Proceedings of the Institute of Acoustics

SPEECH PATTERN PROCESSING: from 'blue sky' ideas to a unified theory?

R K Moore

Speech Research Unit, Defence Research Agency, St. Andrews Road, Malvern, Worcs., WR14 3PS

1. INTRODUCTION

This paper details one person's view of progress in the field of speech technology and its relevance to the field of speech science. Since the readership is likely to be knowledgeable in one or both fields, and since the paper has been prepared in response to the award of the Institute's 1994 Tyndall Medal, I have chosen *not* to present a general survey (of which there are many to choose from [1,2,3,4,5,6,7,8,9]), but a personal account of my own search for a deeper understanding of the nature of speech and of what I like to call '*speech pattern processing*'.

2. A PERSONAL RECORD

It is more than twenty years since, as an undergraduate on a course in Computers and Communications Engineering, I entered the fascinating world of speech research. Faced with having to make a choice from a somewhat uninspiring list of final–year projects, my lab partner and I had a vision of something a little different; perhaps it would be possible to construct a device which could recognise human speech – surely nobody had thought of that before! So, in 1972, was created an 'Electronic Apparatus for Recognising Speech' – EARS[1] – and I was hooked!

Thus, having opted for a career in speech research, I joined the Man–Machine laboratory at the University of Essex under the supervision of Professor Brian Gaines, and my first (rather obvious) question was "what is the state–of–the–art?". I was immediately rather surprised to discover that, whilst the idea of recognising speech was not exactly new, there appeared to be no established methods for measuring and comparing the performance of different recognition systems. Even worse, I came across a paper by John Pierce from Bell Laboratories in the Journal of the Acoustical Society of America entitled 'Whither Speech Recognition?' [10] in which Pierce issued a damning indictment of the quality of contemporary speech recognition research in general:–

> *"... if those who engage in recognition showed more signs of an effective effort to learn something about speech and fewer signs of rapture for computers and for unproven schemes for, and theories of, recognition.*
>
> *We all believe that a science of speech is possible, despite the scarcity in the field of people who behave like scientists and of results that look like science.*
>
> *Most recognisers behave, not like scientists, but like mad inventors or untrustworthy engineers."*

1. For students of the history of UK speech research, the first version of EARS was a real–time small–vocabulary speaker–dependent isolated–word recogniser using a twelve–channel analogue 'crab's eye' filter bank analyser interfaced to a PDP–8 which performed linear time–normalisation on the channel amplitudes/differences and classified the patterns using a statistical classifier. By 1975, EARS had increased to 16 channels, was connected to a PDP–11, used character string encoding (a form of VQ), performed recognition by computing a string–to–string edit distance (which, on meeting John Bridle for the first time, I discovered was called 'dynamic programming') and was integrated into a speech understanding system with a finite–state language model.

I immediately resolved to pursue a personal line of research (somesay, a crusade) which would contribute to mitigating these problems; how could speech recognition systems be evaluated, what was the fundamental nature of speech patterning (and how did it relate to patterns in other domains), what was the relationship between automatic and human speech recognition, and what underlying principles could be established in order to improve significantly the performance of automatic systems?

2.1 Speech Recogniser Assessment

In tackling the first of these, it seemed sensible to evaluate speech recognisers with respect to a *reference* system, and what better reference than the human listener. By 1975 I had established a calibrated model of human word recognition behaviour as a function of the phonetic confusability of words and the signal–to–noise ratio. Thus was born the 'Human Equivalent Noise Ratio' – HENR (the level of noise at which a human listener would perform at the same accuracy as a given machine) – a measure of recogniser 'goodness' which was independent of the confusability of the particular vocabulary with which it had been tested [11,12]. The HENR method is still in use today.

2.2 Speech Understanding Systems

It was around this time (1976) that the majority of workers in the field took the view (to which I have never subscribed) that there was insufficient information in a speech signal to permit the accurate recognition of continuous speech. Thus was born the era of 'speech understanding systems' in which high–level linguistic 'knowledge' was used to help a recognition system *guess* what might have been said. What interested me at the time was, again, the inability to compare such systems, not just in terms of overall performance, but also in terms of their structural component parts.

