The Problem of Capturing Linguistic and Phonetic Knowledge

Mark Tatham

Reproduced from *Proceedings of the Institute of Acoustics*, Vol. 8, Part 7 (1986), pp. 443-450.

INTRODUCTION

There can be no doubt that the success of text-to-speech synthesis systems and latterly the improvements in automatic speech recognition systems owe much to the bringing in of information from linguistics and phonetics. But it seems that a mistake has been made in what has been brought in. This is particularly true in synthetic speech where the ideas have had longer to become entrenched. The speech the latest devices produce is just not good enough. The failure has little to do with the design of the actual synthesiser. Whatever faults synthesiser hardware may have they are patently not the limiting factor. So in speech synthesis by rule it is easy to reach the conclusion: it must be the rules. It is not - at least not in the sense that if we devote more effort to refining the rules all will come right. What is wrong is that the whole system (I use the word in the singular since all the leading systems - JSRU and its derivative BTalk [the British Telecom parallel formant synthesiser], MITalk and its derivatives DECtalk and Prose-2000 - are all similar in type) has been ill-conceived. The fault lies as much with the linguists as it does with the engineers. The same is true of automatic speech recognition systems which employ grammars in a top-down effort to disambiguate the output of low-level bottom-up analysers.

LINGUISTICS

Linguistics is the science which deals with grammars. The theory is exemplified in models which have an exhaustive formal structure which is utterly explicit [1]. In no sense can the argument that linguistics is informal or inexplicit be sustained. The metatheory is equally clear: linguists know what they are doing and why they are doing it. Like any developing science linguistics has known its infighting, but the general principles and goals remain constant and, from our point of view, the main points are established and unarguable.

The theory of linguistics is descriptive and has explanatory aims. It describes the knowledge a person has of his language [2]. It is this which enables him to encode his thoughts for transmission to another person and to decode the thoughts transmitted to him by another person. The carrier which concerns us is sound, though of course there are others. In addition and in common with other branches of cognitive science linguistics seeks through its explanatory power to throw light on the structure and workings of the mind, This latter goal does not for the moment concern us here.

What has been misunderstood in bringing linguistics to bear on speech synthesis and automatic speech recognition research is just what it is that linguistics describes. It characterises the knowledge base human beings have which they access for the purposes of the encoding and decoding procedure. Linguistics has something to say about the structure of the knowledge base and the system constraints placed on it. But in general it has nothing to say about accessing procedures or the encoding and decoding algorithms. Although the knowledge base description may contain rules which delete, add or transform elements they are in no sense intended to be interpreted as elements in some specific encoding or decoding algorithm. It is only in the theoretical sense that the entire knowledge base describes the potential of the entire language, but it is strictly not true that some part of the knowledge base constitutes an algorithm. What a part of the knowledge base might constitute is a description of those rules which might be accessed by some un-prescribed algorithm for the purposes of some encoding or decoding procedure to link a particular concept with a particular soundwave. This is the source of the error researchers in speech synthesis and automatic speech recognition, engineers and linguists alike, have made.

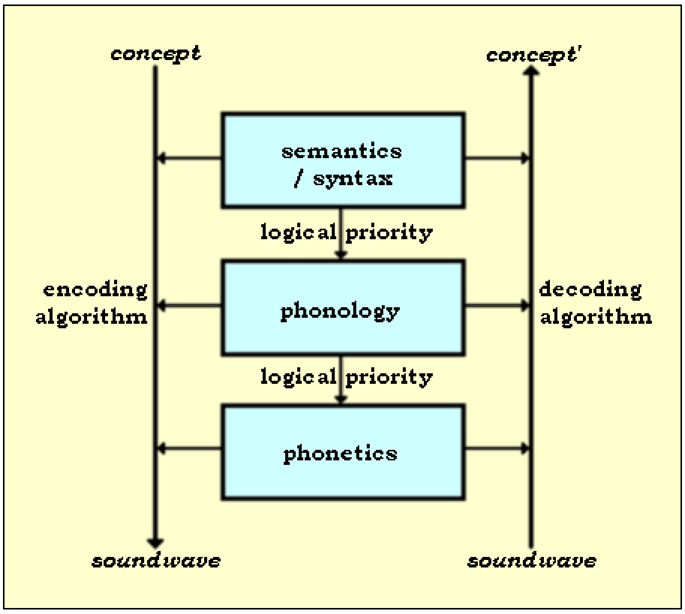

Fig. 1 The general model in linguistics.

