Speech Production and Synthesis - Future Development

Kate Morton

Reproduced from the PhD Thesis *Speech Production and Synthesis*, June 1987.

The work reported here [in this thesis] has had speech production as a common theme. Three separate aspects can be identified:

- gathering data about speech production,
- interpreting the data according to linguistic/phonetic theory of the time,
- developing a simulation of speech production in the form of speech synthesis.

THE DATA AND THE THEORY

Experimental work is always conducted within a theoretical context even if the experiments are primarily to gather data, and no particular theoretical context is stated. I have been interested in establishing facts about speech production within phonetic theory. In the early stages, this involved using electromyography as a technique for the purpose of going deeper into the production system than was possible using surface techniques such as in air flow and acoustic investigations.

THE GAP BETWEEN LINGUISTICS AND PHONETICS

Linguistic theory changed considerably in the 50s and 60s from describing surface events to establishing underlying constructs and relating these to the surface events. At the same time, phoneticians were not looking at language and speech in the same way as linguists; this resulted in a gap developing between linguistics and phonetics. TGG linguistics claimed to be able to say something about the mind, but phonetics remained primarily concerned with surface descriptions of speech processing and how the articulators produced the speech waveform.

In the 60s generative phonology used phonetic data to characterize sound shapes, though its purpose was to say something about the mind. Phonetics was not concerned with properties of the mind; its domain was to describe the physical realization of phonological requirements. In order to bring the two areas together, and since both were concerned with the output of the grammar, it seemed reasonable to try to put phonetics into a linguistic framework. This involved

- making a distinction between competence and performance, and
- constructing primitive objects with rules constraining their co-occurrence.

It became possible to suggest this reorientation following Chomsky's change to the linguistics paradigm.

INTRODUCING COMPETENCE INTO PHONETIC THEORY

Competence was defined as what a speaker knew about his language. Since a phonological requirement could only be realized if the human being knew what was needed carry out that requirement, a phonetics competence was proposed which could describe the knowledge the speaker had about how to realize phonological intentions. So, if phonology wanted /a/, phonetics knew that vocal cord vibration was required, a low back tongue position, etc. Phonetic competence took the form of facts and rules contained in a knowledge base - a characterization of how to physically realize phonological segments within an utterance,

PHONETIC RULES

The knowledge base was not only about the physical aspects of speech production known to the human being. It seemed also that the speaker had to know something about the behavior of segments when they occurred in specific contexts. For example, the speaker needed to know about coarticulatory effects, so that they could be modified systematically if necessary. Thus, in addition to tables of physical segmental specifications, sets of rules were needed to describe physical changes to these canonical phonetic forms in varying contexts. In keeping with linguistic goals, these rules were formulated with maximum generalization.

THE THEORY OF PHONETICS GOES BEYOND COMPETENCE

Chomsky's original proposal for syntax (1957) consisted of a set of hypotheses about how the mind might work, and at the same time characterized a system that provided the knowledge for producing the same output as a human being. In the 60s, linguistics characterized competence, but details of performance were not discussed by linguists. When researchers in other fields tried to adopt linguistics they wrongly tended to see performance as an enactment of competence.

The theory of phonetics, however, had traditionally been about acts of performance. For the phonetician speech production cannot be completely described if the theory concerns itself with a characterization of speaker knowledge only. In applied areas like language teaching and speech synthesis, confining the theory to generalizations, as would be the case with a competence model, was insufficient because applied areas deal with specific events. A model of the relationship between generalizations and these individual acts of performance was needed.

A MODEL FOR SYNTHESIS

A model of speech production should be able to provide some theoretical support for a speech synthesis system. The JSRU text-to-speech system, developed in the early 60s, was based on context sensitive production rules. Current synthesis systems are also rule-based, but in addition make a clear distinction between phonology and phonetics. However, even after twenty years, synthesis systems still produce speech that can sound unintelligible, machine-like and unnatural These deficiencies do not appear to be inherent in the hardware, or due to mistakes in implementing details of linguistic or phonetic theory. The problem may center around the form of the linguistic model used as the basis for what is in fact a simulation of the human speech production system.

SPEECH SYNTHESIS AND SIMULATION

Speech synthesis systems simulate human speech production with varying degrees of success. As has been said before, current linguistic/phonetic descriptions do not provide a particularly good basis for simulation since they are incomplete, there is no explicit procedural relationship between components, and linguistic rules are not stated in such a way that they can easily be translated into algorithms. In addition linguistics does not model cognitive processes. Future production modelling will probably involve cognition at all stages.

It is too early to say just what a simulation based on modelling cognitive processing would look like. I have suggested that if human cognitive processing can be described as sets of reasoned decision taking processes, a linguistics model suitable for simulation might be constructed which includes stages of reasoned decision taking as a means of selecting information from the knowledge bases.

REASONED DECISION TAKING IN THE SIMULATION

One of the ideas to emerge from considering the form of a simulation model involves deciding on what basis knowledge can be retrieved. The knowledge bases contain optional rules which account for a wide range of variability. Selection involves retrieving and

evaluating information from several different sources. The information is not always predictable (for example, ambient noise) so the evaluation process needs to be continuous.

I have been looking at devices capable of reasoned decision taking, and proposed such a device for simulating speech production. This can be regarded as a reasoning device with three types of input (A, B, C) and one output. I shall use input and output to mean both the *channels* along which information flows and the *information* itself (see Fig. 14).

