and the Lyon ERB filter models give comparable classification accuracy but the best overall performance for both feature sets is achieved using Greenwood's model. Highest classification accuracy is once again achieved using the reduced feature method, RCIF, with an average classification accuracy of 89.03%. It gives an improvement in classification accuracy of 2.62%, 2.71%, 5.25%, 13.12%, and 16.01% over the RSIF under clean conditions and at 20dB, 10dB, 5dB, and 0dB SNRs, respectively. As such, the improvement in the classification accuracy increases as the SNR decreases with the most improved results at 0dB SNR, increasing from 53.37% to 69.38%. Similar to the RSIF, the RCIF is not able to match the classification accuracy of MFCCs under clean conditions. However, it gives marginally better classification accuracy at 20dB SNR and significantly higher classification accuracy at 10dB, 5dB, and 0dB SNRs.

IV. CONCLUSION

Cochleagram-based time-frequency representation of a sound signal, which utilizes a gammatone filter, was found to be more effective for feature extraction than the spectrogram image. When compared to the spectrogram features, SIF and RSIF, the corresponding cochleagram features, CIF and RCIF, showed significant improvement in overall classification performance using all three ERB filter models. The performance of the cochleagram image features was seen to be especially better at low SNRs with the most improved results at 0dB SNR for both feature sets.

REFERENCES

[1] 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.

[2] 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.

[3] D. Wang and G. J. Brown, "Fundamentals of computational auditory scene analysis," in *Computational auditory scene analysis: Principles, algorithms and applications*, D. Wang and G. J. Brown, Eds. IEEE Press/Wiley-Interscience, 2006, pp. 1-44.

[4] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory physiology and perception*. vol. 83, Y. Cazals, L. Demany, and K. Horner, Eds. Pergamon, Oxford, 1992, pp. 429-446.

[5] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer, Technical Report 35, 1993.

[6] O. Cheng, W. Abdulla, and Z. Salcic, "Performance evaluation of front-end processing for speech recognition systems," The University of Auckland, New Zealand, Report 621, 2005.

[7] X. Valero and F. Alias, "Gammatone cepstral coefficients: Biologically inspired features for non-speech audio classification," *IEEE Transactions on Multimedia,* vol. 14, no. 6, pp. 1684-1689, 2012.

[8] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America,* vol. 134, no. 4, pp. 3029-3038, 2013.

[9] X.-L. Zhang and D. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," The Ohio State University, Columbus, OH, Technical Report OSU-CISRC-4/14-TR0, 2014.

[10] B. Gao, W. L. Woo, and L. C. Khor, "Cochleagram-based audio pattern separation using two-dimensional non-negative matrix factorization with automatic sparsity adaptation," *The Journal of the Acoustical Society of America,* vol. 135, no. 3, pp. 1171-85, Mar 2014.

[11] B. R. Glasberg and B. C. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research,* vol. 47, no. 1-2, pp. 103-138, 1990.

[12] M. Slaney, "Lyon's Cochlear Model," Apple Computer, Technical Report 13, 1988.

[13] D. D. Greenwood, "A cochlear frequency-position function for several species - 29 years later," *Journal of the Acoustical Society of America* vol. 87, no. 6, pp. 2592-2605, Jun 1990.

[14] M. Slaney, "Auditory Toolbox for Matlab," Interval Research Corproation, Technical Report 1998-010, 1998.

[15] 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.

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

[17] 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.

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