Intelligent Speech Synthesis as Part of an Integrated Speech Synthesis / Automatic Speech Recognition System

Mark Tatham

Reproduced from *The Structure of Multimodal Dialogue*, eds. M. M. Taylor, F. Néel and D. G. Bouwhuis, pp. 301-312.

Copyright © 1989 Elsevier Science Publishers B.V. (North-Holland)

INTRODUCTION

We are concerned with speech synthesis as an integral part of an interactive database inquiry system. Several of the chapters in this book deal with such systems (e.g. Morel, Ch. 24; Waterworth, Ch. 25; Proctor & Young, Ch. 29; Ostler, Ch. 31; Carbonell & Pierre, Ch. 32). Our approach differs from most current synthesis methods in that it permits input both from text and from concepts, and contains within it intelligent elements.

Database inquiry systems usually have an automatic speech recognition input, some intelligent access to the database, and a speech synthesis output. A typical system including its human user is shown in Figure 22-1. In most systems the top level intelligent section outputs text which is then turned into speech using synthesis. Text, however, fails to encode much of the semantic information which might assist in producing a more natural output and allowing a more convincing interaction with the human user. Particularly in the prosodics many deficiencies could be overcome if what is discarded in basing the synthesis on text were retained. This is especially true in subtle areas of communication, such as mood and attitude (Hunt, Ch. 21).

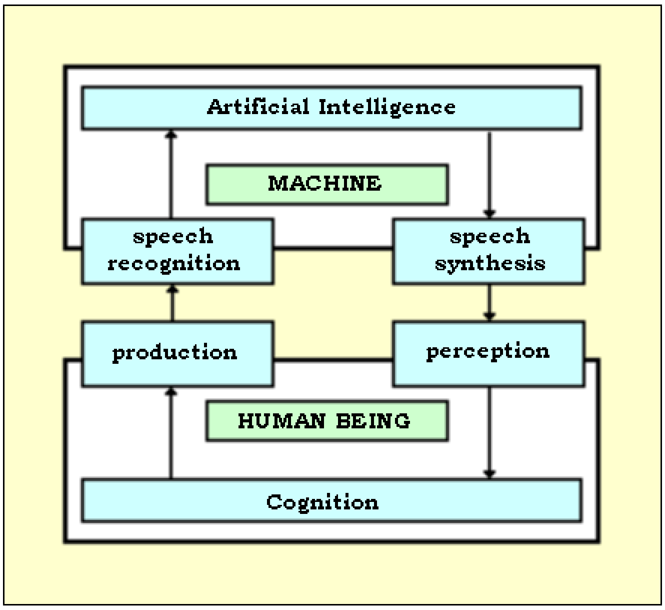

Fig. 22-1. A conventional database inquiry system using speech.

In current linguistic theory speech production from concept in the human being is modelled as a series of procedures successively drawing on knowledge bases to create an appropriate sound output. Transformational Generative Grammar and its derivatives, for example, systematically characterise these knowledge bases, though saying little or nothing about the procedures that draw on them. In the creation of synthetic speech it seems sensible to adopt a similar strategy, paying meticulous attention to keeping procedures separate from the accompanying knowledge bases. Too many of the current systems conflate the two into one continuous rambling algorithm, often opaque to inspection and incapable of easy and sensible modification.

In designing the system it was important for us to remember that the knowledge bases of linguistics are in fact only descriptive models. They result from observations of language conducted by linguists working within the framework of a particular metatheory. The descriptions are, of course, not the original object.

A simulation of the kind we have in synthesis constitutes a second object for observation and description. A perfect simulation would cause the linguist to produce a description identical to that of the original object — natural language. The descriptions of language that linguists and cognitive scientists make may he important in forming the basis of simulations, but it is an error simply to 'program up' a description and call it a simulation.

The synthesis system described later in some detail involves two alternative inputs: text, or the output from an artificial intelligence processor. A cascaded series of procedures results in the creation of an appropriate soundwave. These procedures are: a dictionary search, intelligent parsing, phonological interpretation, intelligent adjustment of the allophonic string, and phonetic constraint of the allophonic string. At distinct levels within the main synthesis algorithm they consult respectively a dictionary, a semantic/syntactic knowledge base, a phonology, a cognitive phonetics knowledge base and a database of phonetic constraints. The knowledge bases are interrogated during the main flow of the algorithm by 'reasoning devices' which intelligently find their way around the knowledge bases to provide local input to the main algorithm.

