to appear in:John A. Bullinaria, David W. Glasspool &
George Houghton (1997). *Proceedings of the Fourth
Neural Computation and Psychology Workshop:
Connectionist Representations*. London: Springer-Verlag.

Extracting Features from the Short-term Time Structure of Cochlear Filtered Sound

Leslie S. Smith
CCCN, Dept. of Computing Science and Mathematics, University of Stirling
Stirling FK9 4LA, Scotland.

Abstract

Auditory modelling uses the architecture of the auditory system to guide early sound processing. The advantage of this approach is (i) time-resolution is better and (ii) many bandpassed channels are available and can be processed in parallel. Good time-resolution allows sophisticated across-time processing to be applied to each channel, resulting in the discovery of features in each channel. Logically each channel can be processed simultaneously. The features discovered can be correlated across channels. We present some early results for processing sound at three different levels of short-term time structure.

1 Auditory Modelling and Spectral Analysis

From the time of Helmholtz, it has been known that the cochlea, the main component of the inner ear, performs a frequency analysis on the pressure wave that we perceive as sound. This has led workers in speech and sound interpretation to perform a similar analysis as a first step in sound interpretation. Much of the work has used a Fourier transform based approach. Such spectral analysis is usually carried out on short sections of sound (perhaps 40ms long), supplying the Fourier transform with a sound pressure/time vector and resulting in a power/frequency vector of the sound during that section. (The phase information is generally discarded.) This produces a representation consisting of an N-element vector (where N is the order of the Fourier transform) per analysis. An alternative technique, the one used in this work, is to use auditory modelling, in which the sound signal is filtered into bands, following what is known of the response of the cochlea. This produces a representation consisting of M channels, each with the same time-resolution as the original signal. Noting that the sections to which the Fourier transform is applied may overlap, it becomes clear that both techniques give rise to representations larger than the original signal. Downstream representations are strongly influenced by the structure of this initial representation.

Table 1 summarises the advantages and disadvantages of each approach. We should point out that Fourier transform based methods can be used to perform auditory modelling by applying the transform to overlapping segments of sound, (e.g. to a new 40ms segment every 1ms), and then regrouping the power vectors to give an appropriate power/frequency distribution. In this way, both accurate time resolution and frequency resolution can be achieved, though at the cost of considerable computation.

Auditory modelling	Spectral analysis
Good time resolution	Poor time resolution
Poor frequency resolution	Good frequency resolution
Near-logarithmic channel distribution	Linear energy/frequency spectrum
Computationally intensive due to presence of multiple channels	Computationally intensive due to initial transform
aVLSI or DSP solutions are possible	dVLSI or DSP solutions are possible

Table 1: Comparison of auditory modelling and spectral analysis techniques

Which approach is preferable depends on what one is trying to achieve. Current commercial systems aim to achieve direct interpretation of a clean incoming sound, and use spectral analysis. However, if one needs to stream the sound to accentuate the source of interest, then we contend that auditory modelling is the better approach because it can permit segregation of features from differing sources prior to any attempt at interpretation. Generally, interpretation of sound (specifically speech) processed using spectral analysis techniques is performed directly, using hidden Markov models or neural networks which constrain the likely interpretation of some input vector sequence using the statistics of the target (phonemic) vocabulary. Auditory modelling based approaches can use the short-term structure of the signal in different channels to define features: signals from a single source tend to share short-term structure, and this can be used to group the features, thus allowing signals from other sources to be ignored. Streaming techniques based on applying correlogram processing to auditory model processed sounds have been used in [15, 26, 5].

In the current work the features used are onsets, offsets, and amplitude modulation pulses: other features are certainly possible [4]. We do not use correlogram based processing, partly because this technique results in even larger volumes of data, and partly for the reasons outlined in section 2.3. The auditory modelling approach to streaming can take advantage of the general characteristics of sound sources, but does not generally use so much high-level information as is used in direct interpretation. Nonetheless, such information can be used later in the interpretation process.

In the rest of this paper, we consider three levels of the short–term time structure of sound, illustrating them graphically and discussing how each of them can be used.

