*Presented at ICASSP-97, Munich, vol. 2 pp. 1307-1310*

# THE WEFT: A REPRESENTATION FOR PERIODIC SOUNDS

*Dan Ellis*

International Computer Science Institute, Berkeley CA 94704
dpwe@icsi.berkeley.edu

## ABSTRACT

For the problem of separating sound mixtures, periodicity is a powerful cue used by both human listeners and automatic systems. Short-term autocorrelation of subband envelopes, as in the correlogram, accounts for much perceptual data. We present a discrete representation of common-period sounds, derived from the correlogram, for use in computational auditory scene analysis: The *weft* describes a sound in terms of a time-varying periodicity and a smoothed spectral envelope of the energy exhibiting that period. Wefts improve on several aspects of previous approaches by providing, without additional grouping, a single, invertible element for each detected signal , and also a provisional solution to detecting and dissociating energy of different periodicities in a single frequency channel (unlike systems which allocate whole frequency channels to one source). We define the weft, describe the analysis procedure we have devised, and illustrate its capacity to separate periodic sounds from other signals.

## 1. INTRODUCTION

Human listeners are highly adept at interpreting the complex soundfield reaching their ears as the superposition of the output of several independent sources; this subtle process of auditory organization has attracted significant investigation both in psychoacoustics [1] and more recently via computer modeling [2,3,4]. Although human auditory scene analysis is the result of sophisticated operations at many levels, modeling work tends to focus on low-level signal processing related to the neurophysiology of the auditory periphery. Given the information observed to be present in the early auditory system, to what extent can the raw signals arriving from different sources be extracted without relying on more circumstantial constraints?

Pitch is perhaps the most important feature for sound organization in listeners. The pseudoperiodicity of many real-world sound sources (notably the human vocal apparatus) gives an extremely powerful basis for the accurate separation of the characteristics of simultaneous sounds, as shown by double-vowel perception experiments [5]. The perceptual segregation of voices with differing pitches is so immediate and so strong as to suggest the operation of special-purpose dedicated neural circuitry. Considerable effort has been expended in order to construct equivalent computer models, with success only for a limited range of phenomena.

The known broad-band nature of cochlea filters in the upper spectrum makes it unlikely that listeners are segregating periodic sounds on the basis of a narrowband analysis into separate harmonics. Consequently, computer models such as the correlogram [6,7] use envelope autocorrelation to reveal the periodicity of the energy in each peripheral frequency channel; this frequency-channel-versus-lag-period display can serve as the basis for isolating the spectra of mixture components.

The correlogram-type model of Meddis & Hewitt [8] predicted listeners' abilities to exploit pitch difference in identifying vowel mixtures very accurately. However, the inputs to the model were essentially 'static spectra' or snapshots, and the model assumed the presence of exactly two periodic sources. The full-blown Computational Auditory Scene Analysis (CASA) system of Brown [4] picked out a single periodic target from a noisy background by segmenting the time-frequency plane into locally-coherent regions he called 'auditory objects'. A search procedure then fused these objects on the basis of consistent local periodicity contours to produce pitch-based entities. Both of these systems were constrained to allocate the whole of a frequency channel to just one source at each instant.

## 2. THE WEFT REPRESENTATION

We developed a new representation, the weft, as part of a more general CASA system [9] which includes other sound elements for noisy and transient sounds. ("Weft" is the Anglo-Saxon word for the parallel fibers in woven cloth, giving the idea of a connected set of threads [10]). Wefts are also based on the correlogram, but have the following advantages over the objects of Brown [4]:

- A single weft represents all the energy associated with a given period, corresponding to several of Brown's objects; the costly (and physiologically implausible) stage of subsequent grouping is avoided.
- The analysis allows a single peripheral channel simultaneously to contribute energy to wefts at several different periods, rather than coarsely allocating entire channels to sources.
- Whereas Brown's resynthesis involved masked refiltering of the original mixture, wefts hold sufficient information to resynthesize the sounds they represent without additional data. This is valuable, for instance, to facilitate restoration or interpolation of corrupted regions.

A weft is defined by two components, illustrated in figure 1. The period-track stores the underlying period exhibited by the weft as a function of time. The smooth-spectrum records the amount of energy reflecting that period in each time-frequency cell of the analysis. Since the system is based around a cochlea filterbank [11], frequency resolution is rather broad and highly overlapped.

