IMPROVED MODULATION SPECTRUM THROUGH MULTI-SCALE MODULATION FREQUENCY DECOMPOSITION

Somsak Sukittanon [1], Les E. Atlas [1], James W. Pitton [2], and Karim Filali [3]

[1] Department of Electrical Engineering, [2] Applied Physics Laboratory (APL), [3] Department of Computer Science & Engineering
University of Washington, Seattle, WA

ABSTRACT

The modulation spectrum is a promising method to incorporate dynamic information in pattern classification. It contains important cues about the nonstationary content of a signal and yields complementary improvements when it is combined with conventional features derived from short-term analysis. Many prior modulation spectrum approaches are based on uniform modulation frequency decomposition. The drawbacks of these approaches are high dimensionality and a lack of a connection to human perception of modulation. This paper presents multi-scale modulation frequency decomposition and shows an improvement over standard modulation spectrum in a digital communication signal classification task. Features derived from this representation provide lower classification error rates than those from a constant-bandwidth modulation spectrum whether used alone or in combination with short-term features.

1. INTRODUCTION

In pattern recognition, conventional feature analysis is usually based on short-time analysis of data, that is, over a short data window. Although short-term features have shown good performance under some assumptions, several researchers, e.g. [1], have raised the question of whether using only short-term features is adequate for nonstationary signal classification. These short-term features cannot sufficiently model time-varying information of nonstationary signals without classifier memory (e.g., Hidden Markov Models) and/or features with expanded temporal extent. To overcome the deficiencies of short-term features, much work [2-4], motivated by psychoacoustic results, has investigated modulation spectrum for long-term signal analysis. The modulation spectrum not only contains short-term information about the signals, but also provides long-term information representing patterns of time variation. Incorporating modulation spectral features into signal classification can provide significant improvement over systems using only short-term features in a broad range of applications [2-5].

Two approaches for generating a modulation spectrum are to take a Fourier [1] or DCT [3, 6] transform of a sequence of short-term magnitude spectrum features. Since this analysis uses uniform frequency decomposition, the resulting modulation frequency resolution is constant. Uniform modulation frequency decomposition may not be appropriate for classification due to the resulting high feature dimension; furthermore, it does not match models of human auditory perception. Recent studies [7, 8] of auditory frequency selectivity for amplitude modulation showed that a log frequency scale best matches human perception of modulation frequency. Accordingly, to overcome these disadvantages, a wavelet transform is applied as the second transform to yield a multi-scale modulation frequency decomposition. The new representation not only yields much lower feature dimensionality compared to the standard modulation spectrum, but also provides high discrimination capability and low sensitivity to distortions. Experiments using real world communication signals show that multi-scale modulation spectrum can provide classification error rates lower than uniform modulation spectrum whether they are used exclusively or in combination with short-term spectral features.

2. MODULATION SPECTRUM

2.1. Previous methods

A conventional joint frequency representation $P_x(\eta,\omega)$ is the correlation function of a Fourier transform, $X(\omega)$, of the time signal, $x(t)$, where ω and η are referred to "Fourier" and "modulation" frequency, respectively, defined as [9]:

$$P_x(\eta,\omega) \quad = \quad X^*(\omega-\eta/2)X(\omega+\eta/2)\,. \qquad (1)$$

This representation is related to a Wigner distribution by a Fourier transform in η. To study the behavior of joint frequency analysis, a simple amplitude modulated signal, $x(t)=(1+\cos\omega_m t)\cos\omega_c t$ where ω_m and ω_c are the modulation and carrier frequencies, respectively, is used. Using (1), the joint frequency representation of this AM signal is illustrated in Figure 1a. Ideally for this AM signal, the desirable representation with reduced cross-terms and good energy compaction should have nonzero Dirac impulse terms at only $(\eta=0,\omega=\pm\omega_c)$ and $(\eta=\pm\omega_m,\omega=\pm\omega_c)$. As shown in Figure 1a, there are also undesirable cross-terms occurring at double modulation frequencies, $\eta=\pm 2\omega_m$, and redundant terms occurring at much higher modulation frequencies, $\eta=\pm 2\omega_c$. The cross-terms can be interpreted as interference due to the quadratic nature of (1).

