SPRUCE: SPEECH SYNTHESIS FOR DIALOGUE SYSTEMS

Mark Tatham
Katherine Morton
Eric Lewis – Computer Science, University of Bristol

Paper was written in 1992-3, and is a revised version of a presentation at the Multimodel Dialogue Workshop held in Maratea, Italy, in August 1991. Selected refereed papers (of which this is one) were later published in book form: Taylor, M.M., Néel and Bouwhuis, D.G. (2000) *The Structure of Multimodel Dialogue II,* Amsterdam: John Benjamins. This paper appears in that book, pp. 271-292, in almost identical form.

ABSTRACT

The version of SPRUCE described here is the one being researched at the time. Subsequent papers describe later versions of our speech production simulation.

The ***SPRUCE*** text-to-speech and concept-to-speech synthesis system is at present under development by the authors at the Universities of Essex and Bristol. It is designed to be fully integrated into multimodal dialogue systems. The characteristics of ***SPRUCE*** which distinguish it from other systems are its large dictionary of words, its syntactic and semantic parsing capabilities, and its inventory of syllables which are used as the units forming the basis of the speech output. These syllables are derived from human speech using parametric analysis and a normalisation procedure. From the outset ***SPRUCE*** has been conceived as a system which integrates fully with other components of a dialogue system. We argue here that such an approach is essential for all aspects of dialogue systems.

The paper focuses on the problem of naturalness in synthetic speech, stressing the importance of basing the model on well founded theory. We address two aspects of the variability found in human speech: cognitively controlled variations which are not due to physical effects, and pragmatically annotated variations of duration and fundamental frequency which human beings use to convey attitudes and feelings in their speech.

Finally, we describe a typical application for natural synthetic speech - computer aided learning, arguing that in the learning environment natural-sounding speech has an important role to play alongside text and graphical interfaces. In this respect we stress the necessity for integration of all modes within a full multimedia system where speech is seen as having a major rôle to play.

PART I – INTRODUCTION

The component parts of a dialogue system

Dialogue systems are generally thought of in terms of their constituent parts. Thus we speak of speech input and output devices, a language processing device, a database accessing and retrieval device, graphics interfaces, and so on - all of which, in this conceptualisation of a dialogue system, are integrated under a central control device.1 The function of the central control is to hold the system together, to send instructions or messages to the various component parts to respond in particular ways on demand. Thus a speech recogniser, acting as an input device to the system, delivers an output (perhaps a sentence) which is directed by the controller to the language processor whose task it is to extract meaning from what the input device has recognised. Once the meaning is determined, strategies are triggered to

consult, say, a database and generate an appropriate response which is then turned into language. In the final stage a speech synthesiser generates output in the form of a sound wave.

Each of these component parts of the overall system is seen as having its own tasks which it knows how to perform, given two inputs - one a simple 'GO', and the other data - which tell the device which tasks from its repertory it should perform. Thus we have a composite system of components which can often be found outside the dialogue environment as stand-alone devices, messages flowing from a central controller in command of task sequencing and general management of the system, and data flowing from one component to another and upon which the individual tasks are performed.

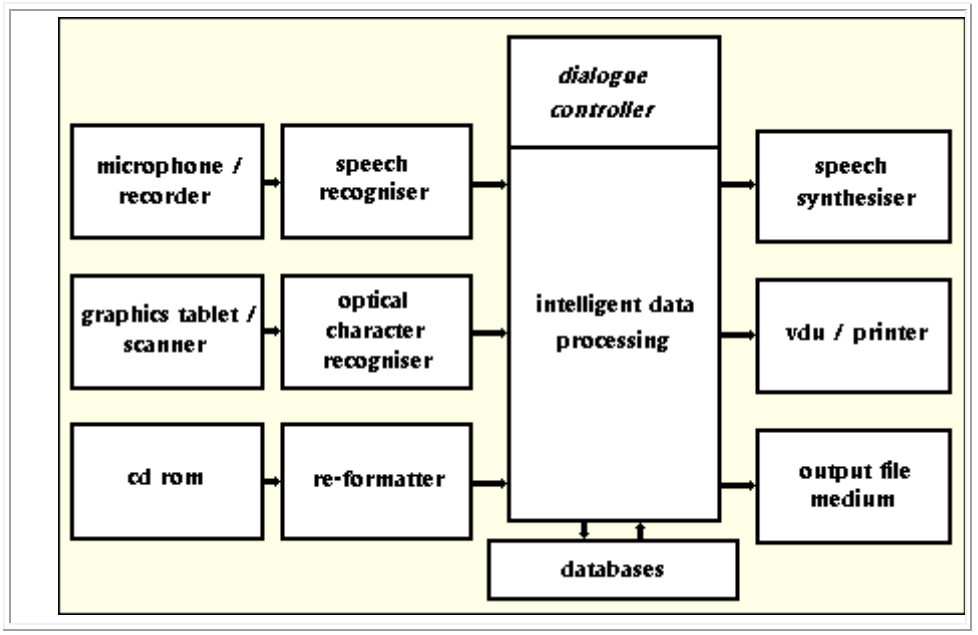

Fig.1 The basic components of a multimodal dialogue system.

Earlier stand-alone components

Each of the components in the dialogue system described above has already been developed to some extent. However, the components have been researched and developed separately usually by researchers working in different disciplines and according to the various metatheoretical principles associated with their areas of work. But a new discipline is emerging - dialogue studies - whose job it is to create a theory of human dialogue and build models simulating such dialogue. As yet the theory is rudimentary and simply characterises the controller responsible for co-ordinating the component parts of the system. But eventually the theory will characterise the entire system and integrate what we now see as separate component parts.

If we look at a human being, these functions appear to operate as an integrated whole. Thus, in the area of speech, there appears to be considerable overlap of the processes of speech production and perception, and they in turn overlap with more central language processing. It is a quirk of the development of our science that we have divided the huge study of human communication using speech into smaller more tractable sub-components. Having done that, many researchers believe that it follows that these areas of study reflect a real separation of processes within the human being. But there is no good reason to suppose this - and what evidence there is points, on the contrary, to large-scale integration of the processes.

Speech production and perception can both be regarded as knowledge based systems, each described as having an input and deriving an output through sets of rules which transform the input by reference to stored knowledge about the nature of speech. It seems

counter-intuitive that much of the knowledge required for the complementary processes of production and perception is not shared. Both processes require at least some knowledge of the nature of speech. Similarly it is difficult to imagine that many of the processes for encoding thought into spoken speech are not in some sense mirror images of similar processes involved in decoding thought from heard speech.