Figure 1 illustrates the general model in linguistics. The overall knowledge base is subdivided into different components. There are formal and substantive reasons for the subdivision which need not concern us here. Let us simply refer to these components as the semantic/syntactic, the phonological and the phonetic knowledge bases. The vertical line linking the components indicates that particular knowledge bases are logically prior to others: they are not temporally prior or procedurally prior since linguistics has nothing to say about timing or procedural activity. The dashed lines are not part of linguistic description. They indicate potential procedural and accessing flow in some system outside the domain of linguistics, and are there to show a relationship between the knowledge bases other than the logical one dealt with by linguistics.

The semantic/syntactic and phonological knowledge bases each contain

- lists of the primitives associated with their particular level in the grammar, and
- sets of rules constraining the co-occurrence of these primitives.
- So at the phonological level [3] the knowledge base contains information on
- the set of phonological features in use in the language and the rules constraining their combination in the formation of phonemic segments in the language;
- the set of phonemic segments actually available in the language and the rules constraining their sequencing in the formation of words;
- rules characterising transformations of the phonemic segments in particular contexts with other phonemic segments;
- a prosodic sub-section comprising primitives and rules for the assignment of prosodic contours to words and word groupings.

Together these primitives and constraints ultimately characterise the extrinsic allophonic patterning used to encode all sentences in the language. It is important to note that it is all sentences, not a particular sentence.

At the phonetic level things are slightly different. Until recently it was believed that there was little in phonetics of interest to linguistics, since apart from a logical entry point accessed by strings of extrinsic allophones the entire component was dominated by constraints of myodynamics, aerodynamics and acoustics. While this was believed to be the case the level was outside the domain of linguistics, since linguistics is a cognitive science and such physical phenomena were anything but cognitive. But we are now coming to believe [4] that although there are many physical constraints on the production of speech sounds it is nevertheless the case that these constraints can be systematically inhibited or enhanced, and indeed are fine tuned under cognitive control for linguistic purposes. The ability to manipulate physical constraints under cognitive control is extremely important to us when we come to consider variability in speech. If cognitive control of physical constraints is possible it must be the case that the nature of those physical constraints must be known to the system. Hence the need for a phonetic knowledge base enumerating those constraints as part of general linguistic knowledge.

It seemed necessary to go into some detail about what linguistics can tell us and what it does not. To repeat linguistics is a descriptive characterisation of the knowledge a human being has to enable him to encode and decode speech. It says nothing about the acts of encoding or decoding. In speech synthesis and automatic speech recognition on the contrary the focus of attention is precisely on encoding and decoding. Synthesis and recognition are not equivalent to descriptive models: they are simulations.

DESCRIPTION AND SIMULATION

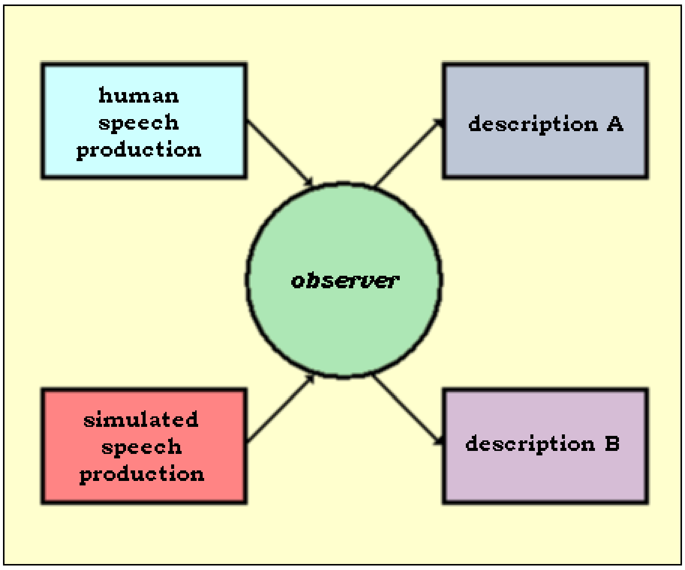

Fig. 2 The relationship between descriptions and simulation.

