Speech Synthesis in Dialogue Systems

Mark Tatham
Katherine Morton

Reproduced from *Spoken Dialogue Systems* (1995). Vigsø: ESCA (European Speech Communication Association), 221-225

ABSTRACT

This paper argues that dialogue synthesis has special requirements. Users expect a high level of naturalness. We argue for good segmental rendering with the addition of pragmatic effects, such as emotion and attitude. We discuss the language model used, emphasising the phonological and phonetic levels essential to handle these effects. We draw on the theories of Cognitive Phonetics and Pragmatic Phonetics.

INTRODUCTION

Dialogue involves a human user and a computer interacting to exchange information. A telephone inquiry system is an example of dialogue based interaction for requesting and providing information. Dialogue systems incorporate several identifiable elements including a dialogue controller [1] and input/output subsystems taking the form of automatic speech recognition and speech synthesis.

In this paper we address a number of questions determining how to provide a synthesis module to meet the demands of dialogue. In particular we consider

- general prerequisites of synthesis,
- special requirements for synthesis in dialogue systems,
- modelling the requirements in the dialogue context,
- implementation of the model.

We make the assumption here that other components of the overall system have determined such matters as

- what is to be spoken,
- how it is to be spoken.

However, the implementation of the output will determine the formulation of these other components of the system.

GENERAL PREREQUISITES

To ensure user acceptability in a wide range of environments the speech produced by a synthesiser must be of high quality. Almost all synthesisers now produce highly intelligible output, and listeners have little difficulty in determining the intended message. Although there are no universally accepted formal evaluation procedures for synthetic speech, there is little disagreement that generally the threshold of good intelligibility has been passed. Much of the achievement is due to improvements in the way the sound wave itself is generated.

To address the question of general requirements satisfactorily we need to make a distinction between low level synthesis and high level synthesis.

Low level synthesis

Low level synthesis refers to the method of generating the actual speech waveform, and high level synthesis refers to the language model used to drive the synthesiser. With low level synthesis it is easier to achieve user acceptability than with high level synthesis, although work continues on developing new techniques like concatenated waveform synthesis [2], and refining older ones like formant synthesis [3, 4]. The acceptability of low level synthesisers is generally demonstrated using resynthesis techniques in which high level synthesis considerations (see below) are removed by using a direct analysis of human speech as the starting point.

Formant synthesisers are now stable devices, producing no unwanted anomalies. Improved methods of conjoining segments of pre-stored waveforms guarantee that intelligibility is no longer a problem for concatenated waveform synthesis. For many purposes it is sure that waveform synthesis will completely displace other methods of generating the sound wave now that means have been found of satisfactorily altering timings and fundamental frequency within stored segments of waveform. A few problems remain, such as changing formant bandwidths to correlate with fundamental frequency changes. But however difficult the solutions to such problems may prove they are refinements rather than major considerations for basic synthesiser design strategy.

High level synthesis

High level synthesis involves the language model used to determine what is to be spoken, and when, and how. When a high quality low level synthesiser is driven by an input derived automatically from a language model the quality of the output falls dramatically compared with that derived from resynthesis. Thus language models need more work.

Present language models enable a spoken output which is generally intelligible, though for sentences of more than a few words or for extended use in dialogue acceptability tends to fall. There are two basic requirements for the language model; it should

- specify segmental production,
- specify supra-segmental production.

These basic requirements come from the phonological part of the model responsible for planning the abstract sound shape assigned to the words and sentences to be spoken. How segmental production is specified will depend on unit sizes and the phonological interactions between them. How supra-segmental production is specified will depend on how the language model interprets the underlying semantic and syntactic bases of any particular phrase or sentence to be spoken.