In studying these issues for my Ph.D., I was struck by the similarity of the processes at different levels in such systems and found that I was able to characterise apparently very different system configurations with a connected graph constructed from only two basic process components: a 'look–up' process for converting from one level of representation to another, and a 'rules' process for handling the sequential relationships within one level of representation [13]. What was interesting about this work was that it unified the apparently very different processes of recognition, understanding and synthesis and, as a by–product, provided a practical design language for the rapid configuration of novel real–time speech understanding systems. I was also able to convey the generality of the underlying principles by means of a demonstration of their use in synthesizing the coarticulated movements of a 'stick' man walking [14].

2.3 The Nature of Speech and Theories of Speech Recognition

It was in the midst of this work that (with Pierce's words echoing in my mind) I felt that further progress could only be achieved by gaining a more formal understanding of the structure and behaviour of speech itself. So, in 1977, I managed to persuade Professor Adrian Fourcin to support my application to join his laboratory in the Department of Phonetics and Linguistics at University College London as an SERC Post–Doctoral Research Fellow – and thereby became exposed to the differing cultural and scientific perspectives of 'speech technology' and 'speech science'.

At the same time, changes were afoot; the best system in the American ARPA sponsored speech understanding project – HARPY [15] – turned out to use, not a traditional knowledge–based approach but, a radically different idea – all the syntactic, lexical and phonetic knowledge could be compiled out into a single data structure in the form of an 'integrated network' [16], and recognition of continuous speech [17] could be achieved using the mathematical search technique known as 'dynamic programming' [18]. The significance of this for theories of human speech perception was immediately clear to me [19].

2.4 Dynamic Time Warping

By lucky coincidence, I was already familiar with dynamic programming from my earlier work with EARS, and from the many valuable discussions with John Bridle of the Joint Speech Research Unit who, inspired by the publications of Vintsyuk [20], had come up with an efficient one–pass dynamic programming based solution to the recognition of *connected* words [21]. The emergent properties of these so–called 'dynamic time warping' (DTW) algorithms (segmentation–by–recognition, delayed decisions, optimal search) seemed so fundamental to *any* speech recognition process – automatic or human – that, funded by JSRU, I began to investigate DTW [22] convinced that it represented an approach with a potential that far exceeded the accepted notion of the technique as simple 'pattern matching'.

Such was the situation in 1980 when I was invited to join the Royal Signals and Radar Establishment (RSRE) in Malvern to head–up a new and small team working on automatic speech recognition – our aim: to develop high–accuracy automatic speech recognition through research into the introduction of sound mathematical modelling techniques. There I was joined by Dr. Martin Russell (a mathematician) and Mike Tomlinson (an engineer experienced in speech coding) and together we edged our way towards the inevitable – the introduction of statistics into DTW [23,24] followed by the full–scale adoption, in 1982, of hidden Markov models (HMMs) for 'whole–word modelling' [25].

2.5 Speech Pattern Modelling

There followed a rapid series of significant developments in the modelling of speech patterns for recognition: self–organised phonetic network structures were derived for accurate discrimination [26,27], hidden semi–Markov models were used to model accurately the timing in speech patterns [28], HMM 'decomposition' was proposed as an optimal approach to the recognition of simultaneous signals [29] (either speech and noise [30] or even speech on speech [31]), sub–word HMMs were used to develop the UK's first real–time phonetically–based large vocabulary recognition system – ARMADA [32], variable–sized context–independent sub–word HMMs were developed for 'vocabulary–independent' recognition [33] which, in turn, lead to the development of the AURIX reconfigurable recognition system and its associated Windows–based CAD tool.