2 Three levels of short-term structure

We discuss three levels of short-term structure, summarised in table 2. Each level corresponds to a different level of the time structure of speech. These are similar to the envelope, periodicity and fine-structure levels discussed in [22], and resemble the three forms of modulation discussed in [21]. Each level corresponds to a different type of activity in the early auditory system, namely firing of onset cells in the cochlear nucleus (CN), firing of chopper cells in the

Timescale	Timing	Application
Coarse	20–50ms Order of movement of vocal tract articulators in animals	Detecting rhythm and basic sound elements (syllables, phonemes) Monaural streaming
Medium	3–10ms Movement period of glottal folds	Detecting voicing Speaker identification Intonation (from fundamental frequency) Monaural streaming Pitch estimation
Fine	0.35-2 ms Order of period of sound in sensitive area of human hearing Auditory nerve spikes	Speech articulation place and vowel quality Direction detection Binaural streaming Pitch estimation

Table 2: Three levels of short-term time structure of sound

CN, and firing of neurons in the auditory nerve [17]. These levels could be extended to longer times as well, as suggested in [30].

2.1 Across-channel processing

Although the features found in each channel can be used independently, it is the ability to correlate features across channels that makes the auditory modelling approach particularly powerful. Signals in different channels which change in similar ways at the same time usually come from the same source. Indeed, human subjects tend to group together signals in different parts of the audible spectrum if they change in similar ways at the same time [3]. We make use of this by concentrating on those features which are supported by similar features in adjacent channels at about the same time. We have applied this technique to the coarse and medium short-term structure of sound, and used this correlation across features to identify features even in the presence of considerable noise. However, we have not been able to use this grouping for the fine scale short term features.

For all of the examples that follow, the Gammatone cochlear filterbank [19] provides the initial bandpass filtering.

2.2 Coarse short-term time structure

The output of the cochlear filterbank was processed channel by channel. It was first rectified, then bandpass filtered to emphasise changes in the envelope of each signal in the 20-50ms window, using a neurally plausible bandpass filter. This corresponds to across-time processing preceding the across-frequency correlation [1]. The across-frequency integration took the form of applying the signals to a one-dimensional network of leaky integrate-and-fire neurons [8].

We used two networks. In both networks, each neuron receives excitatory input from one channel, and excites the neurons in adjacent channels when it fires. In one network, the neural input corresponded to increases in smoothed envelope (onsets), and in the other network to decreases in the smoothed input (offsets). In both networks the result was that when one neuron fired, nearby neurons which were close to threshold fired almost immediately. This results in temporal and tonotopic clustering, giving a volley of spikes across a number of channels in response to an increase (decrease) in signal power. The technique is described in detail in [28]. The onset cell network is loosely modelled on the onset cells of the cochlear nucleus.

The end result is the detection of features corresponding to the start and end of bursts of energy of between 20 and 50 ms duration in the sound. These bursts are characteristic of speech, occurring in plosives, in voiced sounds, even in sibilances, and can be used to endpoint speech elements even in the presence of considerable noise, as can be seen in figure 1. An analogue VLSI implementation of the neural part is under construction [10].

2.3 Medium short-term time structure

There are a number of possible methods for extracting medium short-term time structure. The autocorrelation function (ACF) was suggested originally by Licklider [13], and has been used by many others (e.g.[26, 14, 15]). This provides detailed information of the signal's periodicity, channel by channel, and this information can be combined across channels to produce a summary ACF [14]. We have chosen to use a different technique, one based on amplitude modulation. We do so (i) because of the amplification and classification of amplitude modulated signals which occurs in the cochlear nucleus [12, 17, 11] and (ii) because the only neurobiological evidence for neurobiological ACF computation occurs in very specialised tasks such as echolocation (e.g. [7]) or localisation [6].

Initial processing was as for the coarse time structure, except that the bandpass filter accentuated envelope changes which were in the 3-10ms range. Low frequency bands were ignored, as their envelope cannot change on this timescale. We were particularly interested in sounds generated by the combination of many harmonics of a low-frequency excitation. Voiced speech is one example of this type of sound.