High level synthesis is generally designed to accept an input based on text generated elsewhere. In dialogue applications, it might be desirable to replace the text input by an input representing an expression of concepts. Although some work has been done on this idea [5] it is still difficult to see how we might in practical terms bypass text, although there is complete agreement that an 'interlevel' of text between 'pure' semantic expression and speech is theoretically unsound. Text will prove in the future to be a hindrance because

- human beings when speaking do not first of all output text and then proceed to read it out aloud,
- plain, unmarked text is does not encode most of the important supra-segmental phenomena of speech.

Naturalness

There are several questions concerning naturalness:

- can the voice be confused with being human?
- is the confusion valid over long stretches of speech?
- does the speech seem to have the appropriate 'tone'?

- does this tone change if the content of the message changes?
- does the 'speaker' appear to understand what is being said?
- does the speaker appear to have a view about what is being said?
- does the speaker, by tone of voice, convey feelings which are not part of the meaning of what is being said?

Naturalness involves more than not being able to distinguish between human and synthetic speech. Human speech normally reflects many of the other properties listed on every occasion a person speaks. It involves, therefore, getting the segmental vehicle for the message to sound human. And then it involves getting the supra-segmental properties of timing and intonation right, conveying the non-segmental properties of the message and also the attributes of tone of voice.

Despite the actual words being spoken a person will always betray feelings of attitude or mood. Take away these effects and something of the naturalness of human speech is lost, and the listener will feel uneasy - perhaps without being able to say why.

DIALOGUE SPECIIC REQUIREMENTS

The general requirements for speech synthesis are needed in dialogue systems. There are also additional requirements which make synthesis more critical in applications such as information systems. The needs of synthesis in this context concern long term acceptability and naturalness, and usually depend on the language model used in high level synthesis.

High level synthesis requirements

The dialogue specific requirements are almost all to do with the language model in high level synthesis. Here are some of the main requirements affecting speech production in the planning stage:

- the general tone of voice must suit the dialogue type,
- the 'speaker' should appear confident,
- rate of delivery must vary appropriately,
- tone must vary to reflect changing attitude to the information presented,
- tone must vary to reflect changing attitude toward the human user.

This short list illustrates the kind of requirement which does two things:

- it makes the synthetic speech more acceptable to the user because it reflects the properties of speech and the kind of variability which contribute to naturalness,
- it increases the impact and credibility of the information being communicated.

MODELLING THE REQUIREMENTS

We first outline the general shape of the model, and then suggest a particular approach for handling some of the problems associated with segmental rendering. The general approach adopted also proves a framework for handling tone of voice variations.

Segmental naturalness

All sizes of linguistic unit (and also units which do no receive support in linguistics) have been tried as the 'building blocks' of synthetic speech, and some have been found to work better than others at lower levels of the synthesis procedure. There are two major problems:

- how to conjoin the segments imperceptibly,
- how to time the segments properly.

On both counts the longer the segment the less the difficulty - if the segment is a complete sentence then there are is no problem at all, but if it is allophone sized then the problems on both counts increase, and have no satisfactory solution yet. But just looking for the easiest segment type to make the system work is attempting to tackle the problem of segmental naturalness without recourse to theory.

We base the speech production in our language model on Browman and Goldstein's ARTICULATORY PHONOLOGY [6]. We think however that their theory needs some refinement, and we are sure it needs some adaptation if it is to be used as a basis for the simulation of speech production using speech synthesis. Although overall the theory is very useful, Tatham [7, 8] does discuss some issues, pointing out

- an apparent confusion between what is cognitive and what is physical, without reconciling the two as much as the proponents believe,
- a exemplar coarticulatory phenomenon which the theory handles incorrectly,
- an illustration of inadequately addressing some questions about coarticulation and their manipulation.

