# 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
lss,dsf@cs.stir.ac.uk

## Abstract

We present a neurally inspired technique for detecting onsets and amplitude modulation in sound. This starts with a cochlea-like filter. The outputs from this 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. Amplitude modulation is detected in a way similar to that of onset chopper cells. We present results from some of the TIMIT database.

## 1 Onsets, onset cells, and onset chopper cells

The aim of this work is to provide features for sound source streaming and interpretation. Biological systems far outperform current artificial systems. Thus modeling aspects of the biological system seems a good way forward. Here, we model aspects of the cochlea, auditory nerve 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 auditory nerve responds more strongly, with many neurons in the cochlear nucleus 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 first onset in the received signal will normally be from the direct path, and further onsets caused by reflections will be smaller. Indeed, these are normally ignored by animals when they estimate the sound source location. (This is the Precedence Effect, or Law of the First Wavefront [2].) Other cues such as offsets are severely smeared out in time in reverberant environments.

Onsets are a form of envelope modulation. Some of the cochlear nucleus neurons sensitive to onsets are also sensitive to other forms of envelope modulation, in particular to amplitude modulation (AM) [3]. AM in particular ranges of frequencies is characteristic of animal vocalisations, including speech (as well as musical instruments), and is therefore of ecological importance. AM frequency is unaffected by reverberation, although AM phase and modulation depth can be affected.

### 1.1 Onset detection

Onset detection systems have been used in music transcription [4, 5], for sound segmentation [6], lip synchronisation [7], monaural sound source streaming [8, 9], and determining when to measure interaural time differences for sound direction finding [10]. On-line applications (e.g. real-time speech segmentation or source streaming, real time sound direction finding, or real-time music transcription), may use the sound only up to the time of onset, and the latency of the detector 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 have been used to segment hummed or sung notes [11] to improve the note differentiation capability of early music transcription systems (such as [12]). An alternative is to use first order difference based estimates, [4, 13], which take the maximum of the rising slope of the amplitude envelope as an index of onset. A variant of this is [5] which uses the relative difference, calculating $\Delta I/I$. Another variant is [14] which uses troughs in loudness to segment sung notes. A different approach uses optimal filter based techniques: [7] uses a wavelet based filter and [6, 15, 16] use the difference between a long-term and a short-term average. A related approach uses expectation based techniques [17] 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 for particular reverberation characteristics, and can perform better, but require a convolution, and can have a long latency.

### 1.2 Amplitude modulation and voicing detection

AM is an important source of information about certain types of sounds. For speech, the primary cause of AM in the bandpass filter output is unresolved harmonics of the fundamental frequency. This occurs for voiced sounds. Finding voiced parts of speech has a long history [18]. This source of AM provides a biologically plausible mechanism for voicing detection [19]. Onset chopper neurons (stellate cells) in the cochlear nucleus appear to be part of this process, as they amplify AM in the auditory nerve signal. Although voicing is not necessary for speech intelligibility (whispered sounds can still be understood), it is necessary for intelligibility in noisy situations.

### 1.3 Auditory mechanisms

The detailed mechanism of the auditory system's onset and AM responses are not clear. Onset enhancement starts at the auditory nerve, and this aspect appears to be related to the depletion of the neurotransmitter reserves at the synapse between the inner hair cell (in

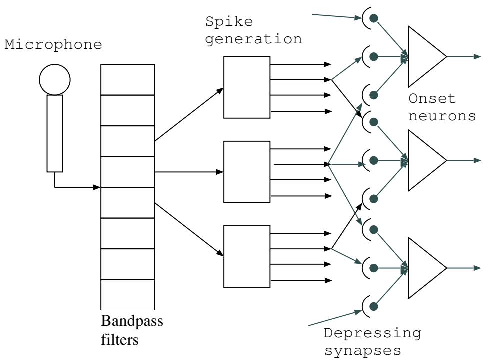


Figure 1: Onset spike generation system. AN-like spike generation is shown for only three bands. Depressing synapses and onset generation are shown for a single sensitivity level for these three bands.

