SOUND FEATURE DETECTION USING LEAKY INTEGRATE-AND-FIRE NEURONS

Leslie S. Smith and Dagmar S. Fraser

Department of Computer Science and Mathematics
University of Stirling, Stirling FK9 4LA, Scotland

ABSTRACT

We present a neurally inspired technique for detecting onsets in sound. The outputs from a cochlea-like filter are spike coded, in a way similar to the auditory nerve. These AN-like spikes are presented to leaky integrate-and-fire (LIF) neurons through a depressing synapse. The spike outputs from these are then processed by another layer of LIF neurons. Onsets are detected with essentially zero latency. We present results from the TIMIT database.

1. ONSETS, AND ONSET (CHOPPER) CELLS

We aim to provide features for sound source streaming and interpretation. Biological systems far outperform current systems. Thus modeling aspects of the biological system seems a good way forward. We model aspects of the cochlea, auditory nerve (AN) and cochlear nucleus, aiming to provide engineering insight into early auditory processing.

Onsets are rapid increases in energy. Different sound sources have different types of onsets. Some are wideband, with sudden co-occurring increases in intensity (e.g. percussive sounds). Others are narrowband, with the increase in energy in some small area(s) of the spectrum (e.g. a note played on a flute). Some sound onsets are very rapid, (e.g. a glass falling on to a stone floor), and others less so (e.g. a note played on a flute). Every sound that starts has an onset, and many have internal onsets (e.g. animal vocalisations, such as human speech, or sequences of musical notes). The energy increase may be anything between 10 and 100dB, and there may be any pre-onset sound level.

Mammalian auditory systems are strongly attuned to onsets. The AN responds more strongly, with many neurons in the cochlear nucleus also spiking strongly, at stimulus start [1]. Ecologically, onsets provide a useful cue. The onset comes at the start of the sound (or of some change in the sound), and is therefore useful for priming a response. Onsets are relatively undamaged by reverberation, since the onset in the received signal will normally be from the direct path, and further onsets caused by reflections will be smaller. Other cues such as offsets are severely smeared out in time in reverberant environments.

Thanks to the UK EPSRC (grant GR/R64654) for funding.

Onsets are a form of envelope modulation. Some cochlear nucleus neurons sensitive to onsets are also sensitive to other forms of envelope modulation, such as amplitude modulation (AM) [1]. By altering the parameters used, the system can detect AM, though this is not discussed here.

1.1. Onset detection

Onset detection systems have been used in music transcription (e.g. [2]), sound segmentation [3], lip synchronisation [4], monaural sound source streaming (e.g. [5]), and determining when to measure ITDs for sound direction finding [6]. On-line applications (e.g. real-time speech segmentation, source streaming, sound direction finding, or music transcription), may use the sound only up to the time of onset, and the detector latency becomes important.

Bandpassing the sound signal into many bands stops onsets in some small part of the spectrum from being overwhelmed by the overall signal strength, unless it is in an adjacent part of the spectrum. Also, it allows onsets found to be characterised, by annotating them with the bands in which they have been detected. This is important for transcription, streaming, and direction finding applications.

The simplest onset detection techniques are based directly on signal energy, and were used to segment hummed or sung notes [7] to improve note differentiation in early music transcription systems. An alternative is to use first order difference based estimates, (e.g. [8]), which take the maximum of the rising slope of the amplitude envelope as an index of onset. [2] uses the relative difference, calculating $\Delta I/I$. Another variant is [9] which uses troughs in loudness to segment sung notes. A different approach uses optimal filter based techniques: [4] uses a wavelet based filter and [3, 10] use the difference between a long-term and a short-term average. A related approach uses expectation based techniques [11] to detect sudden increases in intensity. Simple techniques tend to find only the most prominent onsets, while techniques which rely on finding troughs have a longer latency. Filter techniques can be optimised for particular source types and reverberation characteristics, and can perform well, but require a convolution, and can have long latency.

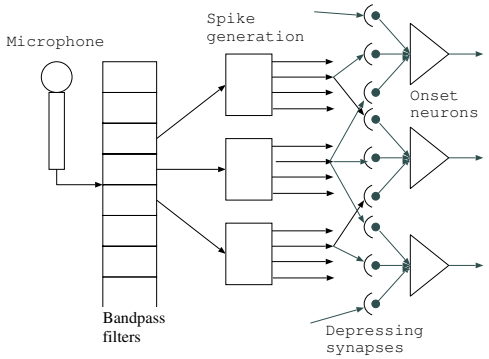


Fig. 1. Onset spike generation system. AN-like spike generation is shown for only three bands. Depressing synapses and onset generation are shown for a single sensitivity level for these three bands.

