SYNTHesis – EXpert

An Expert System for Improving Naturalness in Synthetic Speech

Katherine Morton

Paper for *Expert Systems 87*: Brighton, December 1987 – and for publication in the *Proceedings*.

This paper describes a way to improve naturalness in the quality of synthetic speech. An important feature of human speech is its wide variability. Speech synthesizers do not simulate this feature because there is no simple way to select among optional rules described by linguistics and phonetics models of language processing. An expert system which accesses and evaluates non-linguistic data as well as linguistics can contribute to a better simulation of speech.

INTRODUCTION

There is a certain amount of interest currently in voice input/output systems as a user interface in communicating with a computer system. Speech is the easiest way for human beings to express their thoughts or give instructions to each other; writing or the combination of keyboard and VDU are less acceptable to the ordinary person or to someone who needs to concentrate on another skill at the same time. For example, voice i/o would be very useful to an airplane pilot both for communicating with other people and with aircraft equipment.

Voice input to machines is processed by an automatic speech recognizer, while voice output to the human user is by speech synthesizer. How useful these devices are is assessed by the reaction of the user and his expectation of their performance. Currently, neither device has been developed to the point where there is not some degree of user frustration.

Automatic speech recognizers generally handle isolated words only, with limited vocabularies of a few hundred words, and need to be trained to recognize the voice of their human user. Under laboratory conditions, recognizers can achieve a success rate of over 90%; that is, a word is misidentified on average only once in ten trials. But in the field, where ambient noise or other environmental factors can degrade the speech signal, success rates rarely exceed 50-60%.

Text-to-speech synthesizers are more successful. Text can be turned into intelligible speech with virtually unlimited vocabulary, but synthesizers still sound machine-like and the listener feels disturbed by the quality of speech, without being able to say specifically why. However, researchers have not yet directed their attention to any extent improving naturalness. It has been difficult to identify what features contribute to the liveliness of human speech, but one characteristic that seems to give natural speech its essential humanness is variability.

It is well known in speech science that when asked to repeat a word, the acoustic waveforms a speaker produces will differ even if he tries to say the word identically on each occasion. However, it is becoming clear that variation in human speech is less random than we had thought. At Essex we have identified some patterning within the variability, and are beginning to model its source. If it is possible to separate random variations from structured variations, it should be possible to incorporate variation in synthetic speech to improve its naturalness.

In the case of speech recognition by computer, the situation is reversed. Errors in recognition are caused by this variability in human speech. Mapping from the incoming signal to idealized invariant forms has not been satisfactory, and reliable speech recognition is not yet viable.

Thus while researchers in automatic speech recognition are trying to remove variability from human speech, those in speech synthesis should be adding variability to synthetic speech. In both areas the problem is the same: an insufficiently complete description of human speech variability. However, a specialist speech researcher can hand edit signals presented to a recognizer and the error rate is minimized; similarly hand editing synthetic speech produces an output which is almost indistinguishable from real speech.

Editing consists of adjusting the values the synthesis program assigns to parameters used in synthesizing a speech waveform. Since the results of editing can be very good, this shows that a specialist phonetician has sufficient knowledge to be able to manipulate the relative parameters. What we wish to capture is this expert knowledge. Although I shall be discussing speech synthesis here, the linguistic/phonetic principles underlying the discussion of synthesis are equally applicable to speech recognition.

THE PARAMETRIC DESCRIPTION OF SPEECH SOUNDS

A number of parameters are used by phoneticians in the description of the speech waveform (Fry 1979). In describing how an expert system has been incorporated into our speech synthesis program, I shall discuss only duration and amplitude.

Words can be thought of as being pronounced by stringing together separate sounds: for the purpose of synthesis each sound is specified by a set of acoustic parameters. The boundaries between sounds blend together, depending on the rate of speaking. At the boundaries we can identify transitions between the parametric specifications for individual sounds; timing constraints and the difference between the parametric specification of adjacent sounds will make these transitions more or less abrupt. Using the two parameters selected as examples, we can think of sounds as specified by intrinsic duration and amplitude which later become modified during the concatenation process used to build up whole words or sentences (Holmes *et al*. 1964, Kelly and Gerstman 1961). To produce hand crafted speech, the expert phonetician alters duration and amplitude in a principled way, which results in variability. Standard synthesis systems replicate intrinsic values and calculate the transitions using fixed tables and rules — hence the lack of variability (Fig. 1).