Descriptions and simulations are very different objects and it is important not to confuse the two. Fig. 2 illustrates their relationship. An object, human speech production, is observed by the scientist who produces a description of it. This is description A in the diagram, and is equivalent to the description provided by linguistics together with a characterisation of the general algorithm and procedures for accessing the knowledge bases. A second object, simulated speech production, is also observed by the scientist. He produces description B of the simulated speech. The more like human speech production the simulation becomes so description B approaches description A. Our criterion of success in simulations is the degree of difference between descriptions A and B. The point of this diagram is to indicate that description A of the human speech cannot be substituted for the simulation. That is, the description of human speech is not and cannot become a simulation. A simulation is a different type of object from a description, the latter being a transformation of an observed object.

This is not to say that descriptions of real objects are not useful to the builder of simulations. Caution is necessary to understand the exact nature and purpose of the description and certainly it must not be assumed that the two can be substituted.

To return to the idea that researchers in speech synthesis and automatic speech recognition are in error about their assumptions as the nature of linguistics. I have now described two mistakes. The first is the assumption that the linguist's descriptive knowledge bases are algorithms for speech production or perception, and the second is the confusion between descriptive modelling and building simulations.

Both synthesis recognition systems profit from incorporating linguistic knowledge. But once again caution is necessary. In the work of so many researchers linguistic knowledge seems to mean the knowledge linguists have. It rarely means what it ought to mean: the knowledge base described by linguistics. But supposing we understand correctly what is meant by linguistic knowledge, how shall we incorporate it?

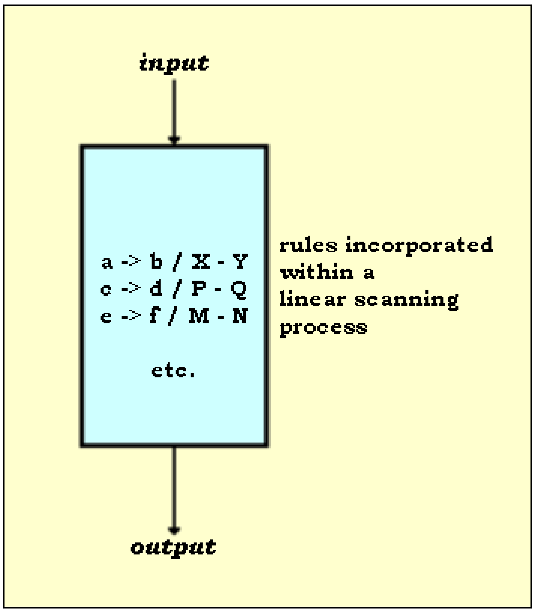

Fig. 3 The phonological level of most synthesis systems.

Fig. 3 diagrams the phonological level of any of the current leading speech synthesis systems. A string from some higher level is input to a process. This process consists of a scan, often linear and therefore unprincipled, of a large set of rules taken directly from linguistics. Each rule describes a transformation which must be applied to a particular phoneme in the input string if contextual conditions are satisfied. So for example,

the string X*a*Y becomes X*b*Y

that is, a becomes b if in the input string it occurs with a left context X and a right context Y, where X and Y can be null or strings of any length.

In linguistics notation we might write $a \rightarrow b$ / X - Y. Often in such systems the entire set of rules has to be scanned for every element in every input string - a seemingly ridiculous procedure. Apart from such unprincipled and wasteful inelegance the procedure is not based on any sound theoretical consideration. The theory of human speech production would suggest a procedure accessing in a principled fashion a knowledge base of suitable rules, where the focus was on the method of access rather than on the knowledge base. I am not just playing with alternative layouts: the two approaches are not equivalent.

I shall not go into theoretical reasons for this assertion. But here is an example which makes the point. There are rules in the phonology and phonetics which describe the varying amounts of precision required in the articulation of speech sounds [5]. In the linguistic description these rules are labelled optional. Assume they are placed within the speech synthesis algorithm. Selection is made by scan of the rules for the item and its context so all contradictory rules are selected. The inclusion of a metarule to the effect that in the case of optional rules only one may apply blocks all of the rules but one. Which? Is the choice to be random? Even a cursory examination of human speech reveals that the choice is not random but based on a reasoned decision.

