ENHANCED MODULATION SPECTRUM USING SPACE-TIME AVERAGING FOR IN-BUILDING ACOUSTIC SIGNATURE IDENTIFICATION

Somsak Sukittanon [*], *Les E. Atlas* [†], *and Stephen G. Dame* [*]

[*] ssukitta@ieee.org, steve@virtual-dsp.com, Virtual DSP Corporation WA United States

[†] atlas@ee.washington.edu, University of Washington, WA United States

ABSTRACT

For most buildings, virtually all subsystems such as fans, generators, and motors generate acoustic energy. This acoustic energy can weakly penetrate walls and pass through hallways and conduits. The propagation paths are complex and, given the typically low energy of received acoustic signals, present challenges for the detection and classification of the subsystems. While conventional approaches would not be expected to work under these difficult conditions, a key observation can be made: these types of subsystems produce line spectra which consist of harmonics of a fundamental frequency. Acoustic propagation effects then strongly affect the relative energy of these harmonics. Modulation spectra, which make use of the frequency spacing instead of the relative energy of the harmonics, are especially insensitive to these frequency-dependent acoustic attenuation affects. When combined with temporal averaging and 3-dimensional spatial (over an array of acoustic sensors) processing, enhanced modulation spectra offer a new approach to the detection and classification of building subsystems which produce sound.

1. INTRODUCTION

Monitoring of mechanical systems within building has maintenance, security, and defense application. However, little work has been done in the area of in-building acoustic signature identification. Simply acquiring large amounts of sampled acoustic data with high resolution time synchronization would not necessarily lead to reliable identification of types of subsystems. The scope and complexity of the type of sounds encountered in a typical building, combined with the multitude of complex and unknown acoustic propagation paths, make the problem difficult. Discovery of a useful and robust solution for field deployment requires a new approach to signal representation. A key observation can be made: mechanical subsystems produce line spectra which consist of harmonics of a fundamental frequency of rotating elements and/or power line frequencies. Acoustic propagation paths strongly affect the relative energy of these harmonics. A signal representation concept called the "modulation spectrum," which makes use of the frequency spacing instead of the relative energy of the harmonics, is especially insensitive to these frequency-dependent acoustic attenuation affects. When combined with temporal averaging and 3-dimensional spatial (over an array of acoustic sensors) processing, modulation spectra offer a new approach to the detection and classification of building subsystems which produce sound.

It can usually be assumed that building subsystems such as operating motors, fans, generators, compressors, centrifuges, lights, and other industrial equipment operate off the power mains. This power supply is typically a steady 60 Hz or 50 Hz signal. There are some exceptions, such as battery-powered inverters which approximate the 50 Hz or 60 Hz signal and aircraft electrical systems which operate at 400 Hz. For all of these cases, there are strong harmonics of the power mains each of which typically rise 20-40 dB above the background sound level of the machine [1]. Even harmonics usually dominate with at least eight and typically many more even and odd harmonics apparent. Also, due to the common use of gear and belt speed reduction systems, energetic lower frequency sub-harmonics are also often present. For remote acoustic observation of a machine, complex acoustic propagation paths will greatly change the relative magnitudes of the harmonics. Some will be fully attenuated yet others will pass with relatively little attenuation. Most importantly, the relative distance in frequency location between harmonic will not be affected by complex acoustic propagation paths, and hence becomes the foundation of our new approach to detection of sub-audible sub-systems.

With the subsystem assumed to be several rooms and floors away from acoustic sensors, acoustic energy can drop substantially to inaudible levels, yet still needs to be detectable. The proposed modulation spectral analysis, called the enhanced modulation spectrum, is capable of exploiting the temporal and 3-dimensional spatial information received from the sensors. Using space-time averaging, the resulting modulation spectrum can reduce the undesirable random effect, for example background noise, and generate useful features for acoustic classification. The acoustic signature output has the potential to extract time-varying information via the nonzero terms (bright colors in Fig. 3d, 4b, and 5b). These nonzero terms are possibly useful for discriminating and classifying rotating machine types.

