to be presented at
1995 International Joint Conference on Artificial Intelligence - Workshop on Computational Auditory Scene Analysis
Montreal, Quebec, August 1995

Mid-level representations for Computational Auditory Scene Analysis

Dan Ellis
Perceptual Computing
MIT Media Lab E15-401B
Cambridge MA 02139 U.S.A
dpwe@media.mit.edu

David Rosenthal
The Agency, Interactive Media
33 Union Street, 5th floor
Boston MA 02108 U.S.A.
dfr@media.mit.edu

Abstract

In this paper we consider representations for use in models of the processing that occurs between the eardrum and our conscious experience of sound. We first list "good" properties for such mid-level representations, then present a framework within which to discuss some examples. We compare in detail two popular schemes — sinusoid tracks and correlograms — and propose a new representation, *wefts*, which seeks to combine their advantages.

1 Introduction: Mid-level representations

Mid-level representation is a term usually associated with computer vision, particularly the ideas of David Marr [1982]. It has since become accepted by many in the computer audition community [Bregman, 1990; Cooke, 1991; Brown, 1992] as an concept useful to models of hearing as well. Auditory perception may be viewed as a sequence of representations from "low" to "high," where low-level representations are (roughly) those appropriate to describing the sound reaching the cochlea, and high-level representations are those to which we have cognitive access, such as "Mary telling John to buy bread," or "Bill playing the trombone with the TV in the background."

Between these two levels we presume there is a network of representations which we label "mid-level," about which we have little direct knowledge; we are only beginning to understand the relevant physiology of the brain, and introspection is unlikely to be useful. Our knowledge of these representations arises chiefly from the constraints imposed on them at the lower and higher levels about which we have more information [Adelson, 1994].

Our view is that despite these challenges, we can understand mid-level representations by building computer models and that this is a fertile area for research in computer hearing. Success in such an endeavor will be rewarded by the construction of greatly improved robot perception systems, and a deepening of our understanding of perception in general.[1]

The relative wealth of knowledge about low-level hearing appears to have imposed an inordinately "bottom-up" orientation on mid-level representations that have been proposed. While low-level processes are a useful source of interesting computations, it is important to retain a focus on the high-level constructs that may be useful for (computer or mammalian) hearing, and then to consider how these constructs may be computed.

1.1 Overview

In the next section we attempt to make a list of abstract qualities considered advantageous for a mid-level representation as we have defined it. Section 3 proposes three dimensions distinguishing different representations, and considers various examples from the literature in this light. Specifically, we consider the strengths and weaknesses of sinusoid tracks and correlograms, which leads us to propose a new representation, *wefts*, in section 4. Wefts address certain limitations of other representation, as we illustrate with examples. Finally in section 5 we conclude by restating our view of the role of representation in complete computational auditory scene analysis systems.

2 Properties desirable in auditory mid-level representations

What criteria should perceptual representations meet? The general requirements of a mid-level hearing representation are that it may be computed efficiently from the input, and that it can readily answer the questions asked of it by the higher levels of processing [Winston, 1984]. If we take the latter to include the full range of computer sound understanding applications, we can list the following desirable properties for mid-level representations:

1. **Sound source separation**. Arguably the *sine qua non* of hearing is the ability to organize sounds reliably according to their independent sources of production, roughly analogous to segmentation in vision. Natural sounds do not occur in isolation; they arrive at the ear from several sources as a complex mixture — meaning that the sounds may overlap in time, frequency, and other representational dimensions. To support a high-level description such as "Bill playing the trombone with the TV in the background," a mid-level representation should decompose sound to a granularity at least as fine as the sources of interest — in this case, pieces that can be labeled as TV noise or trombone.

[1] We acknowledge that some researchers reject this symbolic, explicit approach to representation ([Brooks, 1991] etc.); that debate is beyond the scope of this paper.

2. **Invertibility**. We seek representations in which the original sound can be regenerated from its representation, although according to a *perceptual* rather than bit-wise criterion. That is, we want the regenerated sound be perceptually equivalent to the original, but not to have an identical time-domain waveform. [Knight, 1994] has suggested that bitwise regenerability may be viewed as a failure, since it shows that "nothing unimportant was discarded."

   More important is the *separate invertibility of meaningful parts.* By this we mean that the representation allows us to select a meaningful part of the sound (e.g., the trombone without the TV noise in the example above), which can then be used to regenerate sound (in this case, a noise-free trombone sound).

   We acknowledge that the human system does not usually resynthesize the sounds it represents internally, but the capacity for perceptual invertibility is notionally equivalent to a representation that captures *all* the relevant information. In addition, tractable inversion schemes will be important for applications of computational auditory scene analysis such as advanced hearing prostheses.

