ONSETS: AN ELEMENT OF ECOLOGICAL SOUND INTERPRETATION

Leslie S. Smith and Dagmar S. Fraser

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

ABSTRACT

We justify the usage of onsets in sound processing by appealing to an ecological view of auditory processing. The biological basis for onset processing is briefly discussed, and we describe our biologically motivated approach for a spike-based system for onset detection. This is based on a auditory-nerve like representation (with multiple spike trains per filter-bank band) followed by a leaky integrate-and-fire neuron with depressing synapses. Onsets are detected with essentially zero latency relative to the filter-bank. We show how this can be used to find the starts of certain phonemes in the TIMIT database, and how, by a small variation in the parameters, it can be used to detect amplitude modulation.

1. ECOLOGICAL SOUND INTERPRETATION AND ONSETS

Ecological sound interpretation is direct interpretation of sound in terms of characteristics (affordances, in the sense of [1]) that matter to the perceiver of the sound. For the perceiver, questions such as “do I run away or towards this sound source?” or “does this sound source mean I am about to fall into a river?” need answered. For some specialised animals, there are other questions as well: “should I eat what is in front of me” (for animals with biosonar, such as bats), or “where is my prey going”, for night hunting owls. Additionally, for animals that communicate using sound, questions of the meaning of the sound arise as well. (One can argue that meaning (or at least significance) is often present in other, non-animal sounds: indeed sounds with no significance can safely be ignored!) For humans, the *what* and *where* tasks (what are the sound sources, and where are they located) are often believed to underlie such questions.

From the animal’s point of view, ecological sound perception is direct: certainly humans are not normally aware of rationalisations intervening between the sensation and the perception. Granted, in some circumstances, such rationalisation may occur (e.g. listening to an unfamiliar foreign language, or trying to decide if the fan in the computer really is more noisy than it should be). Yet perception is clearly not direct in a physiological sense. Sound is reflected and refracted by the head and outer ear, passes through the auditory canal, is transduced into a wave in the fluid-filled cochlea, detected (in the sense of being turned into a neural signal) in the organ of Corti, passed to the cochlear nucleus and then transferred to the other nuclei of the auditory brainstem and midbrain before arriving at the auditory cortex where we might expect the perception to actually arise. From a computational viewpoint, we would like to emulate the apparent easy directness of ecological sound perception. How can this be achieved?

Truly direct perception, answering questions such as those in the first paragraph entails knowledge of motivation of the perceiver, and will be different for each perceiver, and possibly even different for the same sound at different times. What we are more interested in supplying is something that can underlie such direct perception, something that can perform the *what* and *where* tasks. Like light, sound interacts with all the surfaces it comes into contact with. Vision systems are generally concerned with surfaces that reflect light. Much of the visual system is concerned with providing invariant percepts independent of the illumination of these surfaces. We suggest that the auditory system is also concerned with providing invariant percepts: however, in this case, the system’s interest is in the sound sources, rather than in the surfaces sound reflects from. Further, sound from most sources reaches the perceiver both directly and by reflections: we suggest that a suitable goal for a computational auditory system would be to provide answers to the what and where questions which are invariant over the surfaces that the sound has reflected from.

This brings us directly to onsets: the onset of a sound at the perceiver will arrive from the most direct source: i.e. the path without reflections. Further onsets caused by reflections will be smaller than the initial onset. Thus, in terms of the where task, onsets provide information that may be masked by reflections later in the signal. Other cues such as offsets are severely smeared out in time in reverberant environments. Onsets are also interesting because, being at the start of (or at the start of some change in) a sound emitted from a source they can provide information rapidly. Fur-

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

ther, the nature of the start of a sound (voicing onset time in particular) has been found to be important in characterising certain phonemes. In addition, as discussed below, onsets are detected in the auditory system, making the use of onsets not only ecologically relevant but also biologically inspired.

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]. The last of these aims to answer the *where* question: here we are more concerned with material that underlies the *what* question. After describing onsets and onset detection, we show how onsets can be used for detecting certain phonemes in continuous speech, and how a similar technique can alternatively detect amplitude modulation.

2. ONSETS, AND ONSET (CHOPPER) CELLS

Auditory onsets are rapid increases in energy in some part of the audible spectrum. 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 above the just perceptible, 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 (stellate, bushy and octopus cells) in the cochlear nucleus also spiking strongly, at stimulus start [7]. Onsets are a form of signal envelope modulation. Some multipolar cochlear nucleus neurons sensitive to onsets are also sensitive to other forms of envelope modulation, such as amplitude modulation (AM) [7, 8]. By altering the parameters used, the system described here can alternatively detect AM.

