Modulation Spectral Filtering of Speech

Les Atlas

Department of Electrical Engineering
University of Washington, Seattle, WA, USA
atlas@ee.washington.edu

Abstract

Recent auditory physiological evidence points to a modulation frequency dimension in the auditory cortex. This dimension exists jointly with the tonotopic acoustic frequency dimension. Thus, audition can be considered as a relatively slowly-varying two-dimensional representation, the "modulation spectrum," where the first dimension is the well-known acoustic frequency and the second dimension is modulation frequency. We have recently developed a fully invertible analysis/synthesis approach for this modulation spectral transform. A general application of this approach is removal or modification of different modulation frequencies in audio or speech signals, which, for example, causes major changes in perceived dynamic character. A specific application of this modification is single-channel multiple-talker separation.

1. Introduction

Zadeh first proposed that a separate dimension of modulation frequency could supplant the standard concept of system function frequency analysis [1]. 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 only defined, but was not analyzed. Kailath followed up nine years later [2] with the first analysis of this joint system function. More recently, Gardner (e.g. [3,4]) greatly extended the concept of joint 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. Cyclostationary analysis does not provide an analysis/synthesis framework.

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 [5], which arose out of optical transfer functions (e.g. [6]), has also been successfully applied to the measurement of speech transmission quality (Speech Transmission Index, STI) [7]. More direct studies on speech perception [8] demonstrated that the most important perceptual information lies at modulation frequencies below 16 Hertz. More recently, Greenberg and Kingsbury [9] 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-Hertz range of modulation frequencies (uniformly for all acoustic frequencies), and emphasized the 4-Hertz range of modulations.

In the remainder of this paper, we shall illustrate how an analysis/synthesis theory of modulation frequencies can be formulated and show some examples of its use in the manipulation of speech signals.

1.1. A Modulation Spectral Model

For further progress to be made in the understanding and applications of modulation spectra, a well-defined foundation for the concept of modulation frequency analysis/synthesis 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 of an amplitude modulated fixed-frequency carrier

$$s_1(t) = m(t)\cos\omega_c t$$

where the modulating signal $m(t)$ is non-negative 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

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

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 cosinsoidal modulator

$$s(t) = (1+\cos\omega_m t)\cos\omega_c t$$

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 non-negative 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 :

$$S(\omega-\eta/2)S^*(\omega+\eta/2)$$

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

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

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 [4].

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 $P_{ideal}(\eta,\omega)$ for $s(t)$ should have significant energy density only at only six impulsive points in the (η,ω) plane

$$P_{ideal}(\eta,\omega)=\delta(0,\omega_c)+\delta(\omega_m,\omega_c)+\delta(-\omega_m,\omega_c)$$
$$+\delta(0,-\omega_c)+\delta(\omega_m,-\omega_c)+\delta(-\omega_m,-\omega_c)$$

Where $\delta(\eta,\omega)$ is the standard Dirac delta function. For the above ideal two-dimensional representation, the desired terms are 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 cosinsoidal modulator $s(t)$ is

$$S(\omega)=F\{s(t)\}=F\{(1+\cos\omega_m t)\cos\omega_c t\}$$
$$=\frac{1}{2}\{\delta(\omega-\omega_c)+\delta(\omega+\omega_c)\}$$
$$+\frac{1}{4}\{\delta(\omega+\omega_c+\omega_m)+\delta(\omega+\omega_c-\omega_m)\}$$
$$+\frac{1}{4}\{\delta(\omega+\omega_c+\omega_m)+\delta(\omega+\omega_c-\omega_m)\}$$

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 of the ideal and actual two-dimensional representation is shown in Figure 1.

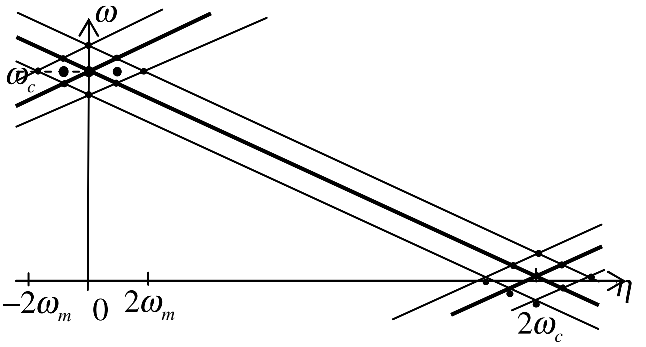

Figure 1: Two-dimensional representation of cosinusoidal amplitude modulation.

The solid lines of Figure 1 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, $P_{ideal}(\eta,\omega)$, of modulation frequency versus acoustic frequency.

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

$$P_{ideal}(\eta,\omega)=\left\{S(\omega-\eta/2)S^*(\omega+\eta/2)\phi_m(\eta)\right\}*\phi_c(\omega)$$

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

In principal, for any signal

$$s(t)=m(t)c(t)$$

where $m(t)$ is non-negative 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.

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, e.g. 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, low pass 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 pass band 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).

