EURASIP Journal on Applied Signal Processing 2003:7, 668–675

Joint Acoustic and Modulation Frequency

Les Atlas
Department of Electrical Engineering, Box 352500, Seattle, WA 98195-2500, USA
Email: atlas@ee.washington.edu

Shihab A. Shamma
Department of Electrical and Computer Engineering and Center for Auditory and Acoustic Research, Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
Email: sas@eng.umd.edu

Received 30 August 2002 and in revised form 5 February 2003

There is a considerable evidence that our perception of sound uses important features which are related to underlying signal modulations. This topic has been studied extensively via perceptual experiments, yet there are few, if any, well-developed signal processing methods which capitalize on or model these effects. We begin by summarizing evidence of the importance of modulation representations from psychophysical, physiological, and other sources. The concept of a two-dimensional joint acoustic and modulation frequency representation is proposed. A simple single sinusoidal amplitude modulator of a sinusoidal carrier is then used to illustrate properties of an unconstrained and ideal joint representation. Added constraints are required to remove or reduce undesired interference terms and to provide invertibility. It is then noted that the constraints would be also applied to more general and complex cases of broader modulation and carriers. Applications in single-channel speaker separation and in audio coding are used to illustrate the applicability of this joint representation. Other applications in signal analysis and filtering are suggested.

Keywords and phrases: Digital signal processing, acoustics, audition, talker separation, modulation spectrum.

1. INTRODUCTION

Over the last decade, human interfaces with computers have passed through a transition where images, video, and sounds are now fundamental parts of man/machine communications. In the future, machine recognition of images, video, and sound will likely be even more integral to computing. Much progress has been made in the fundamental scientific understanding of human perception and why it is so robust. Our current knowledge of perception has greatly improved the usefulness of information technology. For example, image and music compression techniques owe much of their efficiency to perceptual coding. However, it is easy to see from the large bandwidth gaps between waveform- and structural-based (synthesized) models [1] that there is still room for significant improvement in perceptual understanding and modeling.

This paper's aim is a step in this direction. It proposes to integrate a concept of sensory perception with signal processing methodology to achieve a significant improvement in the representation and coding of acoustic signals. Specifically, we will explore how the auditory perception of very low-frequency modulations of acoustic energy can be abstracted and mathematically formulated as invertible transforms that will prove to be extremely effective in the coding, modification, and automatic classification of speech and music.

2. THE IMPORTANCE OF MODULATION SPECTRA

Very low-frequency modulations of sound are the fundamental carrier of information in speech and of timbre in music. In this section, we review the psychophysical, physiological, and other sources of evidence for this perceptual role of modulations. We also justify the need for a theory of and general analysis/synthesis tools for a transform dimension approach often called "modulation spectra."

In 1939, Dudley concluded his now famous paper [2] on speech analysis with "...the basic nature of speech as composed of audible sound streams on which the intelligence content is impressed of the true message-bearing waves which, however, by themselves are inaudible."

In other words, Dudley observed that speech and other audio signals such as music are actually low-bandwidth processes that modulate higher-bandwidth carriers. The suggestion is that the mismatch between the physical nature of the acoustic media (air) and the size of our head and vocal tract has resulted in this clever mechanism: lower-frequency "message-bearing waves" hypothetically modulate our more efficiently produced higher-frequency acoustic energy.

Eleven years later, in a seemingly unrelated paper on time-varying systems [3], Zadeh first proposed that a separate dimension of modulation frequency could supplant

the standard concept of system function frequency analysis. His proposed two-dimensional system function had two separate frequency dimensions—one for standard frequency and the other a transform of the time variation. This two-dimensional bi-frequency system function was not analyzed but only defined. Kailath [4] followed up nine years later with the first analysis of this joint system function.

2.1. Motivation from auditory physiology

In 1971, Møller [5] first observed that the mammalian auditory system has a specialized sensitivity to amplitude modulation of narrowband acoustic signals. Suga [6] showed that for bats, amplitude modulation information was maintained for different cochlear frequency channels. Schreiner and Urbas [7] then showed that this neural representation of amplitude modulation was even seen at higher levels of mammalian audition such as the auditory cortex and was hence preserved up through all levels of our auditory system. Continued work by others showed that these effects were not only observable; they were instead potentially fundamental to the encoding used by mammalian auditory systems. For example, as shown by Langner [8], "... experiments using signals with temporal envelope variations or amplitude modulation ... a mere place model of frequency representation in the central nervous system cannot account for many aspects of auditory signal analysis and that for complex signal processing, in particular, temporal patterns of neuronal discharges are important."