the Organ of Corti) and the spiral ganglion neuron (see [20]). The onset response is much stronger in the auditory brainstem cochlear nucleus, where there are a number of cell types (octopus, and some bushy and stellate cells) which respond specifically to onsets [1, 3]. In addition, some onset cells also enhance AM. How this onset or onset chopping response is mediated is not known: it may be due to their synaptic innervation (many AN fibers converge on these cells), or to the nature of the synapses themselves, or to the form of the leakiness of these cells (i.e. to the particular ion channels expressed in their membrane), or to their morphology, or to some combination of these. Nor is precisely known how the outputs of these cells are used: they appear to innervate the MSO and LSO [3], both implicated in sound direction finding, as well as other auditory brainstem areas. Certainly, sensitivity to onsets and to AM is well documented at higher level auditory processing centers.

## 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 [21]. 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. The precise bands used are described in section 3. An important issue in the design of the filters 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 [9]. Other filters, such as Butterworth have a more constant delay.

The spike based representation we use enables the system to work over a wide dynamic range through the use of 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 (where the cycle is assumed to be at the filter centre frequency). If there are $N$ spike trains, these $E_i$ are set by

$$E_i = D^i E_0 \qquad (1)$$

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 also be generated in all bands $k'$ for $0 \leq k' \leq k$. This technique is similar to that used by in [22], where Ghitza noted that it led to an improvement in 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 which can detect AM as shown in section 3.

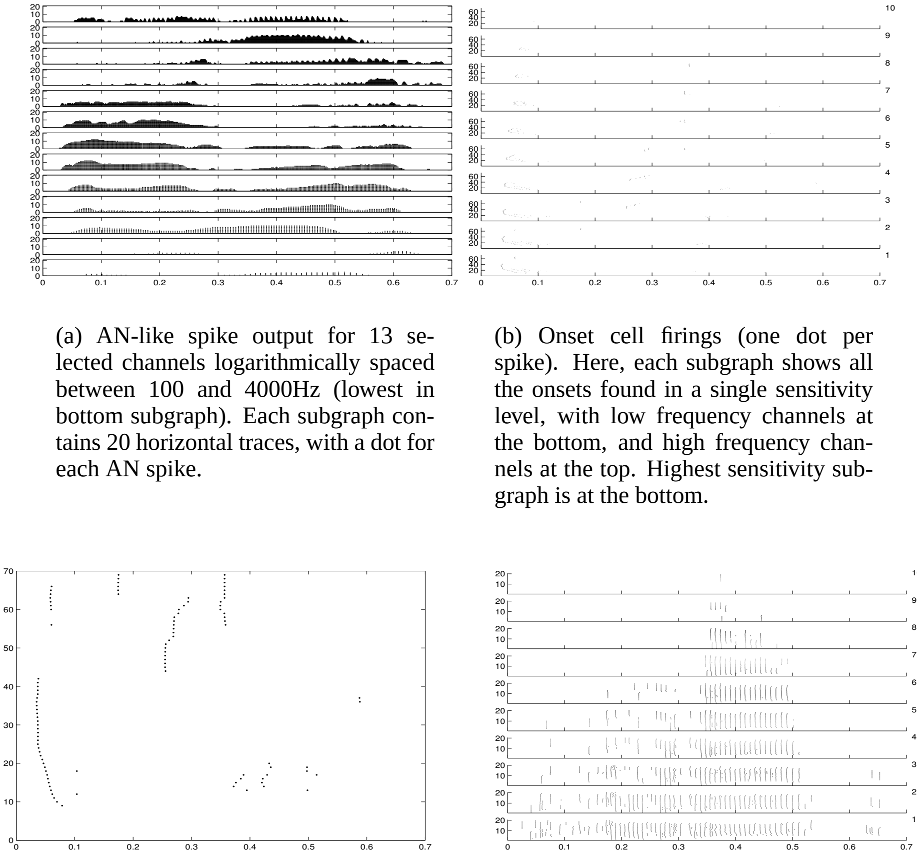
The AN-like spikes are applied to depressing synapses on onset neurons (figure 1) which are leaky integrate-and-fire (LIF) neurons with depressing synapses. LIF neurons are the simplest model neurons which maintain any semblance of the temporal behaviour of real neurons: see [23], chapter 14 for a review. 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, so that 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 if it is too large.