3. **Component reduction**. The initial sound may be regarded as a vast array of individual energy levels in time-frequency. As it is re-represented in successively refined ways, the number of objects in the representation should reduce, and the meaningfulness of each should increase. Note that this does not imply data compression; some of the mid-level representations discussed below require more bits to represent them than the original sound. What is important is that these representations group the elements of the original sound into a relatively small number of pieces, corresponding to meaningful structure in the original sound, suitable for subsequent processing.

4. **Abstract salience of attributes**. The features made explicit by a representation should approach the perceptual attributes of our desired final result. In the interests of robust, modular development, we should define these features according to intended physical characteristics (onset of a new source) rather than specific algorithmic details (first difference of energy in each frequency band).

5. **Physiological plausibility.** Functional physiological knowledge becomes more and more scarce as we progress from the basilar membrane into the auditory cortex, but it still provides many interesting revelations. Since our goal is to understand and model the auditory system, we would be wise to respect this knowledge and not pursue hypotheses clearly inconsistent with physiology. This principle can, however, be difficult to interpret.

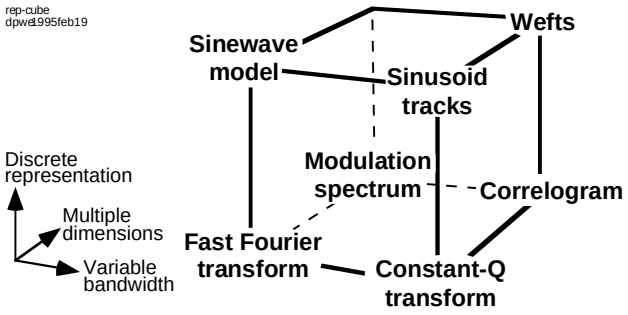

Figure 1: Three dimensions for the properties of sound representations define a cube upon which representations may be placed.

3 An analytic framework for representations

So far we have considered criteria which affect the choice of representation; we turn now to actual candidates for hearing representations, and consider their various merits. We classify hearing representations according to three conceptual "axes":

- the choice between fixed and variable bandwidth of the initial frequency analysis;
- discreteness, corresponding to the degree to which the representation is structured as meaningful chunks;
- the dimensionality of the transform – some representations possess an extra dimension in addition to the usual pair of time and frequency.

By classification along these three dimensions, a given representation may be assigned a position on a cube, as in figure 1. Several candidate hearing representations are now considered in relation to their positions.

In the lower left hand corner is the Fast Fourier Transform (FFT). The FFT is computable by an efficient procedure, and its uses as an analysis tool (for instance, in the spectrogram) are familiar. It cannot, however, be considered an accurate model of the equivalent stage of the auditory system owing, in part, to the fixed bandwidth of its frequency bins. For a given FFT, increased resolution in frequency will come at the expense of resolution in time across all frequency channels, and vice versa. Physiological and psychological measurements of the auditory system indicate that it has the simultaneous ability to resolve frequencies and time variations in a manner not possible with a fixed-bandwidth model.

This limitation is addressed by moving along the "variable bandwidth" axis to the lower right corner in our figure. This position is occupied by the constant-Q transform, implemented as a bank of filters whose bandwidths vary in proportion to their center frequency. This yields an analysis qualitatively similar to that performed by the cochlea. The constant-Q transform retains the simple mathematical formulation of the FFT, though its computational efficiency is not as great [Brown and Puckette, 1992].

Both of these transforms are limited in representational power in that no higher level structure has been imposed on the signal; the signal is not 'chunked' into meaningful parts. In particular, our goal of separating the sound according to its sources requires additional processing.

The representations in the upper half of the cube, along the "discrete representation" axis, convert continuous transforms

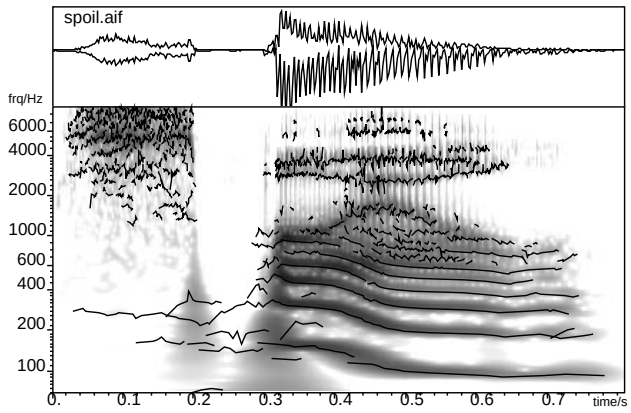

Figure 2: A constant-Q analysis of the word "spoil" with the sinusoid track analysis overlaid. The upper panel shows the time waveform, and the lower panel shows time (left to right) versus log frequency (bottom to top); gray density shows the intensity of the constant-Q transform, and the black lines show the frequency contours of the sinusoidal tracks.