2. THE MODEL

The model we use is illustrated in figure 1. Sound from a microphone (or sound file) is bandpass filtered, using a Gammatone filterbank [5]. The filterbank response is similar to that of the basilar membrane in the Organ of Corti in the cochlea: that is, the 6dB down point bandwidth is approximately 20% of the centre frequency. The filter density provides considerable overlap between adjacent filters. An important issue in filter design is delay: since we will be using the output of each filter in conjunction with adjacent filters, we would like the insertion delay to be similar for all the filters. However, the Gammatone filter delay is proportional to the reciprocal of the bandwidth [5]. Other filters, such as OTA [12] have a more constant delay.

The spike based representation enables the system to work over a wide dynamic range by using multiple spike trains coding the output of each channel. Each spike codes a positive-going zero crossing. Each spike train S_i, for $i = 1 \ldots N$, (where N is the number of spike trains generated from a single bandpass channel) has a minimum mean voltage level E_i that the signal must have reached prior to crossing zero during the previous quarter cycle. If there are N spike trains, these E_i are set by

$$E_i = D^i E_0 \tag{1}$$

for $i = 1 \ldots N$, for some E_0 fixed for all bands. D was set to 1.414, providing a 3dB difference between the energies required in each band. Note that if a spike is generated in band k, then a spike will be generated in all bands k' for $0 \leq k' \leq k$. This technique is similar to that used by in [13], where Ghitza noted that it improved automatic speech recognition in a noisy environment. This auditory nerve-like representation enhances neither onsets, (unlike the real mammalian auditory nerve) nor amplitude modulation. However, the way in which it codes the signal can be used to build a neurally inspired onset detection system as shown in section 3.

The AN-like spikes are applied to depressing synapses on onset neurons (figure 1), leaky integrate-and-fire (LIF) neurons with depressing synapses. LIF neurons are the simplest model neurons which maintain any semblance of the temporal behaviour of real neurons: see [14], chapter 14 for a review. The neurons used here are characterised by their leakiness and refractory period. Each onset cell is innervated by a number of auditory nerve-like spike trains. These arrive from a number of adjacent bandpass channels, but all have the same sensitivity (i.e. value of i in equation 1). Each single post-synaptic potential is insufficient to make the onset neuron fire: a spike on more than one AN-like input is required. The neurons used are leaky, so that these spikes need to be nearly co-incident in time. This tends to reduce the effects of noise (which might result in occasional but uncorrelated firing in auditory nerve-like inputs in adjacent channels). However, as the number of innervating channels is increased, the post onset evoked post-synaptic potential (EPSP) level can result in the onset cell firing.

A number of different models for depressing synapses have been put forward (e.g. [15]). The primary effect is that the first few spikes to arrive have a much larger effect than those that follow soon after. This is a form of onset enhancement. Hewitt and Meddis [16] suggested a form of depressing synapse at the inner hair cell to spiral ganglion dendrite synapse. We are not aware of work suggesting depressing synapses in the cochlear nucleus, but depressing synapses are very common in mammalian neural systems. We use a three reservoir model [15, 16], and this enhances the onsets in each spike train. The three reservoirs are presynaptic (available), cleft (in use), and reuptake (used, but not yet available again). The model parameters (the rates of transfer between each reservoir) are set so that the first few spikes result in near total depletion of the presynaptic reservoir. For a strong enough signal, spikes will arrive at approximately F_c spikes per second, where F_c is the centre frequency of the bandpass channel. However, an EPSP will only be generated for the first few spikes. The recovery time is set by the rate of transfer from the cleft to the reuptake reservoir (which we keep constant), and from the reuptake reservoir to the pre-synaptic reservoir. If this last rate is low, then there will need to be a considerable gap in AN signals before a new onset is marked. By adjusting this parameter, we can change cells from being sensitive purely to onsets to being sensitive to AM as well. If it is set too high, the post onset EPSP (i.e. the EPSP produced by an indefinite train of AN spikes) will be relatively high, resulting in unwanted onset firing. For simplicity, we set the maximal weight on each depressing synapse to the same level.