ARTICULATORY PHONOLOGY makes use of a graphical representation - the gestural score - stacked articulatory parameters unfolding in time. The intention is that the score should constitute a plan of what is to be spoken, but that the representation should be such that the actual results of the execution of the plan can be superimposed. By extension we propose that the gestural tracks superimposed to indicate execution also be used to indicate projected execution. Tatham pointed out that Browman and Goldstein's gestural score was unable to predict some acoustic, intraoral air pressure, and electromyographic data observed in simple stop+vowel sequences in English and French. He proposed the additional mechanism of 'supervision' to account for the difficult data. This notion was based on the theory of Cognitive Phonetics [9] which claims that there are cognitive processes occurring in speech production that cannot be assigned to either phonology or physical phonetics. The supervisory mechanism is used to constrain or enhance phonetically determined effects (coarticulation or coproduction) which are being used by the language to convey meaning.

The claim of Cognitive Phonetics is simple. In addition to using phonetic gestures entirely within a speaker's control as a means of producing an acoustic carrier for meaning, the speech production process also has at its disposal a limited degree of control over spurious effects caused by the intrinsic properties of the production mechanism itself. Thus, in Tatham's examples, the aerodynamically induced effect of delay in vocal cord vibration associated with a stressed vowel immediately following a syllable initial voiceless plosive (as in, for example, car [ka(r)] in English - automobile) can be manipulated deliberately for linguistic purposes (as in, for example, car [kær] in French - because).

Thus in this refinement of non-linear phonology which seeks to relate more closely the traditional areas of phonetics and phonology by introducing a cognitive element in phonetics we find that manipulation of coarticulatory effects occurring within and between the basic syllabic unit add to the range of units able to be deployed for phonological purposes within the language [10]. The hierarchical structure of the syllable is more complex than thought [11], and in Cognitive Phonetics ties between sub-syllabic units are indexed as to their susceptibility to manipulation. Thus the execution of the abstract gestural plan in ARTICULATORY PHONOLOGY is modelled with a supervisor agent responsible for ensuring correct execution of ties between sub-syllabic units and between syllables.

We have found that using this particular phonology with the Cognitive Phonetic enhancements has produced a synthetic speech which contains none of the anomalies present in the more traditional boundaries between allophones.

Prosodic naturalness

Suprasegmental effects which contribute to naturalness are modelled as overlays on neutral prosodic contours [12, 13]. This approach has the advantage of maximising generality, and enabling all effects to be related via the neutral 'parent' contour. Each derived contour inherits its parent's properties.

In the current version of the model, Pragmatic Phonetics, the output of phonological prosodic processing represents a neutral prosody incorporating appropriately placed hooks indicating where and to what extent modifications are possible. Modifications are systematically overlaid on the neutral representation and become incorporated in those parameters of the gestural plan which are appropriate to prosody - initially timing and fundamental frequency. They are able to be systematically overlaid because they are characterised as rule-governed departures from the neutral contours. The rules are activated when particular 'pragmatic markers' are generated by a component in the model which is sensitive to the pragmatic needs of the dialogue as it unfolds.

The very high level areas of the dialogue system generate the pragmatic markers, and the speech production model takes these as input, along with either textual or conceptual representation of what is to be spoken. The pragmatic markers are used to overlay a potentially neutral spoken output plan with corresponding attitudinal or other prosodic variants.

ACHIEVING THE REQUIREMENTS

Speech synthesis for dialogue systems has special requirements which are extra to those needed for synthetic speech in a non-dialogue environment. In our model, the high level synthesis system has two input channels, one conveying text and the other a stream of pragmatic markers synchronised with the text. A textual representation is assumed because we lack a conceptual representation.

There are a number of essential components in the language model. Initially all syllables are located and marked, as well as their external relationships (within words) established, and their internal relationships (ties between sub-syllable units). This phonological information is merged with syntactic and semantic information derived by parsers. The output of these processes is a phonological plan - a parametrically organised representation incorporating neutral prosodics. This is the neutral DYNAMIC SPEECH PRODUCTION SCENARIO [7]. The pragmatic marker stream has been carried through the processing, but has yet had no effect. The detail of speech is modelled against the DYNAMIC SPEECH PRODUCTION SCENARIO - a cognitive projection of a stretch of speech within certain linguistic boundaries - for example, a simple sentence. It is a projection in the sense that it is a plan of what the physical SPEECH EPISODE is intended to be; the plan involves notional jerky and chunked) time. In turn the SPEECH EPISODE will be synthesised to form a time-governed and focussed element within the AUDITORY SCENE [14].