Weft resynthesis is straightforward, according to a traditional source-filter formulation: An impulse-train is generated from the period track. This is broken into frequency channels by the filterbank, and the gain of each channel is modulated before recombination to produce a signal whose energy contour follows the smooth-spectrum. The spread and overlap of the filterbank is accommodated by a non-negative least-squares inversion.

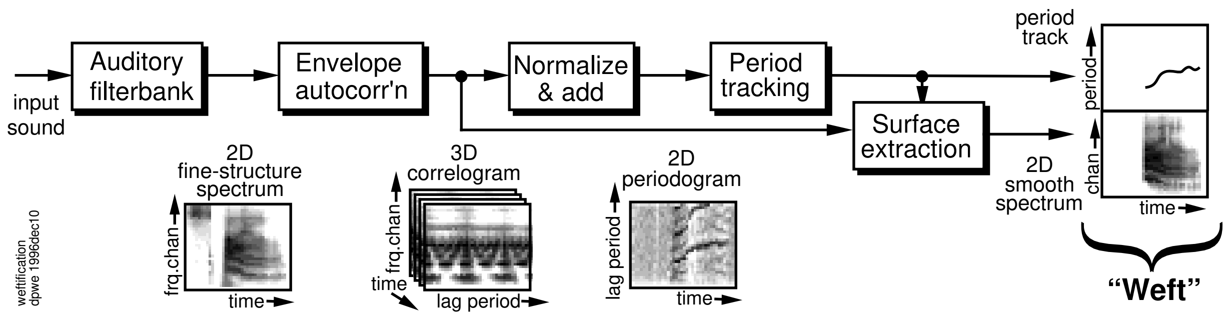


**Figure 1:** Calculation of a weft: The three-dimensional correlogram volume (the short-time envelope autocorrelation of filterbank output channels) is normalized and collapsed across channel to form a summary periodogram. Peaks in this result are extracted to produce the period-track portion of the weft. The correlogram volume is then sampled at that period to extract the smoothed spectrum of the weft for every channel at each time slice, taking care to factor-out interference from other energy in each channel.

## 3. WEFT ANALYSIS

The process by which wefts are extracted from a mixture of periodic sounds is illustrated in figure 1. Successive blocks are discussed below; for greater detail, see [9].

**Correlogram**

The input sound is passed first through a linear filterbank approximating the frequency analysis of the cochlea [11]. Each channel is then half-wave rectified and smoothed with a 1 ms window to remove fine time structure. Short-time autocorrelation is calculated by smoothing the product of an envelope signal with delayed versions of itself. These delays are the samples on the lag period axis, and are spaced logarithmically in time to approximate the resolution of human pitch perception. The correlogram is thus the three-dimensional volume formed of the time-varying short-time autocorrelation of each peripheral frequency channel at each lag period.

**Period tracking**

Broad-band periodically-modulated energy suitable for representation as a weft is detected by looking for features in a summary of the modulations present in the whole spectrum. The *periodogram* is formed by normalizing each autocorrelation function by its overall energy (zero-lag value), then summing across frequency channel to get a two-dimensional function of lag period against time. A peak exceeding a fixed threshold in this domain triggers the creation of a weft element, indicating a modulation period strongly represented in the correlogram, and determining a point on the period track for the new weft.

The search for peaks in the periodogram starts at the shortest period, and detected peaks are iteratively subtracted from the remaining autocorrelation function at multiples of their period to eliminate the subharmonic aliases always produced by autocorrelation. (This makes a shaky but serviceable assumption of linearity which we lack the space to consider). As well as preventing 'ghosts', the subtraction is intended to reduce or eliminate the interference between subharmonics of a short-period modulation and longer modulation periods in the same channel.

Period tracking is guided by the existing wefts, so that a modulation period detected in a previous time frame will be subjected to a lower threshold for continuation, and may even be 'restored' across short gaps in the periodogram.

