# Reasoning Devices in Synthesis - SYNTH-EX

**Kate Morton**

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



An early version of this paper was presented to the Leeds Experimental Phonetics Symposium September 1986. This included a demonstration of SYNTH-EX (Morton 1986).

---

*Early Synthesis - an Example of Simulation* discussed some of the shortcomings of generic synthesis systems. A more successful simulation of speech can be achieved by

- incorporating developments in phonetic theory in particular introducing cognitive modelling,
- understanding the shortcomings of descriptive linguistic theory for building simulations,
- providing a method for selecting among optional rules.

There is a field of study allied to linguistics pragmatics (Levinson 1983), which is about language under varying conditions of usage. I suggest that it might prove productive to base the selection of the options linguistics describes on ideas developed in pragmatics. Throughout the present work I have referred to such considerations as non-linguistic to distinguish them from the traditional domain of linguistics.

Selection among options in situations where many factors need to be considered is a widespread problem. A technique has been developed in artificial intelligence for dealing with such situations. It consists of the principled evaluation of data from a number of different sources in order to select the appropriate option (Hayes-Roth *et al*. 1983, Alty and Coombs 1984).

Synth-ex is an example of the kind of intelligent reasoning device which might begin to overcome some of the shortcomings of current synthesis systems in choosing among options. The intention here is to illustrate that a minimum level of 'reasoning' in a synthesis system can point to a way of dealing with the defects noted.

## WHY AN INTELLIGENT REASONING DEVICE?

So far I have emphasized two points relevant to the synthesis of natural sounding speech for which linguistics provides no support:

- no procedural algorithms are available,
- no method for choosing between options is available.

Synth-ex focuses on the second problem.

The kind of optional rule being looked at (see Fig. 12) can be described as environmentally sensitive. Environmental sensitivity is distinguished from linguistic context-sensitivity in the sense that information external to the linguistic system is used in the decision making process. The environment is described in terms of non-linguistic features such as speaker attitude, mood, world-view, etc.

For the example reasoning device, Synth-ex, I have selected six different optional styles for an utterance: neutral, citation, expectation, surprise, contrast, precision. *Neutral* means that the speaker has a neutral attitude toward what he is saying. *Citation* refers to an intention to highlight a phrase or word. *Expectation* means that the speaker will speak as though his utterance could be predicted by the listener. *Surprise* is how a speaker would communicate surprise to the listener. *Contrast* involves a pronunciation which highlights a difference in meaning with a similar word or phrase. *Precision* means that the speaker pronounces the word

more carefully than usual. These then are six ways of pronouncing an utterance, and for the purposes of the demonstration, of pronouncing a single word *'cat'* at the end of a sentence.

The device is intended to show the use of a decision procedure to select the pronunciation rule corresponding to one of these variants by actively looking for the information required to make the decision. The reasoning device is not passive: it does not just detect a requirement and select an option by using context matching. Instead, it evaluates the relationship between the factors involved in arriving at a final selection. The reason for this is that the use of non-linguistic information concerning environment is not a matter of binary choice. The expression of surprise, for example, is scalar, and does not necessarily exclude expressing other attitudes at the same time. There might be a need to sound a little surprised, while at the same time convey an element of contrast etc.

## GOALS

There are six possible goals for the device to chose between. They are:

- neutral: cat, as in: *Before going out he fed the cat.*
- citation: cat, as in: *The word he said was cat.*
- expectation: cat, as in: *He owned both a dog and a cat.*
- surprise: cat, as in: *Her fur coat was made of cat.*
- contrast: cat, as in: *I did not say cad, but cat.*
- precise: cat, as in: *It was time to feed the cat.* (spoken carefully)

The phonetic parameters selected to differentiate between these goals are word-final [t] released or un-released, and vowel duration. The results of a small experiment showed that, when pronouncing the word *cat*, a speaker made alterations to duration of the vowel [a], and varied the release of the final consonant [t], consistent with achieving the goals. There are other acoustic parameters involved, but these were not selected for this project. In fact in this example, how these distinctions are conveyed was not important since I was interested in whether a reasoning device could select among goals, not the detailed parameter specifications for each.