All of this progress stemmed from the pursuance of a sound mathematical and statistical basis for speech pattern modelling – stochastic modelling [34]; an approach which was expounded by John Bridle and myself in a paper to the 1986 Institute of Acoustics Conference on Speech and Hearing [35] which announced the formal transfer of JSRU to RSRE and the creation of the what we now know as the 'Speech Research Unit':–

> *"... it is highly likely that, in order to achieve high accuracy many talker, large vocabulary automatic speech recognition in a harsh environment, and high quality, high intelligibility, variable talker speech synthesis it is necessary to establish a central 'theory' of* ***speech pattern processing****. Such a theory should be mathematically rigourous, computationally tractable, and should make effective use of available information about the structure and use of human speech and language."*

3. SPEECH PATTERN PROCESSING

To me, speech is essentially a 'process' which relates an acoustic 'pattern' to a corresponding cognitive activity; it mediates the expression and communication of ideas, concepts and information between different physical entities through a regularity of behaviour which is shared (and hence 'understood') by the participants. It is this regularity of behaviour – the patterning – which is the central object of study in all areas of speech research: in speech pattern processing.

Speech pattern processing is thus concerned with *all* aspects of behaviour which relate to speech; it is concerned with the representations of speech, the representations of the structure of speech and the manipulation of such representations – it is a human and a machine activity. Speech pattern processing underpins fields such as speech science (speech production, perception and cognition), speech communications (telephony and semiotics), speech technology (speech recognition, understanding and generation) and speech ergonomics (spoken dialogue), and extends into the physiological, psychological and sociological aspects of speech.

More specifically, speech pattern processing requires a (mathematical) formalism which distinguishes between:–

- the information about the regularities of speech which is to be represented for computation – a–priori information (knowledge, constraints) that is available in the form of descriptive knowledge about speech patterns, linguistic structures and the relationships between different levels of description, in addition to corpora of recorded speech material and linguistic data which serve to exemplify such relationships,
- the representations on which the computations are to be performed – the encoding of the constraints, and
- the computations which are to be performed – constraint satisfaction algorithms.

This leads naturally to an approach to speech pattern processing which is founded on information theory and on 'speech pattern modelling'; information about speech and speech patterns is encoded in a suitable (statistical) model and appropriate algorithms are used to compute the likely output of the model for a specified input (for recognition and synthesis).

3.1 How Are We Doing?

Clearly, hidden Markov modelling (which is a *generative* statistical model of speech) is a powerful contender for a suitable formalism to underpin speech pattern processing, but how far can it be taken and are the appropriate disciplines involved?

Thus far, I would claim that progress in the understanding of speech owes little to the integration of the somewhat independent disciplines of speech science and speech technology. Of course, experimental phonetics has benefited from speech technology in terms of measurement tools and other forms of instrumentation, and speech technology has benefited from speech science in that it has readily adopted a great deal of its terminology. However, as yet, we don't know how to harness the computational skills of the speech technology community with the descriptive skills of the speech science community in order to construct acceptable and meaningful generic models of speech. Speech scientists tend to invoke models which although comprehensive are nevertheless *under–specified*, whereas speech technologists tend to utilise models which are practical but somewhat *over–simplified*.

For evidence of this mismatch between different approaches, we need look no further than a statement by Fred Jelinek in 1985 [36]:–

"every time we fire a phonetician/linguist, the performance of our system goes up!"

Could it be that after years of tinkering with fancy pattern recognition algorithms and high–powered computers that speech technologists (engineers, mathematicians, statisticians, computer scientists etc.) have shown that they have no need for speech science? Or is it that after years of careful study, speech scientists (phoneticians, linguists, experimental psychologists etc.) find that they still don't know enough about speech to influence technological developments?

The truth, of course, lies between these two extremes. However, what is clear is that these fields have only partially converged despite the fact that our knowledge about speech and its implementation in speech systems is in considerable need of further exploration.

What is needed is an alignment between the computational (algorithm rich) disciplines, the descriptive (knowledge rich) disciplines and the experimental (methodology rich) disciplines. In particular, it is essential (in terms of both healthy future funding and efficient scientific progress) that speech science and speech technology should merge into a single scientific discipline underpined by a comprehensive and coherent theory of speech pattern processing.