into discrete objects. We consider a particular example derived from the constant-Q transform, consisting of contiguous regions of local energy maxima in time-frequency called *sinusoid tracks* (fig. 2) ([Ellis, 1992; Ellis, 1994], based on the equivalent representation for FFT analysis introduced in [McAulay and Quatieri, 1986]). Tracks support the goals of component reduction and sound source separation, the latter being simplified to the problem of classifying a relatively small number of discrete objects – provided we are able to construct each track to represent energy from only a single source, which can turn out to be very hard.

We can regenerate the original sound from its representation as tracks by using each track to drive a sine-wave oscillator. This technique can be applied to arbitrary subsets of tracks, addressing our criterion of "separate invertibility of meaningful parts." The resulting regenerated sounds possess a high degree of perceptual fidelity to the original, in spite of being poor approximations in a mean-squared error sense. In other words, they succeed in discarding unimportant information.

The depth axis of fig. 1 is labelled "multiple dimensions." Representations behind the front face of the cube involve another dimension revealing extra information in addition to the time and frequency of a spectrogram. The interesting work of Kollmeier and Koch [1994] on the modulation spectrum falls into this category, although their goals are strictly practical (enhancement for hearing aids) and hence they are satisfied to use the FFT for their frequency analalsys. For a more perceptually-motivated approach, we consider here an example known as the *correlogram* [Duda *et al.*, 1990; Slaney and Lyon, 1993], where the third dimension is lag time of short-time autocorrelations applied to the energy envelopes within each frequency band. Different parts of the spectrum whose intensity is modulated at the same rate will have similar profiles in this dimension, facilitating the goal of source separation. Note that the correlogram is normally calculated from the output of a cochlea model filterbank, although in principle it could be applied to any frequency decomposition.

3.1 Correlograms and Tracks

Why, then, is the extra dimension of correlograms useful? This is best seen by an analysis of a representational failing of tracks.

Periodic amplitude modulation is an important cue for sound source separation, since different parts of a sound sharing common modulation are most likely to have arisen from the same source. We would hope, then, that a sound representation would make such periodicity explicit. Unfortunately, the track representation may display periodicity in two separate ways. Consider a harmonic sound processed by a constant-Q filterbank (fig. 2): At the lower frequencies, the resolution is sufficient to separate the harmonics, which are analyzed to horizontal tracks with the periodicity encoded in their frequency contours. At the higher frequencies, the bandwidth of the filters is broader, and several adjacent harmonics will pass through the same filter. Beating between these harmonics will cause an amplitude-modulated filter output, resulting in the vertical bands in the upper portion of fig. 2. Here, the signal's periodicity is primarily reflected in the *magnitude* variation of the tracks involved, not the frequency. This inconsistent encoding of periodicity in the track representation means that subsequent processing must employ special strategies to group tracks that are related by common period [Cooke, 1991; Ellis, 1994]. (A fixed bandwidth analysis avoids this problem, but at an unacceptable cost in terms of overall time-frequency resolution).

For the same signal, the correlogram effectively tags all of the tracks related to the same amplitude modulation period with a common value. If a particular band-pass filter output has a regular period of energy modulation, the autocorrelation of that channel will have an intensity peak at the lag matching that period. This scheme can be used to group all of the components — resolved and unresolved — resulting from a periodic sound, since they will all share this peak at the fundamental period.[2]

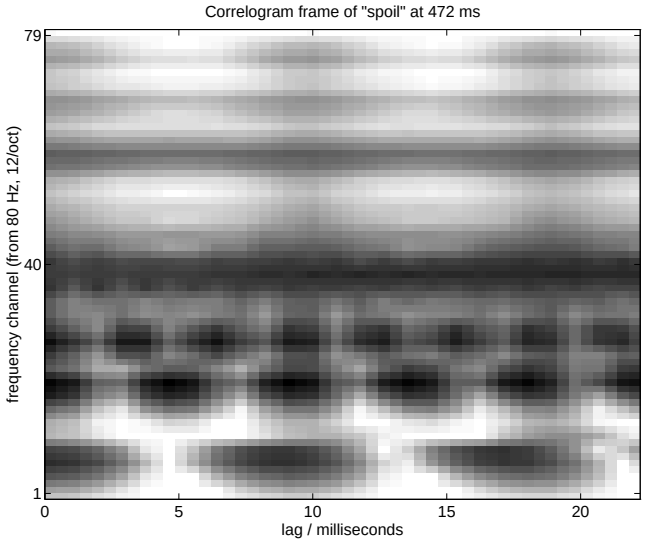

Figure 3: An instantaneous slice of the correlogram of the same word, "spoil", shows that both high and low frequency regions exhibit a common peak at the fundamental lag of 9 ms.