In recent years the physiological evidence has only gotten stronger. Kowalski et al. [9, 10, 11] have shown that cells in the auditory cortex—the highest processing stage along the primary auditory pathway—are best driven by sounds that combine both spectral and temporal modulations. They used specially designed stimuli (called ripples) which have dynamic broadband spectra that are amplitude modulated with drifting sinusoidal envelopes at different speeds and spectral peak densities. By manipulating the ripple parameters and correlating them with the responses, they were able to estimate the spectrotemporal modulation transfer functions of cortical cells and, equivalently, their spectrotemporal receptive fields (or impulse responses). Based on such data, they have postulated that the auditory system performs effectively a multiscale spectrotemporal analysis which reencodes the acoustic spectrum in terms of its spectral and temporal modulations. As we will elaborate below, the perceptual relevance of these findings and formulations was investigated psychoacoustically and applied in the assessment of speech intelligibility and communication channel fidelity.

Finally, Schulze and Langner [12] have demonstrated that pitch and rhythm encoding are potentially separately explained by convolutional and multiplicative (modulation) models and, most importantly, Langner et al. [13] have observed through magnetoencephalography (MEG) that frequency and periodicity are represented via orthogonal maps in the human auditory cortex.

2.2. Motivation from psychoacoustics

The psychoacoustic evidence in support of the perceptual saliency of signal modulations is also very strong. For example, Viemeister [14] thoroughly studied human perception of amplitude-modulated tones and showed it to be a separate window into the analysis of auditory perception. Houtgast [15] then showed that the perception of amplitude modulation at one frequency masks the perception of other nearby modulation frequencies. Bacon's and Grantham's experiments [16] further support this point and they directly conclude that "These modulation-masking data suggest that there are channels in the auditory system which are tuned for the detection of modulation frequency, much like there are channels (critical bands or auditory filters) tuned for the detection of spectral frequency."

The most recent psychoacoustic experiments have continued to refine the information available about human perception of modulation frequency. For example, Sheft and Yost [17] have shown that our perception of consistent temporal dynamics corresponds to our perceptual filtering into modulation frequency channels. Also, Ewert and Dau [18] have recently shown dependencies between modulation frequency masking and carrier bandwidth. It is also worth noting from their study and from [13] that modulation frequency masking effects are indicative that much unneeded redundancy might still be unnecessarily maintained in today's state-of-the-art speech and audio coding systems.

Finally, Chi et al. [19, 20] have extended the findings above to include combined spectral and temporal modulations. Specifically, they measured human sensitivity to ripples of different temporal modulation rates and spectral densities. A remarkable finding of the experiments is the close correspondence between the most sensitive range of modulations, and the spectrotemporal modulation content of speech. This result suggested that the *integrity* of speech modulations might be used as a barometer of its intelligibility, as we will briefly describe next.

2.3. Motivation from speech perception

Further evidence for the value of modulations in the perception of speech quality and in speech intelligibility has come from a variety of experiments by the speech community. For example, the concept of an acoustic modulation transfer function [21], which arose out of optical transfer functions (e.g., [22]), has also been successfully applied to the measurement of speech transmission quality (speech transmission index, STI) [23]. For these measurements, modulating sine waves range in frequency from 0.63 Hz to 12.7 Hz in 1/3-octave steps. These stimuli were designed to simulate intensity distributions found in running speech and were used to test the noise and reverberant effects in acoustic enclosures such as auditoria. More direct studies on speech perception [24] demonstrated that the most important perceptual information lies at modulation frequencies below 16 Hz. More recently, Greenberg and Kingsbury [25] showed that a "modulation spectrogram" is a stable representation of speech for

automatic recognition in reverberant environments. This modulation spectrogram provided a time-frequency representation that maintained only the 0- to 8-Hz range of modulation frequencies (uniformly for all acoustic frequencies) and emphasized the 4-Hz range of modulations.

Based on the premise that faithful representation of these modulations is critical for the perception of speech [17, 21], a new intelligibility index, the spectrotemporal modulation index (STMI), was derived [19, 20] which quantifies the degradation in the encoding of *both* spectral and temporal modulations due to noise regardless of its exact nature. The STI, unlike the STMI, can best describe the effects of spectrotemporal distortions that are *separable* along these two dimensions, for example, static noise (purely spectral) or reverberation (mostly temporal). The STMI, which is based on ripple modulations, is an elaboration on the STI in that it incorporates explicitly the *joint* spectrotemporal dimensions of the speech signal. As such, we expect it to be consistent with the STI in its estimates of speech intelligibility in noise and reverberations, but also to be applicable to cases of *joint* (or inseparable) spectrotemporal distortions that are unsuitable for STI measurements (as with certain kinds of channel-phase distortions) or severely nonlinear distortions of the speech signal due to channel-phase jitter and amplitude clipping. Finally, like the STI, the STMI effectively applies specific weighting functions on the signal spectrum and its modulations; these assumptions arise naturally from the properties of the auditory system and hence can be ascribed a biological interpretation.

2.4. Motivations from signal analysis and synthesis

It is important to note that joint acoustic and *temporal* modulation frequency analysis has not yet been put into an analysis/synthesis framework. The previously mentioned papers by Zadeh [3] and Kailath [4] did propose a joint analysis and, more recently, Gardner (e.g., [26, 27]) greatly extended the concept of bi-frequency analysis for cyclostationary systems. These cyclostationary approaches have been widely applied for parameter estimation and detection. However, transforms that are used in compression and for many pattern recognition applications usually have a need for invertibility, like the Fourier or wavelet transform. Cyclostationary analysis does not provide an analysis-synthesis framework. Furthermore, the foundation that assumes infinite time limits in cyclostationary time averages is not directly appropriate for many speech and audio applications.