To remove these undesirable terms, several approaches can be taken. Synchronized block averaging [10] can be applied if the statistics of its spectrum change periodically with period T_0 when the signal is observed for a long period of time. If the periodicity T_0 of the signal's statistics is known, the cyclic spectrum $\tilde{S}_x(\eta,\omega)$ can be approximated by averaging adjacent joint frequency estimates computed at intervals of T_0. $\tilde{S}_x(\eta,\omega)$, as illustrated in Figure 1b, can be related to (1) as

$$\tilde{S}_x(\eta,\omega)=\left(\sum_{n=-\infty}^{\infty}\delta(\eta-n/T_0)P_x(\eta,\omega)\right)*_\omega 2T_0\,\mathrm{sinc}\left(\omega T_0\right). \qquad (2)$$

For arbitrary signals such as speech or music in which the frequency of periodicity is difficult to estimate, undesirable effects can be significantly reduced by a two-dimensional smoothing function. A simple approach is to exploit the inherent smoothing properties of the spectrogram, which is referred to as the "modulation spectrum [1]." First, a spectrogram or other representation with an appropriately chosen window length 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 $P_x^{SP}(\eta,\omega)$, in which undesirable terms are smoothed and attenuated. Redundant terms are also removed as shown in Figure 1c. $P_x^{SP}(\eta,\omega)$, can be linked to $P_x(\eta,\omega)$ by

$$P_x^{SP}(\eta,\omega) \quad = \quad P_x(\eta,\omega)*_\omega P_h(\eta,\omega) \qquad (3)$$

where h is the window used in computing the spectrogram and $P_h(\eta,\omega)$ is the joint frequency representation of h.

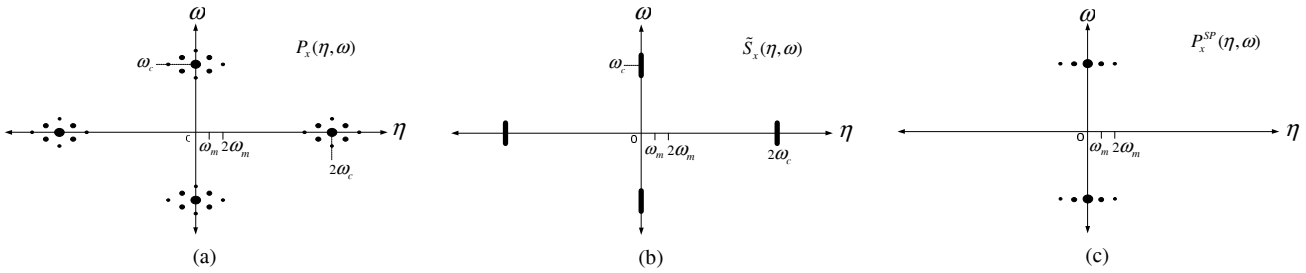

Figure 1: A joint frequency representation of an AM signal using (a) an instantaneous correlation function, (b) synchronized block averaging [10], and (c) a modulation spectrum computed from a spectrogram, as described above.

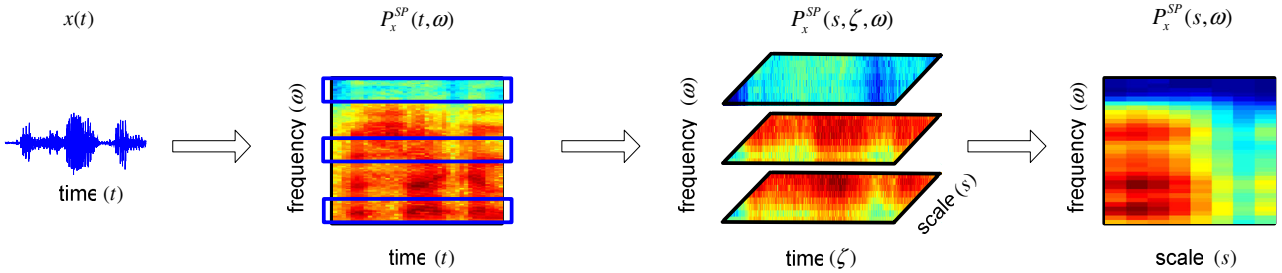

Figure 2: Computing the multi-scale modulation spectrum. The final representation is obtained by a time average for each frequency and scale.

2.2. Multi-scale modulation spectrum

As demonstrated with an AM signal in Figure 1c, modulation spectral analysis has the potential to extract time-varying information via the nonzero terms in the representation. When the analysis is applied to real-world signal, e.g. speech, music, or communication signals, these nonzero terms can represent various quantities such as phonetic information, pitch, tempo, or symbol rate, and they are potentially useful for discriminating signal types in pattern classification. However, using previous modulation spectral analysis as feature extraction still suffered from a fundamental disadvantage, namely that it yields a much larger dimension than traditional short-term spectral estimates. Past research has addressed the method of reducing feature dimension of a two dimensional representation in various ways. Since we are interested in tasks where human auditory signal classification is largely successful, integrating psychoacoustic results into the analysis can possibly provide added advantages in feature design and selection.

