[Moore 77] Moore B.C.J., Introduction to the Psychology of Hearing, Macmillan, London, 1977.

[Patterson and Holdsworth 90] Patterson R., Holdsworth J., An Introduction to Auditory Sensation Processing, in AAM HAP, Vol 1, No 1, June 1990.

[Smith 93] Smith L.S., Temporal segmentation and localisation using onsets and offsets, CCCN Technical Report 16, University of Stirling, Stirling, 1993.

[Smith and Swingler 1993] The filtered associative network, Proceedings of ESANN 93, D Facto, Brussels, 1993.

[Watt 91] Watt R.J., Understanding Vision, Academic Press, 1991.

large number of simple tasks, experimenting with a wide range of possibilities, or one can return to the biological system for ideas. In particular we will consider abstract versions of the computations which appear to be carried out by the neurons of the cochlear nucleus, particularly onset and chopper cells. These will be used firstly in simple problems, e.g. note and musical instrument identification, using simple mapping neural networks. More complex problems will require the use of sequence processing networks (such as those discussed in [Hertz et al 1991], section 7.3, or in [Smith and Swingler 1993]). Used hierarchically, these also allow top-down influences on this otherwise bottom-up system.

References

[Blackwood et al 1990] Blackwood N., Meyer G., Aimsworth W., A Model of the processing of voiced plosives in the auditory nerve and cochlear nucleus, Proceedings IOA, 12, 10, 1990.

[Blauert 83] Blauert, Jens. Spatial hearing : the psychophysics of human sound localization, MIT Press, 1983.

[Bregman 90] Bregman A.S., Auditory Scene Analysis, MIT Press, 1990.

[Brown 92] Brown G., Computational Auditory Scene Analysis, TR CS-92-22, Department of Computing Science, Univesity of Sheffield, England, 1992.

[Diehl et al 91] Diehl R.L., Walsh M.A., Kluender K.R., On the interpretability of speech/nonspeech comparisons, J Acoust Soc Amer, 89, 6, 1991.

[Fowler 86] Fowler C.A., An event approach to the study of speech perception from a direct realist approach, Journal of Phonetics, 14, 3-28, 1986.

[Gibson 79] Gibson, J.J., The Ecological Approach to Visual Perception, Houghton Mifflin Company, Boston, 1979.

[Hertz et al 1991] Hertz J, Krogh A., Palmer R.G., Introduction to the theory of neural computation, Addison Wesley, 1991.

[Hewitt et al 91] Hewitt M.J., Meddis R., Shackleton T.M., A computer model of a cochlear nucleus stellate cell: responses to amplitude modulated and pure tone stimuli, J Acoust Soc Amer, 91, 4, April 1992.

[Koenderink 84] Koenderink J.J., On the structure of images, Biological Cybernetics, 50, 363-370, 1984.

[Meddis 88] Meddis R., Simulation of Auditory-neural transduction: Further studies, J. Acoust Soc Amer, 83, 3, March 1988.

[Meddis 92] Meddis R., Hewitt H.J., Modelling the identification of concurrent vowels with different fundamental frequencies, J Acoust Soc Amer, 91, 1, January 1992.

stimulus information in this case may be is not obvious: it may be periodicity, extracted using autocorrelation [Meddis 92], onset and offset detectors [Bregman 90], [Brown 92].

4 A programme of research

I propose breaking down the task of interpreting sound into two parts: locating segments of sound in time, and then interpreting these segments. Interpreting multiple sources I consider too difficult a problem for the time being.

4.1 Locating segments in time

I believe that the localisation of sound in time is a precursor to the more difficult problem of complex sound segmentation. The first task will therefore be to extract appropriate stimulus information to perform this task, and to demonstrate an effective solution to this task. This can be applied to simple non-overlapping sounds, such as isolated claps, and tones. The next stage is to apply more sophisticated versions of these techniques to the segmentation of a single stream of sound: that is a stream which produces a variety of sounds which abut but do not overlap, such a birdsong or speech.

The primary techniques which will be applied will be the use of onset and offset filters, applied to the filtered sound signal. A very simple form of this is described in [Smith 93]. The work there will need extended to use a range of filters, varying in temporal extent, and in frequency sensitivity.

4.2 Interpreting segments

Although one can assess a particular set of segmentation decisions with reference to the original sound, by simply listening to the segments produced by the system, and deciding whether they are appropriate, the real test of a synthetic system is whether it can perform some useful interpretation of the sound stream. Segmentation allows the stream to be broken down into pieces, each of which can be separately interpreted. (Of course, they need not be interpreted independently.)