2.1. Onset detection

In this work onset detection starts by bandpassing the sound signal into many bands. This 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. In this way, onsets which occur during other sounds can be detected. Further, 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 [9] to improve note differentiation in early music transcription systems. An alternative is to use first order difference based estimates, (e.g. [10]), 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 [11] 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, 12] use the difference between a long-term and a short-term average. A related approach uses expectation based techniques [13] 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. 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 may become important. Long latency is a serious drawback if knowledge of onsets is required instantly (for example, to trigger other processing). In addition, the absence of latency variation (for example due to overall signal strength) is important in applications requiring precise timing, such as direction finding.

3. 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 [14] 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. (There are four such levels in figure 1.) 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,

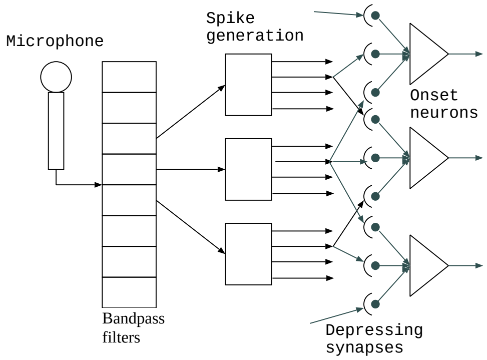


Fig. 1. Onset spike generation system. AN-like spike generation is shown only for three bands. Depressing synapses and onset generation are shown for a single sensitivity level for these three bands.

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 ($\sqrt{2}$), 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 [15], 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, and can be used to enhance amplitude modulation as shown in section 4.

The AN-like spikes are applied to depressing synapses on onset neurons which are leaky integrate-and-fire (LIF) neurons (figure 1). LIF neurons are the simplest model neurons which maintain any semblance of the temporal behaviour of real neurons: see [16], 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. [17]). The primary effect is that the first few spikes to arrive have a much larger effect than those that follow soon after. Where the signal from a continuous input is a train of evenly spaced spikes, this provides a form of onset enhancement. Hewitt and Meddis [18] 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 [17, 18], 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 (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, like onset chopper cells [7]. 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.

4. RESULTS

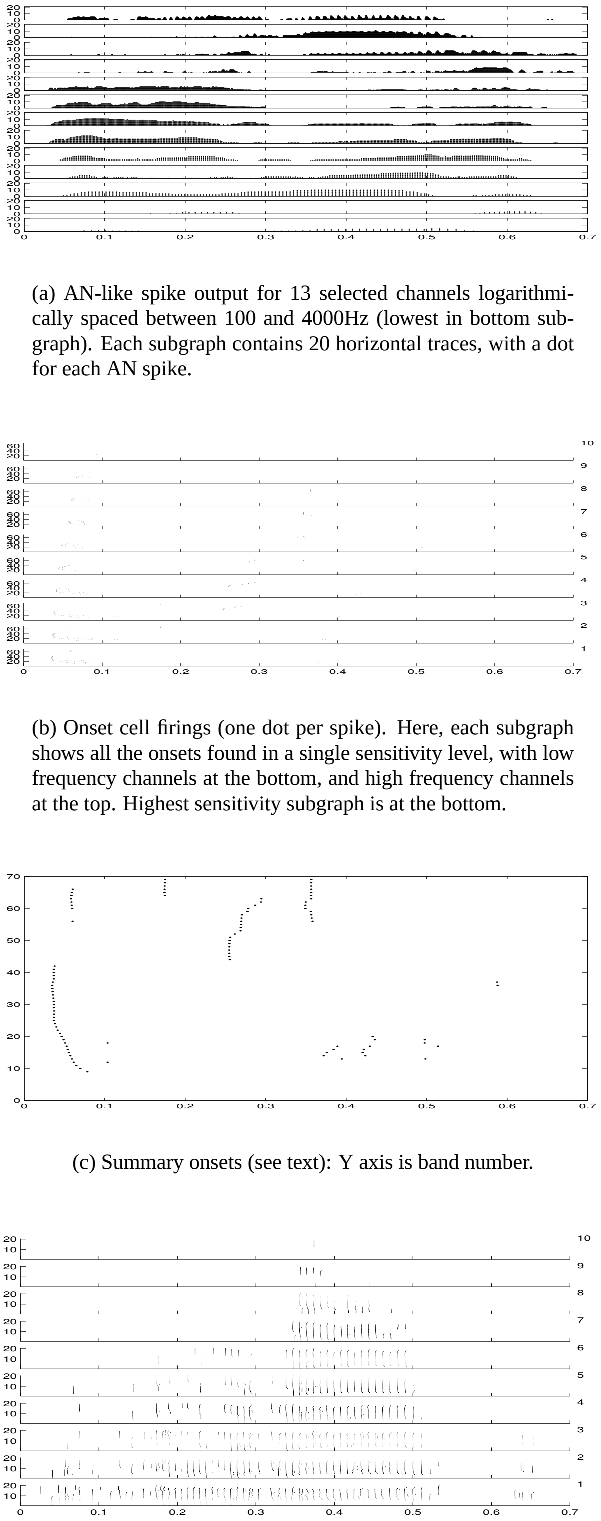
We first present results from a brief section of a TIMIT utterance [19]. We then investigate the relationship between onsets found and the phoneme structure using the TIMIT dataset. Lastly we briefly discuss how the parameters of the system, can be altered to allow AM to be detected.