Using Fourier analysis, or other uniform frequency decompositions, for the modulation frequency transform in modulation spectra results in a uniform frequency bandwidth in modulation frequency dimension; however this approach for modulation decomposition can be inefficient for auditory classification due to the resulting high dimensionality. Furthermore, the uniform bandwidth in modulation frequency does not mimic the human auditory system. Inspired by psychoacoustic results [8], a log frequency scale, with resolution consistent with a constant-Q over the whole range, best mimics human perception of modulation frequency. Auditory modulation filters with frequencies up to 64 Hz are approximated constant-Q, with a value of about 1. Our approach uses a continuous wavelet transform (CWT) to efficiently approximate this constant-Q effect, though we could also less efficiently achieve constant-Q by grouping Fourier coefficients. The multi-scale modulation spectrum representation is the joint representation of the Fourier frequency and modulation frequency with nonuniform bandwidth for the latter. As illustrated in Figure 2, the analysis consists of three important steps. It starts with a standard spectrogram of $x(t)$:

$$P_x^{SP}(t,\omega) = \frac{1}{2\pi}\left|\int x(u)h^*(u-t)e^{-j\,\omega\,u}du\right|^2. \qquad (4)$$

In the second step, for discrete scales s , the wavelet filter $\psi(t)$ is applied along each temporal row of the spectrogram output:

$$P_x^{SP}(s,\zeta,\omega) = \frac{1}{s}P_x^{SP}(\zeta,\omega) *_\zeta \psi^*(-\frac{\zeta}{s}). \qquad (5)$$

The above equation can be viewed as applying wavelet transform on a temporal envelope in each Fourier frequency subband except the scaling term $1/s$ which serves for normalizing the passband magnitude of each filter to be equal. And in the last step, the energy across the wavelet translation axis ζ is integrated:

$$P_x^{SP}(s,\omega) = \int\left|P_x^{SP}(s,\zeta,\omega)\right|^2 d\zeta. \qquad (6)$$

The above equation yields a joint frequency representation with nonuniform resolution in the modulation frequency dimension, as indexed by the discrete scale s .

There are many advantages of using wavelet based transform over Fourier ones in modulation decomposition. For classification purpose, we showed in [5] that a wavelet approach provided better distribution of frequency resolution in modulation frequency by showing correctly distinct nonzero terms of multi-component AM signals. For example, when we compared wavelet and Fourier in term of energy compaction for synthesis and analysis purpose, nonuniform modulation decomposition also achieved higher signal-to-noise ratio in reconstructed speech and music signals for different compression rates.

3. EXPERIMENTS

3.1. Task

In many applications such as interception of battlefield communications, the modulation type transmitted over analog channels is unknown, and identifying the type is a critical first step in monitoring the communication channel. Past research in automatic identification of modulation type has used a combination of short-term spectral features. Benvenuto [11] introduced the second-order moment of a complex envelope of a signal for distinguishing speech from voiceband data. Sewall and Cockburn [12] improved upon Benvenuto's work by discarding the demodulation stage and still achieved comparable performance with less computation. Hsue and Soliman [13] employed zero crossing variance, carrier-to-noise ratio, and carrier frequency features. Later, they proposed the statistical moment of the signal phase [14]. Recently, Azzouz and Nandi [15] proposed a new framework that made a significant contribution to the field of modulation classification. This framework utilized moments of the instantaneous amplitude, phase, and frequency of the signal. A combination of these short-term features with conventional classifiers, such as a decision tree or neural network, showed high performance for both analog and digital modulation classification. Since then these key features have been incorporated in several studies with additional short-term features. In this work, our main goal was to improve the performance of short-term features by incorporating long-term modulation features. We show that multi-scale modulation spectral features can provide lower error rates than conventional modulation spectral features when they are used independently or in combination with other features. Statistical short-term features [15] previously used in this application were chose for comparison.

3.2. Feature Extraction and Classification

The data[1] used in the experiments was collected and labeled by an expert listener. The dataset contained four different modulation classes: FSK (frequency shift keying), MFSK (multilevel FSK), PSK (both binary phase shift keying and multilevel PSK), and MCVFT (multichannel FSK and/or multichannel PSK). These 216 files shown in Table 1 contained several communication modes, such as idle, traffic, or both. The details of each signal file can be found in [5].