3.2 A Way Forward?

Last year, at the Berlin EUROSPEECH conference, I proposed various strategic actions which could perhaps pave the way towards the establishment of such a theory [37], notably, the opening up of an intellectual debate spanning the entire community and the tabling of putative theories which would be subjected to open critical (and constructive) analysis. Also, cultural, social and intellectual gaps need to be bridged; people should feel motivated to attend conference sessions which do not necessarily reflect their mainstream interests, conference organisers should not fall into trap of reinforcing the current divisions, workshops should be organised to address cross–disciplinary issues, supervisors should encourage their students to take a wider interest and courses should attempt to straddle the divisions, not provide them as alternative options.

Earlier this year, at a workshop in Germany, I took a more tactical approach and suggested twenty specific issues which I believe need to be addressed if we are to arrive at a greater understanding of the nature of speech and the mechanisms of speech pattern processing in general [38]. I also circulated the twenty questions by e–mail to the International speech research community, and respondents were asked to rank them in order of their importance:–

- How important is the communicative nature of speech? – ranked 8th
- Is human–human relevant to human–machine? – ranked 7th
- Speech technology or speech science? – ranked 14th
- Whither a unified theory? – ranked 13th
- Is speech special? – ranked 20th
- Why is speech contrastive? – ranked 18th
- Is there random variability in speech? – ranked 12th
- How important is individuality? – ranked 15th
- Is disfluency normal? – ranked 19th
- How much effort does speech need? – ranked 17th
- What is a good architecture? – ranked 10th
- What are suitable levels of representation? – ranked =1st
- What are the units? – ranked 3rd
- What is the best formalism? – ranked 11th
- How important are the physiological mechanisms? – ranked 6th
- Is time–frame based speech analysis sufficient? – ranked 4th
- How important is adaptation? – ranked 5th
- What are the mechanisms for learning? – ranked =1st
- What is speech good for? – ranked 9th
- How good is speech (for what it is good for)? – ranked 16th

Forty–two individuals responded to the survey (including some very well known names in the field). Analysis of the responses into industrial/academic revealed agreement about which were the top five, but complete disagreement about the order of those five. Academics put 'mechanisms for learning' at the top of their list, whereas industrialists put 'levels of representation' at the top of theirs. The ranking of the remaining questions were in general agreement between the two groups apart from 'human–human/human–machine', which academics ranked high but industrialists ranked low, and 'suitable architecture' which was ranked low by academics but high by industrialists.

Respondents also suggested a total of thirty–one additional questions! Of these, I felt that the following were particularly interesting:–

- How good could an automatic system be?
- How do conversations achieve useful results?
- Is language independence possible/useful?
- Is every bit of speech processed?
- How will speech technology change society?

Interestingly, the question relating to a unified theory was ranked 13th just above the issue of the relationship between speech technology and speech science. I suspect that this reflects the dominance of short–term interests over long–term interests in the individuals that replied.

4. WHAT WE DON'T KNOW

My own thoughts on the top ten issues are as follows:–

4.1 The Communicative Nature of Speech

The main purpose of speech is for *communication* between one human being and another. It has evolved over a period of 1,000,000 years for this single purpose and it is likely to be highly optimised in this regard [39,40]. The term 'communication', of course, refers to the transfer of *information* and, in the case of speech, this implicates considerably more than the literal content of the message as described by the words (or even expressed in conceptual terms such as 'ideas'), but a whole range of potentially important aspects of a talker's condition such as their individuality, their emotional state and their degree of involvement in the process.

Likewise the knowledge that is shared by the participants (or that is assumed by one participant to be known by the other participants), the knowledge that participants have of each other (for example, the degree of familiarity) and the social and cultural nature of the interaction (for example, the degree of formality) influences greatly the nature of the communication from its timeliness through to its final acoustic form; a complex interchange between strangers may be needed where in more intimate circumstances a simple grunt might suffice [41,42].

As a key gives access to a room, so speech probes a mind; speech *signals* a message, it is not the message itself.

4.2 Human–Human Speech Communication and Human–Machine Communication

The foregoing refers to speech–based interaction between people; it may or may not be relevant to the interaction between people and machines. Studies of simulated speech interactive system using 'Wizard of Oz' (WOZ) techniques have given some insights into this question – it is clear that people may adopt a simplified linguistic approach to automatic systems whose capabilities are not perceived to be high [43,44] – but, as yet, there is no clear understanding on how to capture and exploit the rich communicative properties in more advanced implementations.