Of particular interest is the cognitive phonetics knowledge base and its usage for adjusting the allophone string output from the preceding phonological process. The cognitive phonetics part of the system is new, both in the theory of language (Tatham, 1984) and in the design of speech synthesis systems (Tatham, 1985). It is based on recent understanding of some aspects of variability in speech. In particular it has been observed that the articulation of segments occurs in some contexts with more precision than in other contexts (Tatham & Morton, 1980). Precision is increased either

- when the semantic loading of a segment is increased, or
- when perceptual confusion on the part of the listener is predicted by the talker.

Cognitive phonetics is therefore sensitive to at least the phonological and semantic systems of the language and is able to do a running prediction of the probable perception of the proposed output.

LINGUISTICS IN SYNTHESIS AND RECOGNITION

There can he no doubt that the relative success of text-to-speech synthesis systems and improvements in automatic speech recognition systems owe much to the inclusion of linguistic and phonetic information. In speech synthesis, since the earliest days of synthesis-by-rule, information from phonology and phonetics has been used in the form of wholesale incorporation of rule sets form those components of linguistics. This can he seen in Figure 22-2, in which the levels used in linguistics are seen as components of the synthesis algorithm.

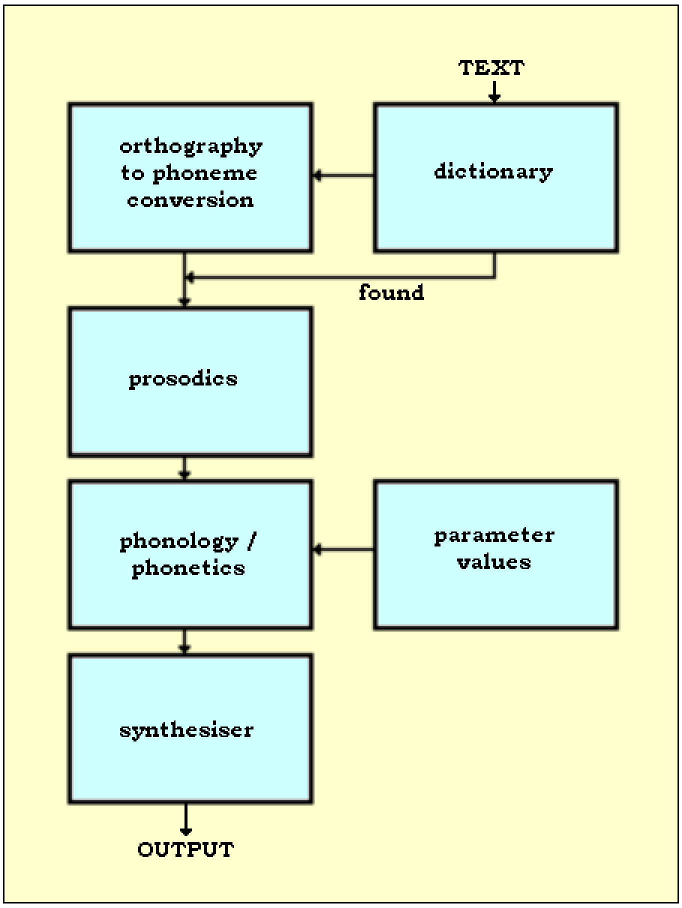

Fig. 22-2. Linguistic levels of description as components of a standard synthesis algorithm

In speech recognition the inclusion of linguistic and phonetic knowledge has come later, as the orthography to phoneme dictionary result of failure to find linguistic conversion units in the soundwave itself no matter how sophisticated the method of acoustic analysis adopted. The rules of phonology and phonetics are regularly used in a top-down prosodics approach to disambiguate the output of the acoustic analyser. The prevalent idea in recognition is that linguistic information should operate in some predictive capacity to limit, on a statistical basis, the possibilities that any one portion of the waveform phonetics will he recognised as one of several different speech elements. For example, the segment sequencing F rules that describe the phonotactics of a language (i.e., the possibilities for linear concatenation of the various sound elements of the language) can aid the recogniser to avoid matching an impossible sequence of elements. The result in recognition has been faster and more accurate matching and an improvement in success rates.