The SPEECH EPISODE is synthesised by interpreting the gestural plan in terms of the requirements of the low level synthesiser. In the case, for example, of a formant synthesiser files are generated which specify values for each parameter usually on the basis of 10ms frames. In the case of a concatenated waveform synthesiser the appropriate stored segments of waveform are marshalled, conjoined, retimed and sent to an A/D converter with the appropriate changes in fundamental frequency. Most of these operations are specific to the particular low level synthesiser in use, and are relatively unimportant compared with the processing requirements of high level synthesis.

As a very simple example of the high level synthesis at work consider the following dialogue:

Q. *When does the plane leave?*

A. *It leaves at 5:30.*

Q. *Could you repeat the time, please?*

A. *5:30.*

Q. *Did you say 9:30?*

A. *No, 5:30, not 9:30.*

In this example the synthesis system is required to generate the utterance 5:30 three times. On the first occasion we can use the neutral pragmatic effect associated with simple delivery of information. On the second occasion, though, the language model responding to the human being's difficulties will generate a pragmatic marker calling for a more precise or emphatic tone of voice. On the third occasion, a general polite tone will be maintained, but the earlier precision will now be supplemented with a further overlay creating a tone indicating firmness.

CONCLUSION

In this paper we have argued that there are special requirements for synthesis used in dialogue systems. Because such systems are often used to deliver important information users have an expectation of a high level of naturalness in the synthetic speech. We have discussed two contributors to naturalness: good segmental rendering and the systematic incorporation of pragmatically derived attitudinal effects. We discussed the language model we use at the phonological and phonetic levels to handle these effects, referring on the theories of Cognitive Phonetics and Pragmatic Phonetics. Finally we outlined how the language model underpins a synthesis strategy for generating the improved natura1ness. We presented a simple example of changing attitudinal effect as a dialogue unfolds.

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. Bouwhis), Amsterdam: North-Holland, pp.375-398.

[2] E. Moulines and F. Charpentier (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication,* Vol.8, pp.453-467.

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

[4] D.H. Klatt (1980). Software for a cascade/parallel formant synthesiser. *Journal of the Acoustical Society of America,* 67, pp.971-995.

[5] 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, pp.685-695.

[6] C.P. Browman and L. Goldstein (1992). Articulatory phonology: an overview. *Phonetica,* 49, pp.155-180.

[7] M.A.A. Tatham (1994). The supervision of speech production - an issue in speech theory, *Proceedings of the Institute of Acoustics,* 16:5, St. Albans, pp.l71-181.

[8] M.A.A. Tatham (1995). The supervision of speech production. *In Levels in Speech Communication - Relations and Interactions* (eds. C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen), Amsterdam: Elsevier, pp.115-125.

[9] M.A.A. Tatham (1986). Towards a cognitive phonetics. *Journal of Phonetics*, 12, pp.37-47.

[10] K. Morton and M.A.A. Tatham (1980). Production instructions. *Occasional Papers,* 23, University of Essex Linguistics Dept., pp.104-106.

[11] J.A. Goldsmith (1989). *Autosegmental and Metrical Phonology: a New Synthesis*, Oxford: Blackwell.

[12] K. Morton (1991). Pragmatic phonetics. In *Advances in Speech, Hearing and Language Processing* (ed. W.A. Ainsworth), London JAI Press, pp.l7-53.

[13] K. Morton (1992). Adding emotion to synthetic speech dialogue systems. *Proceedings of the International Conference on Spoken Language Processing,* Banff, pp.675-679.

[14] A.S. Bregman (1990). *Auditory Scene Analysis.* Cambridge, Mass.: MIT Press.