4.3 Architecture

Interestingly, the areas of speech science and speech technology often make quite distinct assumptions about the nature of a suitable architecture for speech processes. Speech science favours *explicit* levels of representation and layered processing whereas speech technology favours *implicit* representations and integrated processing. The two approaches are not incompatible if it is considered important to be able to define precisely what is to be computed.

It is possible for a layered architecture to be optimal (in the sense that what is computed is guaranteed to be exactly what was required – for example, finding the final representation which has the highest probability) as long as appropriate information is transferred from layer to layer. Unfortunately, such information often requires the use of quite large data structures (lattices, charts etc.), hence the explicit approach tends to be either non–optimal or inefficient and slow. On the other hand, integrated architectures are often very efficient and optimal, but don't readily lend themselves to study and optimisation.

What is not clear is if this is a perpetual dilemma, or whether it will be possible and/or necessary ultimately to arrive at a unified architecture in which the long–term mechanisms (of processing and storage) are more explicit in order to handle the unusual, whilst the short–term mechanisms are compiled–out for efficient processing of the more familiar.

4.4 Levels of Representation

This question is particularly important in an 'explicit' architecture since at each level it is necessary to define the units involved, their relationships with each other and their relationships with the units at other levels. Even in an integrated architecture there are usually similar issues; there is almost always some intermediate representation between the speech waveform and the modelling formalism. For example, in an HMM–based automatic recognition system there is considerable debate about what would constitute a reasonable set of acoustic features.

Such structures are often motivated by phonetic and linguistic priors whereas it may be profitable to view intermediate levels of representation as providing an appropriate interface (analogous to 'impedance matching' in electrical circuits) between the properties of a signal and the assumptions embedded in a model.

4.5 Units

This is the most frequently asked question about the structure of speech, and it usually prompts the generation of a long list of putative answers: features, phones, phonemes, biphones, diphones, triphones, demi–syllables, syllables, morphemes, lexemes etc. etc. However, this successfully sidesteps the more serious underlying question; what constitutes the definition of a unit (any unit)? This may not seem to present any difficulties in the context of the different levels of representation that typify an explicit architectural model, but it becomes much more interesting in an integrated architecture where such 'objects' may simply emerge from the behaviour which arises as the implicit consequence of shared parameters. Such may be the nature of speech patterning itself.

4.6 The Relevance of Physiological Mechanisms

Most speech is emitted and processed by the human biological organism. It is usually generated by the articulatory processes of the human vocal tract, and analysed by the auditory processes in the human ear. Both areas have been subjected to some study, yet there is considerable debate as to the depth of the dependencies on the ultimate structure of speech. Why is the auditory system so over–specified in terms of number of frequency channels? Does the perceptual process derive information relating to the underlying articulations, or can it proceed without hypothesizing the state of the generator? Are speech patterns optimised for speaking, for listening or both?

Should speech technology systems seek to mimic these physiological mechanisms? The answer usually involves an analogy with the observation that aeroplanes don't flap their wings – but they do have wings (the problem lies in the limitations of available construction materials and a difference in the nature of the power source, not with the aerodynamic principles). In practice, models of the auditory system provides so much information that we don't know what to do with it (that is, how to model it). Likewise, models of the articulatory system requires so much computation (for example, using techniques such as 'finite element analysis') that we don't yet have powerful enough machines to cope!

Nevertheless, it would be surprising if more advanced models were not able to take advantage of the high time–frequency resolution provided by auditory–style processing, and that a reference to putative articulatory trajectories could not provide a useful constraint on hypotheses.

4.7 The Sufficiency of Time–Frame Based Speech Analysis

Although we know that speech is a composite acoustic signal arising from multiple sound sources and independent articulator movements, it is hard to break away from analysis techniques which imply a linear frame–to–frame ('beads on a string') time sequence of events. Even speech synthesis has been obliged to adopt concatenative principles in order to generate speech with acceptable quality. Are these techniques sufficient? In the long run, probably not. Alternative views are already emerging but, as yet, it is not clear how to integrate the ideas of non–linear phonology [45], hidden Markov model decomposition [28], parallel model composition [46], temporal decomposition [47] and segmental modelling [48] into a unified and coherent speech analysis and synthesis framework.