2. MODULATION SPECTRUM

2.1. Conventional methods

Given the time signal, $x(t)$, there are various ways to estimate a modulation spectral representation. One of the most widely known methods is related to a Wigner distribution by a Fourier transform in time, t, and time lag, τ, of the local autocorrelation of the time signal [2].

$$P(\eta,\omega)=\frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^{*}\left(t-\tau/2\right)x\left(t+\tau/2\right)e^{-j\eta t}e^{-j\omega\tau}dtd\tau \qquad (1)$$

ω and η are referred to "acoustic" and "modulation" frequency, respectively. As shown in [3], $P(\eta,\omega)$ could generate cross-term interference and poor energy compaction in the estimate. To remove these undesirable terms, smoothing methods have been applied. If the statistics of the signal spectrum change periodically with period T_0 and it is priorly known, the cyclic spectrum [4] $P_{CY}(\eta,\omega)$, can be approximated by temporal averaging of $P(\eta,\omega)$.

$$P_{CY}(\eta,\omega)=E_t\left\{P(\eta,\omega)\right\} \qquad (2)$$

$$=\frac{1}{NT_0}\int_{-\infty}^{\infty}\int_{-\frac{T_0}{2}}^{\frac{T_0}{2}}\left(\sum_{n=0}^{N-1}x^{*}\left(t+nT_0-\tau/2\right)x\left(t+nT_0+\tau/2\right)\right)e^{-j\eta t}e^{-j\omega\tau}dtd\tau$$

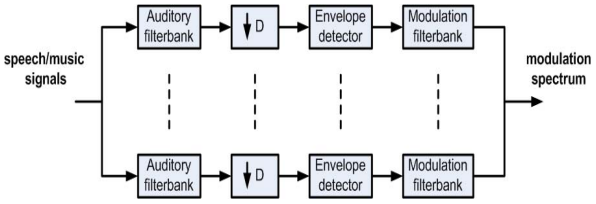

Figure 1: A generic block diagram of modulation spectral analysis

Since this approach assumed the periodicity in the local autocorrelation with a period T_0, i.e. $x^*(\bullet)x(\circ) = x^*(\bullet + nT_0)x(\circ + nT_0)$, a Fourier series expansion is instead applied in Equation (2) causing a major difference between $P(\eta,\omega)$ and $P_{CY}(\eta,\omega)$ in the η dimension. η has continuous values in (1) while η in (2) becomes discrete representing every harmonic of the fundamental frequency, $1/T_0$. As shown in [3], $P_{CY}(\eta,\omega)$ yields smooth estimates but it still provided redundant terms occurring in very high η. This coherent method relies on the accurate estimate of a period T_0 which is difficult to estimate for arbitrary signals such as speech, music, or in-building machine signals. To obtain better cross-terms and redundancy reduction, an incoherent approach which makes use of a two-dimensional smoothing function and does not require prior estimate of a period T_0 is considered. A common approach is to first exploit the inherent smoothing properties of the spectrogram resulting in the "modulation spectrum" [5]. First, a spectrogram with an appropriately chosen window length, $w(t)$, is used to estimate a joint time-frequency representation of the signal. Then, a Fourier transform is applied along the time dimension of the spectrogram, yielding an estimate of the modulation spectrum.

$$P_{SP}(\eta,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(w^*(t-\tau/2)w(t+\tau/2)*_t x^*(t-\tau/2)x(t+\tau/2)\Big)e^{-j\eta t}e^{-j\omega\tau}\,dt\,d\tau \tag{3}$$

In speech or music applications, an auditory filterbank, e.g. mel-spaced filters, is alternatively chosen [6-9] to uniform spaced filters of spectrogram. The generic diagram can be illustrated in Figure 1. First the waveform is assumed to be band-limited, with $f_m^{\min}$ and $f_m^{\max}$ as the lowest and highest modulation frequencies in Hz. The digitized signal, with sample length L and sampling rate F_s Hz, is decomposed by an auditory filterbank. In order to represent $f_m^{\min}$, the minimum length of L is at least $F_s/f_m^{\min}$ samples. For example, at least 100 ms of waveform signal is required to capture a 10 Hz modulation frequency. The decimation step, D or the amount of nonoverlapping for each frame, reduces the amount of data (i.e. sampling rate) in each channel. The upper bound for D to avoid the possibility of subsequent aliasing of the subband signals depends on $f_m^{\max}$ and a type of an envelope detector, e.g. magitude-square, absolute square, or Hilbert. For example, D would be smaller than $F_s/4f_m^{\max}$ when using magnitude square operator. Continuing to work through Figure 1, a modulation filterbank is applied for each channel independently. For implementation purposes, demeaning and a taper window can be used to reduce the sidelobes of the subsequent modulation frequency estimate. The specification of the modulation filterbank, the center frequencies or number of channels, depends on the each application. Normally, the center frequency is up to a few hundred hertz and the number of modulation filters is between 8 and 13.

