AN INTEGRATED TREATMENT OF AUDITORY KNOWLEDGE IN A MODEL OF SPEECH ANALYSIS

M. P. Cooke, M.D. Crawford & G.J. Brown

Department of Computer Science,
University of Sheffield, England

ABSTRACT - This paper addresses the question of how information gleaned about auditory processing from experimental disciplines in hearing may appropriately be incorporated into computational architectures for automatic speech recognition (ASR). Criteria for so doing are developed, based on a software engineering analogy. We present an overview of what we believe to be a coherent system for auditory processing, and illustrate the representations produced at each stage of the model.

INTRODUCTION

In this section we will pursue an analogy between speech research and software engineering in order to demonstrate how auditory knowledge might be applied in a model of speech analysis. The predominant concern in modern software engineering is the reduction of complexity. The key weapon against complexity is the notion of **abstraction**. In software engineering this enables the designer to maintain a conceptual grasp of the whole system during the creative process. In particular, use is made of **procedural abstraction** (**what** a system does rather than **how** it does it) and **representational abstraction** (in which the concrete realisation of some data structure is postponed).

Consider how the problem posed by speech recognition might be viewed in these terms. We might view speech research as adopting an approach somewhere between the two extremes depicted in Figure 1.

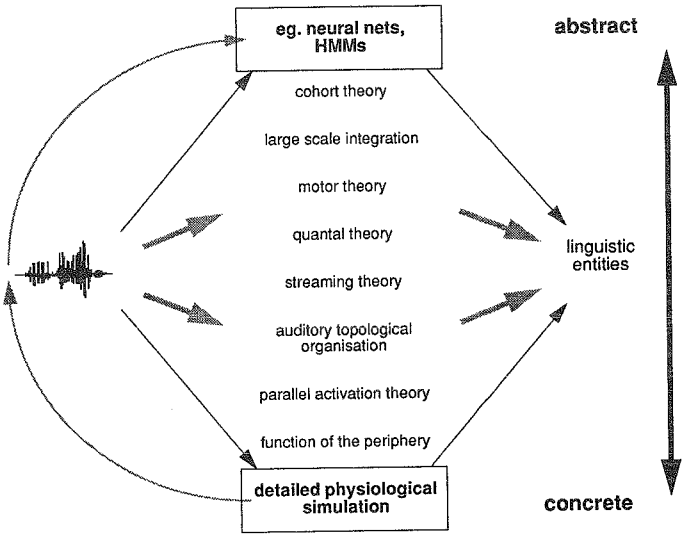

Figure 1: A spectrum of approaches to incorporation of auditory knowledge.

At the abstract extreme, the problem is simply stated: find some supposed or actual linguistic entities (e.g. words) in the incoming signal, without regard for the intervening processes. Much work in ASR can be placed in this category. In this approach, **no** knowledge of how the human accomplishes the task is

required. At the other extreme, we might aim for precise physiological simulation of auditory processes. Of course, such a goal is likely to be unattainable in practice (although some attempts based on the 'neural wiring diagram' of the DCN are being pursued e.g. Pont, 1990).

Neither approach helps us to understand how the auditory system actually solves the problem of sound analysis. It is, we believe, rather unfortunate that virtually all current work in speech technology falls very close to the abstract end of the spectrum. In Figure 1, we have attempted to show that what might appear to be a continuum might instead be viewed as a full circle since there is little real difference in the quality of explanation provided by, for example, a neural network and that available from detailed physiological simulations. As Fant (1989) has said of knowledge-free approaches - "We leave it to the computer to learn what we have failed to understand."

