

A HIGH RESOLUTION AUDITORY-INSPIRED METHOD FOR TIME-VARYING SPECTRAL ANALYSIS

Hilmi R. Dajani

Willy Wong

Hans Kunov

Inst. of Biomaterials and Biomedical Eng., and Dept. of Electrical and Computer Eng.

University of Toronto

Rosebrugh Building, 4 Taddle Creek Road

Toronto, Ontario CANADA M5S 3G9

h.dajani@utoronto.ca willy@eecg.utoronto.ca h.kunov@utoronto.ca

ABSTRACT

Pitch discrimination experiments have demonstrated that human listeners can detect very small frequency changes in stimuli of short duration. Inspired by this ability, an algorithm for high resolution time-varying spectral analysis is proposed. Mathematical analysis, with various types of synthetic modulated signals, demonstrates that the proposed method correctly demodulates these signals. The resulting spectrogram-like display, referred to as a 'Fine Structure Spectrogram', shows the fine structure of the modulations in higher detail than is possible with conventional spectrograms. With recorded speech samples, the fine structure spectrogram detects small frequency and amplitude modulations in the formants of speech. It also appears to identify additional components in speech that are not detected by other methods.

1. INTRODUCTION

The spectral content of natural signals, such as speech and most biological signals, usually varies with time. In the most commonly used method of time-varying spectral analysis (the spectrogram or Short Time Fourier Transform), there is a reciprocal relationship between time and frequency resolution, usually expressed as $\Delta t \times \Delta f \geq 1/4\pi$. Wideband analysis gives good time resolution but poor frequency resolution, while narrowband analysis gives good frequency resolution but poor time resolution. This tradeoff is related to the Heisenberg uncertainty principle of quantum mechanics, and is sometimes cited as a fundamental limitation of time-varying spectral analysis [1]. Recently, much effort has gone into developing algorithms that provide improved joint time-frequency resolution. The most popular of these so-called nonstationary time-frequency representations are bilinear distributions – also known as Cohen's class. A problem, however, is that many of the bilinear distributions suffer from artifacts and regions of negative spectral energy which have no obvious physical meaning [2]. Human listeners are said to regularly beat the uncertainty limit by as much as an order of magnitude [3]. This has been most clearly demonstrated in pitch discrimination experiments involving two short tones. The duration of the tones multiplied by the frequency limen required for their discrimination appears to violate the lowest limit allowed by the uncertainty principle (i.e. $1/4\pi$). Therefore, auditory processing may inspire the development of a method with very high resolution in the time-frequency plane.

2. METHOD

The proposed algorithm tracks the local peaks in the outputs of a bank of many overlapping stages of Filter/Detectors (F/D's), that each consist of a bandpass filter, followed by a rectifier and smoother. Hundreds – or even thousands – of stages may be used, inspired by the thousands of tuning curves of the afferent auditory fibers. A high time resolution may be achieved by having fairly wideband individual filters that respond rapidly to changes in the signal, while a high frequency resolution may be achieved if the filters are separated by a very fine amount. A plot of the local peaks at the outputs of the overlapping stages is referred to as a Fine Structure Spectrogram (FSS).

3. VALIDATION

Mathematical analysis, with several synthetic AM and FM signals, shows that the FSS correctly demodulates these signals. For example, if the input signal is sinusoid that is jointly modulated in amplitude and frequency:

$$x(t) = A_0(1+m_{AM}\cos w_m t)\cos(w_c t + m_{FM}\sin w_m t) \quad (1)$$

Where w_c and w_m are the carrier and modulator frequencies in radians, m_{AM} is the AM modulation index, m_{FM} is the FM modulation index, and A_0 is the amplitude of the carrier. Then if $m \ll \pi/2$ (narrowband condition), a square law device is used for the rectifier, and the cutoff frequency of the smoothing filter is less than $2w_m$, then signal at the output of one F/D is:

$$v(t) = (A_1^2 + A_2^2 + A_3^2 + A_4^2 + A_5^2)/2 + (A_1A_3 + A_2A_3 + A_1A_4 + A_2A_5)\sin w_m t \quad (2)$$

Where A_1, A_2, A_3, A_4, A_5 are the amplitudes at the output of the bandpass filter of components at $w_c - w_m, w_c + w_m, w_c, w_c - 2w_m, w_c + 2w_m$ respectively. If the transfer function of the bandpass filter is chosen to be Gaussian (which can be approximated using an FIR filter), then equation (2) shows that F/D's whose center frequencies are close to the carrier frequency alternate in having the largest output during the modulation cycle. The amplitude of these outputs also varies during the cycle. This allows the peak detector to simultaneously demodulate both the frequency and amplitude components of the modulation.

Fig. 1c shows the resulting FSS for an input signal of this type. The "wiggles" in the curve follow the narrowband FM, while the gray scale of the curve varies with every modulation cycle as it follows the AM. For comparison, Figures 1a and 1b show the Wideband and Narrowband spectrograms of the same signal. The additional detail that is provided by the FSS is clear.