2.2. Enhanced modulation spectrum

For a large in-building subsystem detection and classification, it is desirable to be capable of deploying a three-dimensional sensor network to collect, process and transmit relevant acoustic data to a central server for further analysis and coordination among sensors which are tuned to listen for periodic modulating sounds emitted from localized sources such as operating motors, generators, compressors, fans, centrifuges, lights, and other relevant industrial equipment. A significant problem in such sensor deployments is how modulation spectra can be estimated using data from each sensor within this 3-dimensional space. The enhanced modulation spectral processing should concentrate interacting harmonics into modulation frequency points that are robust to unknown propagation paths, allowing dramatically improved detection capabilities for rotating sound sources. In this paper, we present the new approach, called the enhanced modulation spectrum, exploiting the temporal and spatial averaging which can further increase detection sensitivity.

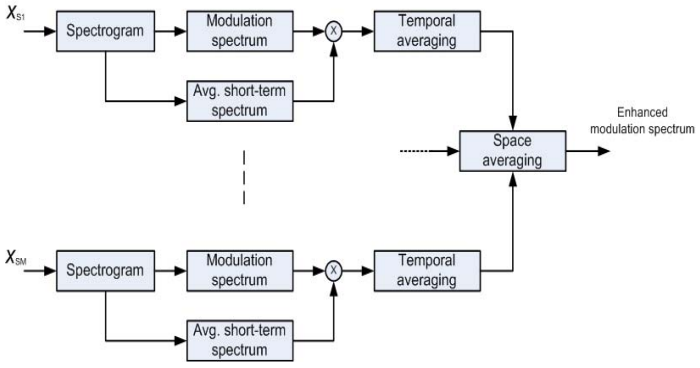

Figure 2: A block diagram of enhanced modulation spectral analysis

As illustrated in Figure 2, the enhanced modulation spectrum starts with a standard spectrogram applied to the data retrieved from the i^{th} sensor, X_{si}. Then, an averaged short-term spectral estimate of the signal is computed by integrating $P_{SP,i}(t,\omega)$ over t.

$$P_{SP,i}(t,\omega) = \frac{1}{2\pi}\left|\int_{-\infty}^{\infty} x_{si}(u)\,w^*(u-t)\,e^{-j\omega u}\,du\right|^2$$
$$P_{SP,i}(\omega) = \int_{-\infty}^{\infty} P_{SP,i}(t,\omega)\,dt \tag{4}$$

After a Fourier transform is applied along the time dimension

$$P_{SP,i}(\eta,\omega) = \int_{-\infty}^{\infty} P_{SP,i}(t,\omega)\,e^{-j\eta t}\,dt, \tag{5}$$

the modulation spectral representation for each time frame, n, is estimated by the product of normalized short-term spectral estimate and modulation spectrum.

$$\hat{P}_{SP,i}^{(n)}(\eta,\omega) = \frac{P_{SP,i}^{(n)}(\omega)}{\left\|P_{SP,i}^{(n)}(\omega)\right\|_1} P_{SP,i}^{(n)}(\eta,\omega). \tag{6}$$

When combined the temporal (N) and space (M) averaging, the enhanced modulation spectral representation becomes

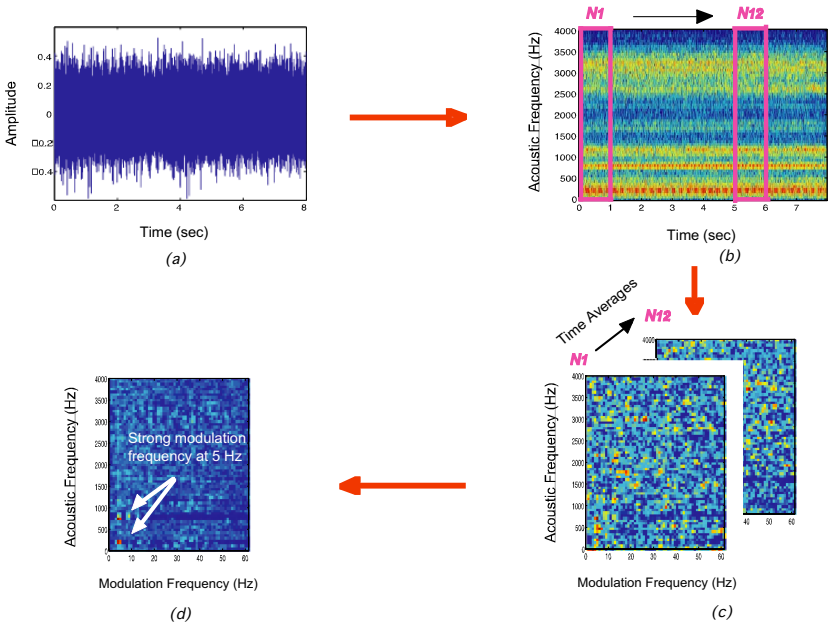

Figure 3: Enhanced modulation spectrum for hotwater pump acoustic sound: (a) time domain; (b) spectrogram representation; (c) modulation spectrum representation for 12 frames; (d) space-time averaged modulation spectrum representation.