4.8 Adaptation

Human behaviour is known to be highly adaptive; the speech of an unknown talker with an unusual accent can be 'tuned' into with relative ease after only a few fragments of speech have been heard, and a talker rapidly adjusts their articulations in order to achieve different effects or to overcome difficult or unusual circumstances. By comparison, automatic systems are fairly static, relying on only minor deviations from the norm being encountered.

In practice, it may be that the exception is the rule, and that it is only continual adjustment, or *normalisation*, to the conditions which pertain, that would allow an organism to keep track of the environment in which it is operating. Interestingly, such a concept of 'tracking' can also be viewed as a kind of recognition – a determination of the conditions which prevail; the objects of relevance and their underlying conditioning variables. In the end it is simply a matter of the (memory) timescales over which such behaviours operate.

Present understanding is limited to tracking surface parameters with only limited recourse to the 'doubly–stochastic' models that would be required to formalise the recognition of, or active adjustment to, important underlying coordinating variables.

4.9 Mechanisms for Learning

This leads on to questions concerning the general nature of learning (adaptation on a longer timescale and with more fundamental structural consequences). Very little is known about mechanisms for acquiring new words, new grammatical constructions, new concepts, new meanings, new interactive strategies. How does the child build up its competence; does it assume the world is full of a wide variety of different stimuli which have to be grouped (clustered) gradually into more meaningful structures, or does it assume that the world is essentially homogeneous only requiring partitioning into alternative categories when a distinction becomes necessary? So far, the majority of automatic schemes take a one–shot approach.

4.10 What is Speech Good For?

Much is discussed about the ergonomics of speech but, as yet, little has been formalised successfully [49]. Speech is only one modality through which an organism may choose to interact with the world (and other organisms). The appropriate orchestration of multiple modalities in an effective *dialogue* is probably key to an understanding of each modality individually. Except on the telephone, speech operates in concert with gesture and touch and is shaped by their co–existence (as witnessed by the intimate interaction between audio and visual cues in speech perception).

The advantages and disadvantages of speech are well established, but designing a multi–modal interface which exploits such properties is still in need of serious study taking into account that the human, at least, often has goals such as 'to be entertained', 'to be interested' or 'to be involved' which overpower more mundane requirements of minimising time and maximising efficiency.

This means that attention needs to be given to *planning* in its widest sense: from the identification of interactive goals and intentions, to the dependent dialogue moves, through message generation and setting of receptive expectancies, to consequent and appropriate realisations in prosodic and segmental forms. The requirements of different scenarios and applications, and the capabilities of all participants will have to be profiled in order to understand and explore the strategies and trade–offs appropriate to communication in a potentially errorful environment; clarification behaviour and error correction will have to be formalised as an essential integral component of any successful interaction.

It is highly likely that progress in this area will point the way to a greater understanding of the intimate integration of segmental and supra–segmental patterning in speech.

5. WHAT WE DO KNOW

Of course it is easy to concentrate on the gaps in our knowledge and not acknowledge the significant progress that has been made in recent years. The methodological issues have become reasonably clear; the success of HMMs is understood to be derived from the use of an (albeit simple) underlying theory of mathematical and statistical modelling. In particular, it is understood how the use of probability theory provides a sound theoretical framework for modelling our uncertainty or *ignorance* [50], and this explains the current heavy reliance on large quantities of speech data for estimating the parameters of such models. When speech is understood more fully, there may be very little residual uncertainty remaining to be modelled, the reliance on vast quantities of data will be reduced and the stochastic approach will have both served and lost its purpose.

Another important aspect of progress in speech pattern processing is the realisation that the process of recognition can be viewed as a 'search'. This has given rise to algorithms with interesting emergent properties derived from the sequential resolution of ambiguity, and which can even accommodate the effective modelling of simultaneous events.