2. Talker Separation

The problem of talker separation is also called "co-channel speech interference." One past approach to the co-channel speech interference problem is blind signal separation (BSS) that approximately recovers unknown signals or "sources" from their observed mixtures [10]. Typically, these mixtures are acquired by a number of sensors, where each sensor receives a different combination of the source signals.

However, a different and perhaps complementary approach can utilize modulation spectra. Figure 2 shows a joint acoustic/modulation frequency transform as applied to two simultaneous speakers. Talker A is saying "two" in English while talker B is saying "dos" in Spanish. This data is from [11].

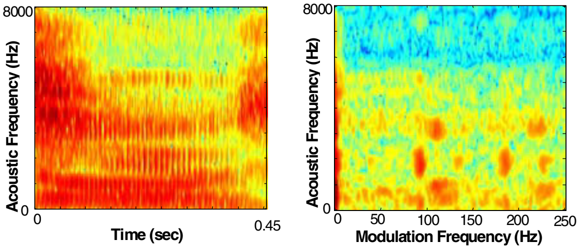

Figure 2: Spectrogram (left) and joint acoustic/modulation frequency representation (right) of the central 450 milliseconds of "two" (speaker 1) and "dos" (speaker 2) spoken simultaneously by two speakers. The *y*-axis of both representations is standard acoustic frequency. The *x*-axis of the right panel representation is modulation frequency, with an assumption of Fourier basis decomposition.

The right side of Figure 2 shows distinct and isolated regions of acoustic information associated with the fundamental pitch and its first and aliased harmonics of the two distinct speakers. These pitch energy locations are both in modulation frequency (at the respective speaker's pitch rate and its harmonics) and in acoustic frequency (quite notably, at the respective speaker's resonant frequencies).

Since it is possible to arbitrarily modify and invert this transform [12] the clear separability of the regions of sonorant sounds from different simultaneous talkers can be used to design talker-separation mask filters.

3. Speech Modification

Psychoacoustic evidence [13] indicates that perceptual modulation filter shapes approximately imitate a constant-Q bandwidth. Also, results in speech recognition studies also point to advantages of a modulation wavelet transform in automatic speech recognition [14]. For audio coding purposes, we recently have proposed the use of an octave-band non-uniform modulation transform to mimic the spacing of modulation filter sub-bands of the human auditory system [15]. This non-uniform second stage transform is depicted in the structure of Figure 3. The use of a non-uniform second transform leads to a resulting representation which generates three dimensions: acoustic frequency, modulation scale, and modulation time-shift. This approach preserves phase, which has previously been found to be important [16], of the modulation spectrum, via a potentially decimated (the amount of decimation depends upon the scale) modulation time-shift waveform.

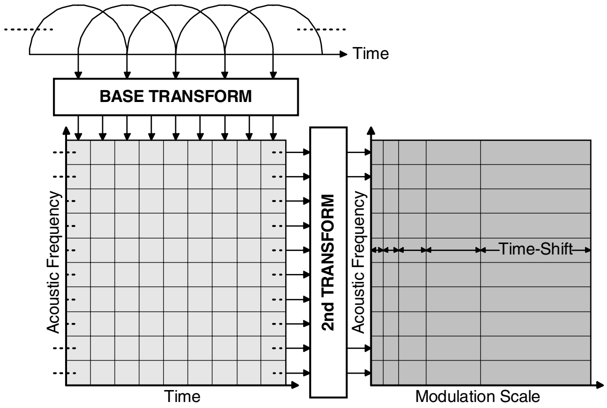

Figure 3: Structure of the proposed non-uniform modulation transform resulting in three dimensions: acoustic frequency, modulation scale, and modulation time-shift.

The structure of Figure 3 employs a time domain aliasing cancellation (TDAC) filter bank [17] as the first or "base" transform. The TDAC filter bank possesses the desirable properties of producing complex outputs and providing 50% overlapping in time while maintaining critical sampling. The TDAC filter bank uses alternating modified-discrete cosine transforms (MDCT) and modified-discrete sine transforms (MDST). Two neighboring MDCT and MDST transform blocks are temporally aligned and combined to form a single complex transform block. The MDCT coefficients are taken as the real part and the MDST coefficients are taken as the imaginary part of the complex coefficients. A magnitude detection operation is performed on the complex transform blocks, which are then arranged into a time-frequency distribution.

A hierarchical lapped transform (HLT) [18] is used for the second stage transform of Figure 3. The HLT is a multi-resolution transform which maintains good time localization for high frequency components and good frequency resolution for low frequency components. The HLT is similar in structure to a quadrature-mirror filter bank (QMF) and the wavelet filter bank. The HLT is applied on the magnitude values in each acoustic frequency sub-band of the time-frequency representation as shown in Figure 3. Note that the HLT second transform is not performed on the phase values from the base transform.

Figure 4 shows an example of the above modulation spectrum applied to the spoken letter "k." Note that the plosive burst is clearly seen at the finer scales of modulation.