Higher-order spectral analysis also has a common formulation called the "bispectrum," which is an efficient way of capturing non-Gaussian correlations via two-dimensional Fourier transforms of third-order cumulant sequences of discrete time signals (e.g., [28]). There is no direct connection between bispectra and the joint acoustic and modulation frequency analysis we discuss.

There have been other examples of analysis that estimated and/or formulated joint estimates of acoustic and modulation frequency. Some recent examples are Scheirer's tempo analysis of music [29] and Haykin-Thomson [30] linking of a joint spectrum to a Wigner-Ville distribution. AM and FM (amplitude modulation—frequency modulation) and related energy detection, and separation techniques are also directed at estimation problems [31, 32, 33, 34]. These techniques require assumptions of single-component or a small number of multicomponent carriers and are hence not general enough for arbitrary sounds and images. All of these examples also lack general invertibility.

Many examples of current sound synthesis based upon modulation grew out of Chowning's frequency modulation technique for sound synthesis [35], as summarized by more recent suggestions of general applicability to structured audio [1]: "Although FM techniques provide a large variety of musically useful timbres, the sounds tend to have an "FM quality" that is readily identified. Also, there are no straightforward methods to determine a synthesis algorithm from an analysis of a desired sound; therefore, the algorithm designs are largely empirical."

Amplitude and frequency modulation-based analysis/synthesis techniques have been previously developed (e.g., [34]), but they are based upon a small number of discrete carrier components. Even with a larger number of discrete narrowband carriers, noise-like sounds cannot be accurately analyzed or produced. Thus, discrete sinusoidal or other summed narrowband carrier models are not general enough for arbitrary sounds and images. For example, while these techniques provide intelligible speech, they could not be applied to high- or even medium-quality audio coding. We are, nevertheless, highly influenced by these models. Simply put, our upcoming formulation is a generalization of previous work on sinusoidal models. As will be justified in the following sections, a more general amplitude modulation or, equivalently, multiplicative model can be empirically verified to be very close to invertible, even after significant compression [36].

In the remainder of this paper, we will illustrate how an analysis/synthesis theory of modulation frequencies can be formulated and applied to the problem of efficient coding and representation of speech and music signals. The focus in this paper will be exclusively on the use of *temporal* modulations, *leaving the spectral dimension unchanged*. This is mostly done to simplify the initial analysis and to explore the contribution of purely temporal modulations to the encoding of sound.

3. A MODULATION SPECTRAL MODEL

For further progress to be made, understanding and applying modulation spectra, a well-defined foundation for the concept of modulation frequency needs to be established. In this section, we will propose a foundation that is based upon a set of necessary conditions for a two-dimensional acoustic frequency versus modulation frequency representation. By "acoustic frequency" we mean an exact or approximate conventional Fourier decomposition of a signal. "Modulation frequency" is the dimension that this section will begin to strictly define.

The notion of modulation frequency is quite well understood for signals that are narrowband. A simple case consists

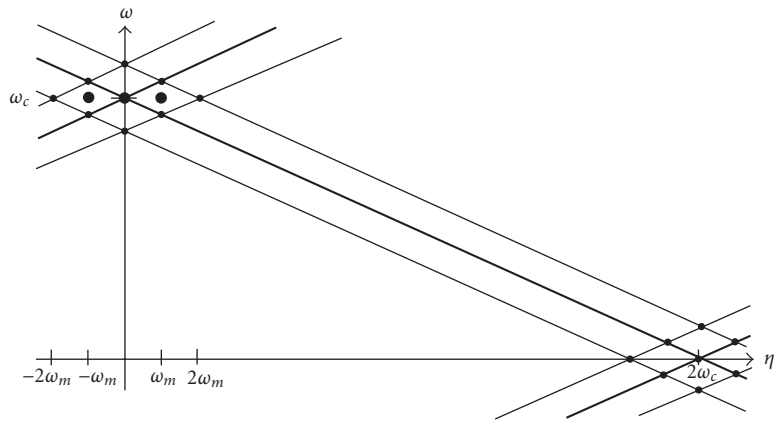

Figure 1: Two-dimensional representation of cosinusoidal amplitude modulation. The solid lines represent the support regions of both $S(\omega - \eta/2)$ and $S^*(\omega + \eta/2)$. Thicker lines represent the double area under the carrier-only terms relative to the modulated terms. The small dots, including the one hidden under the large dot at $(\eta = 0,\ \omega = \omega_c)$, represent the support region of the product $S(\omega - \eta/2)S^*(\omega + \eta/2)$. The three large dots represent the ideal representation $P_{\text{ideal}}(\eta, \omega)$ of modulation frequency versus acoustic frequency.

of an amplitude-modulated fixed frequency carrier

$$s_1(t) = m(t)\cos\omega_c t, \tag{1}$$

where the modulating signal $m(t)$ is nonnegative and has an upper frequency band limit suitable for its perfect and easy recovery from $s_1(t)$. It is straightforward that the modulation frequency for this signal should be the Fourier transform of the modulating signal only:

$$M(e^{j\omega}) = F\{m(t)\} = \int_{-\infty}^{\infty} m(t)e^{-j\omega t}dt. \tag{2}$$

But what is a two-dimensional distribution of acoustic versus modulation frequency? Namely, how would this signal be represented as the two-dimensional distribution $P(\eta, \omega)$, where η is modulation frequency and ω is acoustic frequency?