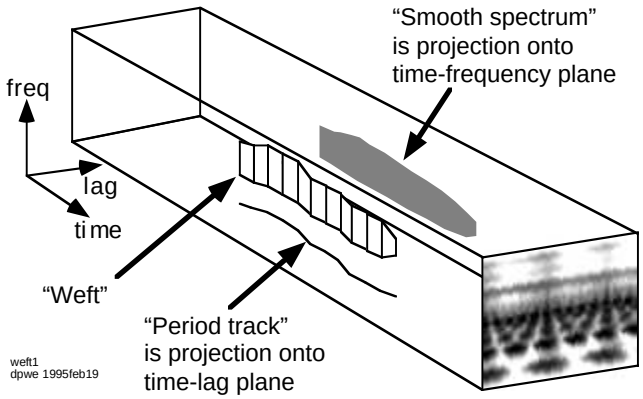

Figure 4: Wefts are formed by making a vertical group of frequency bands that exhibit a common amplitude modulation peak in their autocorrelation lag axes at a given instant, then tracking this set of maxima through time. A weft is completely defined by its two projections, the period track and the smooth spectrum. For clarity, this weft is shown as contiguous in frequency.

4 The Weft representation

The correlogram is in the lower half of the cube in fig. 1, meaning that no discretization process has been performed. We now turn to a discussion of how the structuring advantages of tracks can be combined with the greater dimensionality and improved cues of the correlogram.

The essence of the sinusoid track representation was the tracing of local maxima in the signal energy. Thus, a discrete representation of the correlogram volume (time x frequency x lag) can be constructed on the same basis. To construct tracks from the constant-Q transform, local maxima were picked in the spectrum at a particular instant in time, and these points were grown into tracks by connecting the points in adjacent time frames. In a correlogram, the equivalent of the one-dimensional spectrum is now a two-dimensional surface of frequency versus autocorrelation lag. We could pick the local maxima of this surface in two dimensions and follow them through time, forming a representation of one-dimensional contours extending, tendril-like, down the time axis of our three-dimensional volume.

However, the specific properties of the correlogram allow us to encode common amplitude modulation across frequency bands directly into our basic representation. This provides a satisfying correspondence to the strong perceptual fusion of periodic signals revealed by introspection and experiments.

Thus our approach is to form elements that correspond to entire spectra defined by a common amplitude modulation. We do this by first picking peak lags for each frequency channel in our two-dimensional slice of the correlogram. This selects the dominant modulation periods in each band. We then look for particular periods that occur in a large number of frequency channels, by a mechanism similar to the 'summary autocorrelation' of [Meddis and Hewitt, 1991] and [Meddis and Hewitt, 1992]. The set of frequency bands that reflect modulation at a given period can then be tracked along time to form a kind of 'ragged ribbon' tracing the regions of spectral dominance of a particular modulated signal as it evolves in time. We call these structures *wefts.* [3]

The weft representation partitions the 3-dimensional correlogram space into chunks of sound characterized by common period of amplitude modulation. Wefts appears to combine the important advantages of the various representations we have discussed so far.

4.1 An implementation of wefts

We now describe our initial implementation of the weft representation. Figure 5 shows a block diagram of the system we have constructed. First, the sound (or sound mixture) is analyzed into frequency channels by a conventional constant-Q gammatone filterbank [Patterson *et al.*, 1987]. This two-dimensional representation is then converted to a three-dimensional correlogram as described above by extracting the intensity envelope in each frequency channel (over a 2 ms window applied to the rectified signal), multiplying this envelope by delayed versions of itself, one for each sample of the lag dimension, then smoothing the output of this product over a 20 ms window. This three-dimensional intensity volume (a function of time, frequency and lag) is then reduced to a two-dimensional summary autocorrelation by, for each frequency channel within each time frame, picking the lags corresponding to local maxima in the autocorrelation function for that channel, then superimposing all these lags (convolved with a