Formalism

In the same way that it is easy to regard the human dialogue system as comprising separate parts, each with only input and output connections, so it is easy to believe that a particular formalism used in the model is itself also used by the human being. For example, in the modelling process the knowledge bases referred to earlier are often expressed in terms of sets of rules. As a consequence of this simple and convenient modelling strategy we might begin to think that in the human mind there is a rule which says that in English adjectives usually precede the nouns they go with. So, we might think, we use a rule to make sure we say

The green grass is over there,

rather than

**The grass green is over there.*

Similarly we might use a rule to make sure that in the compound noun *blackbird* we stress the *black* morpheme, whereas in the noun phrase *black bird* we usually stress the noun *bird*. But why should it be the case that just because linguists describe parts of our knowledge of the language in terms of a simple rule formalism we should believe that this is what human beings do?

We might equally use, for example, a neural network as an alternative formalism, in which case we shall have no explicit knowledge base as such. Yet such a formalism will continue to describe quite adequately the observed behaviour of speakers of English. In the case of the neural network formalism, the 'knowledge base' could be regarded as residing in the connections established between neurons and in the strengths of such connections. It has been argued by some that the neural network paradigm is much more plausible than a rule formalism because, they claim, it attempts (although in an elementary way) to model neural processes within the brain.

In the ***SPRUCE*** project described in **Part II**, we have been developing a voice output device for dialogue systems which attempts to fully integrate their component parts.

PART II – THE SPRUCE PROJECT

Introduction

The ***SPRUCE*** project2 currently under way at Bristol and Essex Universities in the UK, while being concerned centrally with the ***simulation of speech production***, takes the integration of all language and speech processes as central to its underlying philosophy.3 In addition it makes no claims concerning the viability, within human beings, of the various formalisms employed in such a simulation model. This is a project which is designed to adhere as closely as possible to current theory in the area of human speech production,4 whilst at the same time meeting the demand for a dialogue voice output device which could be incorporated into future fully integrated systems.

SPRUCE within an integrated dialogue system

As an example of the philosophy of integration, synthesis and recognition (the simulations of human speech production and perception, respectively5) are integrated in as much as the recognition model is available at all times to the synthesiser for predictive

modelling of the perceptual effect of its potential output, just as the recogniser can consult the synthesiser for information as to how a particular sound wave that has been detected might have been produced by a speaker. This is achieved by knowledge base sharing (when rule based subsystems are in use) and by mirror image networks (in those parts of the system using neural networks), as well as by the existence of data channels between the components.

In the ***SPRUCE*** system the strategy goes beyond simply using various sub-components to map an input onto an output. The sub-components which do this mapping are there, but there is more sophisticated communication between them, an example of which is described in **Part III**, where information channels are set up between dialogue control, language processing and acoustic wave production.

Variability in speech

We can cite one or two examples as illustrations of many reasons why, from a theoretical perspective, we have chosen this basic integration premise.

1. Human speech which is part of a dialogue communicates more than the plain meaning of the words or phrases the speaker is uttering. The speaker intentionally or unintentionally communicates much of his or her attitude to or feeling about what is being said by 'modulations' of a 'neutral' prosodic element in the speech which is dictated by the grammatical nature of the utterance.6 By *how* the speaker speaks, rather than by *what* he or she actually says, the listener can become aware of what the speaker feels, or what the beliefs are toward what is being said. The speaker will even convey an attitude toward the listener in general. This could not happen unless the speaker had access *via* a model of perception to the likely effects of such prosodic variations on the listener's decoding process. The basic speech production model incorporating such effects is described in **Part III**.

2. Human speech is characterised by a great deal of variability. Although some of this variability is derived from constraints within the peripheral neuro-physiological, mechanical and acoustic systems, it can be shown that much of it is systematic and under the speaker's cognitive control.7 Thus a speaker will in some way emphasise a word he or she predicts is likely to be misheard (employing the predictive perceptual model) because of semantic ambiguity or for some other reason. Even at the sub-word level a segment (perhaps an individual sound or a syllable) may be articulated with more or less precision dependent upon whether at the phonological level it is predicted that the word itself might be confused with another.

This type of consideration is central within ***SPRUCE*** not only because of the philosophical stance referred to above, but also because the variability of both the prosodic element in human speech and of the precision with which the speech is uttered is the focal parameter leading to a perceived judgement of *naturalness* in the speech output. Speech produced without this variability is simply not perceived to be human - precisely because the variability itself defines to a large extent the humanness of speech.

The perceptual model within *SPRUCE*

No speech synthesis system has yet attempted to capture and reproduce this variability - and for this reason no system yet sounds convincingly natural8 ***SPRUCE*** varies its output dependent on certain criteria, some of which are mentioned above. To do this ***SPRUCE*** models speech and other knowledge in a way which is complementary to how it is modelled in an ideal recognition system simulating human speech perception. ***SPRUCE*** in effect incorporates a model of speech perception which enables it to initially *try out* what it intends to say; an iterative process optimises the output dependent on perceptually-based criteria.

SPRUCE Speech Synthesis

The ***SPRUCE*** synthesis system has a comparatively simple framework (Fig.2) which builds on and extends the tradition of the best text-to-speech synthesis systems.5 Its ability to

accept an alternative concept input9 (only some aspects of which we will be describing here) makes it suitable for incorporating m dialogue systems.