To begin answering this question, we can further simplify the model signal to have a narrowband cosinusoidal modulator

$$s(t) = (1 + \cos\omega_m t)\cos\omega_c t. \tag{3}$$

In order to allow unique recovery of the modulating signal, the modulation frequency ω_m is constrained to be less than the carrier frequency ω_c. The additive offset allows for a nonnegative modulating signal. Without loss of generality, we assume that the modulating signal is normalized to have peak values of ± 1 allowing the additive offset to be 1.

The process of amplitude demodulation, whether it is by magnitude, square law, Hilbert envelope, cepstral or synchronous detection, or other techniques, is most generally expressed as a frequency shift operation. Thus, a general two-dimensional representation of $s(t)$ has the dimensions acoustic frequency versus frequency translation. For example, much as in the bilinear formulation seen in time-frequency analysis, one dimension can simply express acoustic frequency ω and the other dimension can express a symmetric translation of that frequency via the variable η:

$$S\left(\omega - \frac{\eta}{2}\right)S^*\left(\omega + \frac{\eta}{2}\right), \tag{4}$$

where $S(\omega)$ is the Fourier transform of $s(t)$:

$$S(\omega) = F\{s(t)\} = \int_{-\infty}^{\infty} s(t)e^{-j\omega t}dt \tag{5}$$

and $S^*(\omega)$ is the complex conjugate of $S(\omega)$. This representation is similar to the denominator of the spectral correlation function described by Gardner [27].

Note that there is a loss of sign information in the above bilinear formulation. For analysis/synthesis applications, such as in the approaches discussed later in this paper, phase information needs to be maintained separately.

In the same spirit as previous uses and discussions of modulation frequency, an ideal two-dimensional representation $P_{\text{ideal}}(\eta, \omega)$ for $s(t)$ should have only significant energy density at only six points in the (η, ω) plane:

$$\begin{aligned} P_{\text{ideal}}(\eta, \omega) = {} & \delta(0, \omega_c) + \delta(\omega_m, \omega_c) + \delta(-\omega_m, \omega_c) \\ & + \delta(0, -\omega_c) + \delta(\omega_m, -\omega_c) \\ & + \delta(-\omega_m, -\omega_c), \end{aligned} \tag{6}$$

that is, jointly at the carrier and modulation frequencies only with added terms at the carrier frequency for DC modulation, to reflect the above additive offset of the modulating signal. However, going strictly by the definitions above, the Fourier transform of the narrowband cosinusoidal modulator $s(t)$ is

$$\begin{aligned} S(\omega) = F\{s(t)\} &= F\{(1 + \cos\omega_m t)\cos\omega_c t\} \\ &= \frac{1}{2}\{\delta(\omega - \omega_c) + \delta(\omega + \omega_c)\} \\ &\quad + \frac{1}{4}\{\delta(\omega - \omega_c - \omega_m) + \delta(\omega - \omega_c + \omega_m) \\ &\qquad + \delta(\omega + \omega_c + \omega_m) + \delta(\omega + \omega_c - \omega_m)\}. \end{aligned} \tag{7}$$