What then constitutes **auditory knowledge** ? We will use the term loosely here to cover everything from formulae describing the cochleotopic mapping to fully-developed theories of sound perception. Between the two extremes depicted in Figure 1, we have listed some bodies of auditory theory which might be deployed. This is just a selection, and some criteria for determining the applicability of these ideas, both on individual merit, and in terms of how the pieces fit together, are required. Some such conditions are developed below. The important point is that these bodies of knowledge exist and can, in some cases, be applied to the problem of building a speech recogniser. Whilst there are no guidelines as to when some piece of auditory knowledge is of sufficient stature to be incorporated in a speech recognition architecture, it can be argued that useful insights into partially-developed theories can be gained by application to the thorny problem of real speech recognition. We suggest, following the software engineering metaphor, the following criteria to be used in selecting fragments of auditory knowledge for incorporation into speech processing architectures:

Coherence of the interface - few pieces of auditory knowledge cover the whole process from signal to linguistic symbol; hence, we have to ensure that the pieces we select are coherent. As an example of incoherence, a common use of auditory knowledge is the selection of an initial filtering based on supposed auditory filter frequency resolution and an auditory frequency scale (of which there are many; e.g. mel, Bark, ERB, Greenwood). If an accurate measure of auditory resolution is incorporated (e.g. ERB, Moore & Glasberg, 1983), individual harmonics in the region below around 1200 Hz will be resolved. Unless the following processing stages take account of this, recognition rates may decrease. As Beet (1990) suggests, simplistic attachment of a conventional 'back-end' matcher to an auditory 'front-end' is not the answer.
Procedural abstraction - in order to model components, we have to have some meaningful description of their function, i.e. a definition of the transformation being computed. However, this does not require detailed knowledge of how processes operate. It is important to be as abstract as possible.
Representational abstraction - again, at some appropriate level of abstraction, we should be explicit about the sorts of representations which processes require and produce.
Testability - the model should be testable in two senses; conformance of model to theory; and effectiveness of the model (and thus theory) itself. For example, we might test the adequacy of some suggested auditory representation by using it as part of a resynthesis procedure. We reject the 'rush-to-recognition' employed in most ASR research, and instead need to develop alternative testing criteria.

The reader familiar with Marr's computational theory of vision (Marr, 1982) will notice the similarity between the concerns expressed there and here. There is clearly much in common between the two viewpoints and maybe further analogies are revealing. For example, in software engineering, the representational abstraction often dominates (i.e. the way we process depends on what we are processing). Is the same true in computational models of hearing ? Do we allow the terminology of the field (representational abstractions such as formant, harmonic, onset) to dominate the way we think about auditory processing ?

AN INTEGRATED MODEL

Figure 2 shows our first attempt at devising an integrated model of auditory processing in which components have been selected with respect to the above criteria. The central column of the figure describes processes, whilst to the left the input and output representations are specified. A possible testing strategy for each module is indicated below each process.Those bodies of auditory knowledge which

may contribute to the development of each stage are detailed in the right part of the figure. We do not regard this model as the final word in integrated models of auditory processing; however, it does serve to illustrate the attention which we believe should be given to incorporation of auditory knowledge in an integrated system.

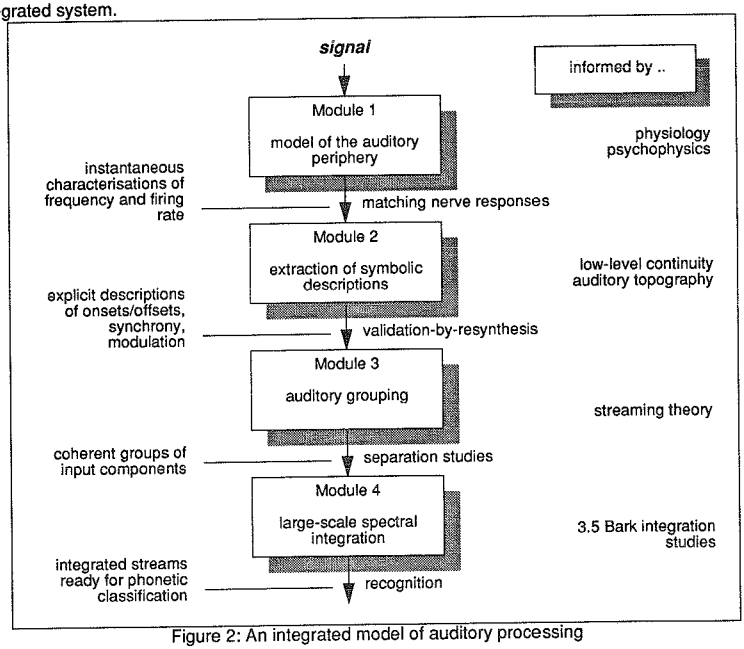

Figure 2: An integrated model of auditory processing

We have been working on various aspects of this model in Sheffield. The various auditory notions implied by this model, together with a description of our modelling progress to date, are reviewed below.