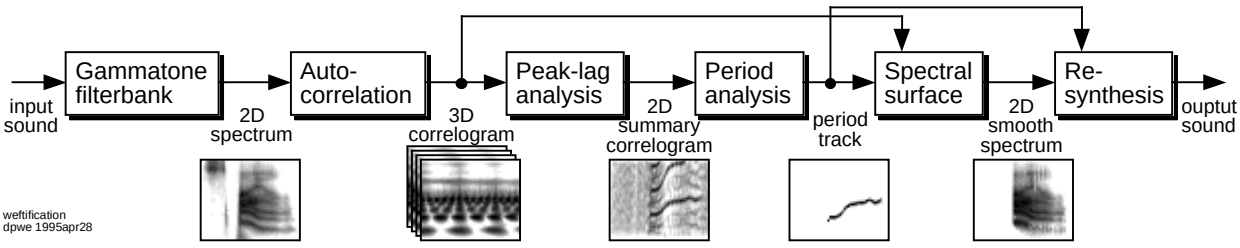

Figure 5: A block diagram of our implementation of a weft analysis-synthesis system for separating mixtures of harmonic sounds.

[2] The fundamental harmonic would naturally be expected to have a peak at the fundamental period. Higher resolved harmonics will have addtional peaks at shorter periods. To implement this scheme, the filtered signals should be half-wave rectified to ensure that resolved harmonics have amplitude modulation period corresponding to their frequency.

[3] From the American Heritage Dictionary: "The horizontal threads interlaced through the warp in a woven fabric," i.e., parallel set of threads.

Gaussian spreading function) for all the frequency channels into a single function of lag for that time frame showing the predominant periods present in the autocorrelations. Since a given periodicity of τ will result in peaks at lags of τ, 2τ etc. (and since resolved harmonics may contribute weak peaks at $\tau/2$, $\tau/3$ etc.), this summary is processed by a period analysis that attempts to explain all the peak lags with a small number of fundamental periods. These periods are tracked along time to produce separate 'period-tracks', encoding the deduced period of the periodic sound element as a function of time. The three-dimensional autocorrelogram is then sampled at these periods; if a given frequency channel has an autocorrelation peak at the period indicated by the period track, the square-root of that intensity peak is copied to the 'smooth spectrum' corresponding to the period track, else that time-frequency cell of the smooth spectrum is left blank (see fig. 4). After pre-compensating for the spreading effect of the filterbank using a non-negative least-squares approximation, a pulse train generated from the period track is filtered by a time-varying filterbank controlled by the smooth spectrum. This results in a resynthesis of the separated periodic signal based on its representation alone.

This algorithm draws heavily on previous autocorrelation-based schemes. The idea of sampling the autocorrelation at the lags indicated by the pitch track was suggested in [Assman and Summerfield, 1990], however, the idea of ignoring frequency channels that do not show a peak at or near that period is more similar to the system of [Meddis and Hewitt, 1992]. [Brown, 1992] makes unique periodicity assignments for each frequency channel, however, these are biased by the dominant periods in the summary autocorrelation, which is therefore largely equivalent to the strategy here of choosing dominant periods from the summary autocorrelation, then recruiting frequency channels to each. Brown's algorithm enforces exclusive allocation of a frequency channel to one period (as his resynthesis requires), whereas the system we describe can detect lag peaks for several different periods within a single channel; it remains to be established if this extra information is of any value. The approach of extracting an entire pitch track for an identified sound object then using it to recover the spectrum, which dates back to [Weintraub, 1985], means that momentary distortions or uncertainties in the tracking can be overcome by interpolation from either side (which is indeed employed by the period-track extraction).

The most original aspect of the weft representation is the emphasis on resynthesis from the information in the representation alone, rather than by recourse to the original signal. The key advantage here is that a representation rich enough to reconstruct a whole signal is a fertile domain for signal modification and restoration in situations where more abstract constraints may be able to improve signal separation performance. A simple example of this is used in the resynthesis block, where small omissions in the recovered 'smooth spectrum' (i.e. time-frequency cells for which no corresponding autocorrelation peak could be located) can be simply interpolated from their neighbors, resulting in a more continuous resynthesis. A reconstruction algorithm that relies on unparameterized input for its fine detail will have difficulty in

