Speech Synthesis in Dialogue Systems

Katherine Morton
Mark Tatham

Reproduced from: Morton, K. and Tatham, M. (1993) Speech synthesis in dialogue systems. In *Proceedings of Eurospeech 93,* 905-910. Berlin: European Speech Communication Association

---

ABSTRACT

This paper deals with the need for speech synthesis in dialogue systems to incorporate tone of voice for cueing in the listener feelings concerning the attitude of the computer-speaker. Dialogue systems intended for different purposes require different global tones of voice to sound completely convincing. But in addition the synthetic speech needs local tones of voice to signify changing and adapting attitudes during the course of the dialogue with the human user of the system. We discuss the format of a tone of voice model, and provide an example using intonation declination.

Keywords: dialogue systems, speech synthesis, pragmatics, intonation, declination.

INTRODUCTION

In dialogue systems where human users and a computer interact for the purposes of information exchange it is essential that the voice input and output interface be of high quality to gain lay user acceptability. An inquiry system operating over a telephone network is a good example of dialogue based interaction for requesting information. Current speech synthesis systems are proving inadequate for high quality voice output in such systems. Failure to provide a quality acceptable for general use lies not with the acoustic model incorporated into good synthesisers [1] but with the language model generating the input to the synthesiser [2].

One area of the language model underlying synthetic speech, prosodics, is particularly important in producing good quality. The reason for this is that it is largely the prosodics of speech – its intonation and rhythm in particular – which convey to the listener information about the speaker's attitude [3].

This information is always present in human speech, but is lacking in synthetic speech. Users of dialogue systems feel uneasy if such information is absent.

Our paper argues that for good quality natural sounding (and therefore convincing) output synthetic speech systems need to capture mood and attitude through adequate prosodic rendering. We see incorporating these effects in the language model which forms the basis of the voice output system as essential for this purpose.

SYNTHETIC SPEECH QUALITY

Speech synthesis systems now produce fairly intelligible output which listeners have no difficulty understanding. But despite the high degree of intelligibility the computer-speaker still sounds unnatural or machine-like. Often the synthetic speech inadvertently communicates to the human user of a dialogue system an impression of irritation or boredom. The point here is that *how* speech is spoken is important to the listener – it must sound appropriate to the dialogue or the listener becomes uneasy and gets the wrong impression concerning the attitude behind what is being heard. Because human speech *always*, it seems, contains pragmatic ‘markers’ to which the listener is sensitive, speech which does not include these markers – such as the usual synthetic speech – sounds unnatural. As the segmental quality of the synthetic speech approaches a threshold of excellence the lack of pragmatic markers becomes even more noticeable; the speech sounds so segmentally natural that it is expected to sound *completely* natural. This has implications for modelling perception of humanness in speech.

The reasons why it is difficult to add ‘tone of voice’ to synthetic speech to provide the listener with pragmatic markers or cues are:

- suitable data is not available for forming the basis of a good model of the phenomenon, and
- suitable applications models for synthesis have not yet been developed in general [4].

TONE OF VOICE IN SYNTHESIS

In developing a model on which to base the synthesis of speech it is important to take tone of voice into account; it is responsible for triggering feelings in the listener which set the context so that the basic message conveyed by the words themselves can be perceived.

We would expect variations in tone of voice, since they are controlled by the speaker, to be obvious in changes in the acoustic waveform. It should be possible in principle to identify and quantify these changes and to express them as departures from some norm. There is, however, so much variability in speech waveforms that it has been difficult to separate out the variations significant for conveying the pragmatic effects from other variability in the waveform.

In order to develop an applications model for synthesis, we distinguish two types of independently varying tone which produce different types of pragmatic effect.

1. Global tone of voice

This is the general or background tone appropriate for the overall dialogue situation For example, in an inquiry system about the weather, the computer-speaker would ideally sound friendly and confident of the facts. In contrast, in some other situation where a synthetic voice is required to announce an emergency, the speaker should perhaps be simultaneously firm, confident and reassuring. In a dialogue forming part of a computer assisted teaching programme, the effect produced in the listener might be that the speaker is being sympathetic as well as instructive. In an aircraft with dialogue information systems a human pilot would expect the synthetic speech heard to be confident and clear and sometimes urgent, but certainly not sympathetic in tone.

2. Local tone of voice

This is a particular tone which varies according to the specific short term requirements of the unfolding dialogue. For example, short term changes may be appropriate in a general telephone-based inquiry system and these would be set against the background of the overall tone which should ideally be firm and friendly.

A specific example of a short term variation of tone might be to sound firm but patient if it becomes apparent to the system that the human inquirer has difficulty understanding the message. In the case of the airline pilot, a long term firm and confident tone might be modified by encouraging and patient instructions if the pilot fails to understand an explanation or instruction about decisions he must make.

Thus the overall model of tone of voice is layered. We begin with a neutral tone which might never exist in real life. This tone is what our standard synthesis system would generate in its basic language model, and is intended to be used for conveying a plain message. It is important to note here that we assume a very high quality voice output system capable of producing truly neutral speech – not having some inadvertent and unsuitable accidentally generated tone of voice.

In turn, local tone of voice is superimposed on the global tone as the dialogue develops. The language model is such that markers indicating these local deviations are generated as appropriate.

Many tone of voice phenomena are characterised physically by constraints on rhythm and f0 movement during a sentence, or by providing a contrasting variation between sentences or parts of sentences using one or both of these parameters of the speech signal. We shall now

examine how a particular intrinsic phenomenon of f0 movement, declination, is manipulated to cue different effects.