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), rather than to onset latency being a function of signal strength. 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 [19] 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. There is a clear correlation between the types of phoneme and the onsets found, and very little variation between male and female. Phoneme onsets may be missed because the onset of this phoneme and the previous one overlap, or because that phoneme does not start with, or contain an onset. Many of the vowel, semivowel and nasals that are missed follow other voiced sounds, but sharper filtering (as suggested recently [20]) may allow these to be recovered. We note that 87% of the starts of sequences of voiced sounds (vowel, nasal and semivowel) are found. The fricatives missed are either just missed by a few milliseconds, or occur just beside a stop. Non-existent onsets may be found because a true onset is broken into multiple onsets. Envelope variations inside a phoneme are sometimes misidentified as onsets. This happens most frequently for vowels and results at least partly from the onset detector being confused by slow envelope modulation inside single vowels. The bulk of false positives, 83%, occur within vowels, with 12% inside sibilances. The remaining 5% occur in stops or at the beginning of the recording (due to extraneous recorded noise). Turning to stops, two particular stops, ’dx’ and ’q’ account for 75% of the missed stops: we believe that these stops are largely not associated with an increase in energy. If we consider the stop consonants (’b’, ’d’, ’g’, ’p’, ’t’, and ’k’) as in [21] the sensitivity of the system is 0.97, compared to their result of 0.93 at 30dB SNR. The overall selectivity (the ratio of useful to total detections) is defined as (true positives)/(true positives + false positives). Here it

Phoneme type	uttered	identified	% correct
affricative	2066	1972	95.4
fricative	21469	16664	77.6
nasal	14122	3789	26.8
semivowel	20179	11123	55.1
vowel	57379	41886	73.0
stop	25377	19312	76.1
total events	140592		

Table 1. Phoneme types in the 4620 TIMIT utterances processed (3260 male and 1360 female), and those detected (within 28ms of recorded onset) by the onset detecting system. Selectivity is defined as (correct)/(correct + false positives).



(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): Y axis is band number.

(d) AM detected by onset-like neurons (same format as (b)) for 20 frequency bands from 2000 to 4000Hz only.

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

has the value 0.75.

Amplitude modulation can be detected by allowing the depressing synapse in the onset LIF neuron to recover more rapidly (that is, by increasing the rate of transfer from the reuptake reservoir to the pre-synaptic reservoir). In addition, we reduce the refractory period to less than the minimum period for the AM frequencies of interest (here we reduce it to 3ms). The results of this processing, for 20 bands between 2 and 4 Khz, is shown in figure 2d: It is clear that the AM has been amplified considerably. The signal detecting the AM is now a pulse train at the frequency of then AM, rather like an onset chopper response.

5. 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 and AM 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.

Real sounds are complex, and often require rapid reactions from the perceiver. Multiple concurrent sounds are the rule, rather than the exception. Yet reactions are necessarily to a single sound source, implying an initial step of grouping the different elements of the sound from the foreground source prior to reacting to it. We suggest that common onset time and common amplitude modulation frequency are features suitable for use in this grouping process [22], and that the form on onset and AM detection here could partially underlie such grouping. We are working on developing this further.

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 [23] 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 [14]. We aim to incorporate this system in a larger real-time system for sound (including speech) detection and interpretation.

Do the features used here also contribute to ecological perception? This is hard to prove or disprove. However, it is the case that the nature of the onset (which bands it occurs in, what the tonotopic and temporal pattern of these onsets), and of the AM (which bands it occurs in, what is its frequency and modulation depth) are available almost instantly. If they are used in grouping, then it would seem reasonable to suggest that they would also be used in interpretation.

6. REFERENCES

[1] J.J. Gibson, *The ecological approach to visual perception*, Houghton Mifflin, 1979.

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

[8] I. Winter, A. Palmer, L. Wiegrebe, and R. Patterson, "Temporal coding of the pitch of complex sounds by presumed multipolar cells in the ventral cochlear nucleus," *Speech Communication*, vol. 41, pp. 135–149, 2003.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

[14] N. Chia and S. Collins, "A spike based analogue circuit that emphasises transients in auditory stimuli," Paper INV-3.5, presented at ISCAS 2004.

[15] 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.

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

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

[18] 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.

[19] 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.

[20] 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.

[21] G. Hu and D. Wang, "Separation of stop consonants," in *Proceedings IEEE Intl. Conf. on Acoust., Speech and Signal Proc.*, 2003, pp. 749–752.

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

[23] 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.