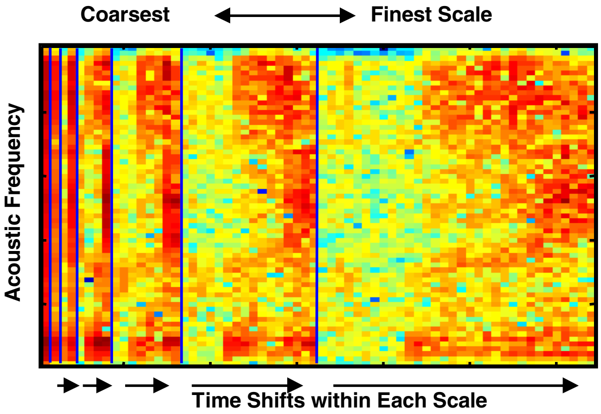

Figure 4: The non-uniform modulation transform.

Filtering in modulation spectra can be effected by simply masking chosen coefficients which result after the above non-uniform modulation transform. If the mask is chosen to only have dependency upon modulation scale (and/or time shift) For example, Figure 5 shows spectrograms of the spoken sound "zero" before and after filtering out all modulation scales except for the finest scale. (All time shifts for this finest modulation scale were retained.) Quite notably, the filtered sound is highly intelligible yet virtually all sensation of pitch is removed. This observation raises questions about the kind of basis functions which are appropriate for the modulation dimension. It also suggests a highly redundant representation across modulation scales.

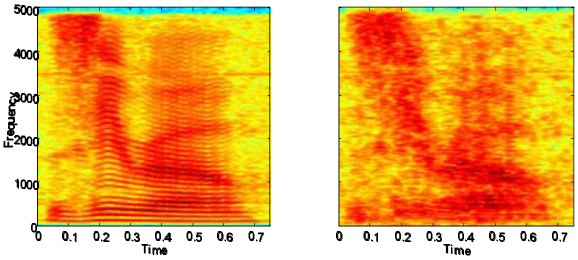

Figure 5: Spectrograms before (left) and after (right) all but the finest modulation scale were removed.

4. Conclusions

We have described the underlying conditions for an invertible modulation spectral transform. This transform, with linear spacing in modulation frequency, was applied to single-channel talker separation. We then extend the transform to a non-uniform (or scale) spacing in modulation and demonstrate its potential to represent and modify speech in new ways.

5. Acknowledgements

This research was supported by the Office of Naval Research. We'd like to thank Drs. Shihab Shamma and Steven Greenberg for their helpful discussions and Somsak Sukittanon, Jeff Thompson, Qin Li, and Chad Heinemann for their helpful discussions and figures.

6. References

[1] Zadeh, L. "Frequency analysis of variable networks," *Proc. IRE* 38(3), 291-299, 1950.

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

[3] Gardner, W. **Statistical Spectral Analysis: A Non-probabilistic Theory.** Prentice-Hall, Englewood Cliffs, NJ, 1987.

[4] Gardner, W. "Exploitation of spectral redundancy in cyclostationary signals." *IEEE Signal Processing Magazine*, 14-36, April 1991.

[5] Houtgast, T. & Steeneken, H., "The modulation transfer function in room acoustics as a predictor of speech intelligibility." *Acustica* 28, 66-73, 1973.

[6] *Optica Acta* 18, Special Issue of Image Evaluation by Means of Optical Transfer Functions, 1971.

[7] Houtgast, T. & Steeneken, H., "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria." *J. Acoust. Soc. Am.* 77, 1069-1077, 1994.

[8] Drullman, R., Festen, J. & Plomp, R., "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.* 95, 1053-1064, 1994.

[9] Greenberg, S. & Kingsbury, B. E. D., "The modulation spectrogram: In Pursuit of an invariant representation of speech," *Proc. IEEE ICASSP*, Munich, Germany, 1647-1650, May 1997.

[10] Lee, T., Bell, A., and Lambert, R., "Blind separation of delayed and convolved sources," *Advances in Neural Information Processing Systems*, Vol. 9, MIT Press, 1997, pp. 758-764.

[11] www.cnl.salk.edu/~tewon/Blind/blind_audio.html

[12] Vinton, M. & Atlas, L., "A scalable and progressive audio codec," *Proc. IEEE ICASSP*, Salt Lake City, May 2001.

[13] Houtgast, T., "Frequency selectivity in amplitude-modulation detection," *J. Acoust. Soc. Am.* 85, 1676-1680, 1989.

[14] K. Okada, T. Arai, N. Kanedera, Y. Momomura, and Y. Murahara, "Using the modulation wavelet transform for feature extraction in automatic speech recognition," *Proc. ICSLP*, Vol. 1, pp. 337-340, 2000.

[15] Thompson, J. & Atlas, L., "A Non-Uniform Modulation Transform For Audio Coding With Increased Time Resolution," *Proc. IEEE ICASSP*, Hong Kong, China, April 2003.

[16] S. Greenberg and T. Arai, "The Relation Between Speech Intelligibility and the Complex Modulation Spectrum," *Proc. Eurospeech 2001*, Scandanavia, pp. 473-476.

[17] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust., Speech, and Signal Processing*, vol. 34, pp. 1153-1161, 1986.

[18] H. Malvar, **Signal Processing With Lapped Transforms**, Boston, Artech House, 1992.