But a number of mistakes have been made in using linguistic information in recognition, and particularly in synthesis. The speech produced by the latest devices is just not good enough, and in seeking to find out why this might he so a reappraisal of our whole approach is called for. The failure has little to do with the design of the actual synthesiser itself. By the actual synthesiser I mean that part of the system which is responsible for producing the waveform rather than the software which determines how the synthesiser is driven. It is true that synthesiser hardware is by no means perfect and we are still making improvements both to our implementation of the acoustic models of speech and to the models themselves. But is also true that the hardware is not the limiting factor. We are no longer pleased with the fact that synthesisers do speak, and are becoming much more critical of the way in which they speak.

What is being reappraised is that it is sufficient to produce speech that is easily understood: the synthesiser should also produce natural-sounding speech. Of course, what is intelligible is not necessarily natural-sounding. It may well be that sometimes the speech should not sound natural, and that the listener should be kept aware that it comes from a machine by the machine-like quality of the speech itself (Hunt, Ch. 21; Mangold, in the discussion, see Ch. 34). But that is a social issue, and the fact that often we do need natural sounding speech puts that issue to one side. Phoneticians and linguists are now faced with the question: what makes human speech sound natural? This turns out to be a rather difficult question, for until there was the possibility of producing unnatural artificial speech the factors that made human speech natural were hardly considered.

Linguistics has virtually nothing to say about what characterises naturalness. The irony is that modem linguistics has gone out of its way to devise elaborate procedures for neutralising many of the characteristics of naturalness in its descriptions. An example might help clarify this point. Whenever an experiment is conducted on the acoustic waveform of speech it is customary to obtain many examples of a particular word or phrase sometimes from many speakers. So if the question has been asked ‘What is the frequency of the second formant in the vowel sound [i] in English?’ then the researcher will obtain recordings for analysis from several repetitions of a speaker saying the vowel in isolation or in words, and perhaps may obtain such recordings from several different speakers. Measurements will be taken and some relatively simple statistic applied to arrive at some abstract number, say 2200 Hz. Two observations are worth making:

- there is a fair chance that not one single one of the measurements made was 2200 Hz, and
- even within the speech of one speaker the variation of values obtained was quite large.

It has become part of the method of conducting experimental research in speech science that this variability, while often being noted, should somehow be normalised. There are two justifications given for this:

- a range of numbers is less manageable than a single number, and
- somehow the range of numbers is considered to be an aberration on the part of the speaker, the idea being that a talker would attempt to produce the same single value for a particular segment every time.

The First justification is of course nonsense. The second has more merit, and indeed is an important and productive consideration in linguistic theory.

Despite the theoretical justification for the normalisation of variability in analysing speech, doing this is of no help in devising a speech synthesis system. If human beings produce variability and synthesisers do not, then synthesisers are not producing natural speech. Although it may well seem obvious that speech would perhaps be more efficient without variability and although it may well be the case that the human perceptual system itself normalises the variability it hears, it is nevertheless the case that a human listener is extremely sensitive to the removal of variability from speech.

Variability is of course only apparent when listening to a comparatively long stretch of speech. Judgments of synthetic speech are rarely made using such a method of scrutiny. Thus a synthesised word may sound perfect and perfectly match the waveform of a human being saying the same thing-but when it is placed in a sentence of equally perfect words, something immediately seems wrong to the human listener. What is wrong is that the variability is not there. Listening to lengthy passages of synthetic speech is surprisingly fatiguing, even if any one portion of the waveform is correct.

LINGUISTIC THEORY

Linguistics is the science that deals with the grammar of languages. The theory is exemplified in models that have an utterly explicit and exhaustive formal structure (Chomsky, 1963). The theory of mathematical linguistics is very well developed. The metatheory is clear as to what it is that linguists are doing and why. As with any developing science there has been a certain amount of infighting in linguistics, but its 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, but has explanatory aims. Its main practica1 aim is to describe the knowledge one has of one's language (Chomsky, 1957). It is this knowledge that enables one to encode one's thoughts for transmission to another person and to decode the thoughts transmitted by another person. In addition, and in common with other branches of cognitive science, linguistics seeks to throw light on the structure and workings of the mind. This latter goal does not concern us here.