REASONED DECISION TAKING

Reasoned decision taking in human beings seems to rely on weighing up evidence or information from a number of sources. The evidence may constitute facts, which may shift in importance depending on circumstances, or beliefs. Reasoned decision taking seems to rest on the balance of probabilities surrounding these facts or beliefs. Use of the balance of probabilities at any one time is one of the mechanisms by which computation can take place when the evidence supplied comes from a large number of sources which may be different in type and when the evidence itself varies with respect to reliability.

It is reasoned decision taking based on evidence which is neither clear cut, nor guaranteed factual or stable and which relies on an assessment of probabilities which to a large extent distinguishes human behaviour from the usual form of machine behaviour. Simulation of reasoned decision taking falls within the domain of artificial intelligence. I am going to suggest that there is room in synthesis and recognition systems research for experimenting with a general artificial intelligence approach to some of the seemingly intractable problems we are coming across. Linguists who engage in simulation modelling rather than descriptive modelling are working within the area of artificial intelligence rather than pure linguistics.

At Essex University we have been experimenting with devices which can perform reasoned accessing of the knowledge bases described in linguistics. The knowledge bases are slightly different because they are intended for simulation rather than descriptive purposes. The kind of device which springs immediately to mind is the so-called expert system. Expert systems are designed to acquire evidence from their surroundings, to conduct a reasoning process and reach some conclusion by the selection of a particular goal from a number of given goals. Such a system for use at the phonological level in speech synthesis has been developed in our laboratory by Katherine Morton [6].

The goals of such a device may be a set of linked optional rules in the knowledge base. The task is to select by reasoning not the correct rule to apply in the circumstances, but the appropriate rule. We are simulating an area of human behaviour where correctness is not the right term. By definition it could easily be the case that all the rules are correct (or they would not be in the knowledge base), but they are not equally appropriate on any one occasion. Each rule has assigned to it within the knowledge base an a priori probability weighting. That is, just as the knowledge base itself describes the native speaker's knowledge of the entire language and all utterances in that language, so these weightings indicate the probability of occurrence of each rule in the entire language. The reasoning device, or expert system, interrogates a number of sources of information. What sort of mood does the speaker (in this case the device determining what the system is to say) wish to convey? Does it believe the listener to be naïve in respect of the subject matter of the conversation? Does it believe the ambient noise or other factors in the environment merit special attention to precision in the utterance? Are there any reasons to suppose the listener will have any special difficulties in understanding the proposed utterance?

Other interrogation is made to sources of information within the synthesis system. Are there any special semantic considerations to be applied? Are the phonological units in the utterance, either separately or in combination, high or low in redundancy? Are there likely to be any special difficulties encountered by the lower-level phonetic component when attempting to execute the projected utterance? And so on.

Before it is answered each question has a predetermined effect on the *a priori* probabilities assigned to the options to be chosen among. Some of the questions will have only a binary possibility for their associated answers. Some questions will have a factual answer falling within a given range of possibility. Other questions will have answers informing of a degree of possibility or probability. All the answers are computed with respect to their influence on the a priori probability weightings assigned within the knowledge base to the range of options. Finally one option will emerge with a resultant probability weighting greater than the others: and that is the one chosen as the most appropriate.

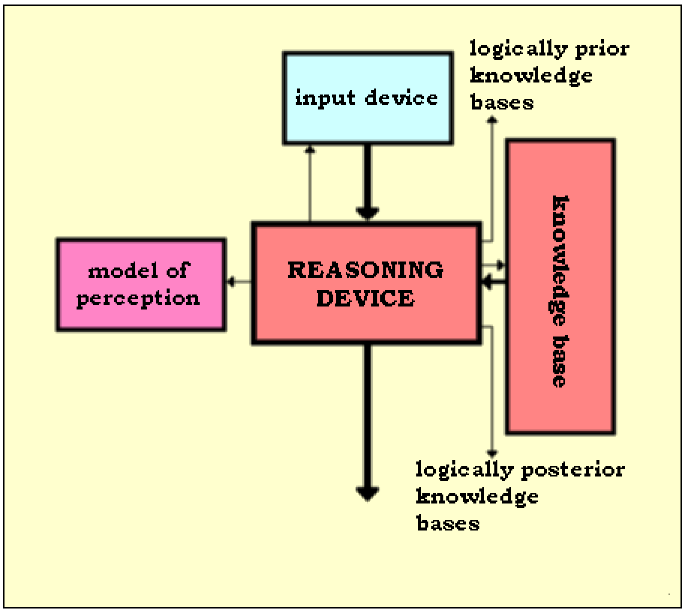
input device
logically prior knowledge bases
model of perception
REASONING DEVICE
knowledge base
logically posterior knowledge bases

Fig. 4 Reasoning in speech synthesis.