The major source of amplitude modulation is unresolved harmonics. The organ of Corti in the cochlea performs a frequency analysis, but the bands are relatively wideband, with a minimum equivalent rectangular bandwidth (ERB)

$$\text{ERB} = F_c/Q + 24.7\text{Hz}$$

where F_c is the centre frequency and Q is the sharpness of the filter in the cochlear model used [19]. Auditory nerve response is complex: at middle and high frequencies, for a constant intensity tone, it consists of a gradual increase as the tone frequency increases, followed by a sharp peak at F_c, followed by a rapid decline [20]. For the cochlear model used, this is best characterised by $Q = 9.265$ [9]. However, this value of Q is for pure tones at low sound pressure level (SPL), and the selectivity broadens (i.e. Q decreases) for higher SPLs, particularly once low–threshold auditory nerve fibres are driven into saturation, and for wideband sounds [16, 24, 23]. The lower value of Q allows unresolved

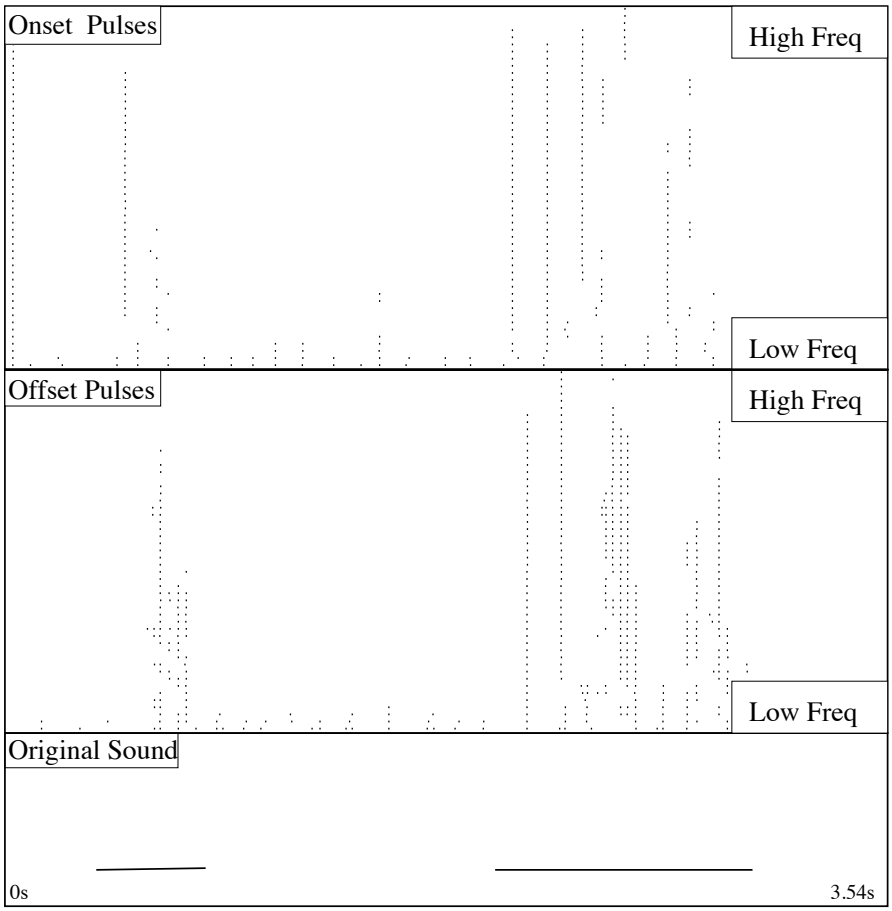

Figure 1: Utterance "Say, that's a nice bike" in motor-bike noise: horizontal lines mark the utterance itself. The clustered onsets and offsets can be seen clearly, even although the SNR is poor, as can be seen from the overall envelope of the sound. The leftmost onset volley marks the start of the wideband motorbike noise.

harmonics to start at a lower F_c. The system is modelled functionally on the chopper cells of the cochlear nucleus, which have been shown to be particularly responsive to AM signals near their best frequency and best AM frequency [17].

Two techniques have been used to combine the channels: the first technique consisted of simply adding up all bandpassed envelopes across the channels, and then detecting amplitude modulation pulses. This requires that the signals in the different bands are accurately time-aligned. To achieve this, it was necessary to compensate for the variations in delay introduced by the filter. This addition loses all information about where in the spectrum the amplitude modulation occurred, so that although the technique can be used to detect voiced sound, it is not useful for streaming, nor for more detailed feature detection. It does, however, provide a technique for detecting voicing which ignores low frequency sound altogether, and can thus achieve good results in the presence of interfering low frequency noise. This is described in detail in [27]. The problem with this technique is that it assumes that the amplitude modulation in different channels is synchronised: this is unlikely to be the case in a real environment.