The ideas discussed above became the basis of linguistics around a quarter of a century ago, by which time it had become quite clear that exhaustive attempts to understand the output of the language process were becoming increasingly less productive if the only object of scrutiny was that output itself. The emphasis shifted therefore from what people produce either as speech or as writing, to what it is that they know of the language when they speak or write. The relationship between what a human being knows about and does with language is by no means obvious or simple.

When one uses linguistics in research on speech synthesis and automatic speech recognition it becomes clear that insufficient attention has been paid to just what linguistics describes. Linguistics and phonetics characterise the knowledge base human beings access to encode concepts into speech and to decode them back from speech. It has much to say about the detailed structure of the knowledge base and something about the system constraints placed on it. But it has nothing to say about procedures for accessing the knowledge base or about the encoding and decoding algorithms themselves. Although the knowledge base contains rules which delete, add or transform primitives or the basic elements used in language, they are in no sense intended to be interpreted as procedures in some specific encoding or decoding algorithm. It is only in the theoretical sense that the knowledge base describes the potential of the entire language, but no part of the knowledge base constitutes an algorithm. The rules of the knowledge base characterise all that can happen in the encoding process, aside from idiosyncrasies of encoding and decoding. In that sense they exhaustively describe or define the language. The language is regarded as being the set of all possible sentences-and that set is infinite. The knowledge base constitutes a description of the rules accessed by some un-prescribed algorithm so that the encoding or decoding procedure can link a particular concept with a particular soundwave. To treat the knowledge base as a set of algorithms is the error researchers in synthesis and recognition, engineers and linguists alike, have made.

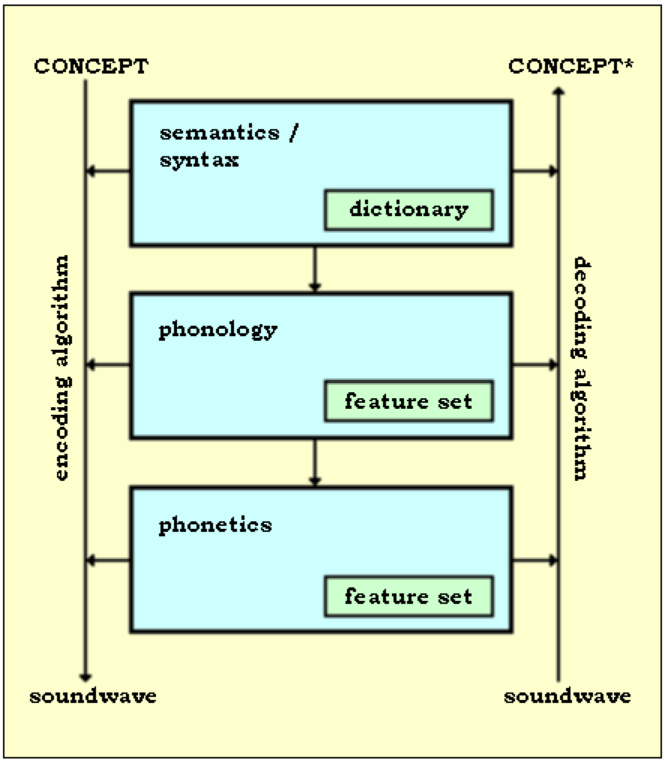

Fig. 22-3. Components of the general model used in linguistics.

Figure 22-3 outlines the general model used in linguistics. The overall knowledge base is subdivided into the different components shown in the centre of the diagram. There are formal and substantive reasons for the subdivision which need not concern us here (e.g., the forms of the rules in the syntax and phonology are different). Syntax is about the arrangement of words in strings, whereas phonology is about the arrangement of sounds within utterances. Let us refer to these components as the semantic/syntactic, the phonological and the phonetic knowledge bases. In simple terms they each consist of sets of elements (the small boxes within the components) and sets of rules manipulating those elements. The vertical arrows linking the components indicate 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, which would be within the domain of the general algorithm. The grey arrows are not part of a linguist's description of language. They indicate 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 that linguistics deals with.

The semantic/syntactic and phonological knowledge bases each contain lists of primitives associated with their particular level in the grammar, and sets of rules constraining the co-occurrence of these primitives.