$$P_{ENSP}(\eta,\omega)=\left(\prod_i\left(\prod_n \hat{P}_{SP,i}^{(n)}(\eta,\omega)\right)^{\frac{1}{N}}\right)^{\frac{1}{M}}. \qquad (7)$$

For the implementation purpose, the subband energy normalization, mean subtraction, and window tapering can be applied to Eq. (5) to improve modulation frequency detection sensitivity.

3. EXPERIMENTS

3.1. Data Collection

The audio data used here were recorded from the in-building subsystems located at the department of electrical engineering, University of Washington, and Virtual DSP Corporation. Each datafile was recorded using Casio DAT player with the sampling rate 48 KHz and stereo channels. The dataset contained 3 different sound classes, i.e. a building's operating hotwater pump, chiller, and fan. For classification purpose, we split the data equally into training and testing sets, without any overlap. A white Gaussian noise at 10dB and 0dB SNR levels was also added to the clean signal on both sets to generate a stereo database, i.e. the data containing clean signal and their corresponding noisy counterparts. In total, about 500 enhanced modulation feature frames were available for training and testing.

First, we showed examples of the proposed technique. In Fig. 3, the acoustic sound collected from hot water pumps was resampled to 8 KHz and shown in Fig. 3a. A spectrogram was then computed using a Hanning window of length 128 samples and a window shift of 64 samples thereby reducing the subband sampling rate to 125 Hz. As illustrated in Fig. 3b, another 1 second window is applied to block the spectrogram data, with 500 ms overlapped between each block, and the corresponding modulation spectrum for each frame was estimated and illustrated in Fig. 3c. The enhanced modulation spectrum which exploited the time (12 frames within 6 seconds) and spatial (2 channels) averaging showed the distinct compaction of high energy occurring low modulation frequencies, 5 Hz, at acoustic frequencies 200 and 800 Hz. We could notice its smoothness compared to the single modulation spectral frame in Fig. 3c. The same technique and parameters applied to different acoustic sounds generated from a chiller and fan, gave us the results as shown in Fig. 4 and 5, respectively. Since they were generated from different sources, we were able to see the significant differences in terms of the high energy location (red color) in outputs of Fig. 4b and 5b. A chiller sound yielded the distinct compaction of high energy occurring at 30 Hz modulation frequency while a fan sound produced the dominant modulation frequency at 45 Hz.

3.2. Feature Extraction and Classification

After generating two-dimensional enhanced modulation spectrum features, $P_{ENSP}(\eta,\omega)$, post processing was applied in order to reduce the large dimensionality and to make the features insensitive to variation, for example, acoustic and time shifts. Past research has addressed various ways of reducing the feature dimensionality of a two-dimensional representation. One method is by viewing these features as an image. The small set of descriptors being invariant to acoustic frequency and time shift can be extracted using singular value decomposition (SVD). As shown in [3], SVD applied to modulation features improved the error reduction in signal interception application. Finally, the 1-second data was represented by 1x63 vectors.

Classification of a test signal was performed after feature extraction using a *k*-nearest neighbor classifier. This method does not assume any prior distribution of data. Given a test feature vector, the decision is chosen to be the class that is most commonly represented in the *k* closest neighbors, assuming equal training samples and prior class probability. In this paper, we experimented with various *k* values (from 1 to 20) and chose the one giving the best accuracy.

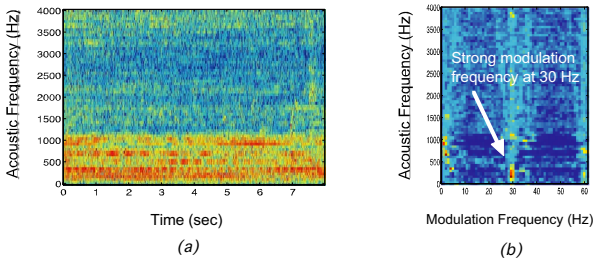

Figure 4: Enhanced modulation spectrum for chiller acoustic sound: (a) spectrogram representation; (b) space-time averaged modulation spectrum representation using the same technique and parameters as in Figure 3.