The second technique maintains channel information, delaying the across channel processing. Each channel is processed to find amplitude modulation pulses, and then pulses which do not conform to the AM expected (i.e. those with AM frequency too high or too low) are discarded. The combination technique used was to retain only pulses "supported" by other pulses: that is, only if there had been a sufficient number of pulses on adjacent channels within a certain length of time. Details are in [29].

The effect of this processing is shown in figure 2. The AM pulses discovered do show the voicing structure of the signal. This is clear both with and without noise (figure 2B-D): retaining only those pulses corresponding to AM between 80 and 140Hz improves the situation (figure 2E-G) for this speaker. The AM detecting technique renders the formant structure of the voiced sections visible: this can be seen both in the absence of noise and when noise was added . Using a simple form of across-frequency processing in which pulses are retained only if supported by sufficient earlier pulses is effective only if the AM extends across many channels. In this case, some of the background noise can be removed, while retaining much of the structure (figure 2F): but if there is much interfering noise, much of the AM structure of the signal may be lost (not shown). In particular, the formant structure is lost.

We can select which pulses should be retained, channel by channel, keeping only those corresponding to some small range of AM frequencies. Figure 3 shows that the frequency of the AM, and hence of the fundamental excitation, F_0, changes in a reasonably smooth way. Detection of amplitude modulation allows the movement of F_0 to be tracked: one can use the inter-pulse interval in each channel as an estimate of the F_0 period, and use the median of these values, thus not requiring the very fine comb-filters which would be needed to find the precise harmonics present.

The example here is of a male speaker, with a relatively low F_0. Female speakers have a higher F_0, so that with the Q used here, unresolved harmonics do not occur until considerably higher F_c. We believe that the wideband nature of speech is such that for normal SPL, the Q will be decreased due to high spontaneous-rate AN fibers becoming saturated [23], and to changes in the action of the outer hair cells. The effect of this would be to reduce the Q of