For example, at the phonological level (Chomsky & Halle, 1968) 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 available 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 interact to 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 that although there are many physical constraints on the production of speech sounds, these constraints are nevertheless systematically inhibited or enhanced, and indeed are fine tuned under cognitive control. The ability to manipulate physical constraints under cognitive control is extremely important when we consider variability in speech, since we now know that much of the variability is controlled. If cognitive control of physical constraints is possible, the nature of those constraints must be known to the system. Hence a recognizer or synthesizer must need a phonetic knowledge base enumerating those constraints as part of its general linguistic knowledge.

Because of the confusion about exactly what it is that linguistics does, it seemed necessary to go into some detail about what it can tell us and what it does not. To restate the point: linguistics is a descriptive characterisation of the knowledge one has of one's own language, as well as of some aspects of language in general, which enables one to encode and decode speech. It says nothing about the actual encoding or decoding.

In speech synthesis and automatic speech recognition, however, the focus of attention is precisely on the encoding and decoding procedures. Synthesis and recognition are not equivalent to descriptive models: they are simulations.

DESCRIPTION AND SIMULATION

Descriptions and simulations are very different objects and it is important not to confuse the two. Figure 22-4 illustrates their relationship. The box in the top left of the diagram depicts an object, in this case human speech, which is observed by the scientist, who produces a descriptive model of it according to the general principles held in the current metatheory of the discipline (Descriptive model 1 in the diagram). Model 1 is equivalent to the model provided by linguistics, together with a characterisation of the general algorithm and the procedures used to access the knowledge bases. Remember, the current descriptions of linguistics are by no means complete; they talk only of the knowledge bases. At the bottom left of the diagram is a second object, simulated speech production, which is also observed by the scientist, who produces descriptive model 2 of the simulated speech according to the same principles and metatheory as before. I have shown an arrow between the two descriptive models, indicating their potential convergence. The more like human speech production the simulation becomes, the more model 2 approaches model 1, the description of human speech itself. Our criterion of success in simulations is the degree of similarity between the two models.

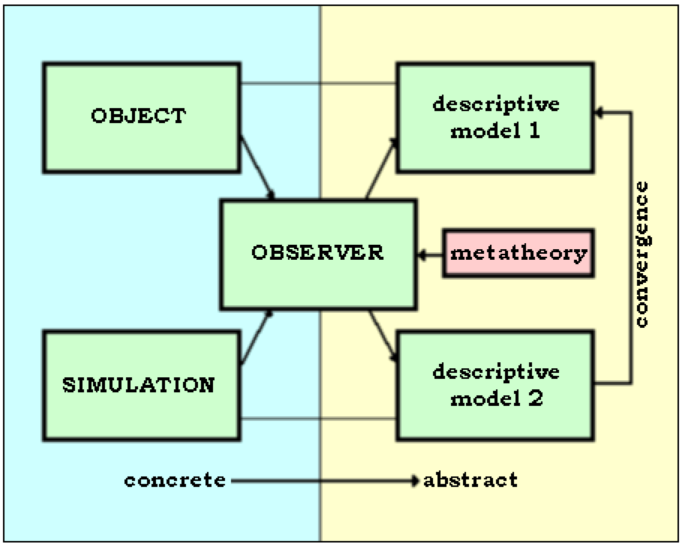

Fig. 22-4. The relation between description and simulation.

One point of this diagram is to illustrate that a descriptive model cannot be substituted for a simulation. The characterisation of human speech is a descriptive model and not a simulation. A simulation is a different type of object from a description and logically stands in the diagram to the left of the grey vertical line that separates the real world from the abstract world of models. Descriptions result from a transformation performed on a real-world object by an observer who is operating under some explicit principles. Simulations share more properties with real-world objects than they share with abstract descriptions of objects (Hutchins, Ch. 2, discusses the use of simulations, as metaphors in the approach to human/computer interaction).

Descriptions of real objects are nevertheless useful to the builder of simulations, since they do enable us to understand the real world by imposing some kind of logical framework on it. Caution is necessary to understand the exact nature and purpose of the description and certainly it must not be assumed that a description of a simulation can be readily substituted for a description of the real object.

INCORPORATING LINGUISTIC AND PHONETIC KNOWLEDGE

Let us return to the idea that researchers in synthesis and recognition have made some wrong assumptions about the nature of linguistics. One is that the linguist's descriptive knowledge bases are algorithms for speech production or perception, and another is the confusion between descriptive modelling and simulation building

It is true that both synthesis and recognition systems profit by drawing on linguistic knowledge. But yet more caution is necessary. In the work of 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?

