COMPUTATIONAL PHONETICS

Roger K. Moore
DRA Speech Research Unit, Malvern, UK

ABSTRACT

Recent impressive advances in the capabilities of systems for automatic speech recognition and automatic speech generation has meant that there is a growing need to unify the emerging theoretical and practical developments in speech technology with the established knowledge and practices in the phonetic sciences. This trigger paper discusses some of the relevant issues and proposes the establishment of a new discipline to be known as *Computational Phonetics*.

BACKGROUND

The idea that an automatic device can be configured to 'recognise' or 'synthesise' human speech not only has the practical benefit of providing human operators with hands-free eyes-free control of equipment or access to information, but could also be said to provide the ultimate test for phonetic theories of human speech perception and production.

Moreover, it is precisely in the area of 'speech technology' (particularly in automatic speech recognition and automatic speech generation) that the experimental and descriptive fields of phonetics, linguistics and psychology meet the computational disciplines of artificial intelligence, computer science and engineering.

Both of these observations raise interesting issues concerning the role of contemporary phonetics in the light of the substantial advances that are currently being made in the capabilities of automatic speech recognition and generation systems.

Automatic Speech Recognition

Automatic speech recognition has come a long way from the first simple attempts back in the 1950s. In the early days, vocabularies were small (usually the ten digits), the words had to be uttered in 'isolation' (that is, with a distinct pause between each word) in a quiet environment, and each user was obliged to 'train' the system by providing a set of example utterances - whole-word 'templates' - against which subsequent words to be recognised would be compared (thereby rendering the process 'speaker dependent').

Forty years on, automatic speech recognition systems can operate with vocabularies containing many thousands of words, the input can be natural 'continuous' speech and, after estimating the parameters of a set of suitable statistical models (for example, 'hidden Markov models' - HMMs) using data from an appropriate spoken language corpus, utterances can be recognised from a wide range of 'independent' speakers operating in more natural environments (such as in an office or over a telephone).

Automatic Speech Generation

Likewise, speech synthesis systems have progressed from manually operated electrical and mechanical devices to automatic text-to-speech reading machines which can be adapted to exhibit the vocal characteristics of a desired target speaker and which can handle abbreviations and acronyms as well as regular textual input.

Also the process of speech generation from first principles using a mathematical analogue of the human production apparatus is being supplemented by an approach based on the concatenation of relevant fragments of natural human utterances which have been extracted from an appropriate spoken language corpus using automatic processes not dissimilar to those used in automatic speech recognition.

Prerequisites for Progress

One might easily imagine that these substantial advances have been caused by the implementation of linguistic and phonetic 'knowledge' in such speech technology systems. However, it can be argued (particularly for automatic speech recognition) that progress has in reality been a direct result of the introduction of rigorous mathematical and statistical modelling paradigms coupled with the development of efficient 'search' and 'parameter estimation' algorithms supported by a phenomenal increase in available computing power and data handling capacity.

It can also be argued that further progress depends on a continued concerted effort to tackle some of the theoretical and practical issues in automatic speech recognition and generation, not the least of which is to arrive at a greater understanding of the structure and regularities of speech signals themselves and of the 'process' which relates an audio-visual speech 'pattern' to it's cognitive counterpart. Such an understanding might be expressed in terms of a *theory* of 'speech pattern processing' [1].

SPEECH PATTERN PROCESSING

Speech essentially 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 central to speech pattern processing and hence to speech recognition and generation. It is the patterning which provides the 'constraint' which allows human behaviour never before encountered to be recognised and interpreted appropriately, and which conditions the generation of novel behaviour never before required.

Speech Patterning

Information about the patterning in speech is derived from two principal sources; (i) the discipline of phonetics (and related areas such as psycho-acoustics, linguistics, psycho-linguistics etc.) which provides descriptive 'knowledge' about the observed regularities in speech, and (ii) annotated speech corpora which provide hard 'evidence' for more detailed speech pattern behaviour.