A number of different models for depressing synapses have been put forward [24–26]. The primary effect of all of them is that the first few spikes to arrive at a depressing synapse have a much larger effect than those that follow soon after. This is a form of onset enhancement. Hewitt and Meddis [20] 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, [20,25] and this enhances the onsets in each spike train. The three reservoirs are pre-synaptic (available), cleft (in use), and reuptake (used, but not yet available again). The model parameters (which are the rates of transfer between each reservoir) are set so that the first few spikes arriving 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 [27] and from a synthetic sound, We then investigate the relationship between onsets found and the phoneme structure of some of the TIMIT dataset.

In figure 2 we show the effect of processing a brief piece of a TIMIT utterance. The speech was filtered into 72 bands between 100 and 4000Hz. There were 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. It is clear that onsets occur at different times in different bands, as is visible in 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 (as measured from the start of the sound to maximum intensity). Figure 2c shows the 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. Figure 2d shows the effect of altering the parameters for the onset cell: here, the AN-like spikes input were from 20 bands between 2000 and 4000 Hz. For this image the reuptake rate for the depressing synapse was increased, so that the depressing synapse recovers more rapidly. Thus the gap between spikes required to cause larger EPSPs is reduced. In addition, the dissipation of the onset cell was increased (so that the onset cell is more of a coincidence detector). This also necessitated an increase in the maximal weight between AN-like spike train and the onset neurons. The overall effect is that the neurons detect the amplitude modulation in the signal.



(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.

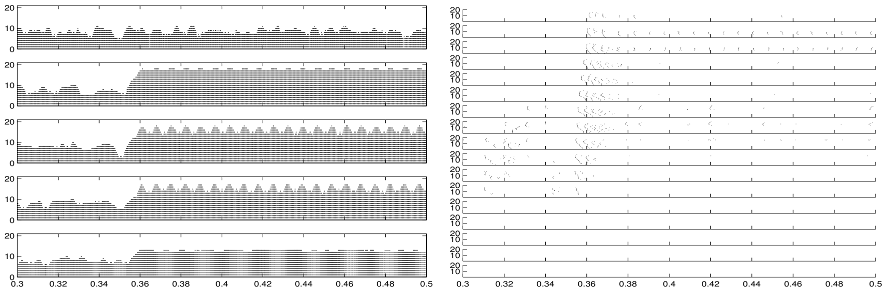
(b) Onset cell firings (one dot per spike). Here, each subgraph shows all the onsets found in a single sensitivity level, with low frequency channels at the bottom, and high frequency channels at the top. Highest sensitivity subgraph is at the bottom.

(c) Summary onsets (see text)

(d) AM detected by onset-like neurons (same format as (b)) for different frequency bands (see text).

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

A short tone complex consisting of 4 neighbouring harmonics of a 140Hz signal (between 2100 and 2520 Hz) on top of a white noise background was analysed the same way: see figure 3. High sensitivity channels show a pure onset response, and lower sensitivity channels (towards top of figure 3b) show both an onset response and a response to the envelope of the signal. During the tone complex, the AM has a modulation depth of approximately 60%. The AM-onset cells in the lowest two sensitivity levels fire reliably at the AM frequency. This behaviour mimics that of the onset chopper cells of the cochlear nucleus.



(a) AN spikes from 60ms before to 140ms after tone complex starts. Only the bands centred at 3.00, 2.48, 2.43, 2.38 and 2.00 Khz are shown.

(b) AM-onset cell spikes. The AM firing can be clearly seen in second and third top subgraphs

Figure 3: Processing of tone complex in background noise.