To perform this interpretation, some form of characterisation of each segment is necessary. Clearly, the particular form of interpretation will strongly influence the characterisation used, depending on what it is that we want to know about the source of the sound. For example, attempting to interpret the melodic line of a flute would need a characterisation which included accurate information about the frequency distribution of the energy, and precise information about the relative lengths of each note and the spaces between them, whereas such accuracy is not important in finding out if the sound had a percussive, turbulent or regularly oscillating source. The appropriate stimulus information is different in each case.

The identification of appropriate stimulus information could proceed either by attempting a

by turbulence, the same technique may work, if it is a sound which has just started. However, if the sound is one from a constant source which the perceiver is travelling towards, then the increase in intensity will be very gradual, and the sound may well be classed with the ambient sound. A longer timescale would be appropriate here. For regular oscillations, if the sound is of short duration, then finding the location of its start in time by onset is again reasonable.

For the end of a sound, it is possible to consider the offset of the sound. It is, however, worth noting that for percussive sounds, such offsets are usually more gradual than the onsets. This is because of reflections of the sounds. For sounds produced by turbulence, offset is again a possible method for finding the end of the sound: again, for the situation where the offset is caused by the movement of the observer, offset will be very gradual. For turbulent and regular sounds, this is less marked. Another candidate for marking the end of a sound is the start of another sound. This is reasonable for sounds from single sources which cannot produce more than one sound at a time: many sources are of this form.

From a biological perspective, there are certainly onset and offset detectors in auditory cortex: e.g. the octopus cell [Blackwood et al 1990], and these cells could act as markers allowing temporal location of sounds. In auditory cortex there appear to be many onset and offset cells, with inputs from different parts of the auditory nerve, so that they can detect onsets in particular ranges of frequency of sound. These could be able to detect both onsets of sounds and changes in the frequency content of a sound.

Temporal location of a sound is a forerunner to the rather more complex problem of sound segmentation: in segmentation, an extended complex sound needs to be divided up into a number of shorter segments. Our belief that speech and sound perception share low-level primitives suggests that the techniques for such segmentation should use the same set of primitives. We refer here only to bottom-up segmentation: that is, we consider that top-down effects (such as the nature of whole words and utterances) have their effect on what has already been started by bottom-up methods.

Given temporal location, one can then consider what useful stimulus information can be extracted from these pieces (or segments) of sound. As perceivers, we are aware of pitch, timbre, non-pitched sound, as well as the duration of sound. There is an enormous psychophysical literature on what we can and cannot perceive from sound. Given this, it would be rash to make particular suggestions for how humans perceive sound. It is perhaps better to consider how one might judge whether some particular measurement was useful or not, for a synthetic sound perception machine. To this end we make the suggestion that such a measurement is useful if it can be incorporated into the synthetic machine, and provide useful discrimination between sounds. This is a circular argument: the measurement is useful stimulus information if it is useful in the machine. However, there is no intrinsic measure of usefulness: it is in the mind of the perceiver, or in the results of the synthetic apparatus.

Sounds do not occur in isolation. They overlap in time, and possibly in location as well. Yet people easily distinguish different streams of sound even when they come from one direction (e.g. distinguishing music from speech coming from a mono radio). A binaural system can distinguish sounds from different directions precise timings of the sounds at the two sensors. But a monaural system, or a system trying to distinguish different sounds from the same direction must use information in the sound signal itself. Exactly what the useful

always wish to ignore these sounds.

Next, consider a percussive sound. One (hard) object strikes another (hard) object, causing both of them to vibrate. The vibration will die away rapidly, in both objects. Thus the sound is necessarily of short duration. Both the way in which the objects start to vibrate, and the way in which the vibration dies away will be characteristic of these particular objects. The intensity of the vibration will depend on the objects and the violence of their clash, but to the perceiver, it will also depend on his distance from them. The way in which the vibration dies away depends entirely on the objects themselves, but the perceived decay of the sound will also depend on the different paths by which the sound reaches the perceiver. Thus, the actual physical sound of a handclap when heard outside in the open is quite different from the same when made in a small room. We do hear both as a handclap, but we can also extract useful information about the environment from the sound as well.

Now consider a sound produced by turbulence. This may be of short duration (e.g. the sound *sss*), or of long duration, as produced by a river flowing. Again, the perceived sound will be affected by the environment. Additionally, if the sound is from a stationary source (like a river), and the perceiver is moving, then the variations in intensity over an appropriate time scale can give useful information about where the perceiver in relation to the source.

