ON THE USE OF CEPSTRAL REPRESENTATION IN SYNTHESIS FROM REDUCED PERFORMANCE INFORMATION

Mark Rudolph
Visiting Researcher
AT&T Bell Labs
600 Mountain Avenue
Murray Hill, New Jersey 07974

Abstract:

The fundamental problem in designing a system for the synthesis of sound from information is to choose an information representation which is capable of capturing the time-varying characteristic features of any arbitrary sound event in a standard form such that actions on distinct subsets of parameters have independent and predictable effects which are of musical interest. One such representation is based on the cepstrum, the inverse discrete Fourier transform of the log spectral magnitude, a non-linear transform which satisfies a generalized principle of superposition. This paper gives a theoretical introduction to the cepstrum and shows how its information representation is very appropriate for use in music synthesis. Finally, the implementation of a system for synthesis from reduced performance information based on cepstral representation is described.

Introduction:

The method of synthesis from information is based on short time transformations between the time and information domains. Sound events are segmented into sequences of time frames on the order of 10-40ms. in length, and each frame is transformed into a corresponding vector of information. In this manner the characteristic acoustic features of an arbitrary sound event can be encoded in a sequence of "instantaneous" information states of the signal over the duration of the event, and these state-vectors can be used to resynthesize the event without distortion, or to synthesize a new event after modifications to the vector sequence. The use of any method of synthesis from information is motivated by the desire to control the generation of sound from a more fine grained level of description than that allowed either by sound sampling or by MIDI. Modifications to the sequences of parameters can precisely control the timbre, articulation and expressiveness of the event, and acoustic features from two or more events can be combined or fused at the level of sound information description rather than at the level of note description. In this way known performance features can be used combinatorially to produce a new event, or the features of a new event can be correlated to known features of an analyzed performance. In addition, the rate of information update in a synthesis can be used to control the rate of time passage relative to a known performance event. Thus the time flow of a performance may be modified with no change in timbre or pitch. Finally, it is usually the case that the time domain representation of an event contains much redundancy, and thus the representation of the event in the information domain can be achieved with a significant reduction in the number of bits needed to reproduce the event. This data reduction allows easier communication in a synthesis system, and makes practical the concept of a cumulative "knowledge base" of reduced performance information consisting of stored information vector sequences corresponding to various chosen performances and performance features. This knowledge base may be used as a source of basic materials for a synthesis or for the real-time processing of a performance input.

There are two major classes of analysis/synthesis systems in frequent use in speech coding which have also been used for music analysis/synthesis. The first is the class of orthogonal transform systems such as the "phase vocoder", first reported by Flanagan and Gold (1966). Additional information on its specific application to music can be found in Dolson (1986) and Moorer (1978). The major weakness in this information representation is the inseparability of the smooth spectrum associated with source resonances or "formants" from the rapidly varying spectral information associated with pitch or fine timbre. In addition there is no direct means of estimating fundamental pitch. Finally, the form of the information, being Fourier magnitudes and phases, is unwieldy with regard to modification with predictable timbral result. The second class of systems is based on linear prediction, and these vary mainly in their choice of representation for the excitation function of the all-pole filter. For an introduction to linear prediction see Makhoul (1975), and for information on its application to music see Moorer (1979). The information representation used in this class of systems succeeds in separating the smooth spectral information captured in the all-pole filter, from pitch and fine timbre information, but the latter is embedded in a second time domain sequence called the "residual", the sequence of errors in the prediction of the event sequence by a weighted sum of previous inputs, which is difficult to analyze and data reduce.

Cepstral Representation:

The cepstrum was first reported by Bogert, Healy and Tukey (1963). It is defined as the inverse Fourier transform of the log of the spectral magnitude function. What this means is that the log of the spectral magnitude function is looked upon as if it were a time domain function, and analyzed in order to determine what constituent "frequencies" are most strongly present. In order to avoid ambiguity these "frequencies" which have as units cycles/Hz., ie. seconds, are termed "quefrencies". A related concept developed by Oppenheim Schafer and Stockham (1968), is the "complex cepstrum", which is the inverse DFT of the complex log of the spectrum. The complex cepstrum requires calculation of the complex log which requires a continuous unwrapped phase, a difficult theoretical problem. However, since the cepstrum is a special case of the complex cepstrum for which all phases are zero, if phase information is determined and stored at some stage of the analysis prior to the inverse DFT, then no information is lost by restricting our attention to the cepstrum. In order to demonstrate the utility of a cepstral representation for synthesis, let us first assume that the signal associated with a given performance event consists of the convolution of a smooth spectral component **x** associated with spectral resonance, with a second component **v** consisting of weighted harmonic series with possible inharmonic components as well. This signal model is precisely appropriate for many classes of instruments and sound sources such as woodwinds, brass, strings and voice, and is quite useful in organizing the spectral information of any sound source into one component associated with broad local maxima (**x**) and another associated with harmonic peaks and other characteristic spectral peaks (**v**). Although the cepstrum is derived from a non-linear transform due to the use of the log, it can be seen from the following step-by-step derivation of the cepstrum of the model that a generalized principle of superposition is obeyed in which the transform of the convolution of two signals is equal to the sum of the two transformed signals. Thus the transform is a homomorphism of vector spaces. Let **F** denote the DFT, **T** the "cepstral transform", and let **X** and **V** and **x'** and **v'** denote the magnitude and cepstra respectively of **x** and **v**.

