TIME-FREQUENCY AND MULTIPLE-RESOLUTION REPRESENTATIONS IN AUDITORY MODELING

Unto K. Laine, Matti Karjalainen and *Toomas Altosaar*
Helsinki University of Technology, Acoustics Laboratory
Otakaari 5A, 02150 Espoo, Finland

Introduction

The human auditory system is known to utilize different temporal and frequency resolutions in different contexts and analysis phases. In this paper we discuss some aspects of using time-frequency representations and multiple resolutions in auditory modeling from an information and signal theoretic point of view. The first question is how to allocate resolution optimally between frequency and time. For this purpose a new method called the *FAM transform* is described. The other question is how to utilize multiple parallel and redundant resolutions to avoid some problems that are faced when using single resolution approaches.

FAM functions, a new class of orthogonal functions

In search for better resolution allocation in time-frequency signal analysis such methods as the *Wigner distribution* and the *wavelet transform* have been used [1]. A new method to compute nonuniform resolution spectra or spectrograms that model auditory features is to use the so called FAM-method [2, 3]. The method is based on a new class of orthogonal functions called FAM functions. These are **F**requency and **A**mplitude **M**odulated sine and cosine functions (or complex exponentials) in which the amplitude modulation a(x) is related to the frequency modulation g(x) by the rule:

$$a(x) = \sqrt{g'(x)} \qquad (1)$$

This rule guarantees the orthogonality of the FAM functions for any well-behaving generative function g(x) which defines the type of the present orthogonal FAM function set (the present member of the FAM class). The FAM functions are defined as:

$$\mathrm{FAMsin}(n, g(x)) = \sqrt{g'(x)}\,\sin(n\,g(x)) \qquad (2)$$
$$\mathrm{FAMcos}(n, g(x)) = \sqrt{g'(x)}\,\cos(n\,g(x)), \quad n = 0, 1, 2, \ldots$$

which can be simply expressed by complex exponential FAM functions

$$\mathrm{FAMexp}(n, g(x)) = \sqrt{g'(x)}\,e^{j\,n\,g(x)}, \quad n = 0, 1, 2, \ldots \qquad (3)$$

It is easy to see that the choice g(x)=x leads to the classical Fourier analysis, and the choice g(x)=arccos(x) to Chebyshev polynomials and g(x)=arctan(x) to frequency transformed Laguerre functions, which all are special cases of the FAM class. By using FAM functions we are able to formulate a nonuniform frequency resolution transform: the FAM transform.

Auditory spectra and spectrograms by FAM transform

In DFT analysis in its simplest form the analysis window in the time domain is a chain of unit impulses which form a rectangular window. Each impulse picks up one signal sample at **n** to be transformed to an exponential of order **n** in the frequency domain. In the auditory FAM transform we want to describe the signal with a nonuniform resolution in the frequency scale, i.e. a uniform resolution in the auditory Bark (or mel) scale. This can be done by choosing the generative function g(x) to correspond to the Hz-to-Bark warping function: g(x) = Bark(x), [4]. When the corresponding FAM exponentials are used in the frequency domain, they can construct spectra related to the auditory critical band resolution. In the time domain the functions corresponding to the FAM exponentials are no longer impulse-like, but somewhat phase distorted versions of them (i.e. inverse Fourier transformed FAM exponentials). In principle the auditory transform can be computed by convolving the signal with the time domain pulses (called Bark pulses) and constructing the complex spectra as a sum of the correlation-weighted FAM exponentials. Presently this is not computationally effective because there is no fast algorithm available. Another method is to construct a complex orthogonal filterbank from the Bark pulses [3].

Our first experiments with the complex orthogonal auditory filterbank have shown that the temporal resolution of the magnitude spectrogram is of high quality: even spectral variations inside a single pitch period of speech can easily be monitored. The phase information gained by the complex Bark bank (each channel 1 Bark in width) also contains detailed temporal information of the signal especially at the lower frequencies. At the high end the phase seems to become more noisy, however. These results are in good harmony with the studies made with phase vocoders [5]. In the near future our aim is to develop fast algorihms for the auditory transform and to use the filterbank in studies on speech perception.

Multiple resolutions in auditory modeling

In addition or contrary to the optimal allocation of resolution in time and frequency there is another general principle in auditory modeling such that multiple parallel and redundant resolutions are used. In fact, a continuous scale of resolutions may be added to the representations. *Scale-Space Filtering* [6] and related methods are good examples thereof. In some sense this resolution dimension or scale corresponds to the frequency scale in time-frequency representations of time signals. The motivation of using computationally expensive parallel and redundant representations is that, because any single representation (and resolution) may lead to erroneous decisions e.g. in speech analysis and recognition, a set of redundant representations is more reliable by deferring the actual decision making.

The sensitivity of multiple-resolution representations on the kernel function was studied by us experimentally [7]. It was shown that in practice we can deviate relatively far from the mathematically ideal Gaussian kernel. Even very crude approximations showed to preserve the topological gross structure of the scale-space. The multiresolution methods can be applied to any signal or feature function, e.g. as postprocessing of auditory spectrograms for speech recognition.

Another general question in multiresolution is the tradeoff between continuous and discrete representations. The scale-space approach hints for approximating continuous functions by some special points that are typically extrema (maxima, minima) or zero-crossings. We have called them *events*: points of special interest or prominence, including related information. Such discretization and abstraction from continuous representations is very efficient computationally and allows for symbolic (e.g. rule-based) methods for further analysis and processing.

A new and interesting question in auditory modeling is how to combine the time-frequency representations, the multiple-resolution methods, continuity and discreteness, and the neural network methods. Although the main objective of auditory modeling is to reflect the physiological and psychoacoustical reality of the human auditory system as well as to find applications for the computational models, there is also need for deeper understanding from the point of view of general signal and information processing principles.

References

[1] Cohen, L., "Time-Frequency Distributions - a Review", Proc. IEEE, 7 7, 941-981, 1989.

[2] Laine, U. K. and Altosaar T., "An Orthogonal Set of Frequency and Amplitude Modulated (FAM) Functions for Variable Resolution Signal Analysis", *Proc. of IEEE ICASSP-90*, Albuquerque, New Mexico, 1615-1618.

[3] Laine U. K., "A New High Resolution Time-Bark Analysis Method for Speech", (to be published in) *Proc. of the XII ICPhS 91*, Aix-en-Provence, France, 1991.

[4] Zwicker, E. and Feldkeller, R., *Das Ohr als Nachrichtenempfänger*, Stuttgart, Hirzel Verlag, 1967.

[5] Flanagan, J. L. and Golden, R. M., "Phase Vocoder", The Bell System Technical Journal, Nov. 1966, 1493-1509.

[6] Witkin, A.P., "Scale-Space Filtering: A New Approach to Multi-Scale Description", *Proc. of IEEE ICASSP-84*, San Diego, 1984.

[7] Altosaar, T. and Karjalainen M., "Event-Based Multiple-Resolution Analysis of Speech Signals", *Proc. of IEEE ICASSP-88*, New York.