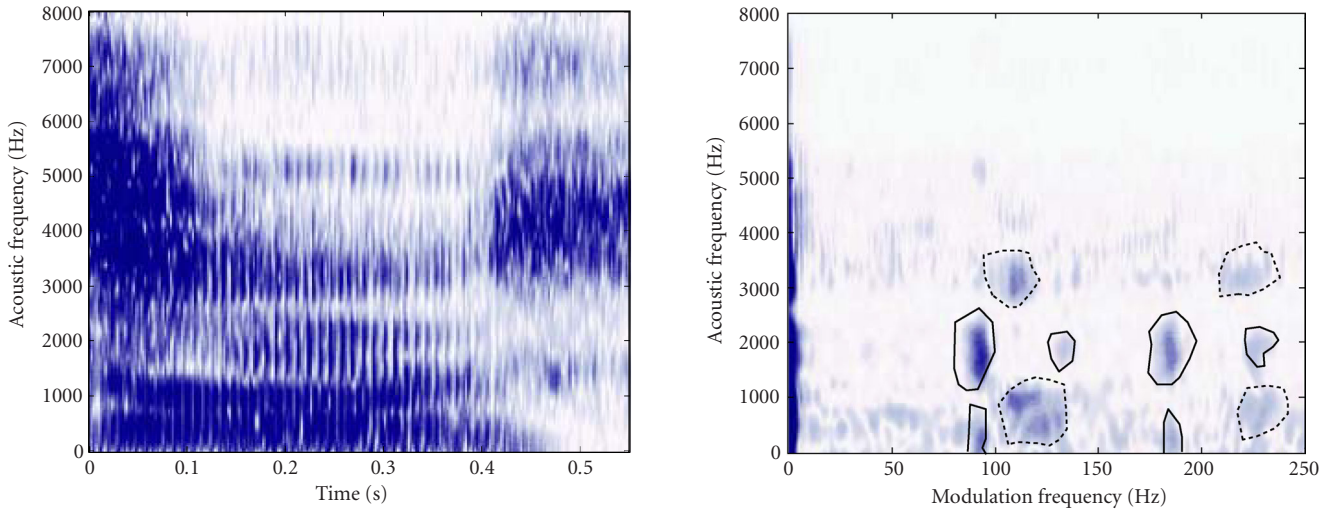

FIGURE 2: Spectrogram (left panel) and joint acoustic/modulation frequency representation (right panel) of the central 450 milliseconds of "two" (speaker 1) and "dos" (speaker 2) spoken simultaneously by two speakers. The y-axis of both representations is standard acoustic frequency. The x-axis of the right panel representation is modulation frequency, with an assumption of Fourier basis decomposition. Solid and dashed lines surround speaker 1's and speaker 2's respective pitch information.

This transform when expressed as a bilinear formulation $S(\omega - \eta/2)S^*(\omega + \eta/2)$ has much more extent in both η and ω than desired. A comparison between the ideal and actual two-dimensional representation is schematized in Figure 1.

It can be observed from Figure 1 that the representation $S_2(\omega + \eta)S_2^*(\omega - \eta)$ has more impulsive terms than the ideal representation. Namely, the product $S_2(\omega + \eta)S_2^*(\omega - \eta)$ is underconstrained. To approach the ideal representation, two conditions need to be added: (1) a kernel which is convolutional in ω and (2) a kernel which is multiplicative in η. Thus, a sufficient condition for the ideal modulation frequency versus acoustic frequency distribution is

$$P_{\text{ideal}}(\eta, \omega) = \left\{ S\left(\omega - \frac{\eta}{2}\right) S^*\left(\omega + \frac{\eta}{2}\right) \phi_m(\eta) \right\} * \phi_c(\omega). \quad (8)$$

It is important to note that the above condition does not require the signal to be simple cosinusoidal modulation. In principal, any signal

$$s(t) = m(t)c(t), \quad (9)$$

where $m(t)$ is nonnegative and band limited to frequency $\omega < |\omega_m|$ and $c(t)$ has no frequency content below ω_m, can have a modulation frequency versus acoustic frequency distribution in the form of the above ideal modulation frequency versus acoustic frequency distribution. No regions will overlap in frequency and, assuming separate preservation of phase, $s(t)$ will be recoverable from $P_{\text{ideal}}(\eta, \omega)$.

An example of an implicitly convolutional effect of $\phi_c(\omega)$ is the limited frequency resolution that arises from a transform of a finite duration of data, for example, the windowed time analysis used before conventional short-time transforms and filter banks. The multiplicative effect of $\phi_m(\eta)$ is less obvious. Commonly applied time envelope smoothing has, as a frequency counterpart, lowpass behavior in $\phi_m(\eta)$. Other efficient approaches can arise from decimation already present in critically sampled filterbanks. Note that the nonzero terms centered around $\eta = \pm 2\omega_c$, which are well above the typical passband of $\phi_m(\eta)$, are less troublesome than the typically much lower-frequency quadratic distortion term(s) at $\eta = \pm 2\omega_m$. Thus, broad frequency ranges in modulation will be potentially subject to these quadratic distortion term(s).

4. EXAMPLES OF APPLICATIONS

4.1. An adjunct to the spectrogram

Figure 2 shows a joint acoustic/modulation frequency transform as applied to two simultaneous speakers. Speaker 1 is saying "two" in English while Speaker 2 is saying "dos" in Spanish. This data is from (http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html.)

As expected, the spectrogram on the left side of Figure 2 offers little to discriminate the two simultaneous speakers. However, the right side of Figure 2 shows isolated regions of acoustic information associated with the fundamental pitch and its first and aliased harmonics of each of the two speakers. These pitch label locations in acoustic frequency also separately segment each of the two speaker's resonance information.