This is the principle behind our simulation of reasoned decision taking in speech synthesis. Fig. 4 illustrates simply what is going on. At the top of the figure there is the input device. It is here that a decision is taken as to what concepts are to be encoded as speech. An example of such a device would exist in an interactive database inquiry system such as the Alvey sponsored VODIS project. At the phonological level the expert system, as this level's reasoning device, accepts an input from the main concept-to-speech encoding algorithm. The task of the expert system is to transform this input depending on information accessed from the associated phonological knowledge base. That accessing is reasoned, as is selection from the knowledge base. The dashed links from the expert system indicate consultation or interrogation paths used in deciding what from other associated knowledge bases is relevant or appropriate for the job of transforming the main input to the main output.

There is one aspect of the system which has not yet been mentioned. Human beings as young children acquire the contents of their linguistic knowledge bases. In linguistic theory this is described as being accomplished by an iterative process performed by a language acquisition device producing successive grammars which progressively approach the mature grammar. Human beings as adults continually adjust their knowledge bases and indeed their strategies for decision taking. In other words they continually learn. The simulation must therefore have learning capability. Bridle 17] has discussed learning machines from the viewpoint of credible modelling of the mechanisms involved in learning. We are concerned with the functional counterpart of such mechanisms So although Bridle's machines are of a somewhat different type from what is discussed in this paper' we would be concerned with how the various weighting factors appearing in his model are arrived at in a principled way, and with the exact functional nature of the nodes or even labelling of nodes within the network.

Our proposed system is rather ambitiously a total one. We see no distinction between knowledge bases for speech synthesis or automatic speech recognition, nor any need for different types of mechanism to access them. The main synthesis and recognition algorithms may well be different, but we believe that better synthesis and recognition will emerge if, as in the human system, they are modelled as different modalities of the same overall device. Such a dual-mode device has many more possibilities internally for continuous updating of the weighting functions I have referred to above. So, for example, the device might ask itself: *Was it the case that the utterance as I produced it evoked in the listener the desired or expected reaction?* If the answer to this question is no, then some adjustment can be made automatically to some aspect of the decision taking processes within the device. In other words the system needs to have the means of detecting its own errors and in addition the means to repair the sources of those errors. In the field of artificial intelligence this kind of strategy is an aspect of what is known as knowledge engineering. That is, the acquisition and structuring of knowledge bases: in this case conducted automatically on a continuous basis.

CONCLUSION

This paper has discussed the nature of linguistic models and what they have to offer research in speech synthesis and a characterisation of the knowledge base to support the encoding/decoding process of relating concepts with speech sounds while saying nothing about the actual procedures involved. Speech synthesis and automatic speech recognition systems are simulations, not descriptions, focussing on the encoding and decoding algorithms The direct substitution of sets of rules characterising a knowledge base for procedures is a mistake, as is the substitution of a description for a simulation. At the present time accessing of the knowledge bases in our simulations of speech production and perception is unreasoned and naïve. I have described an experimental method for reasoned access to the knowledge bases which is proving fruitful in producing a more natural and variable synthesised speech than currently available systems.

References

[1] N. Chomsky, 'Formal properties of grammars', *Handbook of Mathematical Psychology* 2 (R. D. Luce, R. Bush and E. Galanter - eds.), New York: Wiley (1963)

[2] N. Chomsky, *Syntactic Structures*, The Hague: Mouton & Co. (1957)

[3] N. Chomsky and M. Halle, *The Sound Pattern of English*, New York: Harper & Row (1963)

[4] M. A. A. Tatham, 'An integrated knowledge base for speech synthesis and automatic speech recognition', *Journal of Phonetics* 13 (1985)

[5] M. A. A. Tatham and Katherine Morton, 'Precision', *Occasional Papers 23*, Essex University (1980)

[6] Katherine Morton, 'Intelligent speech synthesis', paper and demonstration given to *the Leeds Experimental Phonetics Symposium,* September 1986

[7] J. Bridle, paper presented to this conference.