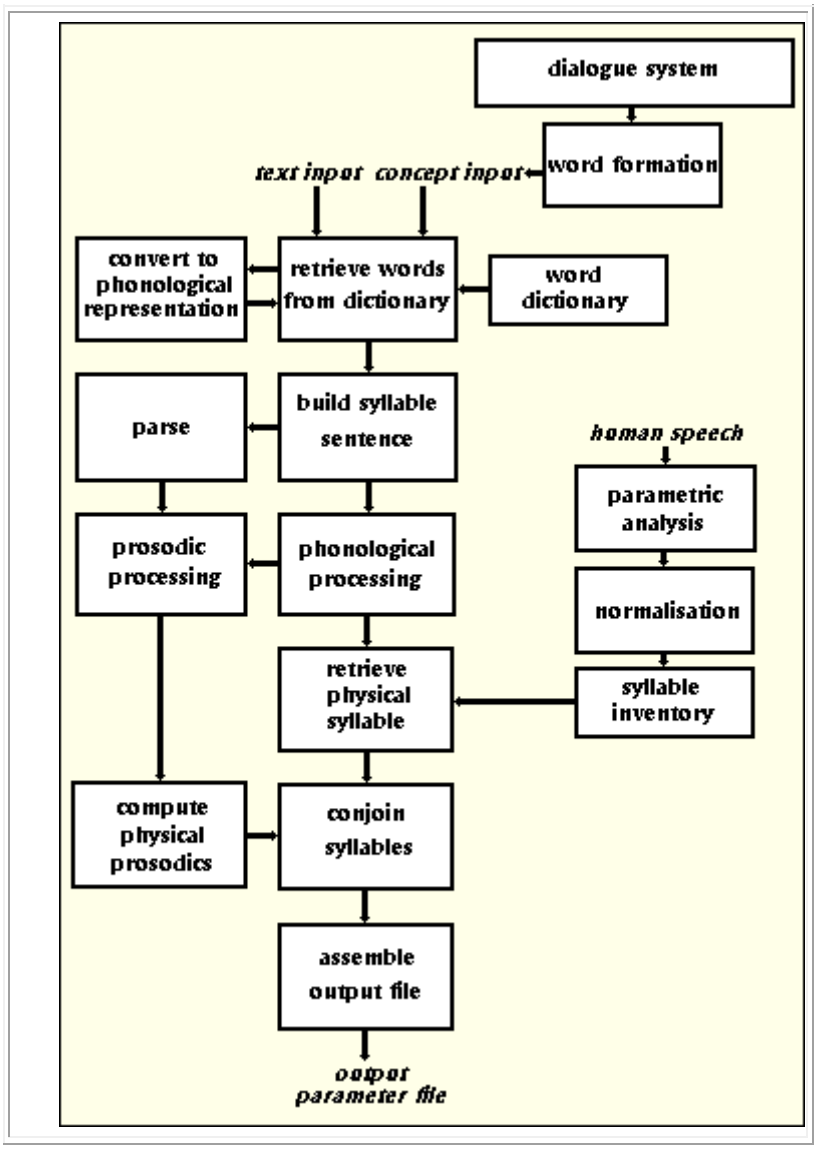

Fig.2 The framework of the *SPRUCE* synthesis system. The pragmatic sub-component and the predictive perceptual model have been omitted.

Text input

The basic ***SPRUCE*** framework can be described quite simply. We begin the text-to-speech system with the assumption that we can input the actual text and assemble it in a form suitable for synthesis processing. This involves having the text in electronic form, either taking it directly from a keyboard or from some other source such as the signals involved in electronic mail transmission or videotex, or from an optical character recognition device scanning a printed page. From such inputs we are able to identify individual words which make up the text and individual phrases and sentences which they form. For the human being the input is by optical character recognition *via* the eyes and optical processing circuitry in the brain. At this point human beings are involved in some language processing to determine the meaning of the text they are reading. We know this because we can observe that reading aloud often goes wrong if the reader does not understand what is being read, or begins a sentence assuming one meaning and then, say, turns a page to find out the assumption is

wrong. The error is most often revealed in an unacceptable rendering of the prosodics of the sentence. In a moment we shall see how ***SPRUCE*** tackles this problem.

Concept input

Defining concept input is more difficult, largely because dialogue systems designed to output concepts rather than words do not yet exist. We envisage a front end to ***SPRUCE*** which forms a processing level between the dialogue system and its voice output system. This processing revel has the task of converting concepts - however they may be expressed - into words concatenated into correctly formed sentences. These sentences can then be input directly into ***SPRUCE***, though they are not represented in normal orthography, but in a phonological representation.

But in addition to objects which can be transformed into words, a concept level in a dialogue system will also embody a pragmatic representation derived from the system's pragmatic processing component. We referred in the introduction to prosodic information in a person's speech which expresses their mood or attitude; such information derives, not from the word structure of a sentence but from its pragmatic inclination. This information is available at the concept stage, ready to be converted in the ***SPRUCE*** front end into markers placed on the sentence representation, and are carried forward for later processing (**see Part III - Pragmatic phonetics and naturalness in synthetic speech**).

This is one of the major differences between concept input and text input; text has no means of encoding these pragmatic and other markers, as we shall now see.

Determining what the input means

SPRUCE needs to determine what the text input means in order to generate among other things the correct prosodic contour. This is not a real problem with concept input since the input itself has come from a sub-component of the system which knows the meaning of what it wants to say. The front end interlevel mentioned above has the task of representing this meaning in such a way that it can be processed by the synthesiser.

In the case of text input, however, the problem is considerable because it involves determining information which is not explicitly encoded in the text itself. Even when a human being reads text aloud success is not guaranteed unless the reader can understand what is being read. Thus in a simulation of the process of reading text aloud (which is what text-to-speech synthesis is) it is necessary to understand the text before we can guarantee correct encoding into a speech waveform. This is not yet fully available, because at the present time complete understanding is impossible since language processing technology is not yet sufficiently developed.

But extraction of some of the meaning of the text is possible, however incomplete. ***SPRUCE*** includes syntactic parsing designed to be sufficient, rather than the full-scale parsing more usual in language technology. But syntactic parsing, though an essential part of understanding the meaning of text, is not enough; we also need what might be called a semantic parser.

Syntactic and semantic parsing are very complex, and determining just how much of each to include in ***SPRUCE*** has only been resolved in a practical way. This section of the synthesis system incorporates an important engineering principle: if the system fails then it must fail gracefully. That is, if the syntactic or semantic parse is insufficient to provide the prosodic element with adequate information to generate the correct prosodic contour, then the resultant contour will not be implausible. The prosodic system is designed to minimise any failure.

The dictionary

The parsing processes are assisted by a dictionary in which words in the text input can be looked up to determine, among other things, what grammatical category they belong to. This is largely unnecessary with concept input because such information is a ready available from the process which turned concepts into the word strings of sentences.

The dictionary also contains information as to how a word relates logically or semantically to other words; this kind of information assists the semantic parse, and as with syntactic information, is largely redundant in the case of concept input.

The ***SPRUCE*** dictionary contains a minimum of 100,000 words. It includes, besides the syntactic and semantic information referred to, phonological and phonetic information to assist in subsequent processes within the system - this information is essential with both text and concept input.

However, no matter how large the dictionary, a word could appear in the text that cannot be found in the dictionary - just as human beings will come across words they are not familiar with. In such a case the system defaults to a process which converts the text's orthographic representation to a phonological representation. Such conversions are notoriously error prone (because of the large number of exceptions to the spelling rules of a language like English). There is clearly a trade off between the size of the dictionary and the number of errors likely to be generated by relying on orthography to phonological representation conversion. The ***SPRUCE*** dictionary is intended to be large enough to minimise the necessity to use the orthography conversion procedure.

Prosodics

If we know roughly what a sentence means and what its grammar is, and if we also know how the individual words are pronounced in isolation, we are in a position to look at the sentence as a whole and work out its prosodics. The ***SPRUCE*** prosodic component is concerned principally with two aspects of sentence prosodics: establishing both a rhythm and an intonation contour for what is to be spoken. These aspects of prosodics interrelate.

In ***SPRUCE*** rhythm and intonation are initially computed as abstract contours which are as yet unrelated to the physical reality of the acoustic signal to be generated later.

The Abstract Representation

The results of all previous sub-components are brought together at this stage to provide an abstract representation of what is to be spoken. We might call this a representation of what the device *intends* to be said rather than a description of what actually will be said.

Or, we could say that at this point ***SPRUCE*** knows what it wants to say, based on the original text or concept input. The input has been transformed by reference to a dictionary and by processing semantics and syntax, and a phonological representation has been assembled suitably annotated with prosodic and other markers. That is, ***SPRUCE*** *know*s what it wants to say in some idealised, abstract sense. What must now happen is that ***SPRUCE*** should determine how this completely specified abstract representation is to be actually spoken - that what has been planned should be actualised.