Thus far, neither source of constraint is sufficient on its own to facilitate high-accuracy automatic speech recognition and generation. However, it is fair to say that it is the extensive use of large-scale speech corpora that has been the key to the success of current automatic speech recognition and generation systems.

Of course it is not sufficient simply to have (even detailed) information about the constraints implicit in speech patterning in order to construct a functional automatic speech recogniser or synthesiser; it is also necessary to define a (set of) 'representation(s)' with which to 'encode' such constraints.

Likewise, the appropriateness of any given representation depends critically on the 'computation' which is to be performed upon it - and such an 'algorithm' needs to be founded on some kind of mathematical 'theory' of recognition or generation.

Speech Pattern Processing Theory

Thus far, the most successful approaches to automatic speech recognition have been based on the theory of 'maximum-likelihood' (or Bayes') classification which defines the interpretation of a sequence of acoustic observations in terms of the most probable explanation taken over all possible interpretations. From this theory it is possible to derive a mathematical and statistical 'modelling' paradigm (such as hidden Markov models) which provides a suitable integrated representation of acoustic, phonetic and lexical constraints together with compatible algorithms for estimating the model parameters from annotated data and for computing the most likely interpretation of an unknown input sequence.

On the other hand, automatic speech generation is founded on less well developed formalisms and, as such, lags behind recognition in it's theoretical sophistication. Low-level processes such as the generation of a spectrum from a parametric representation of a vocal tract are based on solid mathematical principles, but the control of such parameters is often handled in a more heuristic manner. However, the introduction of statistical techniques (more familiar to automatic speech recognition) for control parameter modelling is beginning to take place.

Stochastic Modelling

It is important to appreciate that the use of statistics in speech pattern processing is convenient simply because it provides a rigorous mathematical framework for modelling 'uncertainty' and for characterising the processes of 'approximation', 'interpolation' and 'extrapolation' which are all key components of the requirement to be able to categorise unseen data and to be able to generate novel data.

The value of stochastic models in general, and HMMs in particular, is that the formalism shows no signs of being limited in the extent to which it can be developed to accommodate more complex modelling requirements; the mathematics has already been extended to handle simultaneous asynchronous events (thereby removing the 'single synchronous signal' assumption) and to include dynamic segmental effects (thereby removing the frame-to-frame 'independence' assumption).

Both of these advances point towards a possible unification of HMM structures with the modelling strategies normally employed in speech synthesis and the ideas expounded in the field of 'non-linear phonology' [2]. However, this unification can only be achieved if there is effective communication between the appropriate specialist practitioners involved in the speech pattern modelling and phonetics areas.

THE ROLE OF PHONETICS

Clearly, in principle, the field of phonetics has a great deal to contribute to the design of appropriate annotated speech corpora and to the expression of the phonetic and linguistic 'priors' which might be made implicit in a system's modelling structures.

However, both of these activities must be carried out in full cognisance of the theoretical and mathematical implications involved; *it is not appropriate to propose new representations without considering whether they are compatible with any known scheme for computation.*

It is therefore proposed that the skills and expertise represented by the phonetic science community could be usefully directed *not* towards the construction of better automatic speech recognisers or synthesisers, but towards the exploitation of the theoretical and practical tools and techniques from speech technology for the creation of more advanced theories of speech perception and production (by humans *and* by machines). Indeed it is perhaps now appropriate to begin to think in terms of establishing a new more balanced discipline which could be described as *'Computational Phonetics'.*

Practitioners in this new area should be encouraged to work towards a *unified* theory of speech pattern processing which could answer some of the outstanding fundamental questions about speech [3] to the benefit of *both* speech technology and speech science.

REFERENCES

[1] Moore, R. K. (1993), "Whither a theory of speech pattern processing", Proc. EUROSPEECH'93, pp 43-47.

[2] Moore, R. K. (1994), "Speech pattern processing: from 'blue sky' ideas to a unified theory?", Proc. UK Inst. of Acoustics Conf. on Speech and Hearing, pp 1-13.

[3] Moore, R. K. (1994), "Twenty things we still don't know about speech", Proc. CRIM/FORWISS Workshop on Progress and Prospects of Speech Research and Technology.