Module 1: From signal to auditory-nerve: models of the auditory periphery

Good first-order descriptions of the peripheral auditory system are available, to the extent that building a model of them is a fairly realistic proposition. Indeed, a large number of such models now exist. The model we have developed is described in detail in Cooke (1989). The model consists of a number of parallel, frequency specific channels, each of which models the function of the peripheral auditory system at some point along the basilar membrane, together with proximate structures. A single channel has three stages; a bandpass filter, a saturating static nonlinearity, and a hair cell transmitter depletion stage. The filter has an impulse response which fits data from measurements of auditory nerve fibre tuning curves (de Boer & de Jongh, 1978) and a frequency response which is very close to psychophysical estimates of human auditory filter shape (Patterson et al., 1987). This structure - the so-called gammatone filter - is rapidly becoming a standard tool in auditory modelling (at least in the UK).

The second stage models the nonlinearity involved in transforming basilar membrane displacement to inner hair cell receptor potential. We use an analytic expression from Crawford & Fettiplace (1981) which, when plotted on a decibel scale, resembles the unadapted rate-intensity function describing firing rates in the auditory nerve at stimulus onset (e.g. Smith & Brachman, 1982). The final stage is an attempt to model the transformation between inner hair cell receptor potential and the auditory-nerve fibre response. One of the goals of the model was to incorporate additivity of response, in which a change in response produced by an increment in stimulus level is independent of the amount of prior adaptation. The resulting model is fully analytic for constant input levels and provably additive for certain types of signal.

Module 2: Extracting useful descriptions from the auditory nerve

We have adopted two distinct approaches to this crucial stage of transformation. The first is based on the use of local constraints to group together similar responses, whilst the second is inspired by physiological descriptions of auditory topographic organisation.

Local constraints -There is much redundancy in information flowing along the auditory nerve. Many fibres show temporal responses which are phase-locked to some stimulus component, and neighbouring groups of fibres have similar responses in this respect. Furthermore, stimulus components change relatively slowly over time. Likewise, stimulus onsets and offsets affect a number of channels simultaneously. The goal of the processing here is to create concise descriptions of similar responses. To date, we have worked mainly on descriptions of synchronised activity and on the characterisation of onsets. Specifically, we can compute two forms of representation - *synchrony strands* and *onset groups*. The former is a time-frequency description of the signal, in which channels with similar temporal responses are grouped first across frequency, then in time, to produce the sort of representation depicted in Figure 3. Onset groups are the result of a process in which local peaks in spike rate derived from the hair cell model outputs are grouped across frequency. Both processes are described in more detail in Cooke (1990) and Cooke & Green (1990).

Auditory maps - Rather than rely on intuitive notions of grouping using local constraints, it is of interest to speculate on the kinds of representation which may be formed in the higher auditory system. One possibility suggested by physiological investigation is that information is transformed into the topography of a neural array. Such maps appear to consist of a parallel array of neural processors that are tuned to slightly different values of the same parameter, so that there is a systematic, place-coded representation of that parameter across the map. A variety of such maps have been found, with the parameter mapped including interaural time difference, interaural intensity difference, intensity and best amplitude modulation rate. A simplistic view is to suppose that each of these maps represents an alternative to the usual 'neural spectrogram' and may be useful, individually or jointly, in processing stimuli such as speech. Our initial work on the implementation of a map for amplitude modulation and its use in fundamental frequency estimation is described in Brown & Cooke (1990). An amplitude modulation map is shown in Figure 3.