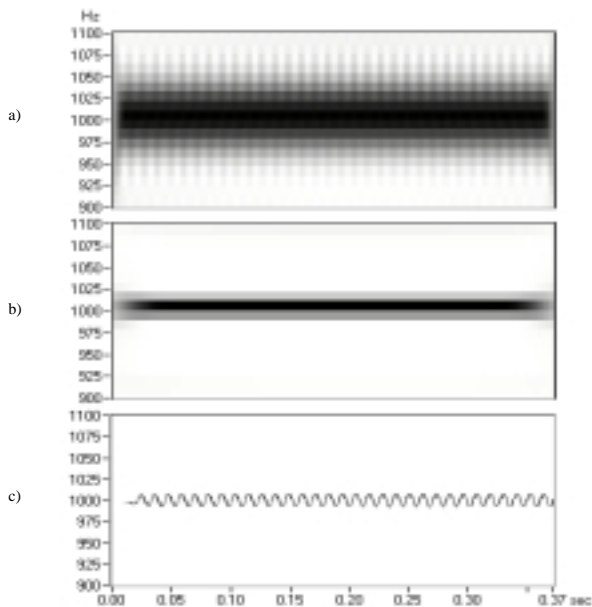


Figure 1. Comparison of methods with a sinusoid simultaneously modulated in amplitude and frequency. Carrier frequency is 1000 Hz, modulation frequency is 90 Hz, peak frequency deviation is 10 Hz, and AM modulation index is 0.25. a) Wideband spectrogram. b) Narrowband spectrogram c) Fine Structure Spectrogram

4. ANALYSIS OF SPEECH

Most current methods of speech processing are based on the source-filter model of speech production [4]. This model assumes that the parameters of the filter formed by the vocal tract vary slowly, and are thus termed quasistationary. Quasistationary analysis misses the fine structure of the modulations of speech. This is illustrated in Fig. 2a that shows the Narrowband spectrogram of the sentence "...an oily rag like that" taken from the TIMIT database.

Recently, there has been increasing recognition of the complexity of the speech modulations. In one study, a model-based approach has been developed to extract details of the frequency and amplitude modulations of speech [4]. This is illustrated in Fig. 2b for the frequency modulations.

The FSS of the same sentence is shown in Fig. 2c. In terms of overall features, the FSS shows details in the modulations that are similar to those seen in Fig. 2b (but with some of the higher frequency formants "hidden" due to the range of the gray scale). More significantly, perhaps, the FSS shows some components not found in Figures 2a or 2b. We believe that these additional components are genuine, and may be of significance for the processing of speech.

5. CONCLUSIONS

The auditory-inspired Fine Structure Spectrogram produces a display with a high joint time-frequency resolution. It detects small frequency and amplitude modulations in the formants of speech, which are not resolved using the conventional spectrogram. It also appears to detect additional components in speech that are usually missed by other methods.

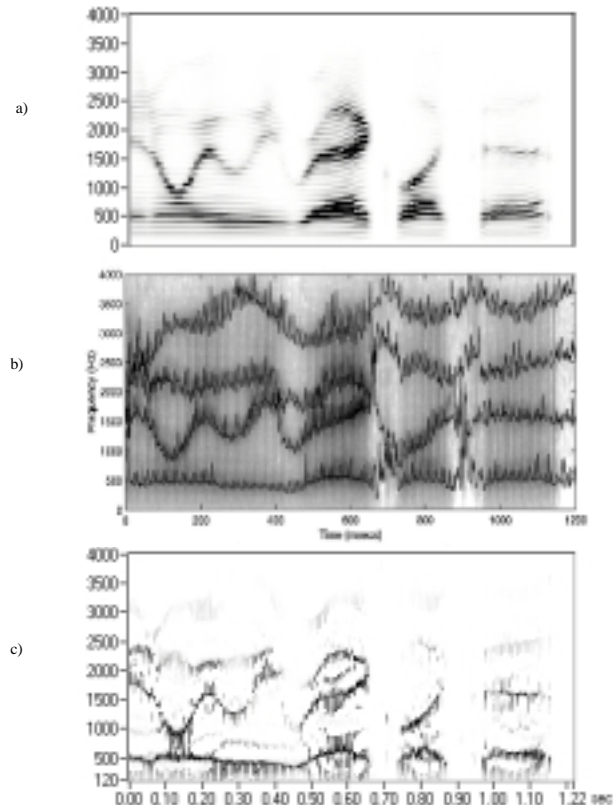


Figure 2. Comparison of methods with a sample of connected speech. a) Narrowband spectrogram. b) Model-based method to decompose speech into modulated components (reproduced from [4]). c) Fine Structure Spectrogram.

6. REFERENCES

- [1] O. Dossing, "Uncertainty in time/frequency domain representations," *Sound Vib.*, vol. 32, no. 1, pp. 14-24, 1998.
- [2] J. W. Pitton, K. Wang, and B. H. Juang, "Time-frequency analysis and auditory modeling for automatic recognition of speech," *Proc. IEEE*, vol. 84, no. 9, pp. 1199-1215, 1996.
- [3] W. M. Hartmann, *Signals, Sound, and Sensation*. American Institute of Physics, Woodbury, N.Y., 1997.
- [4] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 240-254, 2000.