First, each file was resampled to 11025 Hz and all long silences were removed, the resampled audio was partitioned into 3 second windows for long-term feature analysis and 50 ms windows for short-term feature analysis. For every 3-second block with a frame rate 50 ms, modulation scale features were generated using a spectrogram 128 point and Hanning window. A window shift of 21 samples was used to reduce the subband sampling rate to about 512 Hz during the modulation transform. For multi-scale modulation spectrum, biorthogonal wavelet filters, with 8 different dyadic scales, were applied to produce one non-uniform modulation frequency vector in each Fourier subband. After generating two-dimensional modulation scale features $P_{mod}[s_d, k]$, feature normalization was applied.

Table 1: The number of files and feature frames used in the experiments.

Type	*FSK*	*MFSK*	*PSK*	*MCVFT*
Number of files	77	41	72	26
Number of frames	26365	9056	19067	8158

[1] The database is available at http://rover.vistecprivat.de/~signals/.

Due to the method of data collection, the nature of the initial demodulation process may introduce a Fourier frequency shift in the joint frequency representation. Because of this frequency translation, we cannot directly apply modulation scale features to typical classifiers. To reduce this effect, the post processing to modulation features is necessary.

If the signal is shifted by ω_0 in a Fourier frequency dimension, i.e. $y(t) = e^{j\omega_0 t} x(t)$, what results is a shift in the ω dimension or $P_{mod,y}[s_d, k] = P_{mod,x}[s_d, k - k_0]$ for discrete implementation where k_0 is the amount of frequency translation. Equivalently, this effect can be viewed as a vertical shift while the horizontal structure in the joint frequency representation remains the same. When using the SVD, we can estimate the Fourier frequency vectors P^a and modulation scale vectors P^m given $P_{mod,x}$, the feature matrix with rank r, by (where σ is a nonnegative weight)

$$P_{mod,x}[s_d, k] = \sum_{j=1}^{r} \sigma_j P_j^a[k] P_j^m[s_d] = \mathrm{U}_x \Sigma_x \mathrm{V}_x^{\mathrm{T}}. \quad (7)$$

Σ is a diagonal matrix of singular values, U is the matrix of left eigenvectors, and V is the matrix of right eigenvectors. When this representation is shifted vertically, the resulting feature matrix $P_{mod,x}[s_d, k - k_0]$ can be approximated as a row permutation of $P_{mod,x}[s_d, k]$. A row permutation in the matrix results in a row permutation of the left singular vectors implying that the frequency shift affects only P^a values.

$$P_{mod,y}[s_d, k] = P_{mod,x}[s_d, k - k_0] = (\mathrm{I}_r \mathrm{U}_x) \Sigma_x \mathrm{V}_x^{\mathrm{T}} \quad (8)$$

Since P^m (or V) and σ (or Σ) are insensitive to Fourier frequency shifts, they have potential for long-term features that are insensitive to frequency translations. Because P_{mod} can be mostly represented using only one basis vector, we derived multi-scale modulation spectrum features, called MODS, as

$$\mathrm{MODS}[s] = sign\left(\sum_s P_1^m[s]\right) \sqrt{\sigma_1} P_1^m[s]. \quad (9)$$

For conventional modulation spectral features, a Fourier transform was applied instead of a wavelet transform. Note that, both modulation spectral features have the same feature dimension. For comparison, short-term features using 50 ms non-overlapping data window were considered. There are two sets of short-term features. The first set is 8 dimensional high order moment features. These features have been commonly used in modulation classification to describe the spread and peakedness of signals. More details about these moment features can be found in [15]. Five other conventional short-term features, also insensitive to frequency shift, were extracted. They are the modified second-order moment of the real-valued rectified passband signal, the mean and standard deviation of the demodulated baseband spectum, and the entropy and bandwidth of the short-term spectrum. All features were normalized by the standard deviation estimated from all signal classes to reduce their dynamic range. Two parametric classifiers, Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs), were used in the experiments. With GMM classifiers, a diagonal covariance matrix was used for each Gaussian component. To prevent singularities in the model's likelihood function, variances were constrained to have a minimum value of 0.001 in all our experiments. For the HMM classifiers, fully connected topologies that allow transitions between any pair of states were used. From preliminary experiments, these models performed better than left-to-right models. As with the GMMs, the diagonal covariance matrix was used for each Gaussian component. To find the optimal number of states and mixture components, many structures were explored.