Module 3: From descriptions of components to coherent groups: streaming theory

It should be clear that the processes we have described up to this point will provide a relatively rich and complete representation of the incoming signal (in fact, we have demonstrated that synchrony strands alone can form the basis for almost perfect resynthesis of speech). As such, we are still faced with the problem of separating out parts of the signal which belong together. The next stage of our integrated model is an attempt to embody what has been called 'auditory scene analysis' (Bregman,1990) into the process of speech recognition in realistic environments. Bregman and others have suggested, with a good deal of experimental support, that the auditory system groups together sound components if there is some evidence that they have arisen from the same acoustic source. For example, components may be grouped if they share some property such as onset/offset time, amplitude modulation, or proximity in frequency to some later component. The resulting grouped structures are called *streams*. Streaming theory is quite well developed, and we have for some years been building towards a computational model of some of these processes (Cooke, 1986; Williams et al., 1990; Cooke & Green, 1990). Indeed, an adoption of such ideas demands that we construct representations of signals couched in the same descriptive vocabulary as that used by the experimentalist, hence the desire to make the temporal evolution of tonal components and such things as onsets explicit in the way described. To date, we have implemented a single grouping algorithm which attempts to label synchrony strands based on common harmonic spacing and motion. This is described in Cooke & Green (1990). Figure 3 demonstrates the result of applying this algorithm to the strands shown.

Module 4: Speech-specific analysis of coherent groups: large-scale spectral integration

The processing up to this point has, ideally, accomplished the task of separating the complex mixture which reaches the periphery into groups of components which appear to belong together. We still have to determine the linguistic content of any speech-like groups. For reasons described in some detail in Crawford & Cooke (1990), we prefer the notion that the auditory system, at some level, performs large-

scale spectral integration prior to classification into some linguistic unit. Such integration is likely to be a post-streaming process. We are not quite at the stage where integration works on separated groups of strands. However, an example of its action on the strands shown in Figure 3 is depicted at the bottom of that figure. Such integrated strands can be viewed as 'auditory formants' (Karjalainen, 1987).

SUMMARY & FURTHER WORK

We have outlined suggestions for an integrated treatment of auditory knowledge in a model of speech analysis. Work is progressing on modules 2, 3 and 4. We are quite close to the stage where speech can be processed through the whole system (with a limited amount of grouping in module 3) and intend to test the ability of the system to separate speech from structured corruptions (such as other speech). Clearly, the process of building a sophisticated model of auditory processing is a long-term project, but it is necessary to keep the whole system in mind. We have made a start on defining the computational framework within which new results from experimental disciplines in hearing can be incorporated.

ACKNOWLEDGEMENTS

MPC is supported by UK SERC grant GR/E42754, whilst MDC and GJB are supported by SERC CASE awards and British Telecom Research Laboratories.

REFERENCES

Beet, S.W. (1990) *Automatic speech recognition using a reduced auditory representation and position-tolerant discrimination*, Computer Speech & Language, **4**, 1.

de Boer, E., de. Jongh, H.R. (1978), *On cochlear encoding: potentialities and limitations of the reverse-correlation technique*, J. Acoust. Soc. Am., **63**, 115-135.

Bregman, A.S. (1990), *Auditory Scene Analysis*, (MIT Press: MA).

Brown, G.J., Cooke, M.P. (1990) *A computational model of amplitude modulation processing in the higher auditory system*, these proceedings.

Cooke, M.P. (1986) *Towards an early symbolic representation of speech based on auditory modelling*, Proc. Institute of Acoustics Autumn Conf., **8**, 7, 563-570.

Cooke, M.P. (1989) *The auditory periphery: physiology, function and a computer model*, University of Sheffield Department of Computer Science Research Report CS-89-32.

Cooke, M.P. (1990) *Synchrony strands: an early auditory time-frequency representation*, University of Sheffield Department of Computer Science Research Report CS-90-05.

Cooke, M.P., Green, P.D. (1990) *The auditory speech sketch*, Institute of Acoustics Autumn Conf., Windermere, UK.

Crawford, M.D., Cooke, M.P. (1990) *Speech perception based on large-scale spectral integration*, these proceedings.

Crawford, A.D., Fettiplace, R. (1981) *Nonlinearities in the response of turtle hair cells*, J. Physiol., **315**, 317-338.