**Surface extraction**

The result of period tracking is a set of 'active' wefts for the current time step with known periods but unknown spectra. The next stage extracts these spectra by looking back into the three-dimensional correlogram volume at the indicated lag period. For each frequency channel, an autocorrelation peak at that lag indicates unobscured evidence of the modulation period. Subharmonic aliases are also removed from the correlogram at the multiples of tracked periods, so the remaining peak value is approximately proportional to the energy (average squared amplitude) of the periodic modulation

The presence of incoherent noise in a channel is additive in the power domain, and hence has a nonlinear effect on the envelope and the autocorrelation function. A simplified analysis [9] of the situation where the excitation consists of impulses at a period much longer than the peripheral filter impulse response predicts the autocorrelation peak value, $P$, is given by:

$$P = d \cdot M^2 + (1-d) \cdot N^2 \qquad (1)$$

where $N$ is the envelope 'noise floor', $M$ is the envelope level during idealized rectangular excitation bursts (assumed equal to $\sqrt{(N^2+L^2)}$, where $L$ is the noise-free peak), and $d$ is the 'duty cycle', i.e. the ratio between the impulse-response duration and the modulation period of the excitation bursts (see fig. 2). In this case, the average autocorrelation $A$ across all lags beyond the intrinsic channel correlation (i.e. decorrelating the noise) is:

$$A = (d \cdot M + (1-d) \cdot N)^2 \qquad (2)$$

Both $P$ and $A$ may be robustly measured from an autocorrelation function, allowing eqns. (1) and (2) to be solved for $N$ and $M$, thereby giving $L$, the peak amplitude of the periodic component of the envelope. This procedure relies on a value for $d$, which depends on both the modulation period and the filter channel characteristics, and is not defined when the impulse response is long compared to the modulation. However, in the noiseless case (i.e. $N = 0$),

$$d = A/P \qquad (3)$$

In practice, this analysis gives adequate results over most of the

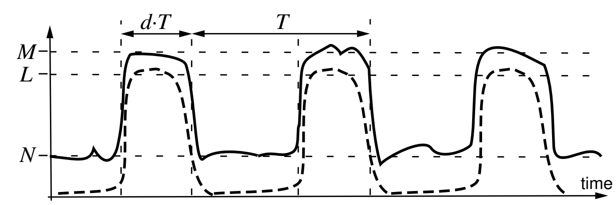


**Figure 2:** The idealized envelope waveform with modulation at period $T$ and impulse response length $d \cdot T$, sketched both with (solid) and without (dashed) additional aperiodic noise.

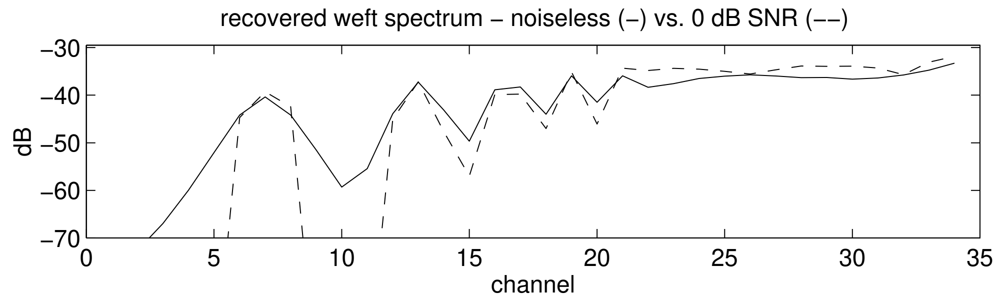


**Figure 3:** Comparison between a slice of a weft 'smooth spectrum' recovered for a periodic signal analyzed in isolation, and the weft produced for the same signal when analyzed in the presence of 0 dB SNR spectrally-matched noise.