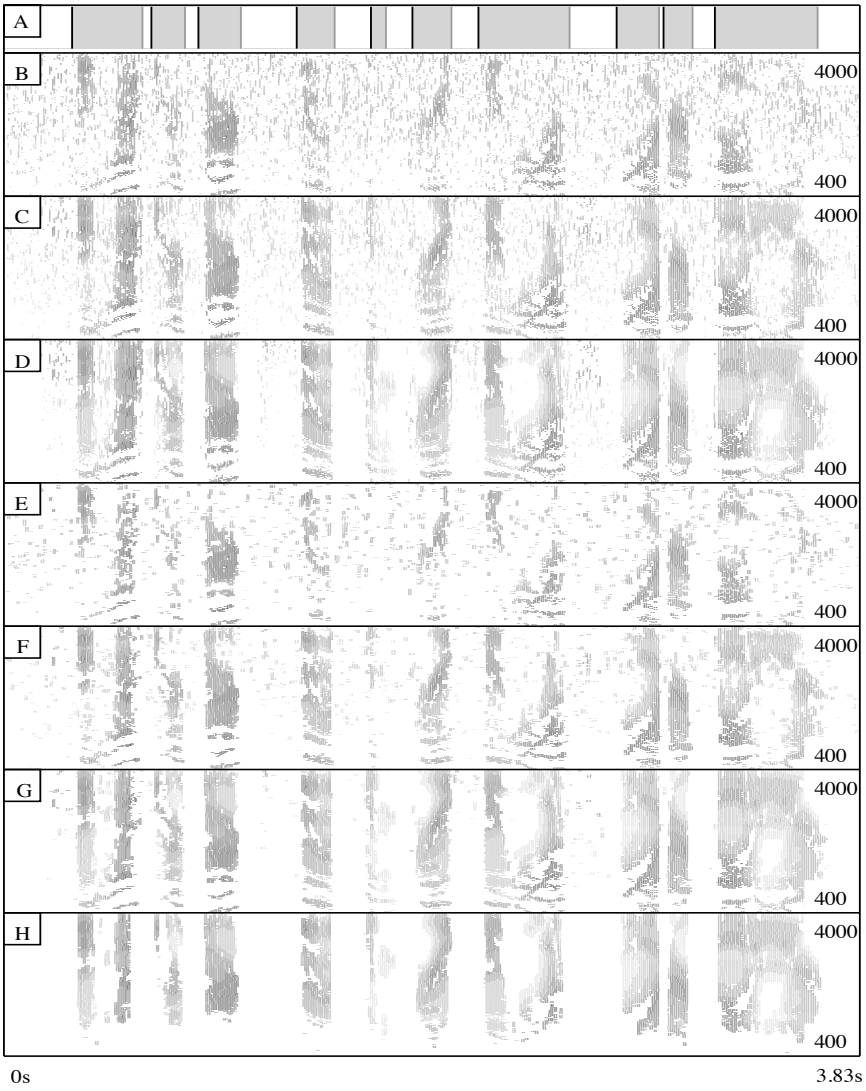

Figure 2: Effect of processing male TIMIT utterance "she had your dark suit in greasy wash water all year", dr3/mkls1/sa1. (A) dark sections mark voiced parts of the utterance. (B-D) Amplitude modulation pulses found, using 141 channels, 400-4000Hz. (B) with white noise added to give 5dB SNR. (C) with white noise added to give 10dB SNR (D) on original TIMIT signal (E-G) Amplitude modulation pulses retained when AM constrained to be in 80-140Hz region. (E) with white noise added to give 5dB SNR. (F) with white noise added to give 10dB SNR (G) on original TIMIT signal. Result of retaining only those pulses corresponding to AM between 80 and 140Hz. (H) Result of retaining only those pulses supported by 10 others within a radius of ± 10 channels for original signal with no noise added.

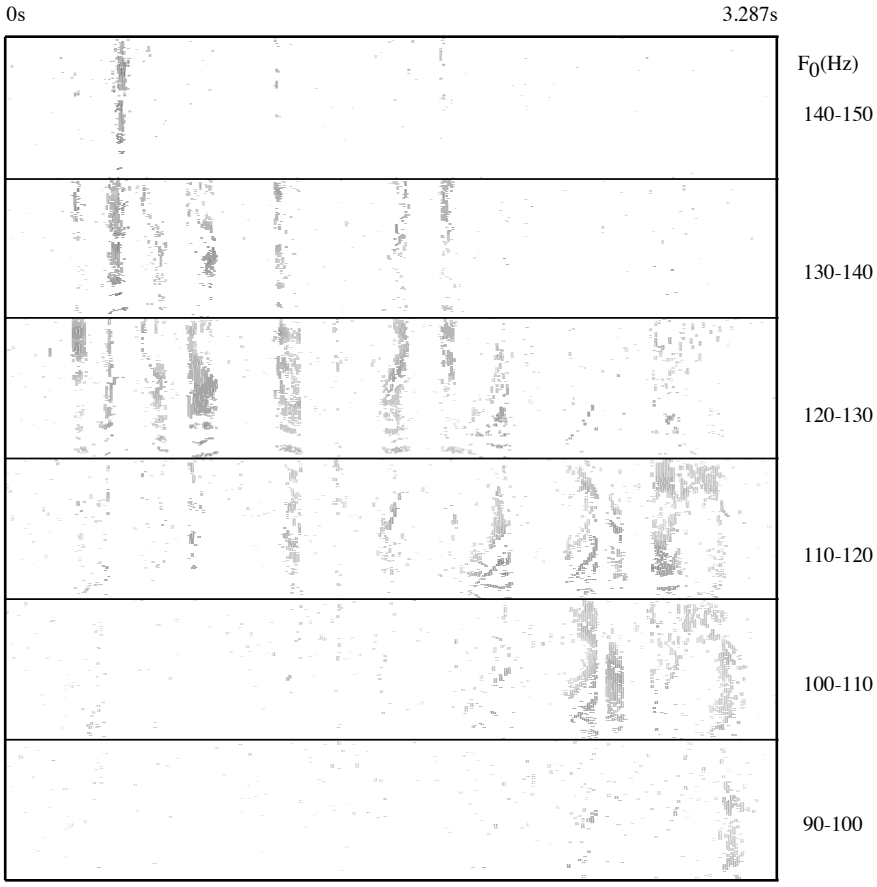

Figure 3: AM pulses retained for different bands of frequency of amplitude modulation. Processing is otherwise as in figure 2c (white noise, 10dB SNR).

those parts of the response in which the sound energy is concentrated, causing AM due to unresolved harmonics to occur at lower F_c. We have not been able to test this directly, although we have found that using a lower Q does permit AM to be found in female speakers speech at lower frequencies [27]. However, using a low Q throughout the spectrum hides the spectral structure of the amplitude modulation.