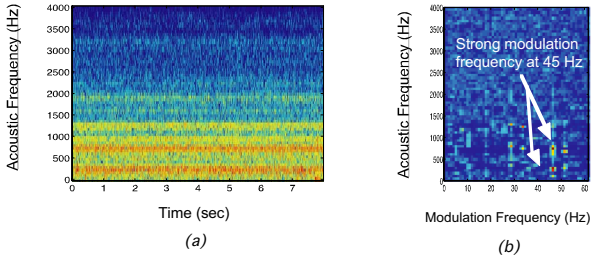

Figure 5: Enhanced modulation spectrum for fan acoustic sound: (a) spectrogram representation; (b) space-time averaged modulation spectrum representation using the same technique and parameters as in Figure 3.

3.3. Results

The first set of experiments was to choose the k value that would be used for other experiments. Mel frequency cepstral coefficients (MFCC) were used as the baseline experiment. These 13 cepstral features were estimated using 40 frames per second. The first coefficient was removed due to energy sensitivity and the first and second order differences were combined to form 36 coefficients.

As shown in Table 1, using 36 dimensional MFCC features yielded the highest accuracy of 55.7% with k = 8. While conventional modulation spectral features, as described by (3), combined with a SVD, provided a much improved accuracy of 89.1%. When applying the same classifier structure with the proposed enhanced modulation spectral features, the accuracy became 98.0%. Note that the output decision was made for every half second for all feature sets. Since 1-NN does not require any parameter, we also used it for comparing the performance of different features. As shown in Table 1, enhanced modulation spectral features provided the higher accuracy results compared to other two features.

Next, the optimal subset of features was searched for comparing the same feature dimensionality. Fisher's discriminant ratio [10] was considered for feature selection. Using k=1, the highest accuracy for MFCC features was 59.5%, using subset 14 coefficients. When subselecting modulation features to the same number coefficients, the improved performance of enhanced modulation spectral features was also seen in this test, as shown in the last three rows of Table 1.

Table 1: The accuracy performance of different features using *k*-NN classifiers.

Features	#dim	1-NN	8-NN
MFCC + Δ + $\Delta\Delta$	36	52.8 %	55.7 %
Modulation	63	89.8 %	89.1 %
Enhanced Modulation	63	97.2 %	98.0 %
MFCC + Δ + $\Delta\Delta$ + Feature Selection	14	59.5 %	62.6 %
Modulation + Feature Selection	14	83.0 %	86.1 %
Enhanced Modulation + Feature Selection	14	89.0 %	91.4 %

4. CONCLUSIONS

We present a new modulation frequency analysis with application to in-building acoustic signature identification. The new approach incorporates temporal averaging and 3-dimensional spatial (over an array of acoustic sensors) processing resulting in the high concentration of interacting harmonics from the spectrogram representation into modulation frequency representation. After post processing using a singular value decomposition, it provides dramatically improved detection and classification for rotating sound sources. When the enhanced modulation spectrum is compared to cepstral features and the conventional modulation spectrum for in-building acoustic signature identification using a k-nearest neighbor classifier, the new approach provided a substantially lower error rate, with or without feature selection for dimensionality reduction.

5. REFERENCES

[1] N. Barkova, "The Current State of Vibroacoustical Machine Diagnostics," www.vibrotek.com/articles/state/

[2] L. Cohen, *Time-Frequency Analysis*. Englewood Cliffs, NJ: Prentice Hall, 1995.

[3] S. Sukittanon, L. Atlas, J. Pitton, and K. Filali, "Improved modulation spectrum through multi-scale modulation frequency decomposition," in *Proc. ICASSP*, vol. 4, 2005, pp. 517-520.

[4] W. A. Gardner, *Statistical Spectral Analysis: A Nonprobabilistic Theory*. Englewood Cliffs, NJ: Prentice Hall, 1988.

[5] H. Hermansky, "Should recognizers have ears?," *Speech Communication*, vol. 25, pp. 3-27, 1998.

[6] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Communication*, vol. 25, pp. 117-132, 1998.

[7] N. Mesgarani, S. Shamma, and M. Slaney, "Speech discrimination based on multiscale spectro-temporal modulations," in *Proc. ICASSP*, vol. 1, 2004, pp. 601-604.

[8] N. Malyska, T. Quatieri, and D. Sturim, "Automatic dysphonia recognition using biologically-inspired amplitude-modulation features," in *Proc. ICASSP*, vol. 1, 2005, pp. 873-876.

[9] S. Sukittanon, L. Atlas, and J. Pitton, "Modulation scale analysis for content identification," *IEEE Transactions on Signal Processing*, vol. 52, pp. 3023-3035, 2004.

[10] D. H. Kil and F. B. Shin, *Pattern Recognition and Prediction with Applications to Signal Characterization*. Woodbury, NY: AIP, 1996.