Take as an example 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 that must be applied to a particular element in the input string when contextual conditions are satisfied. For example, the string X*a*Y becomes X*b*Y: 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* —» *b* / X — Y. Often in such systems the entire set of rules has to be scanned for every element in every input string-a procedure based on no sound theoretical consideration. By contrast the theory of human speech production suggests a procedure that accesses a knowledge base of suitable rules in a principled way. The focus of attention is on the method of access rather than on the knowledge base itself. These are not just alternative layouts: the two approaches are not equivalent to each other.

I shall not go into theoretical reasons for this assertion. But here is an example that makes the point. Phonology and phonetics have rules that describe the varying precision in the articulation of speech sounds. In the linguistic description these subsets of rules are labelled optional, which means just one member of the subset is selected. Assume they are placed within a trivial speech synthesis algorithm. Selection is made by scanning the rules for the item and its context. The result of such a procedure would be to select all the subset, that is, several rules that are mutually exclusive and may well be contradictory. The inclusion of a metarule to the effect that in the case of sets of optional rules only one may apply blocks all of the rules except one. But we cannot tell which, and the choice cannot be random. Even a cursory examination of human speech reveals that choice between optional ways of saying things is not random but based on a reasoned decision.

REASONED DECISION TAKING IN SYNTHESIS

Reasoned decision taking in human beings relies on weighing up evidence or information from a number of sources. The evidence may constitute facts that shift in importance depending on circumstances, or even depending on beliefs. Reasoned decision taking rests on evaluating the probabilities surrounding these facts or beliefs. In our simulation, we use the balance of probabilities at any one time as one of the mechanisms by which decisions can be computed. This become especially relevant when the evidence supplied comes from a large number of sources of different types and when the evidence itself varies with respect to reliability.

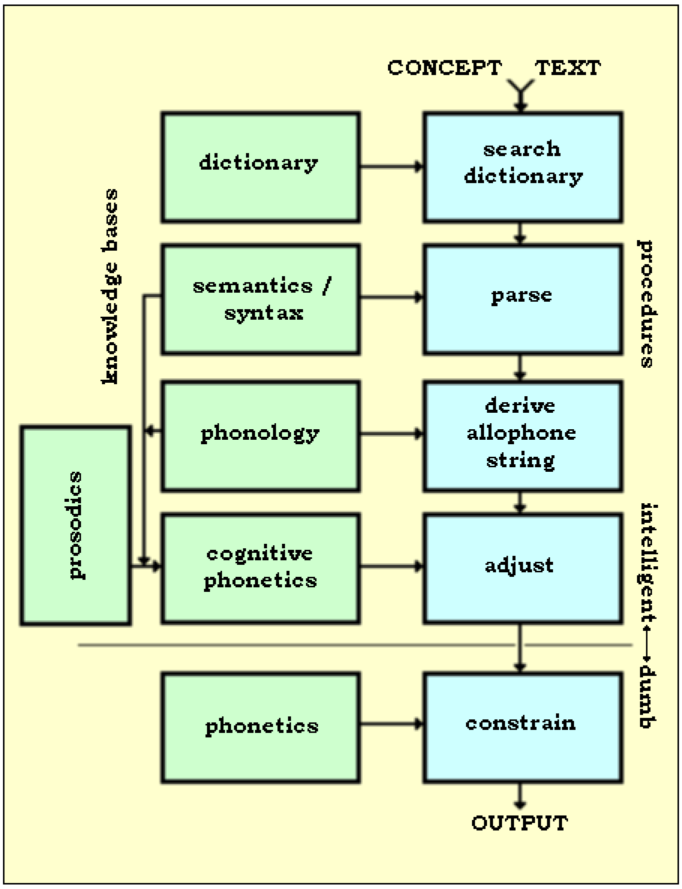

Fig. 22-5. A proposed model for a speech synthesiser.

Figure 22-5 represents the general outline of a more sensible approach to speech synthesis than many of the current models. The model makes correct use of information from linguistics and has several components not explicitly recognised in most systems. The main algorithm proceeds from top to bottom to the right of the diagram. Input takes the form either of concepts (suitably encoded) or of text. How we deal with concept input is not something I want to discuss here; suffice it to say that such an input is easier to deal with than text input since the latter does not properly encode some of the semantic subtleties so characteristic of human speech. These need to be assigned by the parser when the system is restricted to textual input.