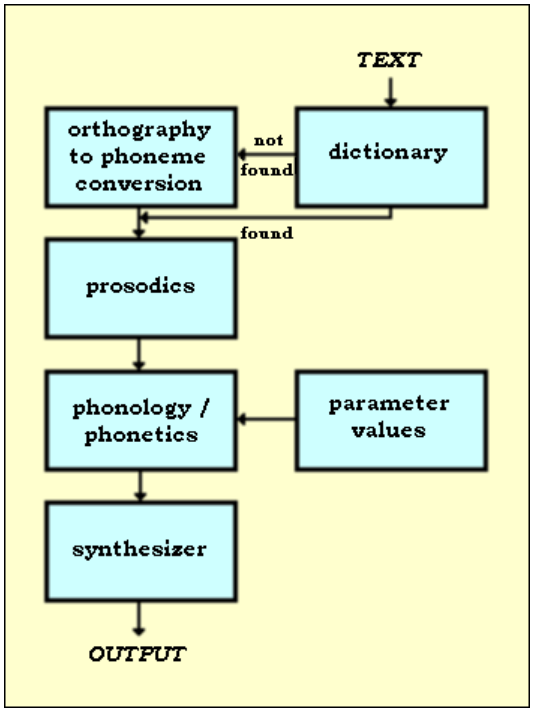

Fig .1. Text-to-speech synthesis system.

THE INPUT TO A SYNTHESIZER

The input to a text-to-speech synthesis system consists of ordinary plain text. The synthesizer's task is to convert this to spoken language, simulating a human reading text aloud. Not all the information needed to speak a text is encoded in the text itself. For example, information about the speaker's attitude is not in the text unless explicitly part of the linguistic message. In general, text encodes the linguistic message only, and leaves conveying subtleties of expression to interpretation by the reader. In order to do this, the reader must understand the meaning of the text. Therefore a computer simulation of reading aloud requires some understanding or intelligence.

Linguists, phoneticians, and cognitive psychologists do not yet know what processes of understanding are required to convey mood, etc., in speech (Levinson 1983). However, we are beginning to model some of the variations in the speech waveform itself which seem to encode attitude. The expert phonetician doing the editing asks himself what attitude on the part of the speaker is required and alters duration and amplitude values. The system reported in this paper is intended to simulate the editing, and therefore needs to determine the required attitude before encoding the message.

It is also possible to produce speech that does not have overlaid attitudes: for example; television and radio newsreaders are trained to minimize these effects so as not to communicate their own personal attitudes toward news items.

In principle it is possible to externalize the knowledge that an expert phonetician applies when editing synthetic speech to enable the non-expert to do the editing, and we wish to build that expert knowledge into the synthesizer system. Some of the knowledge is described in linguistics and phonetics models, but a way to access and evaluate this knowledge in a working synthesis system has not yet been devised.

LINGUISTIC KNOWLEDGE BASE

Linguistic models characterize a speaker's knowledge of his language by describing sets of production rules grouped together to enumerate all possible combinations of meanings, words and sounds to form grammatical sentences of the language (Chomsky 1965). All sentences of the language and all possible pronunciations are represented in the database.

In phonetics, this involves characterizing alternative ways of pronouncing words. So to describe pronunciation of a word ending in a sound like [t] (*cat,* for example) two rules describe the fact that a speaker knows he can say the word with or without fully pronouncing the final [t] sound. The purpose of linguistics and phonetics is to describe this knowledge; the theory does not provide a mechanism which can choose between the two rules.

Using the same word, *cat*, we can easily see that the word would be pronounced differently in respect of amplitude and duration in the following three sentences:

He owned both a cat and a dog. (neutral)

He fed the *cat,* not the dog. (contrastive)

The mouse chased the *cat*!? (surprise)

In speech synthesis, however, where the object is to simulate human performance in speaking, the choice must be made so as to produce a single output for a single utterance. The choice involves linguistic information and non-linguistic factors such as attitude, evaluation of rules selected so as to arrive at the most appropriate utterance for the occasion.

THE BASIS OF LINGUISTIC AND PHONETIC SYSTEMS

Linguistics has been called a cognitive model. But it is not a model of the cognitive processing involved in language production. Linguists are concerned with describing the potential sentences of language, and have formulated a descriptive characterization of the

knowledge base required to produce the totality of language, rather than describe what happens on any one occasion (Chomsky 1965). A linguistic description could constitute a set of hypotheses about actual processing in human beings, but this is more the province of cognitive psychology. Linguistic descriptions systematize linguistics knowledge without making claims concerning human language processing, and in fact can be said to describe the linguist's knowledge of his description of the language (Chomsky and Halle 1968). Therefore, in speech synthesis we are in effect simulating the knowledge of linguists and phoneticians expressed by their models, rather than the knowledge and selection processes of human beings.