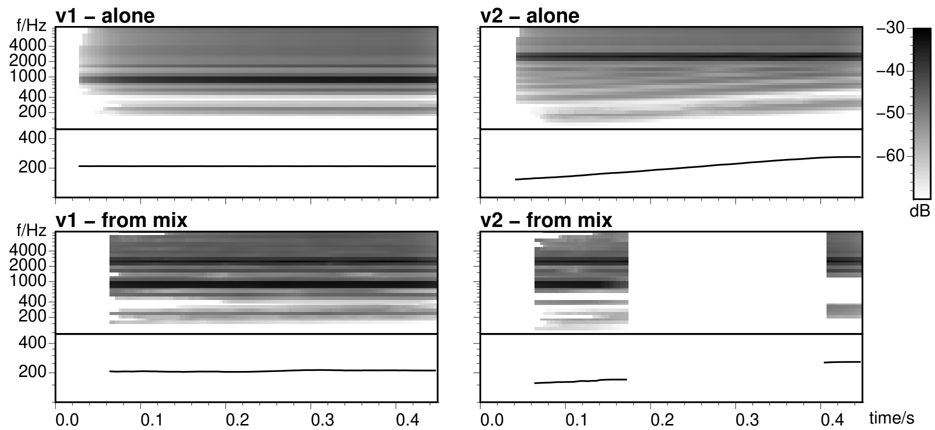


**Figure 4:** Wefts for a pair of vowel-like signals analyzed separately and mixed together. Each panel illustrates a weft with two frames: the smooth spectrum as a function of time (x-axis) and frequency (y-axis), and the pitch-track, also drawn on a log-frequency y-axis.

range of interest. It is applied using a table of values for $d$ measured according to eqn. (3) for all 40 frequency channels over all 240 quantized values of the lag period. When the measured average-to-peak ratio for some autocorrelation line is larger than the appropriate $d$, additive noise is presumed present and factored-out from the weft energy by these equations.

If a periodic signal is being masked by a stronger signal in the same frequency band, its autocorrelation peak may be hidden. To accommodate this situation, even non-peak autocorrelation values are recorded as an 'upper-bound' on the energy of the modulation in that channel. This can constrain estimates of the masked envelope based on temporal interpolation or other sources of information.

## 4. RESULTS

Wefts are intended to be used in separating periodic signals from added interference. Questions we might ask include: how well does the weft reject aperiodic noise? How well can wefts separate a combination of two periodic signals? And how do they perform on real signals, such as mixtures of speech?

Figure 3 shows a slice from the smooth spectrum of a weft recovered from a broadband periodic signal. The dotted line shows how the extracted spectrum changes when noise of the same power is added. Some 'bleed-through' is noticeable in the upper channels, but most of the spectrum is nearly unaffected. Around the fundamental (in channel 7), the weft energy estimate falls to zero when noise is added because the periodic modulation in these dips is completely masked.

Figure 4 illustrates the particularly difficult case of two periodic signals colliding in pitch. Signal "v1" (left column) has a period of 200 Hz and a formant-like spectral peak around 1 kHz. "v2" (right column) has a pitch sweeping from 140 to 280 Hz under a spectral prominence at 2.2 kHz. The top two panels show the wefts generated when each signal is analyzed in isolation. The lower two panels are extracted from the *mixture* of the two signals. Most obviously, there is a large gap in the middle of v2 where the analysis has lost the second modulation period—irretrievable in the situation of a pitch collision without exploiting additional context (for instance, by extrapolating the obscured signal until it reappears, for which wefts are well suited). Unfortunately, the spectral envelopes of both wefts extracted from the mixture show evidence of both formants; the factorization of periodic combinations needs additional study.

Figure 5 shows the weft analysis of a mixture of a male and a female voice (taken from [4]). The points to note are:

- The continuously-voiced male utterance gives a single weft. Stops and fricatives interrupt the female speech resulting in a sequence of four wefts.
- Most time-frequency cells have contributed to both signals; factoring-out the interactions within each channel gives each voice a different spectrum.
- The periodogram shows octave-collisions between the voices at several points. The predict-and-remove analysis strategy has successfully disentangled this mixture.