To the left of the diagram are the various knowledge bases consulted by the main algorithm. Their contents are supplied by linguistics, though in a form suitable for usage in a simulation model rather than in a descriptive model. The horizontal grey line near the bottom of the diagram divides intelligent processes from unintelligent or dumb processes. Another way of saying this is that the processes down as far as cognitive phonetics are active processes that model cognitive activity in the human being. Below the line the processes are passive, modelling physical activity and constraints. Much has been written about the deceptively simple arrow in the main algorithm which crosses this cognitive / physical separator (Tatham, 1980), but this is not the place to rehash such philosophical arguments. In some sense the knowledge bases shown to the left of the diagram interconnect, though they are depicted here separately for clarity.

Text input to the system initiates a dictionary scan The dictionary itself is more comprehensive than those currently associated with synthesis systems, and aims to be relatively exhaustive in terms of the vocabulary of the language For words not found in the dictionary, there are orthography-to-phoneme conversion rules which are not shown here. Items found within the dictionary are indexed in a number of ways: indication of their underlying phonological shape is retrieved, along with syntactic category markers, some semantic information and other indexing which assists the various processes throughout the algorithm. Some of this information, together with a consultation of the rules of semantics and syntax, assists in parsing the input text. The parsing information enables the later assignment of suitable prosodic contours, and permits selection among the variants possible at the phonetic level.

Phonological procedures consult the phonological rules of the language to map the input string onto an extrinsic allophonic string. Extrinsic allophones are a symbolic representation of abstract sound variants and represent the lowest level of representation of the utterance normally dealt with in linguistic theory.

The main points of Figure 22-5 are the separation of the procedures of the main algorithm from the knowledge bases, and the active, intelligent nature of each of the levels in the algorithm itself. The output at each level is the result of reasoned decision taking with respect to information coming from the level's associated knowledge base and together with information coming from elsewhere when appropriate. That information is obtained by interrogation performed by the reasoning device.

It is reasoned decision making based on evidence which is neither clear cut, nor guaranteed factual or stable, and which relies on an assessment of probabilities, that 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 suggesting that there is room in synthesis and recognition research for experimenting with a general artificial intelligence approach to some of the seemingly intractable problems. Linguists dealing with these problems and who engage in simulation rather than descriptive modelling are working within the area of artificial intelligence rather than pure linguistics.

In the Advanced Speech Technology Laboratory in the Centre for Cognitive Science at Essex University we have been experimenting with devices which can perform reasoned accessing of knowledge bases derived from linguistics. The knowledge bases are slightly different because they are intended for simulation rather than descriptive purposes. The class of device which springs immediately to mind includes the so-called expert Systems. Expert systems are designed to interrogate their surroundings for evidence to enable them to conduct a reasoning process and reach some conclusion by the selection of a particular goal from a number of given possible goals. Such a system for use at the phonological level in speech synthesis has been developed in our laboratory by Katherine Morton (1986).

The goals of such a device may be a set of linked optional rules, of the kind I described earlier in this paper, 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 in which correctness in 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 at the same time they are not equally appropriate on any one occasion. Each rule has assigned to it within the knowledge base an a priori probability weighing. 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 in the simplest model these weightings indicate the probability of occurrence of each rule in the entire language.

The reasoning device interrogates a number of sources of information. What sort of mood does the speaker (in this case the device determining what they system is to say) wish to convey? Does it believe the listener to be naive in respect of the subject matter of the conversation? Does it believe the ambient noise or other factors in the environment merit special precision in the utterance? Are there any reasons to suppose the listener will have difficulty understanding the proposed utterance? And so on.

Other sources of information within the synthesis system are interrogated. 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? Is the lower-level phonetic component likely to encounter any special difficulties in 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 available for selection. Some of the questions have only yes-no answers. Some questions have a factual answer falling within a given range of possibility. Other questions have answers informing of a degree of possibility or probability. All the answers are computed together 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; it is chosen as the most appropriate.