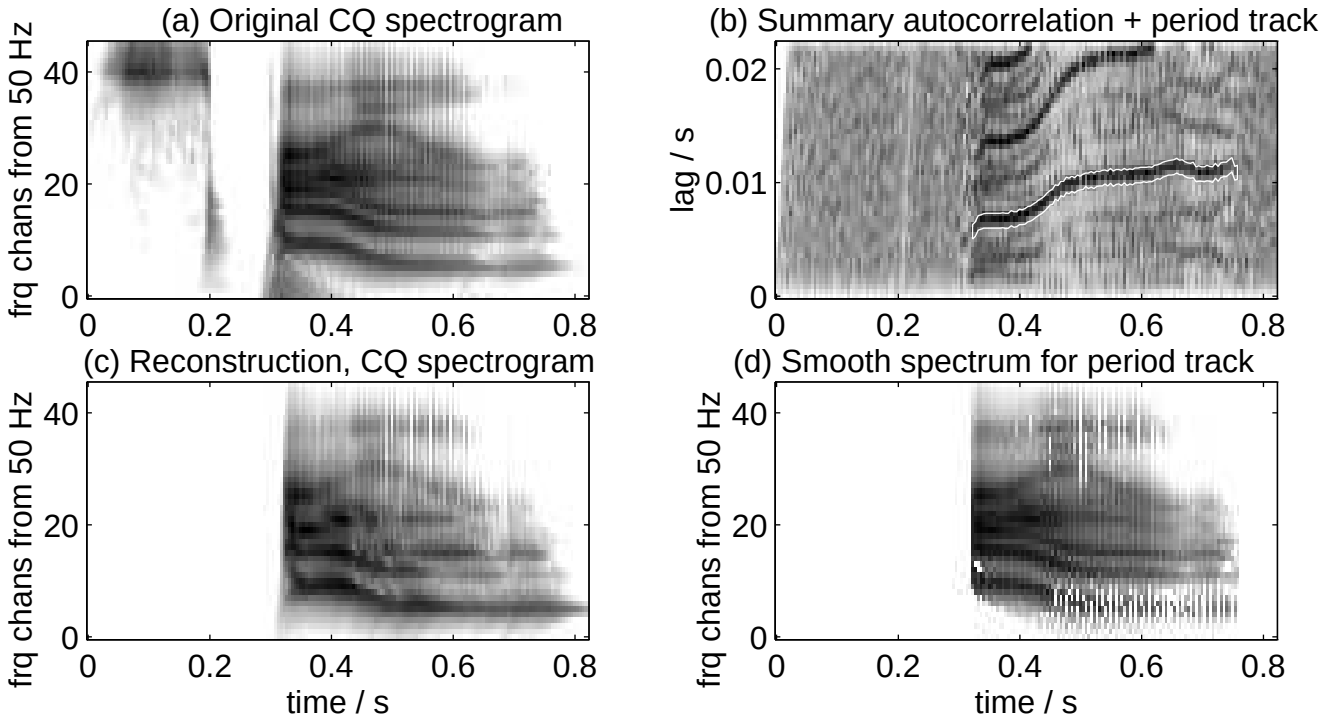

Figure 6: Analysis of the single word "spoil". Panel (a) shows the constant-Q spectrogram of the original sound, analysed by a 46-channel Gammatone filterbank with Q=8, 6 channels per octave, covering 50 Hz to 10 kHz. Panel (b) shows the summary autocorrelation for the entire sound. Note that the aperiodic fricative /s/ initial does not have any pronounced period. The period-track for the main weft is shown in outline. Panel (c) shows the constant-Q spectrogram for the resynthesis of that weft (i.e. the periodic portion of the original speech). Panel (d) shows the smoothed energy surface extracted from the correlogram volume for the period track indicated in (b); together, the period track and the smoothed surface completely specify the weft.

the case of this kind of corruption. A wealth of perceptual evidence suggests that high-level inference is central to the success of auditory perception [Warren, 1970; Ellis, 1993; Slaney, 1995].

Figures 6 and 7 give a graphical representation of the results of our process. Figure 6 considers the single word, "spoil", used as an example before. The first panel is the 'spectrogram' derived from the constant-Q analysis filterbank. The next panel shows the summary autocorrelation function derived from picking the peak lags across frequencies from the three-dimensional autocorrelogram (which, unfortunately, is difficult to visualize, especially on paper!). The pitch track that has been extracted is also marked, forming half of the definition of the weft, its projection onto the time-lag plane. The third panel shows the smooth spectrum extracted from the autocorrelation volume, the other half of the weft definition, its projection onto the time-frequency plane. The final panel shows the spectrogram representation of the resynthesized voice. Notice that only the periodic vowel portion of the speech has been successfully reconstructed.

Figure 7 deals with the analysis of a more interesting case of the mixture of two voices. In this case, the sound is the mixture of male and female speech used as example "v3n7" in [Brown, 1992]. We see that two main pitch tracks have been formed over the summary autocorrelations, and there is a reasonable distinction between the spectrograms of the two reconstructed, isolated voices. These sound examples may be heard, and the Matlab code for the analysis may be obtained, by visiting the World-Wide Web page for this paper,
<http://sound.media.mit.edu/~dpwe/ijcai95.html>.

