TWENTY THINGS WE STILL DON'T KNOW ABOUT SPEECH

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

July 1994

Abstract

Human language has been the object of study since ancient times [BLO73]; it is 130 years since Alexander Melville Bell started on the road which was to lead to the modern-day field of phonetics [BEL67]; it is 50 years since the invention of the sound spectrograph [PKK66]; and it is 40 years since the first electrical speech recognition and synthesis systems were constructed [FLA72]. How then is it that, in the final years of the twentieth century, we have yet to witness a complete and worthwhile unification of the science and technology of speech?

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.

This paper identifies a number of themes which the author believes may be important to the greater understanding of the nature of speech and the mechanisms of speech pattern processing in general - *twenty things we still don't know.*

1 How important is 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 [CHE70, FRY77]. 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.

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

2 Is human-human speech communication relevant to human-machine communication by speech?

All of 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 [MOO92] - but, as yet, there is no clear understanding on how to capture and exploit the rich communicative properties in more advanced implementations.

3 Speech technology or speech science?

Thus far, 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 benefitted from speech technology in terms of measurement tools and other forms of instrumentation, and speech technology has benefitted 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*.

These divisions are further enhanced by the recent dramatic growth in speech research and the consequent specialisations which have arisen. Unless individuals are encouraged (or educated) to span the broad array of speech related themes, then communication and understanding will drop to a level below even that which it has reached today.

4 Whither a unified theory?

As yet, the different speech research communities do not benefit from anything which might be called a common underlying *theory* of speech [MOO93]. Notwithstanding the cultural problems involved in bringing the different disciplines together, is it possible to conceive of such a theory - a functional model of integrated speech production and reception which would serve as both an explanation and a prescription for the organisation of speech patterns and provide a language for describing the totality of speech related behaviours?

Is there any reason to believe that such an objective would be ultimately unattainable or, indeed, undesirable? Certainly the speech sciences have amassed a wealth of information about speech and human speech behaviour which, if combine with the practical experience of the speech technologists, ought to be able to be used to construct a first-order attempt at unification. Perhaps the real question is: can such an activity be encompassed within the short-term interests of the different communities because of a perceived near-term benefit, or must we wait until progress begins to grind to a halt (or until one community is obliged to re-invent the knowledge of the other)?

5 Is speech special?

An important question in speech science is whether speech is in some way privileged in terms of the mechanisms it evokes, or can it be considered alongside other complex acoustic signals. Evidence from the technologies would perhaps suggest that speech is by no means special, just highly structured, man-made and communicative. Morse-code, or indeed sign-language, is as capable of engaging the full human linguistic apparatus as speech.

6 Why is speech contrastive?

A fundamental tenet of speech science is that speech is said to be 'contrastive' in nature. That is, minimal phonemic contrasts serve to distinguish one lexical form from another, and acoustic-phonetic featural contrasts serve to distinguish one phonetic form from another. Speech perception is said to be 'categorical' and humans exhibit enhanced discriminative behaviour at category boundaries.

Some of this has impacted on the structures of speech technology systems, but often only contributes to the enumeration of the categories and does not exploit any contrastive properties of the patterns. In fact, most effort (by researcher and by modelling algorithm) is commonly expended on capturing the norm rather than the exception - the distribution mean rather than the class boundary. Clearly this puts undue pressure on the quality of the variance estimation and thereby calls for even greater amounts of data to be fed to the modelling engine.

Could it be that this is where 'artificial neural networks' [RUM86] signal a different view? MLP-style discrimination and categorical behaviour could both be seen to be aspects of the same important process of minimising the number of parameters in a models in order to achieve maximum generalisation (interpolation) to novel forms. This would constitute a practical as well as a theoretical benefit to speech reception and generation.

7 Is there random variability in speech?

It is often said that the problem with speech is its inherent variability . This is certainly true; if speech signals did not exhibit such a high degree of variation, then speech science would have closed-down its business years ago! The statement is also underlined by the recent successes in automatic speech recognition being almost entirely attributable to the use of powerful statistical modelling techniques.