The inventory of phonetic elements

An inventory of the basic phonetic elements from which the final acoustic signal will be constructed is central to a reinterpretation of the abstract intention. Phonetic elements in ***SPRUCE*** are syllabic in size. This contrasts with most speech synthesis systems which use segment sized units, either in the form of allophones or diphones.

[We shall not consider diphones here since they are units which are not used in linguistics or phonetics. Further discussion relates only to those systems which use allophones as their speech building blocks.]

These units are designed to match up with the phonological information contained in the dictionary earlier in the system: part of the function of the dictionary is to identify the syllabic structure of words.

Just as in earlier systems the objects in the inventory are stored as parametric representations. And as with other synthesisers ***SPRUCE*** terminates in a parallel formant synthesiser identical in concept with its predecessors. The formant synthesiser used is the Loughborough Sound Images Ltd. implementation of the well-known Holmes design.5

Although the inventory representation is parametric in form, the specification as a type is different from that used in other systems. In a standard system, the inventory usually contains an abstract or static representation of each allophone. A single set of values, one for each of a dozen or so parameters, is given, along with a value indicating the segment's duration. The duration marker is used to expand the segment by repetition of the set of values for the given time.10 An allophone so derived is constant with respect to all parameters throughout its duration - unlike real speech.

By contrast the representation of syllable sized units in ***SPRUCE*** is *dynamic* and real. Every syllable is stored as a number of 10ms frames each of which contains a value for each parameter. The dynamically varying parameters throughout the duration of the unit are thus captured in the representation. In the construction of the inventory, each unit has been obtained by a process of excision and normalisation from parametrically analysed recordings of real human speech.

The essential point of this approach is that the variability within syllable sized portions of human speech is faithfully captured and stored in the inventory. It is partly the inclusion of this variability which makes ***SPRUCE*** speech output so natural-sounding and contributes to the improvement in quality which characterises the system. It should be noted that this particular variability is not the variability referred to elsewhere in this paper. Here the variability is low-level and not cognitively dominated, it is a property of constraints in the human speech production system imposed by the neuro-physiological, aerodynamic and acoustic properties of the system. What we have called variability elsewhere in the paper is determined cognitively and carries information about the attitude and emotion of the speaker (see **Part III - Pragmatic phonetics and naturalness in synthetic speech**).

Conjoining

The conjoining procedure accesses and copies the required inventory object in turn, and performs smoothing at the boundaries between concatenated units.

For research purposes the ***SPRUCE*** inventory is currently several parallel inventories each containing units of different sizes. The standard system uses the syllable sized units just described, but we have phrase and word sized units also available for use in restricted domains where these would be more appropriate. A set of the static allophone sized units referred to earlier is also included for use where words or syllables are 'unknown' to the system. As a rough guide we could say that it would take around 250 allophones, or 10,000 syllables, or 100,000 words, or an infinite number of phrases to synthesise the entire language. The loner the unit the better the quality of the synthetic speech, but the longer units can only be used on a practical basis in a restricted domain. We have chosen syllable sized objects for the standard ***SPRUCE*** (rather than the larger word or phrase sizes) since it is intended for use in unrestricted domains.

Informal experiments have shown that listeners are more sensitive to errors in conjoining the smaller units used to make up the speech output. For allophone sized units, conjoining is critical in producing a natural sounding output,10 with syllable sized units mild errors are

tolerated, and so on. With sentence sized units simple abutting with no attempt at smoothing will usually go unnoticed. Thus in ***SPRUCE*** syllable based synthesis, errors are more tolerated than with earlier allophone systems. We have found that the smoothing algorithms for the optimal joining of syllable sized units are not the same as those needed for conjoining allophones. To restate this in phonetic terms: coarticulatory effects at syllable boundaries are not the same as those allophone boundaries.

Putting it all together

Once a sequence of speech units has been determined there remains the task of marrying this with the prosodic contour calculated earlier in the system. The basis of rhythm m speech is the sequencing of syllables, so a system which is syllabically based automatically specifies the necessary rhythmic units. In contrast an allophone based system needs to identify the rhythmic syllables within the stream of sounds.

In ***SPRUCE***, durations of syllables are adjusted to match the rhythm required by the abstract representation and according to phonetic models of rhythm. At the same time the intonation requirements generated in the prosodic component of the system are reinterpreted as a numerical string. This output is linked as a new parameter to the parameter stream already derived by conjoining inventory units. This process of reinterpretation of an abstract intonation representation is as yet not satisfactory in any text-to-speech system,4 and is too complex to discuss here. However the new algorithms show promise by sensing errors and minimising their effect.

Natural-sounding synthetic speech

As far as the listener is concerned, natural-sounding synthetic speech is, by definition, indistinguishable from real speech. This does not mean that the synthetic speech is exactly the same as real speech. Current theories of language and speech are not sufficiently detailed to enable us to replicate speech production. The goal therefore is to produce a simulation of the human output which is perceptually accurate by employing a system which is as good a simulation as we can manage of the human processes which derive that output.

SPRUCE incorporates two properties of human speech not found in text-to-speech or concept-to-speech systems. These are

1. variability over sketches longer than a single unit, and

2. a pragmatic interpreter.

1. Variability in the production of units in stretches of speech is characteristic of all human speech. Current synthesis systems do not make provision for this kind of variability, with the consequence that repetitions are always rendered identically. A listener detects this error and consequently feels the speech to be unnatural. The phenomenon is beginning to be modelled in Cognitive Phonetic Theory.7 The explanation lies in the fact that a human speaker varies the precision of articulation depending on a predictive assessment of the listener's difficulty in understanding what is being said: if the speaker predicts the listener will encounter ambiguity or lack of clarity then the precision of articulation (and hence of the sound wave) will be increased and *vice versa*. In a synthesis system this would mean a continual adjustment to the 'accuracy' of the units retrieved from the inventory before conjoining them, dependent on the semantic, syntactic and phonological context of the units.

This ongoing adjustment is one of the tasks undertaken in ***SPRUCE***. It does this by incorporating a model of human speech perception against which it tests every utterance it intends to make, and continually adjusting the variability of the projected speech output.11

2. Pragmatic effects are characteristic of every utterance in human speech. They are subtle effects overlaid on a normal neutral speaking 'tone' which convey to the listener such things as the mood of the speaker, his or her attitude to what is being said or attitude

toward the listener. In general such effects are most often encountered in changes to the prosodic element in human speech. ***SPRUCE*** attempts to generate these effects with the result that the listener feels he or she can detect the speaker's feelings. Characterising the prosodic effects which communicate a speaker's feelings has proved difficult, and the best results have been obtained from training a neural network to learn the effects for itself by presenting it with many examples of human speech. The neural network is then used to modify the otherwise pragmatically neutral output of the text-to-speech system.12 **Part III** of this paper discusses adding these pragmatic effects.