5 Summary and conclusions

Since computational models of auditory scene analysis constitute a relatively new endeavor, there is a temptation for each researcher to set out to solve the entire problem; in a sense, we don't understand the problem well enough to identify and disentangle more manageable mouthfulls. We should recognize that, like computer vision, there will be myriad different aspects to this area of research, and limiting focus to particular subproblems will probably prove rewarding. This paper has attempted to move in that direction by reducing the scope to considering just the representation to be employed, its desirable properties, and what we can learn by analyzing the different representations that have been used in the past. But we have failed to reduce the problem by very much, since representation is intimately involved with aspects of scene analysis not addressed in this paper, such as hierarchic abstraction and grouping rules.

Our introduction of wefts serves mainly to illustrate what we mean by a mid-level representation, and also follows our critique of sinusoidal tracks and autocorrelation to its logical next step. Certainly, our present system leaves much to be desired: At a detailed level, the extraction of the period track, and particularly the smooth spectrum, could benefit from a more careful analysis and implementation (for instance to establish the best way to estimate, from the autocorrelation, the magnitudes of two periodic energy bursts mixed in a single channel). More generally, a representation that relies on the concept of a period-track cannot handle the large aperiodic portion of our acoustic environment, usually categorized as noisy and/or impulsive. A hypothetical 'ultimate' mid-level representation will encompass these categories, though it is

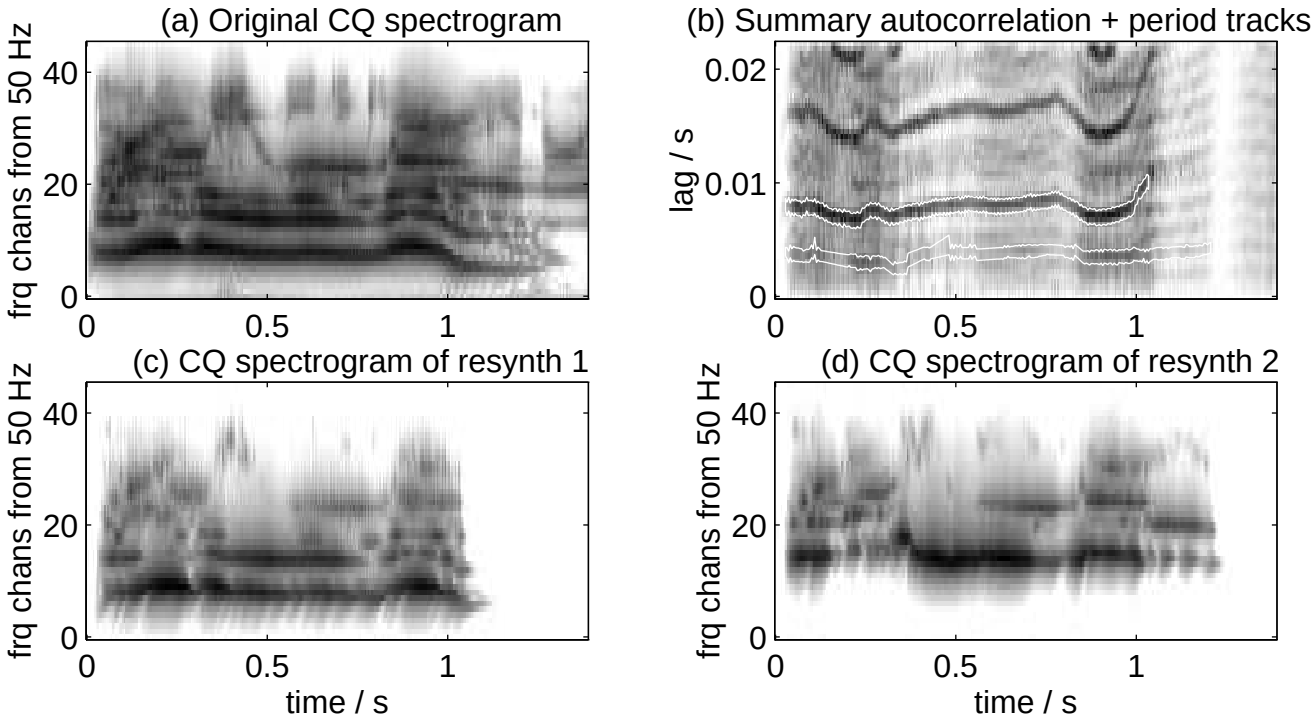

Figure 7: Another example of weft analysis, this time for the mixture of male and female voices called "v3n7" in [Brown, 1992]. (a) and (b) show the spectrogram of the original and the summary autocorrelation, as in figure 6. (c) and (d) show the constant-Q spectrograms of the two resynthesized voices, from the two wefts whose period tracks are outlined in (b).

more likely that a variety of representations should be employed, each more or less appropriate for different kinds of sound.