The most similar system is the stabilised auditory image [18]: however, our system works channel by channel, seeking amplitude modulation (partially a biologically motivated exercise), rather than buffering and triggering channel outputs to produce a visually inspectable image displaying the medium short-term structure of the sound.

Further work is required to find more effective ways of combining the information across multiple bands: we would particularly like to retain the formant structure while discarding isolated AM pulses caused by noise.

2.4 Fine short-term time structure

Fine time structure is used in binaural sound direction-finding [2]. This appears to be based on the phase locking of auditory nerve spiking for signals below about 2.5kHz in humans. Although exactly how one might extract useful information from the timing of AN spikes at this level is not clear, Rosen [22] identifies specific applications for this level of short-term time structure in speech interpretation.

The use of this level of time structure in the interpretation of sound or speech in unclear: nonetheless, applying auditory modelling does give some interesting pictures!

Figure 4 shows the result of applying cochlear filterbank followed by simple (i.e. half-wave) rectification to some sounds. The branching structure is characteristic of sounds with many pure sinusoidal frequency components: a simple tone results in a sequence of near-vertical stripes which are strongest (darkest) where the filter responds to that frequency. The precise structure is determined by the way in which the strongest frequency component in the bandpass filtered signal changes as the centre frequency of the bandpass filter changes. For wideband noise-based sounds with a flat spectrum, the strongest frequency component of the bandpass filtered signal will always be at the centre frequency of the bandpass filter. In this case, the pattern of branching will appear random. For voiced speech sounds, the branching reflects the strength of each harmonic: thus the branching has a regular structure and this structure is determined by the formants of the vocal tract. For bandpassed unvoiced sounds, like /s/, the branching has some degree of regularity, as can be seen from the bottom part of figure 4.

3 Discussion

Auditory modelling techniques retain more information about the short-term time structure of sound than techniques based on Fourier transforms. We have shown how certain features we believe to be useful in interpretation or streaming can be adduced from the coarse and medium scale short-term time structure. We have not identified how such features may be adduced directly from the