One branch of phonetics, physical phonetics, on the other hand, is considered to be a physical model of the workings of the neurophysiological, aerodynamic and acoustic systems employed in speaking. More recently a branch of phonetics — cognitive phonetics — has been developed which attempts to model some cognitive aspects of speaking (Tatham 1984). It claims specifically that a speaker knows something of the physical properties of the system, so that he is able to anticipate and therefore influence the effects of these properties. This ability to manipulate the system's physical properties involves selecting rules from the knowledge base which describe the properties and the limits of the physical system. The cognitive phonetics component added to the linguistic model accesses linguistic rules which describe what to say and also phonetics rules which describe how to say it.

THE BASIS OF AN EXPERT SYSTEM TECHNIQUE IN SPEECH SYNTHESIS

We have built a small example reasoned decision system to simulate activity of the phonetician editing synthetic speech. The technique is derived from general principles described by researchers in knowledge engineering techniques, such as Alty and Coombs (1984), Hayes-Roth, Waterman and Lenat (1983), Shortliffe (1976). The system is based on processes modelled in cognitive phonetics (Tatham 1985) which access several knowledge bases. For example,

a. *linguistic and phonetic* knowledge bases, to ascertain what options are available for each utterance; in our example

Rule 1 : all sounds have intrinsic durations and amplitudes,

Rule 2: the intrinsic duration of vowels may be shortened in certain linguistic contexts (e.g. when they occur before /p, t or k/),

Rule 3: the intrinsic amplitude of vowels and consonants within a syllable is increased when linguistic stress falls on the syllable, etc.

Linguistic rules do not manipulate scalar values: vowels are said to either short or long (a relative perceptual concept), stressed or not, etc.

b. knowledge bases about *non-linguistic features*, such as the speaker's attitude toward the content of what he is saying and toward the person he is talking to; in our example

Rule 1: duration of stressed sounds is increased if contrastive emphasis is placed on the word they occur in,

Rule 2: if the speaker suspects that the listener does not easily understand him, amplitude and precision of articulation in general are increased on key words in a sentence,

Rule 3: if the speaker wishes to focus attention on a key concept in what he is saying, then the appropriate word has increased duration of vowels and final consonant, the amplitude of the stressed vowel increased and overall precision of articulation increased, etc.

Non-linguistic rules expressing mood, intention, etc., often introduce scalar values to parameters. Thus, under Rule 3 above, duration and amplitude are increased proportionally depending on how much attention is be drawn to the word.

c. a predictive model of *perceptual processes*, so that he knows whether or not he will he speaking clearly or loudly enough for the hearer; in our example

Rule 1: if the ambient noise is high, a correlating increase in perceptual difficulty leading to potential decoding error will be experienced by a listener,

Rule 2: if the amplitude and duration of a word are suddenly increased during a sentence, then decode that the speaker's intention is to highlight the particular word,

Rule 3: if a speaker bounds a word or phrase on either side by a brief pause, and increases durations of sounds during the word or phrase, then it is intended to be decoded as emphasized in some way, etc.

Notice that these rules are not instructions, but explanations of departures from expected norms in the acoustic waveform being heard.

d. the *environment*, to retrieve data on features such as the level and nature of ambient noise; in our example

Channel 1: monitor relative ambiant noise,

Channel 2: monitor spectral distribution of ambient noise, etc.

THE SYNTHESIS SYSTEM

General architecture

The system is constructed in such a way that its output constitutes a message to a standard synthesizer to produce appropriate parameter values which generate a soundwave corresponding to a particular word — but with the mood variations described above. Fig.2 shows a block diagram of the system.