The TIMIT database [27] 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 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 [28]). The few fricatives that are missed are either just missed by a few milliseconds, or occur just beside a stop. Non-existent onsets may be found because a single onset breaks into more than one due perhaps

| Phoneme | Male | | | Female | | |
|---|---|---|---|---|---|---|
| type | uttered | identified | % correct | uttered | identified | % correct |
| affricative | 3 | 3 | 100 | 4 | 4 | 100 |
| fricative | 74 | 68 | 91.9 | 78 | 71 | 91.0 |
| nasal | 41 | 16 | 39.0 | 52 | 22 | 42.3 |
| semivowel | 124 | 64 | 51.6 | 148 | 66 | 44.6 |
| vowel | 219 | 153 | 69.9 | 331 | 232 | 70.1 |
| stop | 113 | 85 | 75.2 | 148 | 105 | 70.9 |
| false pos've | 99 | | | 114 | | |
| selectivity | 0.797 | | | 0.814 | | |

Table 1: Phoneme types in the 24 TIMIT utterances processed (12 male and 12 female), and those detected (within 28ms of recorded onset) by the onset detecting system. Selectivity is defined as (correctly identified)/(correctly identified + false positives)
.

to slow rise times, or because envelope variations inside a phoneme are misidentified as onsets. Two particular stops account for 75% of the missed stops: we believe that these stops are largely not associated with an increase in energy. Of the 213 false positives, 176 occur inside vowels, and 25 inside sibilances. The remaining 12 occur in stops. The starts of almost all sequences of voiced sounds (vowel, nasal and semivowel) are found.

## 4 Conclusions and further work

The system modelled has some similarity to the biological system, and 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 and amplitude modulation (even for relatively low modulation depth) over this wide 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 [10]). 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 almost all detected, as are the starts of voiced sequences. In addition, we have shown that the onset and onset chopper behaviours in the cochlear nucleus can be mediated either by relatively small changes in the parameters of the onset neurons, or simply by altering which AN-like spike trains innervate the model neurons. We believe that by using both onset and AM-onset neurons we can improve on the detection of vowel onsets in [29] in terms of level dependence: this requires further investigation. Further by using both the spectral areas in which onsets occur, and the AM-onset information we believe we will be able to characterise fricative, voiced and stop onsets. In addition, the improved onset time estimation should assist onset qualified ITD and IID estimation, improving sound source direction finding in reverberant environments. The model is currently implemented entirely in software: work on VLSI implementation is ongoing.

### Acknowledgments

The support of the UK EPSRC (grant number GR/R64654) is acknowledged.

## References

[1] J. O. Pickles. *An Introduction to the Physiology of Hearing*. Academic Press, 2nd edition, 1988.

[2] J. Blauert. *Spatial Hearing*. MIT Press, revised edition, 1996.

[3] E.M Rouiller. Functional organization of the auditory pathways. In G. Ehret and R. Romand, editors, *The Central Auditory System*. Oxford, 1997.

[4] J. Bilmes. Timing is of the essence. Master's thesis, Massachussetts Institute of Technology, 1993.

[5] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. In *International conference on acoustics, speech and signal processing*, pages 3089–3092, 1999.

[6] 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.

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

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

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

[10] L.S. Smith. Phase-locked onset detectors for monaural sound grouping and binaural direction finding. *Journal of the Acoustical Society of America*, 111(5):2467, 2002.

[11] 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.

[12] J.A. Moorer. On the transcription of musical sound by computer. *Computer Music Journal*, 1:32–38, 1977.

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

[14] 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*, pages 171–174, 2002.

[15] L.S. Smith. Sound segmentation using onsets and offsets. *Journal of New Music research*, 23(1):11–23, 1994.

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

[17] 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*, pages 199–209, 1999.

[18] O.O. Gruentz and L.O. Schott. Extraction and portrayal of pitch of speech sounds. *Journal of the Acoustical Society of America*, 21(5), September 1949.

[19] L.S. Smith. A neurally motivated technique for voicing detection and $f_0$ estimation in speech. Technical report, Centre for Cognitive and Computational Neuroscience, University of Stirling, Stirling UK, 1996.

[20] 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*, 90(2):904–917, 1991.

[21] 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.

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

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

[24] M.V. Tsodyks and H. Markram. The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability. *Proc. Nat Acad Sciences*, 94:719–723, 1997.

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

[26] R. Bertram. Differential filtering of two presynaptic depression mechanisms. *Neural Computation*, 13:69–85, 2000.

[27] 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.

[28] 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*, 99:842–846, 2002.

[29] 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*, 99(2):1185–1199, 1996.