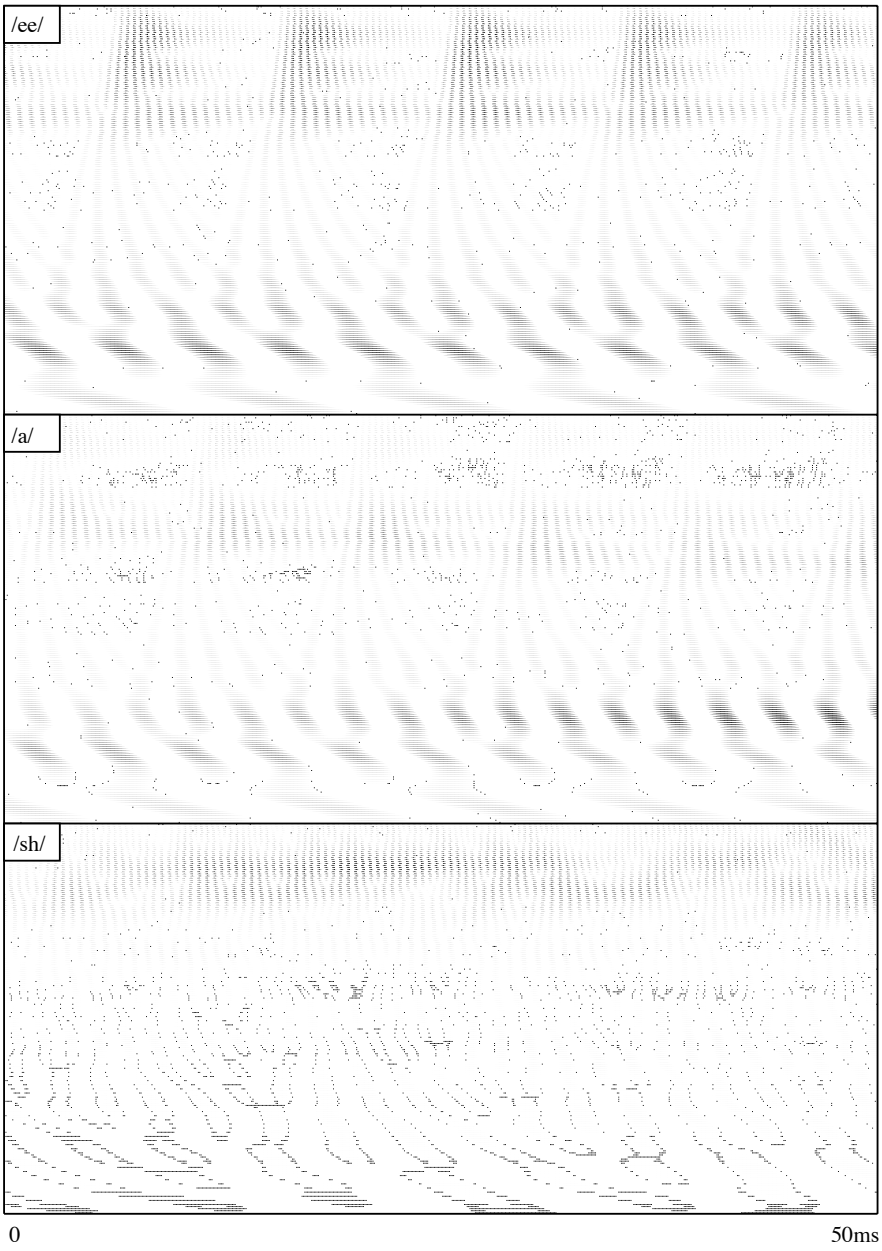

Figure 4: Fine time structure of two vowels and a sibilance. Channel centre frequencies range from 100Hz (bottom) to 6kHz (top).

fine scale short-term time structure. We note that Rosen [22] identifies more applications for each level of short-term time structure of speech.

For the coarse-scale time structure, we have shown that performing across-time processing before across-frequency processing can permit us to perform temporal and tonotopic clustering of onsets and offsets. Future work should consider exactly which channels are bracketed by the volleys of onset and offset firings, and concentrate attention on those channels during that segment. We note also that even when the spectral information is compressed down to 3 or 4 bands, envelope temporal cues suffice to permit high levels of word recognition [25], emphasising the importance of this level of structure.

For the medium-scale time structure, we have shown the usefulness of across-time processing on a channel by channel basis. However, we have yet to produce an effective method for across-frequency integration, and this is still under investigation: one possibility is to use lateral inhibition to sharpen the response profile. Using a cochlear filter/auditory nerve model whose response depends on the distribution of the energy of the sound in a more biologically realistic way would we believe, allow sounds with a higher frequency of fundamental to be processed in the same way as we have processed sounds with lower fundamental frequency. In particular, the widening of the peak level response of auditory nerve fibers with high spontaneous rates (reviewed in [23]) would allow unresolved harmonics to generate amplitude modulation pulses at lower F_c's. We are also interested in combining the features from different time-scales, with a view to performing feature-based sound streaming and interpretation.

Processing multiple channels produced by cochlear filtering is time consuming when performed with software. We are currently working with the department of Electrical Engineering at the University of Edinburgh to transfer some of the processing into analogue VLSI, in order to be able to perform these algorithms in real-time [10].

Acknowledgements

I am indebted to one of the referees, Professor Stuart Rosen, for his comments.

References

[1] J.B. Allen. How do humans process and recognize speech. *IEEE Transactions on Speech and Auditory Processing*, 2(4):567–577, 1994.

[2] J. Blauert. *Spatial Hearing*. MIT Press, 1983.

[3] A.S. Bregman. *Auditory scene analysis*. MIT Press, 1990.

[4] G. Brown. Computational auditory scene analysis:a representational approach. Technical report, Department of Computer Science, University of Sheffield, Sheffield, UK, 1992.