Regular oscillation is more rarely produced in nature, except by living animals. When it is produced by animals, it is usually of relatively short duration, or is part of a varying string of sound produced by the animal. Indeed, animal sounds are frequently characterised by being a mixture of short regular oscillations, interspersed with short sounds produced by turbulence. Again, the perceived sound will depend on the the reflections and diffractions introduced by the environment, but the regular element of its nature (and the nature of the mixture of regular and turbulent airflows) will be undisturbed.

3.1 Extracting useful stimulus information from sound

The first useful piece of information to extract is that there has been a sound. To characterise it, one would like to locate it in time and space from the ambient sound, and then interpret it. We shall not consider spatial location here: this is a problem which has been analysed from a physical point of view by [Blauert 83], and on which considerable work has been done in neuroanatomy. This is not to minimise its importance.

Locating a sound in time can be done at a number of scales. For example, at a long timescale, one might locate a complete birdsong or utterance, and at a shorter timescale, one might locate particular notes of the birdsong, or the constituent sibilances, vowels, etc. of an utterance. This possible range of scales is reminiscent of the situation in vision, where blobs at a low spatial resolution break down into sets of blobs at higher resolutions [Koenderink 84] [Watt 91] page 180. The appropriate timescale will depend both on the sound, and on what matters about the sound to the listener.

For locating a single percussive sound in time, one can consider that the start of the rapid increase in intensity of the sound marks the start of the sound in time. For a sound produced

Instead, we are saying that the human interpretation of the sound will depend on the associations between the character of the sound (gleaned from the extraction of appropriate stimulus information), and the associations of the sound itself. Knowledge of the exact production mechanism is only rarely useful. On the other hand, the ability to associate stimulus information which reflects sound production and transmission with the multi-modal associations of the sound itself is frequently useful, since it allows sounds produced in similar ways to be similarly interpreted.

2.2 Approaches to extracting appropriate stimulus information

There are two ways of considering this problem: one can look at how biology has solved the problem (via anatomy and neurophysiology), or one can consider the problem itself, in terms of the sounds produced. A very considerable amount of work has taken the former approach, most notably [Patterson and Holdsworth 90], [Blackwood et al 1990], [Meddis 88], [Hewitt et al 91] in the UK. The latter approach has usually only been taken with respect to speech. There, workers have considered both the spectral composition, and also the effects of the movements of the speech making apparatus (motor theory), reviewed by [Moore 77], page 223. For non-speech sounds [Diehl et al 91] does discuss this, but points out that the range of possibilities can be indefinitely large. In fact, top-down influences will be strong here: one might identify the sound of a hammer crashing on to an anvil, or the sound of birdsong. The former describes the actual process of the generation of the sound, whereas we are (generally) not interested in the production mechanism of the latter. Yet the process of initial stimulus information extraction is the same.

We propose merging the two approaches. Firstly, we propose considering the way in which ordinary (i.e. non-speech) sounds are made, and using the resulting analysis to inform processing methods, and secondly using the work on cell responses in the cochlear nucleus to develop appropriate filters for extracting useful information. We believe that the lowest levels of speech analysis and the lowest levels of non-speech sound analysis are the same, and that the techniques used in the ear reflect a compromise between providing a useful characterisation of the sound as a physicist might, and what is possible in biological terms, strongly influenced by the extraction of useful stimulus information.

3 The stimulus itself: sound

Consider first, constant ambient sound, such as might be found in a quiet office or other room. Unless the room is remarkably well insulated for sound, there will be quite a considerable amount of ambient sound, from fans, distant noise from cars, creaking of furniture, wind, rain etc. To a physicist, such noise is not constant sound, even if it has only a low intensity. Its general character is either that it is near constant in intensity (e.g. fans), changing only gradually, or else that is is not constant at all (e.g. wind, or creaking furniture), but is sound that we are so used to hearing that we have learned to ignore it almost entirely. It is possible to devise a synthetic system which can ignore the former type, but it is neither possible nor desirable to devise a low level system which can ignore the latter, since the perceiver may not

help produce a synthetic method for interpretation of sound, we need to consider how one may interpret the stimulus information in sound towards these ends. This leads away from a directly physics-based method of interpretation, where it is the precise characteristics of the pressure waveform which are analysed, and then hopefully connected to the interpretations we would like to use towards an approach where we hope to extract more directly the information from the sound which will be important for answering these questions. It can be summarised by asking *What can the sound tell us about the source?*

We do not, however, adopt a direct realist stance [Fowler 86]: clearly how the sound was produced, reflected, diffracted, etc., will characterise the sound, but this is not necessarily what we want to know about the sound. As [Diehl et al 91] comment, it is the event associated with the sound that generally matters: 'For example, on hearing a lion's roar...direct acquaintance with a lion's vocal tract structures seems completely irrelevant'. A similar situation applies in vision: on looking at a picture, we are not interested in how the pigments reflect the light, but in how the colours group together to form an image. The question of the most appropriate level of interpretation cannot, we suggest, be decided purely from the incoming stimuli, but needs consideration of higher-level tasks at hand. Be that as it may, the actual initial extraction of stimulus information is not itself task-dependent.