$$
\begin{aligned}
T(x*v) &= F^{-1}\log\{F(x) \cdot F(v)\} \\
&= F^{-1}\log\{X.V\} \quad \text{(plus phase)} \\
&= F^{-1}\{\log X + \log V\} \\
&= F^{-1}\log X + F^{-1}\log V \\
&= x' + v'
\end{aligned}
$$

Furthermore, the log magnitude of **x** varies very slowly in the spectral domain, whereas the log magnitude of **v** varies rapidly and periodically with frequency. Thus, the contribution to the cepstrum due to **x** is found at the low index quefrencies, whereas the contributions due to **v** are found at multiples of the fundamental. It should now be clear that the cepstrum separates the two convolved components into two distinct regions in the cepstral domain. The region near the origin concentrates the information concerning the smooth spectrum **X**, and modifications to this region can be made to correspond to variations in the shape and size of a resonating cavity, either real or idealized, through which passes the signal **v** containing pitch and timbre information. Modifications to the positions of the high quefrency peaks affect the perceived pitch of the resynthesized sound, and modifications to their relative amplitudes changes the perceived harmonics and timbre. Thus two distinct cepstral information vector sequences are available for modification and combination. By summing the low quefrency sequence of one signal with the high quefrency sequence of a second signal a "cross-synthesis" effect is achieved. Furthermore, two sets of cepstral values may be interpolated with time-varying weights in order to achieve fusions of distinct features. In addition to its use in re-synthesis, the cepstrum can be used directly for deriving information about fundamental frequency and harmonic structure. If the signal is periodic there is likely to be a strong peak at the period of the fundamental frequency, and perhaps other peaks at periods of the strongest harmonics. Also, the amplitude of the peak is a good measure of the degree of periodicity of the signal, and absence of a peak indicates that the signal is non-harmonic. The cepstral peaks are better indicators of fundamental period than, for example, the auto-correlation function since the peaks in the auto-correlation function are distorted due to the convolution of the harmonic series with the smooth spectrum. Finally, the utility of the cepstrum for pitch detection is not diminished by the absence of a strong fundamental, and hence the cepstrum can be used to detect pitch in cases where the first few harmonics may dominate in strength over the fundamental such as for telephone speech. Thus the cepstrum provides a direct parametric representation of the smooth spectrum contour associated with "formants", and also a direct parametric representation of the fundamental frequency and harmonic series, as well as an indication of the degree of periodicity of the signal. For further information on cepstral pitch detection see Noll (1964) . For further information on deconvolution of a composite signal by use of the cepstrum, see Oppenheim (1968).

Computational Considerations:

Analysis/synthesis using cepstral representation is a calculation intensive procedure. However there are a few techniques which can be used to lessen the computation. First, the deriving of the low quefrency cepstral coefficients can be achieved by first solving for the LPC filter coefficients and then deriving the cepstral coefficients recursively from these (see Rabiner and Schafer 1978). This procedure requires, for example, the calculation of a small number of auto-correlation coefficients, solving a set of normal equations by using Durbin-Levinson recursion, and then solving for the cepstrum recursively as follows:

$$c_0 = 1/N \sum_{k=0}^{N-1} \log M_k; \qquad c_n = a_n + \sum_{k=1}^{n-1} k/n \, c_k \, a_{n-k} \qquad 1<=n$$

Also, in order to eliminate the squares, square root, division and inverse tangent calculations associated with transforming the complex Fourier trig values into magnitude and phase, it is possible to even extend the current frame. Then the double length frame is real and symmetric so can be solved with a quarter length DFT, and the even extension yields purely real Fourier values whose absolute values are the spectral magnitudes, and whose phases are zero or pi depending on the sign of the respective real part. An added feature is that in this case the phases can be stored as single bits, and any modifications to the phases such as those required in time scale dilation require only one multiplication instead of the normal one per phase value.