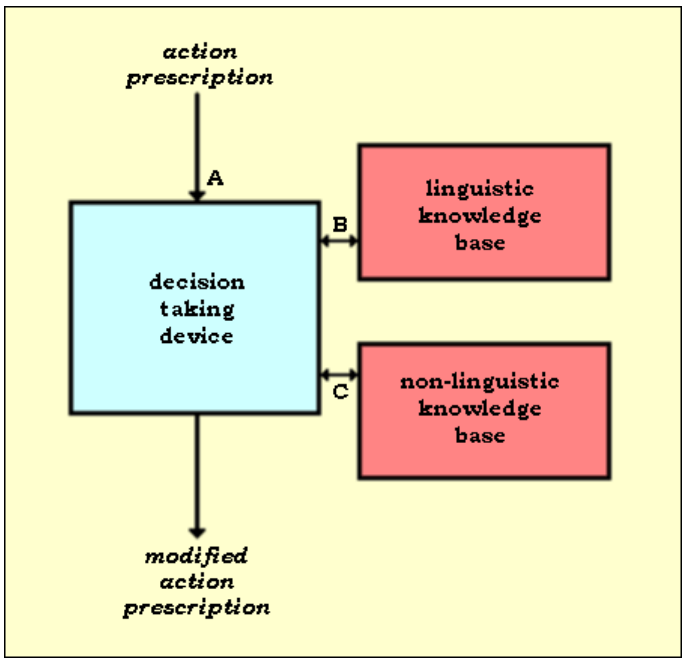

Fig. 14 Transformation of an action prescription following reasoned decision taking.

Input A is from the main algorithm which encodes concepts as sounds. For example, consider a phonetic reasoning device which can determine phonetic realization of phonological requirements: the requirement enters the device at input A as an action prescription. An action prescription is a generalized specification of what is to be done, but not how to achieve it.

The output of the device is the result of deciding how to realize the phonetic requirement: it too is an action prescription directed at the next stage of processing, which may involve further reasoning.

The device works toward achieving a goal , which is to provide another action prescription. Input A is the result of a higher level decision and specifies a requirement without alternatives. The output goal is selected from a number of different possibilities by the reasoning procedure.

Input B is an access channel to a general knowledge base of possible goals or output action prescriptions for this device. They are grouped as a logical set, brought together at this point. Thus, at the phonological stage there is a knowledge base specifying features and their combinatorial restrictions, and another specifying extrinsic allophones and the constraints on phonological contextual variability. At a phonetic stage there will be knowledge bases specifying co-ordinative structures and equations of constraint governing their activity and methods of marshalling.

Thus, the device has an input action prescription (Input A), a channel to an associated knowledge base (Input B), and an overall constraint to make a decision which will produce an output action prescription.

Input C is a bundle of channels which allow the device to access other knowledge bases within the overall system. Some will be analogous to competence components in linguistic

descriptions. Others will consist of non-linguistic information about the speaker's mood, intentions and awareness of the hearer.

THE MECHANISM FOR DECISION TAKING

One characteristic of the proposed device is that it can deal with scalar information by incorporating a probability algorithm in its inferencing mechanism. Replies to queries can be in terms like 'perhaps' 'possibly', 'somewhat'. It is essential to build in this kind of non-binary reply.

For example, in English there is a phonological rule "Devoice voiced consonants in utterance final position", but a human being actually producing speech is more likely to produce slightly devoiced voiced consonants which depend on how much the speaker wants to emphasize the word, how fast and how loudly he wants to speak, and how he assesses the listener's attention and interest in what is being said. The device evaluates this information to respond with a consonant which is not necessarily either fully voiced or fully devoiced.

The most useful evaluation would be of data that may be imprecise, variable, defective, or missing. A decision making device should evaluate the total available information in the light of rules and available options for solutions (goals). It should then output a preferred solution together with a confidence metric attached to it and, if required, rank order other possible solutions together with confidence ratings.

Reasoning proceeds according to rules held by the decision making device. It knows what weighting to give to information from different sources So if information is to be given a high weighting in the decision yet arrives indexed with low confidence then this information will play a significantly less important role in selecting the goal. The probabilities of the goals themselves are brought into the decision equation. All these variables are thus brought together to provide a unique solution: the selection of an appropriate goal. Since on some other occasion it will almost certainly be the case that the indices on Input C information will be different, solutions are unique to the information environment which gave rise to their selection.

Because the model I am proposing is a cascade of such reasoning devices, the information reaching any one device takes the form of the typical output described above. That is, it has a probability and/or confidence rating attached to it.

THE DEVELOPMENT OF INTELLIGENT SYNTHESIS SYSTEMS

In the early 60s the researchers in speech synthesis turned to linguistics and found a well thought out, clearly structured descriptive system presented in formal terms. I have tried to show that the form of the linguistics was not quite right for their purposes. However, we are now more aware of what the theoretical basis should look like for the new generation of simulations.

If simulation of human speech production requires simulation of decision-oriented cognition linguistics cannot provide it. Psychologists have yet to take linguistic hypotheses about language and relate them to cognitive processing models. Researchers in artificial intelligence who deal with language are concerned with formal languages and with problems in syntax and semantics. In general, artificial intelligence has not been interested in speech, and has nothing to say about how to apply its techniques to model speech production. Additionally, within linguistics pragmatics is comparatively new and is being developed in the traditional descriptive way.

I am claiming that reasoning devices can form an essential part of the next generation of synthesizer, moving towards better simulation of speech production. I have presented one possible way of doing this.