PART III – PRAGMATIC PHONETICS AND NATURALNESS IN SYNTHETIC SPEECH

User reaction to poor synthetic output

In dialogue systems using speech mode, current synthesis systems often produce voice output which sounds monotonous, unnatural and is tiring to listen to. The speech produced cannot be listened to easily over periods of time even as short as a paragraph span. In an interactive dialogue situation users become irritated with the system, and in other situations such as where the system is giving instructions, the user can become bored or uninterested.

Good speech output is important for dialogue systems because user awareness is heightened in dialogue mode: the listener's attention is focused on the speech output, since the task of decoding speech requires concentration. In addition to the plain message, all of the information about the thoughts, ideas and feelings that are being communicated is encoded in the speech waveform, and the range of variability in natural speech is narrow.

In contrast, the speech recognition mode is, from the listener's point of view, concealed within the first stage of the automatic communication system. In human speech systems, errors in recognition can usually be repaired by the human system. Therefore simulation of human dialogue systems needs to take into account both error repair for recognition and high information content for synthesis.

Common errors in current synthesis systems are: poor quality, limited bandwidth, inadequate segment conjoining, monotonous and inappropriate intonation, poor stress assignment, inability to disambiguate homophones, etc. The conclusion is clear that the majority of speech synthesis is not practical at the present time for voice output in dialogue systems without some improvement being made.

Lack of naturalness in synthetic output

There are a number of factors which contribute to the lack of naturalness in the speech output from speech synthesis systems.

a. Intonation and rhythm

Errors of intonation and rhythm lead to monotonous or incorrect output, or can contribute to a misunderstanding of the meaning of what is being said. Intonation errors arise from inadequately modelling intonation generation, incorrect assignment of prosodic markers at a higher linguistics revel, and incorrect interpretation of these markers at lower levels. Errors of rhythm arise from failure to model adequately the way in which segmental durations vary during an utterance by failing to set an appropriate range of acceptable variation.

b. Variability along the prosodic parameters

Another source of error involving the prosodic parameters of duration and fundamental frequency is the failure to take into account the fact that human speakers intend to vary these

parameters for specific effects. So, for example, slowing down the overall rhythm is often used to focus the listener's attention on a specific word or phrase. A speaker may pitch the overall fundamental frequency a little lower to indicate that the current piece of information is confidential between speaker and listener, and not intended for anyone else (even if no one else is currently present); this is often accompanied by an overall drop in acoustic amplitude.

c. Incorrect segmental rendering

Errors generated in the phonological processing within a synthesis system an lead to an incorrect choice of segments for rendering part of a particular word. There are, for example, occasions in human speech where vowel reduction under stress conditions or in slow speech is not correct. In fast speech, on the other hand, there may be occasions where greater vowel reduction is called for in unstressed syllables, or even total deletion of these syllables.

d. Paragraph prosodics

It is clear that to correctly render prosodic elements when simulating human speech production the domain over which the prosodic contour is computed needs to exceed the single, isolated sentence. Not only should the domain be paragraph size, often there are prosodic and pragmatic effects w-which occur in a particular paragraph that depend on what has been said in the preceding paragraph. This is particularly important in dialogue where the meaning of what one speaker says influences the reply in terms of word choice, etc., but also influences how it is spoken - a factor affecting the prosodics and pragmatics of the reply.

Pragmatic features and variability

As mentioned earlier, there are a number of sources of variability in speech production. Some of the variability the listener detects and decodes in natural speech is due to pragmatic factors generated at a linguistic level higher than the phonology and phonetics normally used in speech production simulation. Pragmatic features characterise information about attitudes and feeling the speaker wishes to convey to the listener. They are realised by changes in duration of words and syllables, and by changes in fundamental frequency and amplitude.

In the study of human speech production the influence of pragmatic factors on aspects of speech production is studied under the heading *pragmatic phonetics*. In synthesis the purpose of this level is to generate the means of expressing the attitudes, beliefs, emotions and intentions of speakers where these are not directly encoded using words, but are encoded in the manner of speaking. The speech produced is derived by overlaying the pragmatic requirement on the otherwise neutral plain message.

Other sources of variability are introduced as a consequence of choices dictated by the discourse model which is managing the system, and some from context supplied by previous utterances.

Modelling variability

There are currently two ways of obtaining data for building models useful in simulating voice output.

a. Standard scientific data gathering

This consists of building a database derived from natural speech. Normalising sets of utterances from many speakers has not proved successful; therefore most synthesis systems are based on the speech of one speaker.

In this approach, measurements are made of the formant frequencies, amplitudes and durations of relevant segments. The notion of segment usually refers to a unit at the phonological level, including dialect variations, but short duration stretches of speech called

'acoustic phonetic elements' such as burst frequency on stops are included in the segment tables or inventory of basic speech building blocks found in synthesis systems.

Sample durations are derived for segments and acoustic elements, again usually from a single speaker. A decision must be made as to whether durations from words spoken in isolation, in lists, or in contexts provide suitable data for duration values.

Intonation patterns are usually simple, thus creating the effect of monotony. However, a relatively new approach, based on work by Pierrehumbert13 and Silverman14 is promising. Higher level information, such as grammatical category, is required in this method; algorithms can then be developed which assign a varying intonation pattern to sentences.7

b. Neural networks.

A second approach involves the use of neural networks as a data reduction device. This consists of training a network to associate sample abstract prosodic patterns with real phonetic data about human durational and intonation contours. In the work by one of the authors this involved deriving an abstract intonation pattern from a phonological description and matching it with fundamental frequency changes obtained by measurement of human speech.6

When, after training, an intonation pattern is presented to the network input layer, the correct fundamental frequency contour will be output from the network (Fig.3).

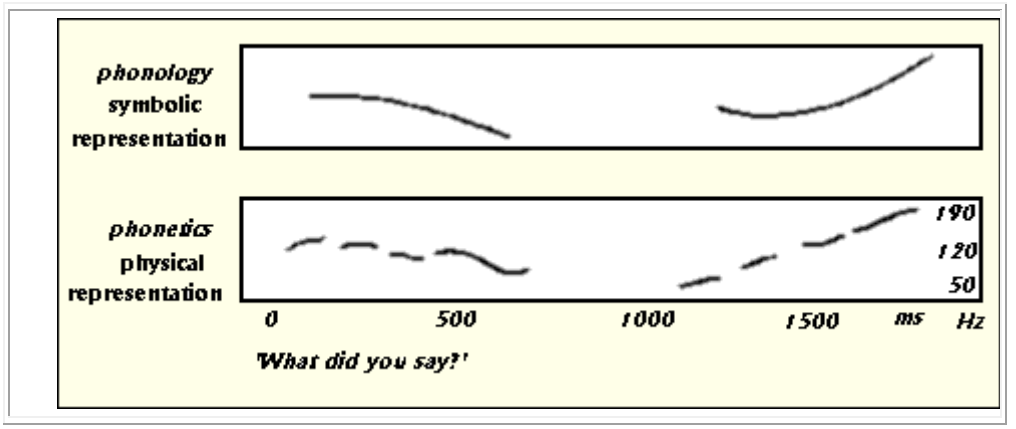

Fig.3 The phonetic representation shows the actual fundamental frequency change with time. The phonological representation shows the speaker's underlying intention in accordance with traditional phonological descriptions.