Fant, G.(1989) *Speech research in perspective*, Proc. EUROSPEECH '89, Paris, 3-4.

Karjalainen, M. (1987) *Auditory models for speech processing*, Proc. 11th Int. Cong. Phonetic Sciecnes, Tallinn, USSR, paper PL 2.1.1.

Marr, D. (1982) *Vision*, (Freeman).

Moore, B.C.J., Glasberg, B. (1983) *Suggested formulae for calculating auditory filter bandwidths and excitation patterns*, J. Acoust. Soc. Am., **59**, 750-753.

Patterson, R.D., Nimmo-Smith, I., Holdsworth, J., Rice, P. (1987) *An efficient auditory filterbank based on the GammaTone function*, Meeting of the Speech Group of the Institute of Acoustics, RSRE.

Pont, M.J. (1990) *The role of the dorsal cochlear nucleus in the perception of voicing contrasts in English stop consonants: A computational modelling study*, Ph. D. Thesis, University of Southampton, UK.

Smith, R.L., Brachman, M.L. (1982) *Adaptation in auditory nerve fibres: A revised model*, Biol. Cybern., **44**, 107-120.

Williams, S.M., Nicolson, R.I., Green, P.D. (1990) *Streamer: Mapping the auditory scene*, Institute of Acoustics Autumn Conf.

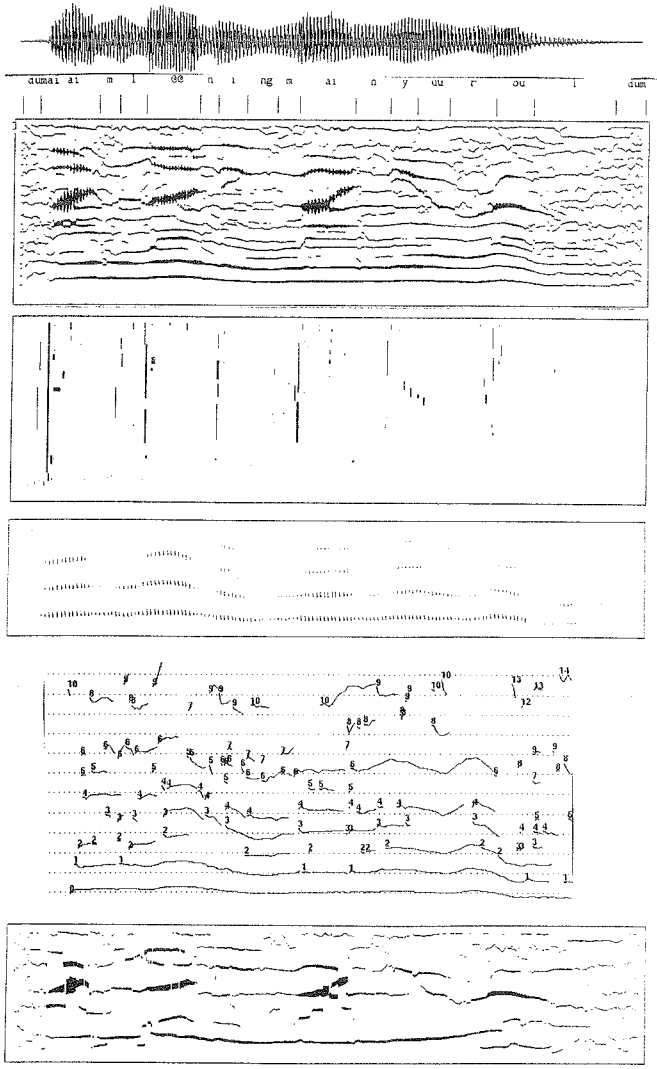

Figure 3: Representations produced at various stages of the model. From the top, waveform for the utterance "I'm learning my new rule" (male speaker, 1.64 s); transcription; synchrony strands; onset groups; amplitude modulation map; grouped harmonics; integrated strands. All frequency scales are linear in ERB-rate and cover the range 0-30.33 ERB (0-6500 Hz), except for the grouped harmonics and AM rate map which are linear in Hz and cutoff at 1300 Hz and 500 Hz respectively.