In conclusion, we hope to have presented a useful theoretical framework, conceptual emphasis, and perhaps a practical tool for the representation of sounds in computational auditory scene analysis systems. While our common goal of automatic models of the human perceptual system may still be some little way distant, we are at least on the path of what promises to be a fascinating journey.

Acknowledgments

This paper would not have been possible without coffee.

References

[Adelson, 1994] Adelson, E. "Layered representations in vision," in Proceedings of the Abstract Perception Workshop, Kanazawa, Japan, January 1994. <http://dfr.www.media.mit.edu/people/dfr/apw-summary/apw-summary.html>

[Assman and Summerfield, 1990] Assman, P.F. and Summerfield, Q. "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," JASA 88(2), 680-697, August 1990.

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

[Brooks, 1991] Brooks, R.A. "Intelligence without reason," MIT AI Lab memo 1293, presented at the Intl. Joint Conf. on Artif. Intell., 1991. <ftp://publications.ai.mit.edu/ai-publications/1000-1499/AIM-1293.ps.Z>

[Brown and Puckette, 1992] Brown, J.C. and Puckette, M.S. "An efficient algorithm for computing the constant-Q transform," JASA 92(5) 2698-2701, November 1992.

[Brown, 1992] Brown, G.J. "Computational auditory scene analysis: A representational approach," Ph.D. thesis CS-92-22, CS dept., Univ. of Sheffield, 1992.

[Cooke, 1991] Cooke, M.P. "Modeling auditory processing and organisation," Ph.D. thesis, CS dept., Univ. of Sheffield, 1991.

[Duda *et al.*, 1990] Duda, R.O., Lyon, R.F. and Slaney, M. "Correlograms and the separation of sounds," Proc. IEEE Asilomar conf. on sigs., sys. & computers, 1990.

[Ellis, 1992] Ellis, D.P.W. "A perceptual representation of audio," MS thesis, EECS dept., MIT, February 1992. <ftp://cecelia.media.mit.edu/pub/dpwe-ms-thesis.ps.tar.Z>

[Ellis, 1993] Ellis, D.P.W. " Hierarchic models of hearing for sound separation and reconstruction," Proc. IEEE Workshop on Apps. of Sig. Proc. to Acous. and Audio (Mohonk), October 1993. <ftp://sound.media.mit.edu/pub/Papers/dpwe-waspaa93.ps.gz>

[Ellis, 1994] Ellis, D.P.W. "A computer model of psychoacoustic grouping rules," Proc. 12th Int. Conf. on Pattern Recognition, Jerusalem, October 1994. <ftp://sound.media.mit.edu/pub/Papers/dpwe-ICPR94.ps.gz>

[Knight, 1994] Knight, T.F. "Lessons in perception from mammals," in Proceedings of the Abstract Perception Workshop, Kanazawa, Japan, January 1994. <http://dfr.www.media.mit.edu/people/dfr/apw-summary/apw-summary.html>

[Kollmeier and Koch, 1994] Kollmeier, B. and Koch, R. "Speech enhancement based on physiological and psychoacoustical models of modulation perception and binaural interaction," JASA 95(3), 1593-1602, March 1994.

[Marr, 1982] Marr, D. *Vision*, Freeman, 1982.

[McAulay and Quatieri, 1986] McAulay, R.J. and Quatieri, T.F. (1986). "Speech analysis/synthesis based on a sinusoidal representation," IEEE Tr. ASSP-34, 1986.

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

[Meddis and Hewitt, 1992] Meddis, R. and Hewitt, M.J. "Modeling the identification of concurrent vowels with different fundamental frequencies," JASA 91(1), 233-245, January 1992.

[Patterson *et al.*, 1987] Patterson, R.D., Nimmo-Smith, I., Holdsworth, J. and Rice, P. "An efficient auditory filterbank based on the gammatone function," Institute of Acoustics Speech Group Meeting on Auditory Modelling, RSRE, December 1987.

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

[Slaney, 1995] Slaney, M. "A critique of pure audition," Proc. Intl. Joint. Conf. on Artif. Intel, Workshop on Computational Aud. Scene Anal., Montréal, August 1995. <ftp://ftp.interval.com/pub/papers/malcolm/PureAudition.psc.Z>

[Warren, 1970] Warren, R.M. "Perceptual restoration of missing speech sounds," Science 167, 392-393, January 1970.

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

[Winston, 1984] Winston, P.H. *Artificial Intelligence (2nd edition)*, Addison Wesley, 1984.