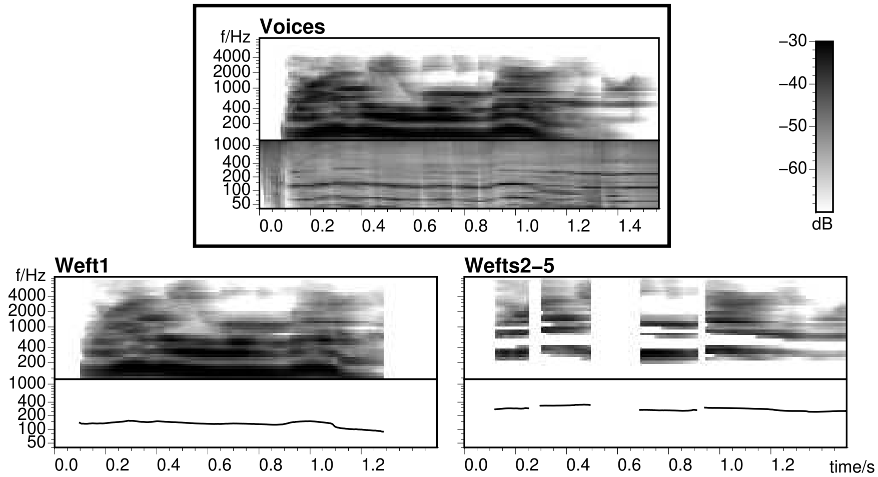


**Figure 5:** Weft analysis of a voice mixture: The top panel shows the cochlea-model spectrogram above the periodogram for the mixture of male and female voices from [4]. (The periodogram is displayed with an upside-down, logarithmic time axis, labeled by frequency). The lower panels show the smooth spectrum and period tracks of the five wefts extracted for this sound. The continuously-voiced male utterance is extracted as a single weft (left). The female voice is recovered in four voiced segments, showing considerable spectral overlap with the male voice (right).

The subjective quality of the separation and resynthesis has been assessed via listening tests. Nine subjects compared the original to the resyntheses and rated the quality on a percentage scale from identical (100%) to unrecognizable (0%). Ratings were 44% for the male voice and 30% for the female voice; Brown's [4] resynthesis of the male voice scored 37% in the same test (his system did not extract a second voice).

These resynthesis examples may be heard on the Web at http://www.icsi.berkeley.edu/~dpwe/research/icassp97/ .

## 5. CONCLUSIONS

As part of a CASA system's vocabulary encompassing both periodic and aperiodic sounds, wefts are a useful low-level representation that plausibly reflects the special-purpose pitch-based sound-organization mechanisms apparently possessed by human listeners. Wefts are derived from a spectral representation approximating the known function of the cochlea, and are intended to reflect the information used in the auditory system. Peculiarities of the envelope-autocorrelation domain require novel mechanisms to compensate for the interaction between signals; to the extent that this resembles the processing in the auditory periphery, its limitations may be less noticeable to human listeners that would otherwise be expected.

Future developments of this work will improve the resynthesis quality through a more accurate inversion of the spreading effect of the filterbank, and improve the estimation of component signal energies in channels reflecting multiple periodicities through a detailed analysis of the autocorrelation of envelopes of multi-periodic mixtures.

## 6. ACKNOWLEDGMENTS

Portions of this work were performed while the author was at the Machine Listening Group of the M.I.T. Media Lab, whose support is gratefully acknowledged. We also thank the European Union, through their basic research grant to project SPRACH (2077).

## 7. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis*, MIT Press, 1990.

[2] M. Weintraub, *A theory and computational model of monaural auditory sound separation*, Ph.D. dissertation, Stanford Univ., 1985.

[3] M. P. Cooke, *Modelling auditory processing and organisation*, Cambridge University Press, 1993.

[4] G. J. Brown, *Computational Auditory Scene Analysis: a representational approach*, Ph.D. thesis, CS dept., Sheffield Univ, 1992.

[5] P. F. Assmann & Q. Summerfield, "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acous. Soc. Am. 88(2), pp. 680-697, 1990.

[6] R. O. Duda, R. F. Lyon & M. Slaney, "Correlograms and the separation of sounds," Proc. IEEE Conf. on Signals, Systems and Computers, Asilomar, 1990.

[7] M. Slaney & R. F. Lyon, "On the importance of time – a temporal representation of sound," in *Visual Representations of Speech Signals*, ed. M. Cooke, S. Beet & M. Crawford, John Wiley, pp. 95-116, 1993.

[8] R. Meddis & M. J. Hewitt, "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acous. Soc. Am. 89(6), pp. 2866-2882, 1991.

[9] D. P. W. Ellis, *Prediction-driven Computational Auditory Scene Analysis*, Ph.D. dissertation, EECS dept., M.I.T., 1996.

[10] D. P. W. Ellis & D. F. Rosenthal, "Mid-level representations for computational auditory scene analysis," workshop on Comp. Aud. Scene Analysis at the Intl. Joint Conf. on Artif. Intel., Montréal, pp. 111-117, 1995.

[11] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Tech. Report #35, Apple Computer Co., 1993.