All six goals occur within English speech, but not equally frequently. To reflect this fact each goal was given an *a-priori* index corresponding to its frequency of occurrence in speech. In future, part of the research into synthesis systems which can handle style will involve large scale examination of the statistics of variability of this kind. For the moment, I took an educated guess to arrive at *a-priori* probability ratings for the goals:

| GOAL | A-PRIORI PROBABILITY |
|---|---|
| Neutral | 0.4 |
| Citation | 0.06 |
| Expectation | 0.2 |
| Surprise | 0.06 |
| Contrast | 0.08 |
| Precise | 0.2 |

## QUESTIONS

In order to make a reasoned choice from among the six goals, the device needs information from three sources:

- from the overall input device to the system (see Fig. 13),

- from the semantic knowledge base, and
- from the phonological and phonetic knowledge bases.

The input device to the synthesis system determines what is to be said. This is an artificial intelligence (AI) device delivering concepts which are to be linguistically encoded. An example of such a device would be the AI unit in an interactive database inquiry system. A human being asks the machine a question: *When is the next plane to Paris?*. The AI unit responds by looking up the timetable and formulating the concepts necessary for a reply. The concepts are passed to the language encoding system, and from there to the synthesis subsystem. Formulating these concepts involves certain presuppositions about who the message is intended for and the attitude the device wishes to take. The speech production reasoning device needs this information to select the most appropriate goal.

The questions in this example are:

1. Questions to the input device:
   - Is it the case that the standard values for the phonetic parameters are to be used as the basis for synthesizing the sentence? [Non-standard values would be used for a different voice or different dialects.]
   - Is it the case that you wish to convey a neutral attitude when this sentence is spoken? [A yes/no answer is required.]
   - Do you wish to convey surprise? [yes/no]
   - Do you wish to convey contrast? [yes/no]
   - Is it the case that the person you 're talking to speaks English well? [This calls for an assessment: a scalar answer is required.]
   - Is a technical term involved that the person you're talking to is familiar with? [assessment: scalar answer]
2. Question to the semantic knowledge base:
   - What is the semantic confusion rating for this word? How likely is the concept expressed by this word to be confused with others? [The answer depends on how close the concept (*cat-ness*) is to other concepts (e.g. *dog-ness*) in the semantic space: a scalar answer is required.]
3. Questions to the phonological knowledge base:
   - What is the confusion rating at the word level? [How near in the phonological space is the word as a whole to other words (e.g. cat *vs*. dog)? A scalar answer.]
   - What is the confusion rating at the segmental level? [For the segments in the word, how near are they to other similar segments (e.g. / . .t# / *vs*. / . .d# /)? A scalar answer.]
   - What is the confusion rating at the feature level? [For features of the final segment, how critical is it to get a proper phonetic realization (e.g. [+release] in the final /t/)? A scalar answer.]

Not all questions are asked on any one pass through the system. For example, if a neutral style is needed, then there is no point in asking questions about contrast and surprise. Questions are indexed as to whether a certain response blocks asking other questions.

## RULES

Responses to questions are taken to a rule network which evaluates them according to pre-established relationships holding between the questions. In this case, the relationships involve blocking unnecessary further questions. For example, a decision to speak neutrally can be pre-empted by a belief that the listener does not speak English very well (the word will have to be spoken carefully); or the desire to speak precisely will be weakened by the belief that the listener is familiar with the word if it is a technical term.

Finally, for each goal the critical questions and rules are identified by an index on the goal itself. The index has two parts, one indicating the degree of logical sufficiency of the particular question or rule for the selection of this goal, the other indicating the degree of logical necessity.

The overall system works therefore as follows. There are six possible outcomes for the pronunciation of the word *cat*. Before starting, each has an index of likelihood of occurrence. The object of asking the questions and using rules to evaluate the answers is to modify this index to obtain a final value for each goal. The goal with the highest final index is the one selected as the basis for synthesizing the word cat.

## THE PROGRAM

The program for the *cat* example is written in the advice programming language required to drive the probability evaluation shell called MicroExpert (Isis Systems Ltd). In effect, the program constitutes a knowledge base for evaluation by the mechanisms built into the shell The shell itself

- picks up the questions and asks them,
- takes in the answers and evaluates them by rule,
- brings all results together to select an appropriate goal.