However, how much of this endemic variability is actually random, as opposed to the view that the high level of observed variability is simply a manifestation of our lack of understanding of the underlying behaviour? In the end, statistics is just a sound mathematical approach for modelling uncertainty or *ignorance* [MAK84]. When speech is fully understood, there may be very little residual uncertainty remaining to be modelled and the stochastic approach will have both served and lost its purpose.

8 How important is individuality?

The classic source of variability in speech is that exhibited between different individuals. However, despite the assumptions made by most speech technology systems, such variability is not just random variation between individuals but stylistic effects and different preferred strategies for both production and perception. It has become clear to speech scientists that there is not only one way of performing or interpreting a given behaviour; a talker appears able to exploit

a number of degrees of freedom in the speech system in order to achieve a desired effect, and individuals can aquire alternative strategies for listening which may be measurably more or less effective in different circumstances.

Understanding such effects is likely to be crucial for dealing adequately with so-called 'normal' speech as well the speech that is found in more unusual situations. For example, it may be that the difficulty associated with characterising the effects of emotional or workload stress on speech, or the more exotic claims of the Lombard effect [LOM11], arises from the fact that physiological and psychological reactions to unusual environments are themselves highly individual in nature.

9 Is disfluency normal?

Linguists have spent a considerable amount of time and effort studying the written form and hence there is a view that the disfluencies of spontaneous speech are a manifestation of a poor linguistic *performance* (conditioned by practical problems associated with organising the speech organs, or simply failing to apply sufficient effort) overlaid on, and therefore obscuring, an underlying well-formed linguistic *competence*. An alternative view is that disfluencies reveal the underlying organisational and planning processes involved.

10 How much effort does speech need?

It is well established that, depending on the circumstances, a human talker or listener is able to apply an appropriate amount of effort in order to arrive at the most useful outcome. Speech may be mumbled using a low degree of effort but understood perfectly using a high degree of effort, or it may be hyper-articulated for clarity allowing a listener to perceive with ease but placing a consequent high physical demand on the talker. Such effects even condition the structure of language itself [ZIP49] and are posited to offer a more significant and active explanation of effects such as coarticulation than purely passive dynamical behaviour is able to provide [LIN90].

Such effects suggest that the long-term and short-term organisation of speech may owe a considerable amount to the existence of active plan-based mechanisms which could only operate in an environment where speech generation and reception are linked intimately together.

11 What is a good 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 tranfered 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.

12 What are suitable 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 in a model.

13 What are the 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.

14 What is the best formalism?

Some areas of speech science are so deeply enmeshed in a descriptive framework that relevance to computable forms are often not seen as being of primary interest. Nevertheless, first-order predicate logic (rules) are inevitably favoured because of their obvious simplicity when dealing with explicit representations and structures. The fact that such a formalism would fail to operate as an adequate functional explanation of real events is often not understood by the practitioners.

Linguistic formalisms , in general, have been considerably richer with a range of classes of formal grammar of differing properties and complexity having been identified (regular grammars, context-free grammars, context-sensitive grammars, unification grammars etc.) [LYO68]. Unfortunately, such schemes also appear to falter when faced with the variable reality of speech.

As a result, speech technologists have developed their own models using variable-sequence modelling formalisms such as finite-state automata and their statistical variants from simple n-grams to the more powerful hidden Markov models (HMMs) [LEV83]. Interestingly, such formalisms have direct equivalents in formal grammars (a hidden Markov model is also a stochastic regular grammar and the powerful Baum-Welch parameter re-estimation technique [BAU70] can also be applied to a stochastic context-free grammar) [BAK79]. Also interesting relationships

are being established between such models and powerful signal modelling techniques such as Kalman filtering. The functional effectiveness of patently simple-minded statistically-based formalisms is both a cause of some concern (for those who are committed to an alternative approach) and of some excitement (for those who see the implications of such results for even more powerful and interesting models of speech).

Artificial neural networks might on the surface appear to offer a radically alternative approach but, in practice, success has usually been dependent on their contribution to the calculation of conditional probability as embedded in a more traditional hidden Markov framework. Nevertheless, the fact that such schemes can operate with significantly fewer parameters points strongly in the direction of even further integration of these formalisms as an understanding of the need for non-linear and discriminative modelling processes grows.

15 How important are the 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 hypothesising 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.