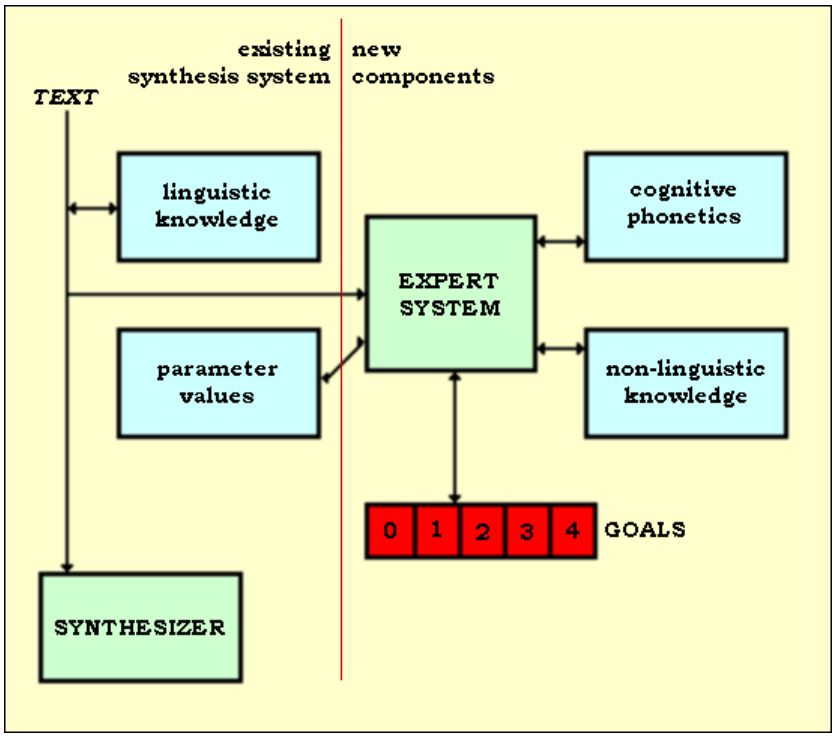

Fig. 2. Block diagram of the general system architecture.

To compute the output the system consults

- a database containing the parameter values for the sounds within the word,
- a knowledge base enumerating the options for synthesizing the sounds. In our exemplar system, called **SYNTH**esizer-**EX**pert, we have five optional

pronunciations represented as variations overlaid on the standard neutral pronunciation. The options are:

- neutral
- citation
- expectation
- surprise
- contrast
- precision.

It is important to remember that these are not variations in the wording of a sentence, but variations in the style of pronunciation of the same neutral word or sentence.

The system assigns to the set of options found in the knowledge base the status of **goals**, indexing each as to its intrinsic *a priori* probability of occurrence.

Consultation of linguistic sources

Next, the expert system determines factors influencing the production of a speech waveform. It scans the linguistic knowledge bases to see what special circumstances to take into account in selecting between the goals. It asks questions such as: is the patterning of sounds within the word such that this word can be confused with some other word unless carefully pronounced? If so, then a combination of assessing that chance, together with rating the importance of avoiding confusion, is computed to make differential adjustments to the original *a priori* probability ratings of various optional goals.

The information consulted in this initial scan is found in linguistic knowledge bases which state, for example, the statistical distribution of the various sounds in the language. Standard synthesis systems do not incorporate such knowledge, although it is described by linguistic theory, because their requirement to produce only neutral pronunciations makes it redundant.

Consultation of non-linguistic sources

The system next scans sources of non-linguistic information to determine such pragmatic factors as speaker mood and attitude. Each factor rated according to relative importance, and assessed on how necessary it is; the goal probability ratings are then recalculated.

For example, a pragmatic factor might involve the speaker's assessment of his relationship with the person he is addressing. The same sentence would probably be spoken differently when addressed to the prime minister rather than one’s own child. Altering speaking style might be considered very important in such circumstances. A high importance indexing would equate with a value from a scalar rating to produce a certain influence on the balance of the original goal *a priori* probability values.

Selecting between options

The overall system contains an addition to normal procedures in speech synthesis. Unlike the usual synthesis system which is based on a single fixed option, this system potentially includes all options and can select the most appropriate one (though only six are included at the moment).

Selection from among options is done by progressively shifting the balance of probabilities between options away from that determined by the *a priori* probabilities, toward one more suited to what is currently required of the synthesizer.

The information required to change the balance between goals derives from two types of source: linguistic and non-linguistic. Within these two types, the system seeks information according to pre-established procedures. Each category of information is given an index establishing degree of influence it has *as a category* on the final outcome, and then for each category a scalar value is chosen to enter into the probability calculation.

Responses to consultation queries are taken to a rule network. Rules within the expert system evaluate responses according to pre-established relationships holding between the set of queries. To avoid unnecessary or repeated consultation the relationships include blocking further queries if the response to one pre-empts another. The reverse procedure is also included: under certain conditions a response may need expanding, so the device is sent back. to the knowledge base for further querying.

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 precisely), or the desire to speak a technical term precisely will be weakened by the belief that the listener is thoroughly familiar with the subject matter. Probability calculation is according to the Bayesian method. Results of the calculation are sent to the synthesizer driver in terms of amplitude and duration. Shapes of transitions by which linguistic and non-linguistic knowledge has been encoded are changed and the final calculated parameters are sent to the synthesizer driver.

In standard speech synthesis systems, style of speech is preselected be as neutral as possible. Systems do not include other options because no reliable method has been established for selecting between styles. The resultant speech is said to be idealized or based on the most usual way of speaking as determined by the designer of the system.

The system I have described here demonstrates a method of allowing some of the options which need to be included to provide a variable output. The system includes a reasoned decision making device, simulating a phonetician, designed to consult as appropriate within the overall system in order to make a reasoned choice between the various options. The method of selection between the options is to progressively adjust *a priori* probability indices on each option until the system stabilizes at the conclusion of the consultation session. The probability indices indicate the most appropriate option.