[The program is available on request.]

This is an example of a simple reasoning device whose task is to decide which version of a sentence involving the word *cat* is appropriate, One small example was chosen because it would not have been possible to set up an entire artificial intelligence input device, a semantic knowledge base and a phonological knowledge base. What I have done therefore is substitute a human being for these areas of the system. Thus all questions are addressed in fact to the user of the system for the purposes of running the program. The user is a linguist who is able to provide answers similar to those which would be provided automatically by the machine in a complete system.

## A MORE COMPLETE SYSTEM

In a fuller system, semantic, syntactic, phonological and phonetic knowledge bases would be included, linked by a set of reasoning devices which draw on them, and which if necessary access more than one knowledge base at the same time (see Fig. 13). This system would replace the simple algorithms of current models which draw on knowledge bases sequentially, and which cannot access more than one component at a time to modify the information flow within the system.

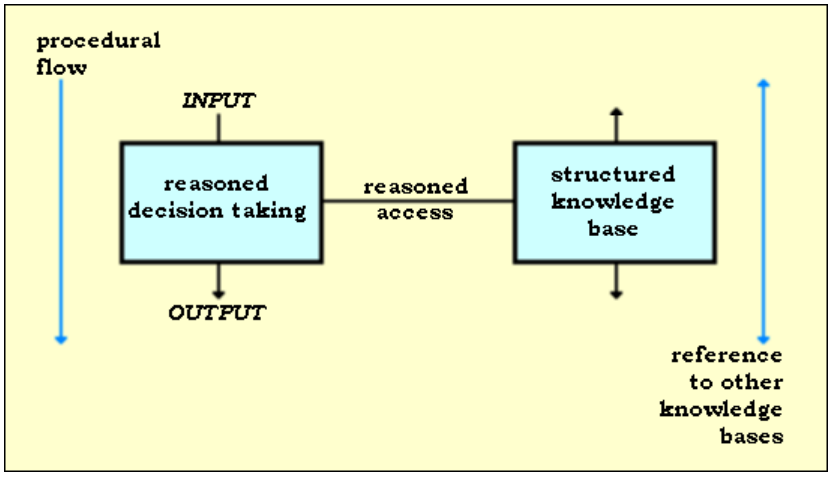


Fig. 13 Reasoned decision taking at one level in the algorithm. Components are cascaded to form a multi-level structure with cross referencing between components.

COGNITIVE PHONETICS is central to this system, because unlike existing components which characterize either abstract knowledge or physical processes (but not both together in any one component), COGNITIVE PHONETICS is able to describe both types of knowledge in a single component. Additionally, COGNITIVE PHONETICS accesses information throughout the system, including non-linguistic information about the environment, to supplement the basic linguistic encoding process.

In principle, the simulation of expression could be implemented by multiple parallel algorithms each generating an alternative version of each possible sentence. But the computational load would be considerable. In contrast, the proposal here treats variability as a modification of idealized speech rather than as a set of completely specified alternatives. The reasoned decision taking devices do not select from among entire algorithms and their associated knowledge bases, but from among algorithms subordinate to them. The purpose of these subordinate algorithms is to constrain the performance of the core procedures.

The same principle of core procedures, together with principled modification, underlies action theory. The claim is similar: a co-ordinative structure automatically assembles a muscle configuration appropriate to a canonical version of an articulation, but the facility of tuning adjusts (on each occasion) the actual configuration achieved. Here again, in a descriptive model, COGNITIVE PHONETICS organizes the information which is needed for tuning signals to adjust elements in the peripheral muscle systems.

Therefore, in the descriptive model, I propose a system which provides a general specification, established at higher levels, of what is to be output, which is later adjusted for expression by a decision device in the cognitive phonetic component. These adjustments constitute instructions for tuning the way in which elements within co-ordinative structures cooperate.

The implications for simulation are that a main algorithm generates a general linguistic message, COGNITIVE PHONETICS accesses non-linguistic information, and evaluates it taking into account how the information can be realized. In the human system, the result is that the elements within the co-ordinative structures are reorganized to produce the appropriate speech. In a simulation system, the result is a re-computation and reorganization of pre-defined segment parameter values.