Pragmatic markers and neural networks

As an extension to this work, a network has been trained to associate the prosodic patterns embodying attitude or emotion changes in sets of sentences. The training stage consisted of presenting the abstract phonological pattern together with a pragmatic marker to the network, associated with the measured phonetic sample of a speaker's attitude, or emotion as encoded in that person's duration and intonation contours. The emotions presented were: happiness, gloom, excitement, surprise, disappointment, neutrality, questioning. In training the network was able to form an association between the phonological representation with an appropriate pragmatic marker and the known phonetic samples.

The trained network is then presented with pragmatic markers, and is able to compute the correct phonetic output in terms of duration and intonation. The training details are reported elsewhere.6

The emotions marked by the pragmatic markers for the purposes of this experiment with neural networks were considered too extreme for application to usual discourse as we find in, for example, timetable enquiry systems. It would be unusual for such an interactive database system to convey its information with gloom or excitement! But one of the essential features of dialogue is to verify information received, and in this case the pragmatic effects of *contrast* and *emphasis* may be used by the listener, and *questioning* used by the speaker.

For example, a passenger inquiring about the time a plane is scheduled to leave may not entirely understand the information given by the system, and the following fragment of a dialogue may occur:

passenger: What time does the London plane leave Tuesday morning?
*airline:*9:30.
passenger: Sorry, I didn't hear that.
airline: (with emphasis) It leaves at 9:30.
passenger: Was that 9:50?
airline: (with contrast) No, it leaves at 9:30, not 9:50.

The phrase *9:30* will be spoken here in three different ways: neutrally, with contrast, and with emphasis.

A network was trained to associate the phonological representation (intonation pattern together with pragmatic markers for neutral, contrast, or emphasis) with the acoustic representation of fundamental frequency and duration for a set of time phrases for a series of simple plane timetable dialogues. The following example (Fig.4) shows the graphed patterns for the phrase *9:30*.

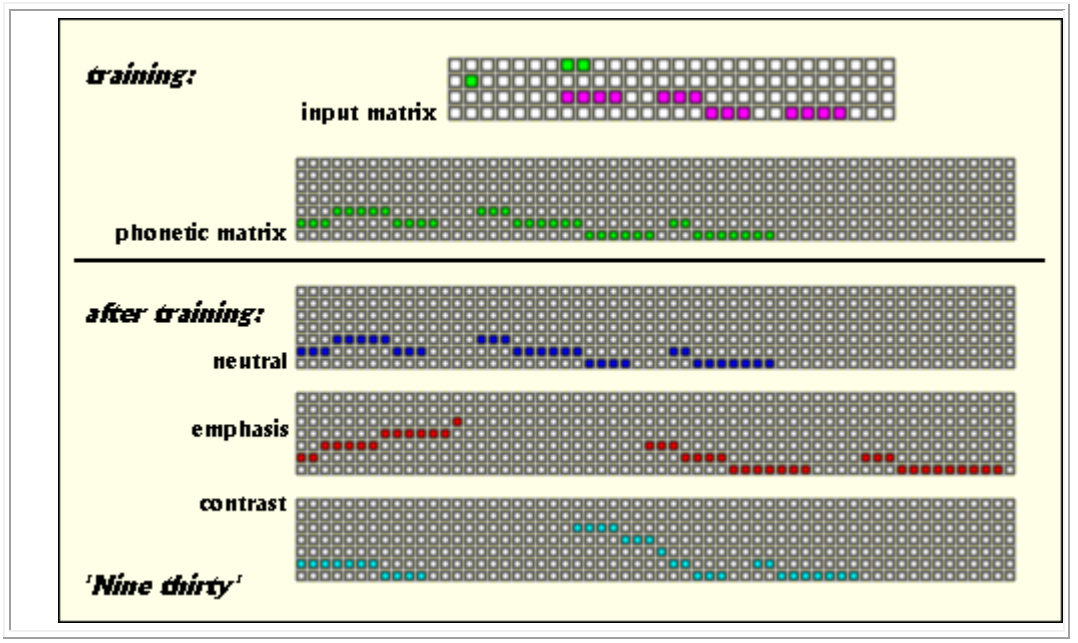

Fig.4 The neural network was trained to associate a phonological representation with an acoustic representation of fundamental frequency and duration.

By using data obtained by standard methods and also by neural network, we are building a rule-based system to generate natural-sounding intonation and rhythm contours. The trained network seems to be a useful way of using phonological information overlaid with pragmatic information. The aim is to improve intonation and duration patterning of computer speech by reducing some of the disadvantages of the effects of monotony and fatigue. It should be more acceptable because it sounds more natural.

In order to produce good computer speech, it will also be necessary to clarify the relationship between speech research and language processing in dialogue simulation. Although database inquiry systems are obvious users of good computer speech, such clarification will be useful in any general human-computer interaction device. Another problem which confronts all speech research concerns the constraints imposed by the nomenclature and descriptive vocabulary available to us. Because information can be conveyed by pragmatic information it is thought that computer speech will be more

acceptable to the general public if these pragmatically derived prosodic effects can be communicated in the correct dialogue setting.

Applications for dialogue systems can be seen in human-computer interactions such as database inquiries, or mixed-mode interfaces for information systems such as natural language query as well as tabular query, and as confirmation and error correction in development of dialogue systems. one important application is in computer aided learning (see **Part IV - An application - speech synthesis for computer aided learning**). It is essential in all applications systems that voice output can convey clearly all the information required.

Applications

Speech and language technology environments where good computer speech is essential might be in education, medical diagnosis, safety monitoring and voice controlled appliances. In of office environments database inquiry systems might be more flexible in their need for good speech, but they must still be reliable and the information they give must still be capable of verification by the listener.

The addition of appropriate pragmatic effects obviously lends a more friendly feeling to communication, but friendliness is not the objective. There are subtle effects that could be essential: a pleasant neutral tone when dealing with children or anxious patients, a firm tone in dealing with emergencies, a neutral but not boring style in the office, and so on.

The technology is developing rapidly for good voice output (see **Part II - The *SPRUCE* Project**) but for fully flexible use, tone of voice and expression of attitude enhance considerably the generalised use of computer speech systems (see **Part V - Multimedia and spoken dialogue systems**).