4.2. Applications to audio coding

When applied to signals, such as speech or audio, that are effectively stationary over relatively long periods, a modulation dimension projects most of the signal energy onto a few low modulation frequency coefficients. Moreover, mammalian

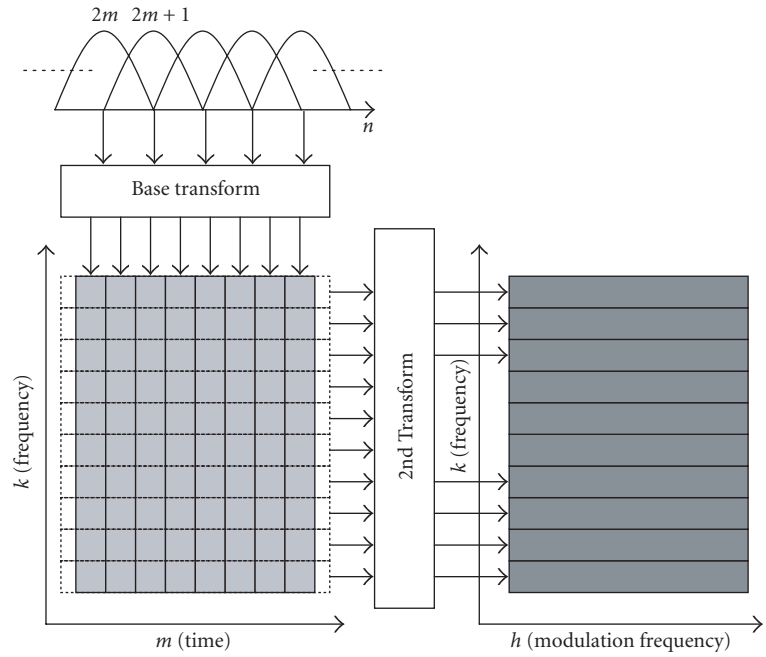

Figure 3: Simplified structure of the two-dimensional transform used in the new approach to audio coding [36]. The left matrix represents a magnitude of a perfect reconstruction and critically sampled filterbank. The detection operation previously mentioned was inherent in the magnitude operation. Signal phase was encoded separately and did not undergo the second transform.

auditory physiology studies have shown that physiological importance of modulation effects decreases with modulation frequency [19, 20]. While these traits suggest an approach for ranking the importance of transmitted coefficients and coding at very low data rates, this past work has provided an energetic yet not invertible transform. We have recently devised a transform, which after modification to a lower bit rate is invertible back to a high-fidelity signal [36].

This result confirms that there are modulation frequency transforms that are indeed invertible after quantization. Moreover, the energy compaction provided by the transform allows significant added compression. Our design, which is schematized in Figure 3, allows for essentially CD-quality music coding at 32 kilobits/second/channel and provides a progressive encoding which naturally and easily scales to bit rate changes.

Simple subjective tests were performed [36] and, as seen in Figure 4, the results suggested that the proposed algorithm performed significantly better quality coding at 32 kilobits/second/channel than MPEG-1 layer 3 (MP3) coding at 56 kilobits/second/channel. Furthermore, the proposed algorithm was shown to be inherently progressively scalable, lending itself well to the widely increasing range of applications where bandwidth cannot be known prior to coding.

This result represents only a first attempt for using joint acoustic and modulation frequency concepts in analysis/synthesis. The result does not just confirm the expected tolerable quantization of perfect reconstruction, it also demonstrates how quickly this approach has been able to come close to state-of-the-art performance in audio coding. We thus expect continuing improvements in quality for these bit rates.

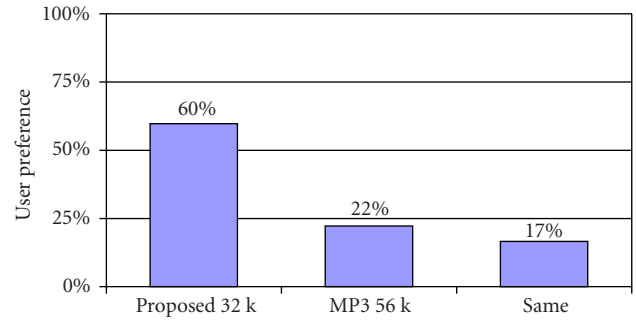

Figure 4: Listener preferences between the proposed algorithm at 32 kbit/s and MP3 at 56 kbit/s. Note that while the proposed algorithm was used at almost 1/2 the bit rate of MP3, a large majority of listeners preferred it to the higher bit rate MP3.

5. OTHER APPLICATIONS

While our previous examples have focused on single-channel talker separation and audio compression applications, there are many other potential opportunities for joint frequency analysis. These future applications are divided into analysis-only and analysis/synthesis systems. Some key examples are given below.

5.1. Analysis/synthesis systems

Both audio and images compression, as suggested by our preliminary results, could gain efficiency and flexibility (e.g., fine-grained scalability) by compaction in modulation frequency dimensions. Furthermore, as justified earlier, human perception is less sensitive or insensitive to high modulation frequencies. Also, as demonstrated by previous researches [12, 13, 14, 15, 16], psychophysical models indicated limited resolution and significant masking in modulation frequency. Joint acoustic and modulation frequency also provides a framework for investigations into human perception. For example, it cannot necessarily be assumed that psychoacoustic masking in the two dimensions of joint frequency can be accurately predicted from only the product of one-dimensional functions of standard acoustic frequency masking and modulation frequency masking. Thus, a framework for two-dimensional masking studies could provide a new viewpoint.