3. RESULTS

We first present results from a brief section of a TIMIT utterance [17]. We then investigate the relationship between onsets found and the phoneme structure using the TIMIT dataset.

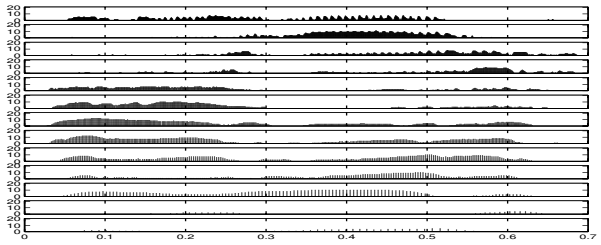
In figure 2 we show the effect of processing a section of a TIMIT utterance. The speech was filtered into 72 bands between 100 and 4000Hz, with 20 AN-like spike trains for each band, with a 3dB energy difference between each. Some of these spike trains are shown in figure 2a. Onsets occur at different times in different bands (see figure 2b). From this image it is also clear that the onset is generally found later in lower sensitivity bands (tracing the spikes in a single channel generally results in a line with positive gradient). This is due to the finite length of actual onsets (from the start of the sound to maximum intensity). Figure 2c shows a summary of these onsets. This was produced by merging together those onsets from the same channel but from different sensitivity bands which were judged to come from the same source by virtue of occurring at approximately the same time. This results in a considerable reduction in the total number of onset spikes, and is easier to use for analysing what in the signal is causing the onsets.

The TIMIT database [17] is a database of short read utterances in many US English dialects, and includes phonetic transcriptions. We have correlated the onset times found with the starts of the phonemes, and the results are shown in table 1. Phoneme onsets may be missed because the onset

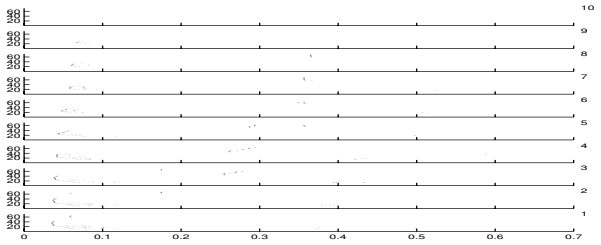
Phoneme type	uttered	identified	% correct
affricative	1227	1179	96.1
fricative	12494	9596	76.8
nasal	8302	2153	25.9
semivowel	11852	6438	54.3
vowel	33410	24691	69.9
stop	15047	11494	73.9
total events	82556		
false pos've	19022		
selectivity	0.745		

Table 1. Phoneme types in the 2700 TIMIT utterances processed (30% female), and those detected (within 28ms of recorded onset) by the onset detecting system. Selectivity is defined as (correct)/(correct + false positives).

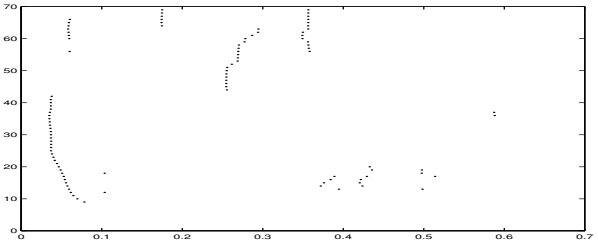
of this phoneme and the previous one overlap, or because that phoneme does not start with an onset. Many of the vowel, semivowel and nasals that are missed follow other voiced sounds: finding the frequency modulation that marks the change would require more or sharper filtering (as suggested recently [18]). 82% of the fricatives that are missed are v, th, f or dh. False positives may be found because



(a) AN-like spike output for 13 selected channels logarithmically spaced between 100 and 4000Hz (lowest in bottom subgraph). Each subgraph contains 20 horizontal traces, with a dot for each AN spike.



(b) Onset cell firings (one dot per spike). Here, each subgraph shows all the onsets found in a single sensitivity level, with low frequency channels at the bottom, and high frequency channels at the top. Highest sensitivity subgraph is at the bottom.



(c) Summary onsets (see text)

Fig. 2. Effect of processing a 0.7 second long extract from male utterance MJWT0SA1 from TIMIT dataset (2.57-3.27seconds).

a single onset breaks into more than one due to slow rise times, or because envelope variations inside a phoneme are misidentified as onsets. Two particular stops (dx and q) account for 68% of the missed stops: we believe that these stops are largely not associated with an increase in energy. Most of the false positives occur inside vowels, with the remainder inside sibilance or stops. The starts of almost all sequences of voiced sounds (vowel, nasal and semivowel) are found.