PART IV – AN APPLICATION – SPEECH SYNTHESIS FOR COMPUTER AIDED LEARNING

Introduction

As soon as computers were introduced into universities in the early 1960s they began to be used for teaching as well as research, and that process has continued to the extent that the use of computers for teaching has permeated the entire teaching environment. For the purpose of this article the use of computers for training is equated with their use for teaching, and the acronym CAL (Computer Aided Learning) will be used here as a generic term to cover all uses of computers as a tool to enhance the learning environment.

Since the dialogue between user and computer is so dependent on the components of the computer system it is helpful to consider the development of the hardware and software used for CAL over the last 20 to 30 years. Initially the equipment available was rather basic, consisting usually of a teletype attached to a minicomputer. These terminals were slow (10 characters per second) and noisy, and since they were time-shared each terminal had to compete for the available resources. In the early 1970s graphics terminals, mainly Tektronix, appeared on the scene, but though they provided CAL developers with considerable more scope in the design of their software these terminals were very poor for handling text, racked facilities for selective erasure and required subdued lighting for viewing. In the late 1970s and early 1980s the introduction of the microcomputer began the transformation of the CAL environment into what is available today. The typical CAL environment now is one in which users have their own personal computer consisting of a powerful microprocessor, medium to high resolution colour monitor, keyboard and mouse together with a human-computer interface that makes the computer altogether more 'friendly'. High-quality animated graphics is now possible which enables CAL software to provide material that could only have been provided on film before.

Continuing enhancement of the CAL environment is providing even greater advances with the advent of multimedia systems which combine computers with computer controlled videodiscs and CD-ROMs, thus enabling high quality moving video with accompanying sound or speech to be displayed on the screen. If required, text messages and computer generated graphics can be superimposed upon these images by the computer.

Why CAL with speech?

In the early days of CAL there was no option but to conduct the dialogue using text, but to what extent has this changed as the interface has become more sophisticated? The answer to this question is *hardly at all.* Dialogue with current CAL systems is generally conducted by means of a keyboard and mouse for input, and the screen for output. This inevitably means that text is the most important medium both for instructions and explanations being given to the user, and also for responses from the user.

This situation contrasts strongly with most other teaching environments involving human teachers, where the emphasis is usually on spoken communication. Even courses which rely heavily on tuition by correspondence such as those provided in the UK by the Open University endeavour to provide some face-to-face contact. It would appear, therefore, that spoken communication is considered to have considerable advantages in teaching situations.

Aside from the fact that humans communicate by speech when teaching, are there other justifications for using speech in CAL? Current CAL systems rely very heavily on the visual sense. A large part of a CAL lesson consists of the user reading text instructions from the screen or an associated manual and then acting upon these instructions by giving commands to the computer via the keyboard or mouse. The computer responds by displaying more text or some graphics image. In contrast traditional teaching frequently involves the teacher in explaining some subject to the pupil by means of blackboard or overhead projector or sides so that the teacher is able to *simultaneously* talk about what is appearing before the pupils' eyes.

There are many situations in CAL where one wishes to do the same, namely comment upon what is happening on the screen. In simulations of various physical phenomena one wants to draw the pupil's attention to some particular feature being displayed. Although this can be done using flashing colours or some similar attractive ploy the explanation has to be provided by text which means using a pop-up window or split screen or similar effect to show the text. Clearly this distracts the pupil's attention from the feature being described.

Tutorial packages for spreadsheets, word-processors, etc., are another example where the addition of speech would greatly enhance the naturalness of communication. In describing an example of using a spreadsheet one would want to draw attention to data entry commands and show how columns can be manipulated, and one would want to show the effects of commands as they are described. Clearly, the ideal situation is to *talk* about what is happening on screen. An extra channel of communication is needed and speech is the obvious solution.

Distance learning and open learning are now having an increased profile, and in situations where students have little face-to-face contact with the lecturer a speech enhanced CAL system has the added advantage of personalising the teaching a little. The Open University already makes extensive use of cassette tapes on its courses - which is indicative of the importance it gives to the use of speech in communicating with the student. Now that more use is also being made of computers at home it is only logical that speech should be added to the human-machine interface.

Indeed the Open University is already investigating the provision of a purpose built computer called the Thought Box. Its architecture is described in an article by Alexander and Lincoln15 in which they state that 'the combination of speech and visuals is likely to be the most powerful computer based learning medium of the short to medium term'. They propose to provide speech by having an integrated cassette recorder. A speech synthesis card would be a much more satisfactory solution.

Indeed, it has long been recognised that speech is an essential component of the human-computer interface. Tandberg were marketing computer-controlled cassette recorders in the

early 1980s, and the BBC micro was also able to control the switching on and off of a tape recorder.

Of particular relevance are those applications where the users need to use their hands for some specialised activity e.g. learning to wire up some electrical circuit. In such a situation the users can first of all be shown how to do the task while *simultaneously* listening to a commentary, and then *instructed* on how to perform the task such that they can concentrate exclusively on using their hands without having their attention distracted by needing to read some text. There are many practical skills for which training can be given using computers and which would benefit by having a speech interface, not least of which is the skill of being able to use the computer in the first place.

Other categories of people for whom CAL would be transformed by the introduction of speech would be the blind and the handicapped. In particular for these groups speech input would be as important as speech output.

The emphasis so far has been on extolling the virtues of getting the computer to speak to the user. This can be justified by the fact that in most teaching situations the majority of speech communication is from teacher to pupil and, therefore, in CAL it is more important to have speech output rather than speech input. However, that is not to say that speech input does not have a valuable role to play.

For example, data input and screen movement provided by voice would be a very useful mode of communication for spreadsheets. Requesting repeated explanations and/or instructions could also usefully be done under voice control. And, as has already been mentioned, voice input would be very welcome for blind and certain physically handicapped people. The situation at present, however, is that recognition of *any* speaker is extremely difficult for continuous speech although there are systems capable of this in some research laboratories.

How should speech be provided?

Once having accepted the case for speech to be an integral part of the CAL interface the question arises of how such speech should be provided. Although users are tolerant of machine like speech for limited periods of time future CAL systems must surely have natural sounding speech if the dialogue is going to be successful.

Tape recorders are undoubtedly cheap and provide good reproduction of the human voice. However, they are not very convenient to use. The inconvenience of loading tapes, checking connecting cables, rewinding to repeat sections, etc., means that speech will not become readily available using tape recorders. This procedure has already been tried and been shown to be unsuccessful.

A better alternative would be to use digitally recorded speech. Computers such as the Apple Macintosh and NeXT workstations are equipped with facilities for the recording and playback of speech. However, digitally recorded, speech even though it can be compressed, is expensive on disc space. Creating and editing speech enhanced software is also more time consuming and cumbersome than creating text files.

Another alternative is, of course, computer controlled videodisc as provided by multimedia systems. However, this is more complicated than the previous alternative since a disc has to be created and pressed, and any subsequent editing could require the reassembly of the team of people responsible for the original disc.