[5] G.J. Brown and D.L. Wang. Modelling the perceptual segregation of double vowels with a network of neural oscillators. Technical Report CS-96-07, Department of Computer Science, University of Sheffield, Sheffield, UK, 1992.

[6] C.E. Carr and M. Konishi. A circuit for the detection of interaural time differences in the brain stem of the barn owl. *Journal of Neuroscience*, 10:3227–3246, 1990.

[7] S.P. Dear and N. Suga. Delay-tuned neurons in the midbrain of the big brown bat. *Journal of neurophysiology*, 73(3):1084–1100, 1995.

[8] W. Gerstner. Time structure of the activity in neural networks. *Physical Review E*, 51(1):738–758, 1995.

[9] B.R. Glasberg and B.C.J. Moore. Derivation of filter shapes from notched-noise data. *Hearing Research*, 47:103–138, 1990.

[10] M.A. Glover, A. Hamilton, and L.S. Smith. Analogue VLSI integrate and fire neural network for clustering onset and offset signals in a sound segmentation system. Submitted to 1st European Workshop on Neuromorphic Systems, 1997.

[11] M.J. Hewitt and R. Meddis. A computer-model of amplitude-modulation sensitivity of single units in the inferior colliculus. *Journal of the Acoustical Society of America*, 95:2145–2159, 1994.

[12] G. Langner. Periodicity coding in the auditory system. *Hearing Research*, 60:115–142, 1992.

[13] J.C.R. Licklider. A duplex theory of pitch perception. *Experentia*, 7:128–133, 1951.

[14] R. Meddis and M.J. Hewitt. Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America*, 89(6):2866–2882, 1991.

[15] R. Meddis and M.J. Hewitt. Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America*, 91(1):233–245, 1992.

[16] A.R. Moller. *Auditory Physiology*. Academic Press, 1983.

[17] A.R. Palmer and I.M. Winter. Cochlear nerve and cochlear nucleus responses to the fundamental frequency of voiced speech sounds and harmonic complex tones. *Advances in the Biosciences*, 83:231–239, 1992.

[18] R. D. Patterson and J.W. Holdsworth. Generation of stabilised waveforms. UK patent GB 2212801 A, December 1990.

[19] R.D. Patterson, M.H. Allerhand, and C. Giguere. Time-domain modelling of peripheral auditory processing: A modular architecture and a software platform. *Journal of the Acoustical Society of America*, 98:1890–1894, 1995.

[20] J.O. Pickles. *An introduction to the physiology of hearing*. Academic Press, second edition, 1988.

[21] R. Plomp. The role of modulation in hearing. In R. Klinke and R. Hartmann, editors, *Hearing–physiological bases and psychophysics*. Springer-Verlag, 1983.

[22] S. Rosen. Temposal information in speech: acoustic, auditory and linguistic aspects. *Philosophical transactions of the Royal Society of London B*, 336:367–373, 1992.

[23] M.A. Ruggero. Physiology and coding of sound in the auditory nerve. In A.N. Popper and R.R Fay, editors, *The Mammalian Auditory Pathway: Neurophysiology*. Springer-Verlag, 1992.

[24] M.B. Sachs, H.F. Voigt, and E.D. Young. Auditory nerve representation of vowels in background noise. *Journal of Neurophysiology*, 50(1):27–45, 1983.

[25] R.V. Shannon, F.G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid. Speech recognition with primarily temporal cues. *Science*, 270:303–304, 1995.

[26] M. Slaney and R.F. Lyon. On the importance of time–a temporal representation of sound. In M.Cooke, S.Beet, and M.Crawford, editors, *Visual representations of speech signals*. John Wiley and Sons, 1993.

[27] L.S. Smith. A neurally motivated technique for voicing detection and f_0 estimation in speech. Technical Report CCCN–22, Centre for Cognitive and Computational Neuroscience, University of Stirling, Stirling UK, 1996.

[28] L.S. Smith. Onset-based sound segmentation. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems 8*, pages 729–735. MIT Press, 1996.

[29] L.S. Smith. A noise-robust auditory modelling front end for voiced speech. Accepted for ICANN97, Lausanne, October 8–10, 1997, 1997.

[30] N.P.McA. Todd. The auditory primal sketch: A multiscale model of rhythmic grouping. *Journal of New Music Research*, 23(1), 1994.