Implementation on a Multi-Processor:

A system for analysis/synthesis from reduced performance information based on a cepstral representation is in the process of being implemented on a multi-processor system consisting of a high-fidelity over-sampling A/D and D/A, AES/EBU interface to RDAT, and a four DSP board with shared and local memory. The system uses the AT&T DSP32C floating point device which operates at 50MHz. and is highly pipelined for very efficient signal processing. The multi-processor communications consist of shared memory usage routines using a hardware lock mechanism for "monitor" accesses to test and modify a software system of buffer states and software-locked "semaphores". In addition, some communications are achieved by serial DMA and by interrupts, and in future a message passing system will be implemented using DMA on a parallel bus connecting all processors on multiple four-DSP boards. At present the parallel DMA has only been simulated at very slow speed on the host computer which is also used for code development and loading, and for system messages and error monitoring. Software tools for the system include assembler, an extensive simulator for flexible code diagnosis and debug, C-compiler, C-library and assembler applications library. In their present form the cepstral analysis/synthesis routines do real-time signal analysis and modification, but in future an entire system for synthesis from a knowledge base of reduced performance will be implemented. Most of the software has been written in DSP32C assembler for maximum speed, but some has been written in C and compiled for use on the DSP32C, and some has been written in C++ to run on the host only. The cepstral routines have two forms. One is an array processing form (SIMD), and the other is a partitioned pipeline processing form (MIMD). The first form is less communication bound, but has greater overall latency, whereas the second has less system delay and greater use of concurrent computation, but requires more shared memory communications. At present sequential applications can be partitioned by hand and made to run on the multi-processor simply by choosing an appropriate communications topology, and including in the source code at specific points a few header files which enable the use of shared memory communications without direct programmer intervention other than to invoke one of three system functions for shared state update, read and/or write. All four processors are required for the cepstral analysis/synthesis system to run in real-time which requires on the order of thirty five ms. processing time per each 10.67 ms. frame advance.

References:

Flanagan, J.L.,& R.M. Golden. 1966."Phase Vocoder." Bell System Tech. J. 45:1493 -1509.

Moorer, J.A. 1978. "The Use of the Phase Vocoder in Computer Music Applications." J. of the Audio Eng. Soc. 26(1/2): 42 - 45.

Dolson, M. 1986. "The Phase Vocoder: A Tutorial." Comp. Mus. J. 10(4):14 - 27.

Makhoul, J. 1975. "Linear Prediction: A Tutorial Review." Proc. of the IEEE. 63(4):561-580.

Moorer, J.A. 1979. "The Use of Linear Prediction of Speech in Computer Music Applications." J. of the Audio Eng. Soc. 27(3):134 - 140.

Bogert, B. P.,& M. J. R. Healy,& J.W.Tukey. 1963."The Quefrency Alanysis of Time Series for Echoes: Cepstrum, Pseudo-autocovariance, Cross-Cepstrum, and Saphe Cracking." in Time Series Analysis. M. Rosenblatt Ed. New York. Wiley. Ch. 15. pp 209- 243.

Oppenheim, A.V.,& R.W. Schafer,& T.G. Stockham. 1968. "Non-Linear Filtering of Multiplied and Convolved Signals." Proc. of the IEEE. 56(8):1264-1291.

Noll, A.M. 1964. "Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection." J. of the Audio Eng. Soc. Vol 36(2):296-302.

Oppenheim, A.V. 1969. "Speech Analysis-Synthesis System Based on Homomorphic Filtering." J. of the Audio Eng. Soc. 45(2):458-465.

Rabiner, L.R.,& R.W. Schafer. 1978.Digital Processing of Speech Signals New Jersey Prentice-Hall. p442.

AT&T

AT&T Bell Laboratories

subject: **Mark Rudolf**

date:

from: **T. G. Grau**
WH 46414
14J-219 x3644

N. S. Jayant

I would like to express my appreciation for the contributions that a member of your organization, Mark Rudolf, made in putting together an audio application that was used by us in an important customer demonstration on Tuesday, February 20, 1990.

Mark went out of his way and worked long hours to port his application to the Aspen DSP3. In addition, by his accompanying us to Washington, DC, he brought a great deal of subject matter expertise to our presentation and was very well received by our customers.

Again, thank you for the support, and we look forward to continuing interaction with your organization.

T. G. Grau

WH-46414-EBM-cdw

Copy to
J. L. Flanagan
B. T. Fought
M. Rudolf ✓