Analysis/synthesis approaches could also be used to generate other novel realistic sounds and images for psychoacoustics, hearing and vision science, audiometry and optometry, and entertainment. For example, a music modification system, based upon this form of analysis/synthesis, could generalize the standard notion of an acoustic frequency equalizer to a two-dimensional joint frequency equalizer. This joint equalizer could potentially accentuate, attenuate, or remove musical instruments within ranges of joint frequency. Also, polyphonic combinations of instruments with acoustic frequency overlap but different rhythmic structure could be separated. This concept of polyphonic separation has interesting generalizations to images and video signals. Thus, we expect that success in joint frequency could help bridge representations of natural sounds and images to the structural modeling proposed in MPEG-7 standards.

5.2. Analysis

Joint frequency features can also be used for novel representation of signals and images. For an acoustic example, the work of Kingsbury [37] suggests that modulation spectrogram features could be useful for correcting multiplicative reverberant distortions in speech. Other acoustic applications include speech recognition in noisy environments, music and speech enhancement, audiology and optometry testing, and audio fingerprinting [38]. Some image and vision applications of joint frequency analysis include segmentation and classification under nonuniform lighting conditions. The segmentation and general conversion of naturally produced material to structural models might also be facilitated, opening up new possible areas for standards like MPEG-7 and MPEG-21.

6. SUMMARY AND CONCLUSIONS

Previous work in modulation spectra justifies the importance of this concept in auditory physiology, psychoacoustics, speech perception, and signal analysis and synthesis. There is a remaining need for analysis/synthesis tools which provide a transform to and from a modulation spectral representation. Modifications of this representation can thus affect a novel and general form of filtering which goes well beyond conventional linear time-invariant filters.

An analysis/synthesis approach ideally requires invertibility and perfect reconstruction. A joint acoustic/modulation frequency model was outlined along with a set of minimum attributes for invertibility. This model was validated via high-quality and efficient performance in audio coding. It also shows potential for single-channel multiple-talker speech separation. Other applications were suggested for acoustic and multimedia signals.

A key future extension of this theory would involve a *combined* (or a two-dimensional) spectrotemporal modulation transform. This is intuitively analogous to combining the modulation spectrum with the well-known cepstral representation widely used in speech recognition. A more versatile approach might utilize a two-dimensional wavelet transform of the time-frequency representation [19, 20]. While it is critical that this representation be invertible in coding applications, this restriction may not be necessary in many other applications such as for the recognition of speech where robustness in noise or utility for segregation and streaming of competing speech signals might be more important.

REFERENCES

[1] B. Vercoe, W. Gardner, and E. Scheirer, "Structured audio: creation, transmission, and rendering of parametric sound representations," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 922–940, 1998.

[2] H. Dudley, "Remaking speech," *Journal of Acoustical Society of America*, vol. 11, no. 2, pp. 169–177, 1939.

[3] L. Zadeh, "Frequency analysis of variable networks," *Proc. IRE*, vol. 3A-8, no. 3, pp. 291–299, 1950.

[4] T. Kailath, "Channel characterization: time-variant dispersive channels," in *Lectures on Communication System Theory*, E. Baghdady, Ed., pp. 95–123, McGraw-Hill, New York, NY, USA, 1961.

[5] A. Møller, "Unit responses of the rat cochlear nucleus to tones of rapidly varying frequency and amplitude," *Acta Physiol. Scan.*, vol. 81, pp. 540–556, 1971.

[6] N. Suga, "Analysis of information-bearing elements in complex sounds by auditory neurons of bats," *Audiology*, vol. 11, pp. 58–72, 1972.

[7] C. Schreiner and J. Urbas, "Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF)," *Hearing Research*, vol. 21, pp. 227–241, 1986.

[8] G. Langner, "Periodicity coding in the auditory system," *Hearing Research*, vol. 60, no. 2, pp. 115–142, 1992.

[9] N. Kowalski, D. Depireux, and S. Shamma, "Analysis of dynamic spectra in ferret primary auditory cortex: I. Characteristics of single unit responses to moving ripple spectra," *Journal of Neurophysiology*, vol. 76, no. 5, pp. 3503–3523, 1996.