16 Is time-frame based speech analysis sufficient?

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 [CLE83], hidden Markov model decomposition [VAR90], parallel model composition [GAL93], temporal decomposition [ATA83] and segmental modelling [RUS93] into a unified and coherent speech analysis and synthesis framework.

17 How important is 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.

18 What are the 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 stimulae which have to be grouped (clustered) gradually into more meaningful structures, or does it assume that the world is essentially homogenous only requiring partitioning into alternative categories when a distinction becomes necessary? So far, the majority of automatic schemes take a one-shot approach.

19 What is speech good for?

Much is discussed about the ergonomics of speech but, as yet, little has been formalised successfully [TAY89]. 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 multimodal 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.

20 How good is speech?

The foregoing leads directly on to the lack of worthwhile metrics or 'meta-models' which can be employed to predict how systems and users would behave. Assessment is the flavour of the decade (typified in the US ARPA speech technology programme) but it is essentially empirical. There are no calibrated models of individual components; there are models for human behaviour and models inside speech technology systems, but there are no meta-models of speech technology systems as a whole - for example, the only way of testing the intelligibility of a speech synthesis system is to play its output to a panel of listeners, and the only way of testing the 'goodness' of a recogniser is to give it input from a panel of talkers. This lack of suitable analytical tools simply reflects our lack of understanding of speech itself.

References

[ATA83] 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.

[BAK79] J. K. Baker. Trainable grammars for speech recognition. D. H. Klatt and J. J. Wolf, *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, pp 547-550, 1979.

[BEL67] A. M. Bell. *Visible Speech: The Science of Universal Alphabetics.* Simkin, Marshall and Co., London, 1867.

[BAU70] L. E. Baum, T. Petrie, G. Soules and N. Weiss. A maximisation technique occuring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, Vol.41, No.1, pp 164-171, 1970.

[BLO73] L. Bloomfield. *Language.* George Allen and Unwin Ltd., London, 1973.

[CHE70] C. Cherry. *On Human Communication.* The MIT Press, 1970.

[CLE83] G. N. Clements and S. J. Keyser. *CV-Phonology* MIT Press, 1983

[FLA72] J. L. Flanagan. *Speech Analysis, Synthesis and Perception.* Springer-Verlag, Berlin - Heidelberg - New York, 1972.

[FRY77] D. Fry. *Homo Loquens.* Cambridge University Press, 1977.

[GAL93] M. J. F. Gales and 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.

[LEV83] S. E. Levinson, L. R. Rabiner and 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.

[LIN90] 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.

[LOM11] E. Lombard. Le signe de l'elevation de la voix. *Ann Maladiers Oreille, Larynx, Nez, Pharynx*, vol.37, pp 101-1199, 1911.

[LYO68] J. Lyons. *Introduction to Theoretical Linguistics.* Cambridge University Press, 1968.

[MOO92] R. K. Moore and A. Morris. Experiences collecting genuine spoken enquiries using WOZ techniques. *Proc. 5th DARPA Workshop on Speech and Natural Language*, New York, 1992.

[MOO93] 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.

[MAK84] J. Makhoul and R. Schwartz. Ignorance based modelling. J. Perkell and D. H. Klatt (eds.), *Ignorance Modelling, Invariance and Variability in Speech Processing*, Erlbaum, 1984.

[PKK66] R. K. Potter, G. A. Kopp and H. G. Kopp. *Visible Speech.* Dover Publications, 1966.

[RUM86] D. E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing.* MIT Press, 1986.

[RUS93] M. J. Russell. A segmental HMM for speech pattern modelling. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, ICASSP'93, Minneapolis, 1993.

[TAY89] M. M. Taylor, F. Neel and D. G. Bouwhuis. *The Structure of Multimodal Dialogue.* North-Holland, 1989.

[VAR90] A. P. Varga and R. K. Moore. Simultaneous recognition of concurrent speech signals using hidden Markov model decomposition. *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, ICASSP'90, Albuquerque, 1990.

[ZIP49] G. K. Zipf. *Human Behaviour and the Principle of Least Effort.* Addison-Wesley Publishing Co., Inc., Cambridge, Mass., 1949.