6. CONCLUSION

This paper has presented a personal view of the progress that is being made towards a greater understanding of the nature of speech. I believe that the most effective research directions will involve, and are dependent upon, a merging of the methodologies in speech science and technology, and that this will result in the establishment of a coherent, comprehensive and scientific theory of *speech pattern processing*. With our current state of knowledge, I see no restrictions on our ability to start down this path; the only thing that can hold things back is a lack of vision!

7. REFERENCES

[1] W A AINSWORTH, **Speech Recognition By Machine**, Peter Peregrinus, 1988.

[2] M ALLERHAND, **Knowledge-based Speech Pattern Recognition**, Kogan Page, 1987.

[3] G BRISTOW, **Electronic Speech Recognition**, Collins, 1986.

[4] J N HOLMES, **Speech Synthesis And Recognition**, Van Nostrand Reinhold (UK) Co. Ltd., 1988.

[5] A S HOUSE, **The Recognition Of Speech By Machine – A Bibliography**, Academic Press Ltd., 1988.

[6] W A LEA, **Trends In Speech Recognition**, Prentice-Hall Inc., 1980.

[7] J J MARIANI, 'Recent advances in speech processing', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp 429–440, 1989.

[8] C ROWDEN, **Speech Processing**, McGraw-Hill Book Company, 1992.

[9] D B ROE & J G WILPON, **Voice Communication Between Humans and Machines**, National Academy Press, Washington, 1994.

[10] J R PIERCE, 'Whither Speech Recognition?', JASA, Vol.46, No.4, 1969.

[11] R K MOORE, 'The evaluation and optimisation of a basic speech recogniser', M.Sc. Thesis, Dept. Elec. Eng., University of Essex, 1975.

[12] R K MOORE, 'Evaluating speech recognisers', IEEE Trans. Acoustics, Speech and Signal Processing 25, pp 178–183, 1977.

[13] R K MOORE, 'A descriptive technique for the analysis and design of speech understanding systems', Ph.D. Thesis, Dept. Elec. Eng., University of Essex, 1977.

[14] R K MOORE, 'A multilevel approach to pattern processing', Pattern Recognition 14, pp 261–265, 1981.

[15] B T LOWERRE, 'The HARPY speech recognition system', Ph.D. Thesis, Dept. Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1976.

[16] J K BAKER, 'The DRAGON system – an overview', Proc. IEEE symposium on Speech Recognition, pp 22–26, 1974.

[17] F JELINEK, 'Continuous speech recognition by statistical methods', Proc. IEEE, Vol.64, pp 532–555, 1979.

[18] R E BELLMAN, **Dynamic Programming**, Princeton University Press, 1957.

[19] R K MOORE, 'Speech recognition systems and theories of speech perception', **The Cognitive Representation Of Speech**, Myers, Laver and Anderson (eds.), North Holland, pp 427–441, 1981.

[20] T K VINTSYUK, 'Speech discrimination by dynamic programming', Kibernetica, Vol.4, pp 81–88, 1968.

[21] J S BRIDLE & M D BROWN, 'Continuous connected word recognition using whole word templates', The Radio and Electronic Engineer, Vol.53, pp 167–175, 1983.

[22] R K MOORE, 'Dynamic programming variations in automatic speech recognition', Proc. Inst. Acoustics Spring Conf., Newcastle, April 21–24, pp 269–272, 1981.

[23] R K MOORE, M J RUSSELL & M J TOMLINSON, 'Locally constrained dynamic programming in automatic speech recognition', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Paris, May 3–5, pp 1270–1273, 1982.

[24] M J RUSSELL, R K MOORE & M J TOMLINSON, 'Some techniques for incorporating local timescale variability information into a dynamic time warping algorithm for automatic speech recognition', Proc. IEEE Int. Conf. on Acoustic s, Speech and Signal Processing, Boston, 14–16 April, pp1037–1040, 1983.

[25] S E LEVINSON, L R RABINER & M M SONDHI, 'An introduction to the application of the theory of probabilistic functions of a Markov process', Bell System Technical Journal, vol.62, pp 1035–1074, 1983.

[26] R K MOORE, M J RUSSELL & M J TOMLINSON, 'The discriminative network; a mechanism for focusing recognition in whole–word pattern matching', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Boston, 14–16 Ap ril, pp1041–1044, 1983.