2 Extracting appropriate sound stimulus information

2.1 What does *appropriate* stimulus information mean?

Appropriate stimulus information should allow the sound to be useful to the perceiver. That is, it should help the perceiver to learn about their environment, to communicate, to find out about the presence of predators or food, etc. It is unlikely that any single set of parameters about sound will fill all of these categories but one can make some observations.

1. So far as predators and food are concerned, it will be changes in the sound that are likely to be important, either as the food or predator moves, or as the perceiver moves.
2. Localisation of sound is clearly important if the source of the sound is to be fled from, or to be eaten.
3. For finding out about the environment identification of the different sources is important. Thus the stimulus information should be able to be used to separate out different types of source. Clearly localisation is important here as well.
4. For communication, the important stimulus information depends entirely on how what was to be communicated was modulated on to the source. However where it is a biological system which performs the modulation, it is unlikely to require methods of stimulus information extraction completely unrelated to identification and localisation. This need not be true for synthetic modulation systems (e.g. frequency shift keying).

Although we are saying that the exact nature of the sound will depend on the way in which the sound was produced and transmitted, we are not arguing for direct perception of this.

1 Towards an ecological approach to sound perception.

In J.J. Gibson's seminal work [Gibson 79], he develops what he calls an *ecological* approach. He carefully distinguishes the stimulus (i.e. the light energy falling on the retina) from the stimulus information, the information contained in the variation in brightness (and presumably colour) in the ambient light falling on the retina. Visual perception is *not a response to a stimulus but an act of information pickup* ([Gibson 79] page 56-7). In particular he cites light arriving through dense fog as an example of stimulus without stimulus information[1].

The ecological approach is more than this: it asks the question: what is the (ecological) niche of vision (or sound)? Perception is about perceiving events, events which take place outwith the perceiving organ. Gibson attempts a classification of such events as they affect vision. We consider the same problem, but for events as they affect sound perception. Thus, we can ask: what are the events that cause sound to be perceived?

These are (and the list is sure to be incomplete)

1. Percussion: two objects collide and vibrate, the vibrations are imparted to the medium (air), causing sound energy to be radiated.
2. Turbulence: gas or liquid moving in a turbulent way, so that rapidly moving eddies are created, and these movements are imparted to the medium (air) causing sound energy to be radiated.
3. Regular oscillation: an object moves in a regularly repeating fashion, and these vibrations are imparted to the medium (air), causing sound to be radiated.

These are often combined with each other: e.g. in a piano, a hammer percussively strikes a wire, which then oscillates in a regular way.

In addition to these primary sources of sound, there are reflections. Solid surfaces reflect sound well, and sound also diffracts round corners. Thus, the sound that reaches the hearer will come from passive reflective and diffractive sources as well as the active types of source listed above.

In the same way that Gibson considers ambient light, there is ambient sound: however, there are differences from vision, since people actually generate sound even when not trying to (due, e.g. to blood flow in the ear), so that a room with no other sources of sound and no sound leakage from outside will not be totally silent once a listener enters it. It is worth noting that such rooms are unusual, and need to be specially built, whereas excluding ambient light is straightforward.

The other side of the question of the ecological niche of sound, is what is (the perception of) sound useful for? We use sound for many purposes, for learning about our environment, for communicating (via speech and music), for learning about the presence of other people and animals and for many other purposes. If we wish to use the ecological approach to sound to

[1] What is the aural equivalent of such a dense fog?

Position Paper: Processing sound for interpretation.

Dr. Leslie S. Smith
Centre for Cognitive and Computational Neuroscience
Departments of Computing Science and Psychology
University of Stirling
Stirling FK9 4LA, Scotland, UK

CCCN Technical Report CCCN-17

September 24, 1993

Abstract

By taking an ecological view of sound perception strongly influenced by J.J. Gibson, we consider what the sound itself can usefully tell us about the source. We are therefore interested in extracting appropriate stimulus information. Since we also believe that low-level human auditory processing is independent of the nature of the sound, be it speech or not, we are interested in how the low-level human auditory system extracts such information. The eventual aim of this work is synthetic sound interpretation systems which can perform the simple tasks which we take for granted as humans. A brief description of a programme of research is included.