The problem in trying to introduce speech into the CAL interface is that it appears to lack the ease and flexibility which text files possess. Such files can be edited and linked into programs at a moment's notice and, moreover, do not have to be maintained by the person who created the speech in the first place. It is clear, therefore, that what is required is the facility to convert text into speech. Fortunately there are now systems available of doing exactly this, namely converting text to speech. Such systems, generally referred to as text-to-speech (TTS) synthesis systems, can operate in real-time and run on inexpensive computers. However, the systems which are commercially available still have problems in producing

totally natural sounding speech, partly due to the fact that the computer doesn't 'know' what it is about to say.

The development of the ***SPRUCE*** system as described earlier will resolve many of the current problems with TTS systems. This system has the advantage that the original voice is always available, or even several different voices if sufficient tables are provided. It will also allow teams of people to work on the same CAL package without creating different voices. CAL developers will then be able to take advantage of this system to produce multi-modal dialogues involving speech with no more difficulty than they now experience in producing text-based programs.

It is conceivable that the superior quality of professionally produced CD-ROMs will win the day for commercially produced CAL material, but consider the vast amount of CAL software produced for local consumption. Indeed even for CD-ROM software there would be great advantages in developing the package using TTS and only going to CD-ROM when all the teething problems have been sorted out. One doesn't want the cost of recalling an original production team to recreate a disc when flaws in the presentation are found during usage.

A study of the viability of CAL systems as proposed above is currently in progress at Bristol University and provisional work on enhancing a tutorial spreadsheet package with text-to-speech is encouraging.

PART V – MULTIMEDIA AND SPOKEN DIALOGUE SYSTEMS

There is a difference between communication and the transfer of information. Communication involves an appeal to the listener as a person, whereas the transfer of information consists of transmitting facts to the listener's intellect. Good multimedia is a seamless integration of text, audio, speech, video and data communications - all within one system. The task for applications developers is to design a user interface which allows access to the different components and simple control of the different functions.

Until recently it has been difficult to exploit the potential of the multimedia technique because of problems associated with storage, programming and speed. Recent advances in technology have brought forward the possibility of realistic high quality multimedia systems. Furthermore, some of these systems can 6e used by computationally competent users in many fields. For example, manipulating storage and display features has become practical for the reasonably adept.16

The value of multimedia in a dialogue system lies in the ability to communicate with the whole person through sound and vision, accompanied by printout if wanted. Thus feelings and beliefs can be brought into a computer based environment, as they are in face-to-face or telephone human communication systems.

One of the uses of good speech synthesis in a multimodal system is that information can be transferred through the speech medium, but also can enhance accompanying information and ideas communicated by video. In order to avoid distraction natural sounding synthesis is essential both to convey plain messages and also to help to convey attitudes and feelings.

PART VI - CONCLUSION

In this paper we have been discussing ***SPRUCE*** - a new speech synthesis system under construction that accepts either text or concept input. A major consideration underlying the work has been the development of a speech production theory which lends itself to building an acceptable simulation of the processes involved.

In order to improve the usefulness of speech synthesis we have been developing ***SPRUCE*** with a view to applying it in specific man-machine interface environments. There are many such uses including general interactive information systems. We have illustrated an

application in the area of computer aided learning, an area of research of particular interest to one of the authors **(EL).**

One of the main requirements of a synthesis system is that its output should not only be intelligible but that it should be as natural as possible. In the development of ***SPRUCE*** we have paid considerable attention to modelling what causes a listener to decide whether what is being heard is natural or artificial.

Another of the authors **(KM)** has worked in the area of pragmatic phonetics. Much of the naturalness of ***SPRUCE*** speech comes from incorporating this work into the simulation model. Part of what makes human speech sound human is that it conveys to the listener more than just the basic meaning of the words being spoken: something of the speaker's attitudes, emotions and beliefs are conveyed by *how* words are spoken.

The third author **(MT)** has been modelling other aspects of naturalness in human speech, in particular systematic variability effects. Then these are coupled with the pragmatic effects the increase in naturalness in the resultant synthetic speech is appreciable. We are beginning to experience text-to-speech and concept-to-speech synthesis which, to the lay user in the kind of application we have described, is indistinguishable from human speech.

REFERENCES

[1] C. Proctor and S. Young (1989) Dialogue control in conversational speech interfaces. In *The Structure of Multimodal Dialogue* (eds. M.M. Taylor, F. Néel and D.G. Bouwhuis). Amsterdam: North Holland

[2] E. Lewis and M.A.A. Tatham (1991) ***SPRUCE*** - a new text-to-speech synthesis system. *Proceedings of Eurospeech '91*. Genova: ESCA

[3] M.A.A. Tatham (1989) Intelligent speech synthesis as part of an integrated speech synthesis / automatic speech recognition system. *In The Structure of Multimodal Dialogue* (eds. M.M. Taylor, F. Neel and D.G. Bouwhuis). Amsterdam: North Holland

[4] M.A.A. Tatham (1990) Preliminaries to a new text-to-speech synthesis system. *Proceedings of the Institute of Acoustics* 8. London: Institute of Acoustics

[5] J.N. Holmes (1988) *Speech Synthesis and Recognition.* Wokingham: Van Nostrand Reinhold

[6] Katherine Morton (1992) Pragmatic phonetics. *In Advances in Speech, Hearing and Language Processing* Vol. 2 (ed. W.A. Ainsworth). London: JAI Press

[7] M.A.A. Tatham (1990) Cognitive Phonetics. *In Advances in Speech, Hearing and Language Processing* Vol. 1 (ed. W.A. Ainsworth). London: JAI Press

[8] M.A.A. Tatham (1992) Generating natural-sounding synthetic speech from text. *Proceedings of Voice Systems Worldwide - United Kingdom.* New York: Media Dimensions

[9] S.J. Young and F. Fallside (1979) Speech synthesis from concept: a method for speech output from information systems. *Journal of the Acoustical Society of America* 66

[10] J.N. Holmes, I.G. Mattingly and J.N. Shearme (1964) Speech synthesis by rule. *Language and Speech* 7 (3)

[11] P.D. Green, G.J. Brown, M.P. Cooke, M.D. Crawford and A.J.H. Simons (1990) Bridging the gap between signals and symbols in speech recognition. In *Advances in Speech, Hearing and Language Processing* Vol. 1 (ed. W.A. Ainsworth). London: JAI Press

[12] Katherine Morton (1991) Improving naturalness in speech synthesis using a neural network. *Proceedings of NeuroNimes '92* . Nanterre: Editions Colloques et Conseil

[13] J. Pierrehumbert (1981) Synthesizing intonation. *Journal of the Acoustical Society of America* 70

[14] K.E.A. Silverman (1988) *The Structure and Processing of Fundamental Frequency* . PhD thesis, University of Cambridge

[15] G. Alexander and C. Lincoln (1989*) Mindweave: Communication, Computers and Distance Education* . Oxford: Pergamon Press

[16] W. Rash (1992) Multimedia moves beyond the hype. *Byte* Vol. 17 (2). New York: McGraw Hill