[27] R K MOORE, M J RUSSELL & M J TOMLINSON, 'Introducing phonetic discrimination into whole–word pattern matching', Presented at 10th Int. Congress of Phonetic Sciences, Utrecht, 1–6 August, 1983.

[28] M J RUSSELL & R K MOORE, 'Explicit modelling of state occupancy in hidden Markov models for automatic speech recognition', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Tampa, 26–29 March, 1985.

[29] R K MOORE, 'Signal decomposition using Markov modelling techniques', RSRE Memorandum No.3931, July, 1986.

[30] A P VARGA & R K MOORE, 'Hidden Markov Model Decomposition of Speech and Noise', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Albuquerque, 1990.

[31] A P VARGA & R K MOORE, 'Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition', Proc. ESCA EUROSPEECH conference, Genova Italy, September, 1991.

[32] M J RUSSELL, K M PONTING, S M PEELING, S R BROWNING, J S BRIDLE, R K MOORE, I GALIANO & P HOWELL, 'The ARM continuous speech recognition system', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 1990.

[33] R K MOORE, M J RUSSELL, P NOWELL, S N DOWNEY & S R BROWNING, 'A comparison of phoneme decision tree (PDT) and context adaptive phone (CAP) based approaches to vocabulary–independent speech recognition', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Adelaide, 19–22 April, 1994.

[34] R K MOORE, 'Recognition – The Stochastic Modelling Approach', **Speech Processing**, C Rowden (Ed.), Mc Graw–Hill, 1992.

[35] R K MOORE & J S BRIDLE, 'Speech Research at RSRE', Proc. UK Inst. Acoustics Conf. on Speech and Hearing, 1986.

[36] F JELINEK, Quote during IEEE ASSPS workshop on Frontiers of Speech Recognition, Arden House, 1985.

[37] R K MOORE, 'Whither a theory of speech pattern processing', Proc. 3rd European Conference on Speech Communication and Technology, EUROSPEECH'93, pp 43–47, Berlin, 1993.

[38] R K MOORE, 'Twenty things we still don't know about speech', Proc. CRIM/FORWISS workshop on Progress and Prospects of Speech Research and Technology, Munich, September, 1994.

[39] C CHERRY, **On Human Communication**, The MIT Press, 1970.

[40] D FRY, **Homo Loquens**, Cambridge University Press, 1977.

[41] G K ZIPF, **Human Behaviour And The Principle Of Least Effort**, Addison–Wesley Publishing Co., Inc., Cambridge, Mass., 1949.

[42] B LINDBLOM, 'Explaining phonetic variation: A sketch of the H and H theory', W. J. Hardcastle and A. Marchal (eds.), **Speech Production and Speech Modelling**, pp 403–439, Kluwer Academic Publishers, 1990.

[43] N M FRASER & G N GILBERT, 'Simulating speech systems', Computer Speech and Language, Vol.5, No.1, pp 81–99, January 1991.

[44] R K MOORE & A MORRIS, 'Experiences collecting genuine spoken enquiries using WOZ techniques', Proc. 5th DARPA Workshop on Speech and Natural Language, New York, 1992.

[45] G N CLEMENTS & S J KEYSER, **CV–Phonology**, MIT Press, 1983

[46] M J F GALES & S J YOUNG, 'HMM recognition in noise using parallel model combination', Proc. 3rd European Conference on Speech Communication and Technology, EUROSPEECH'93, pp 837–840, Berlin, 1993.

[47] B S ATAL, 'Efficient coding of LPC parameters by temporal decomposition', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'83, pp 81–84, Boston, 1983.

[48] M J RUSSELL, 'A segmental HMM for speech pattern modelling', Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP'93, Minneapolis, 1993.

[49] M M TAYLOR, F NEEL & D G BOUWHUIS, **The Structure Of Multimodal Dialogue**, North–Holland, 1989.

[50] J MAKHOUL & R SCHWARTZ, 'Ignorance based modelling', J. Perkell and D. H. Klatt (eds.), **Invariance And Variability In Speech Processing**, Erlbaum, 1984.