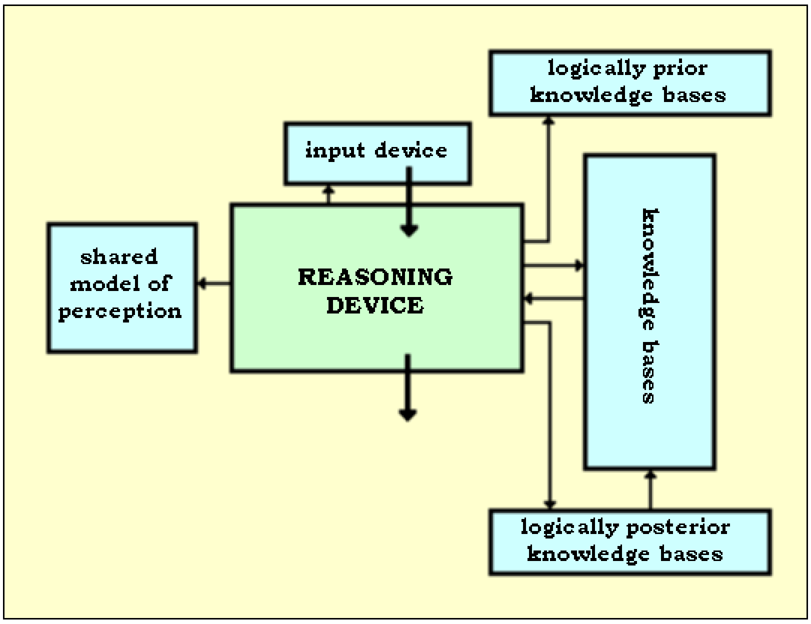
logically prior knowledge bases
input device
shared model of perception
REASONING DEVICE
knowledge bases
logically posterior knowledge bases

Fig. 22-6. The general structure of a level in the synthesis device.

The combining of probabilities is the principle behind our simulation of reasoned decision taking in speech synthesis. Figure 22-6 illustrates in a simple way what is going on at each level in the general case. In a synthesis system there would be several such units all having the same general properties. 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. Examples of such devices in use in interactive inquiry system were mentioned earlier. At any one level the expert system (Reasoning Device in Figure 22-6) accepts an input from the main concept-to-speech encoding algorithm. The expert system transforms this input depending on information accessed from the immediately associated knowledge base shown to the right of the diagram. Access to the knowledge base and selection of information from it is reasoned. The grey links from the expert system indicate consultation or interrogation paths used in deciding what information from other associated knowledge bases is relevant or appropriate to transform the main input to the main output.

Notice the box to the left of the diagram. This synthesis system incorporates a model of perception. The reason is that decisions regarding some of the articulatory and acoustic precision variants in speech production are determined by a predictive assessment of the likely success of perception of the intended utterance. A person who is concerned about being misheard because, say, of high ambient noise, will markedly increase the precision of some portions of the utterance.

This inclusion of a model of perception within the synthesiser reveals that our proposed system is, rather ambitiously, for recognition as well as for synthesis. We see no distinction between knowledge bases for speech synthesis and for speech recognition, nor any need for different types of mechanism to access them. The main synthesis and recognition algorithms may well be different and areas of the knowledge bases may be focussed or accentuated differently depending on whether synthesis or recognition is currently in progress, but we believe that better results will be obtained if, as in the human system shown in Figure 22-3, they are modelled as different modalities of the same overall device.

Such a dual-mode device has many internal possibilities for continuous updating of its weighting functions. 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 adjustments can be made automatically to some aspect of the decision taking processes within the device. The system has ways of learning, of detecting its own errors, and can repair the sources of those errors (Tatham, 1986). In the field of artificial intelligence this kind of strategy is an aspect of knowledge engineering. the acquisition and structuring of knowledge bases, in this case done automatically and continuously.

CONCLUSION

This chapter has discussed the nature of linguistic models and what they have to offer research in speech synthesis and automatic speech recognition in interactive database inquiry systems. Linguistics provides a descriptive characterisation of the human knowledge base to support the encoding / decoding process of relating concepts to speech sounds, while saying nothing about the actual procedures involved. Speech synthesis and automatic speech recognition systems are simulations, not descriptions, focusing 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 access to the knowledge bases in our simulations of speech production and perception is unreasoned and naive. 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 of the kind now needed in interactive database inquiry systems.