[10] D. Depireux, J. Simon, D. Klein, and S. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[11] S. Shamma, "Auditory cortical representation of complex acoustic spectra as inferred from the ripple analysis method,"

Network: Computation in Neural Systems, vol. 7, no. 3, pp. 439–476, 1996.
[12] H. Schulze and G. Langner, "Periodicity coding in the primary auditory cortex of the Mongolian gerbil (*Meriones unguiculatus*): two different coding strategies for pitch and rhythm?," *Journal of Comparative Physiology A*, vol. 181, no. 6, pp. 651–663, 1997.
[13] G. Langner, M. Sams, P. Heil, and H. Schulze, "Frequency and periodicity are represented in orthogonal maps in the human auditory cortex: evidence from magnetoencephalography," *Journal of Comparative Physiology A*, vol. 181, no. 6, pp. 665–676, 1997.
[14] N. Viemeister, "Temporal factors in audition: A systems analysis approach," in *Psychophysics and Physiology of Hearing*, E. Evans and J. Wilson, Eds., pp. 419–427, Academic Press, London, UK, 1977.
[15] T. Houtgast, "Frequency selectivity in amplitude-modulation detection," *Journal of Acoustical Society of America*, vol. 85, pp. 1676–1680, 1989.
[16] S. Bacon and D. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," *Journal of Acoustical Society of America*, vol. 85, pp. 2575–2580, 1989.
[17] S. Sheft and W. Yost, "Temporal integration in amplitude modulation detection," *Journal of Acoustical Society of America*, vol. 88, pp. 796–805, 1990.
[18] S. Ewert and T. Dau, "Characterizing frequency selectivity for envelope fluctuations," *Journal of Acoustical Society of America*, vol. 108, pp. 1181–1196, 2000.
[19] T. Chi, Y. Gao, M. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *Journal of Acoustical Society of America*, vol. 106, no. 5, pp. 2719–2732, 1999.
[20] M. Elhilali, T. Chi, and S. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," to appear in Speech Communication.
[21] T. Houtgast and H. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, vol. 28, pp. 66–73, 1973.
[22] "Special issue of image evaluation by means of optical transfer functions," *Optica Acta*, vol. 18, pp. 81–163, 1971.
[23] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of Acoustical Society of America*, vol. 77, no. 3, pp. 1069–1077, 1985.
[24] R. Drullman, J. Festen, and R. Plomp, "Effect of temporal envelope smearing on speech reception," *Journal of Acoustical Society of America*, vol. 95, no. 2, pp. 1053–1064, 1994.
[25] S. Greenberg and B. E. D. Kingsbury, "The modulation spectrogram: In pursuit of an invariant representation of speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, pp. 1647–1650, Munich, Germany, April 1997.
[26] W. Gardner, *Statistical Spectral Analysis: A Non-Probabilistic Theory*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1987.
[27] W. Gardner, "Exploitation of spectral redundancy in cyclostationary signals," *IEEE Signal Processing Magazine*, vol. 8, pp. 14–36, 1991.
[28] C. Nikias and M. Raghuveer, "Bispectrum estimation: A digital signal processing framework," *Proceedings of the IEEE*, vol. 75, no. 7, pp. 869–891, 1987.
[29] E. Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.
[30] S. Haykin and D. Thomson, "Signal detection in a nonstationary environment reformulated as an adaptive pattern classification problem," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2325–2344, 1998.
[31] A. Bovik, P. Maragos, and T. Quatieri, "AM-FM energy detection and separation in noise using multiband energy operators," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3254–3265, 1993.
[32] W. Torres and T. Quatieri, "Estimation of modulation based on FM-to-AM transduction: Two-sinusoid case," *IEEE Trans. Signal Processing*, vol. 47, no. 11, pp. 3084–3097, 1999.
[33] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech, and Audio Processing*, vol. 8, no. 3, pp. 240–254, 2000.
[34] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.
[35] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *Journal of the Audio Engineering Society*, vol. 21, no. 7, pp. 526–534, 1973.
[36] M. Vinton and L. Atlas, "A scalable and progressive audio codec," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '01)*, Salt Lake City, Utah, USA, May 2001.
[37] B. Kingsbury, *Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments*, Ph.D. thesis, University of California, Berkeley, Calif, USA, 1998.
[38] S. Sukittanon and L. Atlas, "Modulation frequency features for audio fingerprinting," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '02)*, pp. 1773–1776, Orlando, Fla, USA, May 2002.

Les Atlas received a Ph.D. degree in electrical engineering from Stanford University in 1984. He joined the University of Washington in 1984, where he is a Professor of electrical engineering. His research is in digital signal processing, with specializations in acoustic analysis, time-frequency representations, and signal recognition and coding. His research is supported by DARPA, the Office of Naval Research, the Army Research Lab, and the Washington Research Foundation. Dr. Atlas received a National Science Foundation Presidential Young Investigator Award and a Fulbright Research Award. He was General Chair of the 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing, Chair of the IEEE Signal Processing Society Technical Committee on Theory and Methods, and a member of the Signal Processing Society's Board of Governors.

Shihab A. Shamma obtained his Ph.D. degree in electrical engineering from Stanford University in 1980. He joined the Department of Electrical Engineering at the University of Maryland in 1984, where his research has dealt with issues in computational neuroscience and the development of microsensor systems for experimental research and neural prostheses. Primary focus has been on uncovering the computational principles underlying the processing and recognition of complex sounds (speech and music) in the auditory system, and the relationship between auditory and visual processing. Other researches include the development of photolithographic microelectrode arrays for recording and stimulation of neural signals, VLSI implementations of auditory processing algorithms, and development of algorithms for the detection, classification, and analysis of neural activity from multiple simultaneous sources.