4. CONCLUSIONS AND FURTHER WORK

The system modelled resembles the biological system, and has some of the qualities of that system. The spiking AN-like representation provides an effective early representation over a wide dynamic range, enabling onset detection over this range. Because of the spiking nature of the system, the latency is essentially that of the filterbank: indeed, the onset pulses are essentially phase locked (see [6]). The onsets detected fit with an informal definition of an onset. We have investigated how this model's onsets correspond to phonemes in the TIMIT dataset: fricatives and affricatives are largely detected, as are the starts of voiced sequences. We believe that by using both onset and AM-onset together neurons we can improve on the detection of vowel onsets in [19] in terms of level dependence: this requires further investigation. Further, using the spectro-temporal onsets structure, and the AM-onset information we believe we will be able to characterise fricative, voiced and stop onsets. The model is currently implemented entirely in software: work on VLSI implementation is ongoing [12].

5. REFERENCES

[1] E.M Rouiller, "Functional organization of the auditory pathways," in *The Central Auditory System*, G. Ehret and R. Romand, Eds. Oxford, 1997.

[2] A. Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *International conference on acoustics, speech and signal processing*, 1999, pp. 3089–3092.

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

[4] C. Tait, *Wavelet analysis for onset detection*, Ph.D. thesis, Department of Computing Science, University of Glasgow, 1997.

[5] M. Cooke, *Modelling Auditory Processing and Organisation*, Distinguished Dissertations in Computer Science. Cambridge University Press, 1993.

[6] L.S. Smith, "Phase-locked onset detectors for monaural sound grouping and binaural direction finding," *Journal of the Acoustical Society of America*, vol. 111, no. 5, pp. 2467, 2002.

[7] R.J. McNab, L.A. Smith, D. Bainbridge, and I.H. Witten, "The New Zealand Digital Library MELody inDEX, http://www.dlib.org/dlib/may97/meldex/05witten.html," May 1997.

[8] M. Goto and M. Muraoka, "A real time beat tracking systems for audio signals," in *Proceedings of the 1995 international computer music conference*, 1995, pp. 171–174.

[9] L.P. Clarisse, J.P. Martens, M. Lesaffre, B.De Baets, H.De Meyer, and M. Leman, "An auditory model based transcriber of singing sequences," in *Proceedings of ISMIR*, 2002, pp. 171–174.

[10] M. Marolt, A. Kavcic, and M. Privosnik, "Neural networks for note onset detection in piano music," in *Proceedings of ICMC 2002*, 2002.

[11] J. Huang, T. Supaongprapa, I. Terakura, F. Wang, N. Oshnishi, and N. Sugie, "A model-based sound localization system and its application to robot navigation," *Robotics and Autonomous Systems*, pp. 199–209, 1999.

[12] N. Chia and S. Collins, "A spike based analogue circuit that emphasises transients in auditory stimuli," Submitted to ISCAS 2004.

[13] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," *Computer Speech and Language*, vol. 1, pp. 109–130, 1986.

[14] C. Koch, *Biophysics of Computation*, Oxford, 1999.

[15] M. Giugliano, M. Bove, and M. Grattarola, "Fast calculation of short-term depressing synaptic conductances," *Neural Computation*, vol. 11, pp. 1413–1426, 1999.

[16] M.J. Hewitt and R. Meddis, "An evaluation of eight computer models of mammalian inner hair-cell function," *Journal of the Acoustical Society of America*, vol. 90, no. 2, pp. 904–917, 1991.

[17] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993.

[18] C.A. Shera, J.J.Guinan Jr., and A.J. Oxenham, "Revised estimates of human cochlear tuning from otoacoustic and behavioral measurements," *Proceedings of the National Academy of Sciences*, vol. 99, pp. 842–846, 2002.

[19] R.W.L. Kortekaas, D.J. Hermes, and G.F. Meyer, "Vowel-onset detection by vowel strength measurement, cochlear nucleus simulation and multilayer perceptrons," *Journal of the Acoustical Society of America*, vol. 99, no. 2, pp. 1185–1199, 1996.