Table 2: The classification error rates of different features using GMM and HMM classifiers where the leave-one-out approach was used to evaluate the performance.

Features	Gaussian Mixture Models	Hidden Markov Models
Short-term spectral	25.5%	25.5%
Modulation Spectrum	31.9%	27.8%
Modulation Spectrum + Short-term Spectral	21.8%	22.7%
Multi-scale Modulation Spec	30.1%	27.3%
Multi-scale Modulation Spec + Short-term Spectral	19.4%	19.0%

3.3. Results

Due to the small amount of data, a leave-one-out approach was employed to evaluate all experiments. For each test file, the class giving the maximum posteriori probability was chosen. In order to obtain reliable estimates, the ratio of training data to the number of model parameters was considered to be at least ten. From Table 1, there are about 8000 frames of the MCVFT class corresponding to 15 mixtures for GMM classifiers. The minimum error rates for each feature using GMMs was chosen from 15 experimental results and summarized in Table 2. Using only modulation spectrum or multi-scale modulation spectrum does not provide error rates lower than using only conventional short-term features. These effects were consistent with the results in [2]. However, when the short-term features were combined with long-term features, the error rate was significantly reduced. Multi-scale modulation spectrum combined with short-term features yielded an error rate of 19.4% which was lower than the error rate of modulation spectrum combined with short-term features, 21.8%, an 11% reduction in error rate over standard modulation spectral analysis using GMM classifiers. For HMM classifiers, to find the optimal number of states and mixture components, a ratio of the number of training data to the total number of parameters of at least ten was still applied. The maximum number of states and mixture components were 10 and 15, respectively. The minimum error rate for each feature was chosen from 80 experimental results and summarized in Table 2. In testing with dynamic classifiers, the inclusion of multi-scale modulation spectrum into the feature extraction also yielded an error rate lower than combining modulation spectrum. Multi-scale modulation spectral analysis achieved 16% reduction in error rate over standard modulation spectral analysis using HMM classifiers.

4. CONCLUSIONS

We present an improvement to the modulation spectrum using multi-scale decomposition of modulation frequency. The multi-scale modulation spectrum incorporates recent knowledge about human perception in modulation dimension in the design. The fundamental advantage of this approach is reduced dimensionality. When compared to a uniformly spaced modulation spectrum in digital communication signal classification, the multi-scale approach provided error rates lower than the standard modulation spectrum approach. In a study using real-world data with long-term (i.e., modulation) features combined with short-time features, the multi-scale method achieved an 11-16% reduction in error rate compared to a uniform-resolution modulation spectrum, and a 24-25% reduction in error rate when compared to conventional short-term features alone. These results were confirmed with both static (Gaussian Mixture Model) and dynamic (Hidden Markov Model) classifiers.

5. REFERENCES

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

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

[3] V. Tyagi, I. McCowan, H. Misra, and H. Bourlard, "Mel-cepstrum modulation spectrum (MCMS) features for robust ASR," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 399-404.

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

[5] S. Sukittanon, "Modulation Scale Analysis: Theory and Application for Nonstationary Signal Classification," Ph.D. Dissertation, University of Washington, Seattle, 2004.

[6] B. Milner, H. T. Bunnell, and W. Idsardi, "Inclusion of temporal information into features for speech recognition," in *Proc. ICSLP*, vol. 1, 1996, pp. 256-9.

[7] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *Journal of the Acoustical Society of America*, vol. 85, pp. 1676-1680, 1989.

[8] S. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *Journal of the Acoustical Society of America*, vol. 108, pp. 1181-96, 2000.

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

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

[11] N. Benvenuto, "A speech/voiceband data discriminator," *IEEE Transactions on Communications*, vol. 41, pp. 539-43, 1993.

[12] J. S. Sewall and B. F. Cockburn, "Voiceband signal classification using statistically optimal combinations of low-complexity discriminant variables," *IEEE Transactions on Communications*, vol. 47, pp. 1623-7, 1999.

[13] S.-Z. Hsue and S. S. Soliman, "Automatic modulation classification using zero crossing," *IEE Proceedings of Radar and Signal Processing*, vol. 137, pp. 459-464, 1990.

[14] S. S. Soliman and S.-Z. Hsue, "Signal classification using statistical moments," *IEEE Transactions on Communications*, vol. 40, pp. 908-916, 1992.

[15] A. K. Nandi and E. E. Azzouz, "Algorithms for automatic modulation recognition of communication signals," *IEEE Transactions on Communications*, vol. 46, pp. 431-436, 1998.