DECLINATION

Many speech researchers have observed a progressive lowering of f0 during the utterance of a sentence. The phenomenon is usually attributed to falling sub-glottal air pressure from a high point at the beginning of the sentence to a low point toward the end. The phenomenon is phonetic rather than phonological. It would be right to say that f0 lowering during a sentence should not be modelling in phonology [5]. If it is true that the gradual fall in f0 frequency is a physiologically determined event it must be true of all languages – that is, it must be a phonetic universal, constrained only by a particular aerodynamic effect.

Although some researchers have observed the phenomenon, naming it *declination,* it does not always occur as much as the aerodynamic explanation would predict, nor does it occur all languages. This variation in the manifestation of declination has been sufficient for some researchers to deny its existence [6] [7] [8].

We believe, though, that we have a clear example here of cognitive phonetic manipulation of an intrinsic event [9] [10], which can be constrained to disappear altogether. We are saying that the intrinsic aerodynamic constraints on f0 control over sentences longer than just a couple of words is universal, but it can be cognitively constrained. Furthermore it is important to model the phenomenon as a two-stage process to make transparent the interplay between physiological (or in this case aerodynamic) constraints and constraints which are imposed by cognitive intervention.

Declination provides a useful example of our explanatory format, for the way in which, under cognitive control, rhythm and intonation *in general* are manipulated – enhanced or constrained – to deliberately alter certain parameters of the resultant waveform with a view to triggering certain perceptual effects. For, if we say that rhythm and intonation are 'used' to overlay subtle pragmatic effects on the general meaning of a sentence or part of a sentence, we must suggest a mechanism by which this could happen. The general premise of the theory of cognitive phonetics says that within the *phonetics* of production intrinsic phenomena can be manipulated which are *not* of linguistic origin; and that this manipulation results in an output which is linguistically significant and provides perceptual cues.

Observations of declination

As an informal by-product of an experiment being conducted in our laboratory for another purpose we have been able to assess for ourselves some of the extent to which declination as a physical feature is expressed in English speech. In addition we now have some idea of the kind of circumstances under which it can be manipulated to the point of apparently disappearing or even reversing to produce what we term *inclination.* Inclination can occur, for example, as a high f0 ending to a sentence where a low f0 might normally be expected. This, if produced in error by a synthesis system, may create an unwanted perceptual effect in the listener such as an impression of aggressiveness on the part of the computer-speaker.

Results illustrating the global layer

We obtained some informal results concerning the extent of declination and its modification with speech of different types spoken by several human subjects. Here are the results from three such conditions:

a. *reading text out aloud* – apparently no declination, except perhaps a little on the last sentence of a read paragraph;

b. *giving out information* (as in an inquiry system) – some declination;

c. *spontaneous conversational speech* – greater declination;

d. *in all three cases* – if more than one sentence spoken then declination, if any, was greatest on the last sentence.

Results varied with different speakers, but the above observations represent a general pattern.

Our model accounts for these sample, exemplar results by stating that:

1. In the above cases every sentence potentially underwent declination of its f0 parameter. That is, f0 would fall in frequency as each sentence progressed, and that this fall is an intrinsic aerodynamic consequence of the way the production of f0 occurs over stretches of utterance of sentence length. On this point the model concludes that data concerning subglottal air pressure, glottal impedance, vocal cord tension, etc. would enable accurate prediction of the tendency for f0 to fall in frequency, given a certain starting frequency and a certain sentence length. The detail here is not important. These processes would occur as predicted unless deliberately interfered with to enable constraint or enhancement.
2. The universal progressive fall in frequency of f0 during a sentence can be constrained or enhanced for the purposes of conveying pragmatic effects as an overlay on the basic meaning of the sentence. In reading text out aloud the declination effect is completely or almost completely inhibited; in putting across information the declination effect is allow to a certain extent, and in spontaneous speech the declination effect is relatively unconstrained.

Declination is modelled as an example of global tone of voice – layer one of the tone of voice overlay. However, in actual fact it is the constraint (or lack of it) applied cognitively to intrinsic declination which is conveying the tone of voice information.

Results illustrating the local layer

e. *clause boundaries* – declination (if any) interrupted, and reset somewhat higher before resuming;

f. *at words bearing sentence focus (nuclear stress)* – declination interrupted by a reset.

These are two examples of local effects to cue pragmatic overlay in the listener. They both use a deliberate change in the rate of declination to draw the listener's attention, either to a clause boundary or to the intended prominence of a particular word in the sentence; the same physical phenomenon is being used to signal two different events. The human perceptual system responds particularly well to change in an observed phenomenon and much less well to continuing steady state. Here the speaker deliberately manipulates intrinsic declination to bring about a change in the expect drift down in frequency as the sentence progresses to jolt the listener into noticing something.

DIALOGUE SYSTEMS

An ideal dialogue system would not simply accurately recognise, understand and act on what it hears from a human being but, like a human being, would react to his or her tone of voice. This reaction would include an appropriate adjustment of its tone of voice. This would be in addition to an appropriate tone of voice to match the anticipated use of the system mentioned above.

Thus the artificially intelligence dialogue manager would be sensitive to input from the system's speech recogniser informing that there had been a departure from the predicted state of parameters such as rhythm and f0 pattern. Extending the declination example from above, the recogniser, anticipating a steady fall of f0 through a sentence is 'jolted' when the fall suddenly resets to a higher value on an adjective preceding a noun. A linguist might say that 'contrastive emphasis' had been detected, and the dialogue manager would interpret the reset as the speaker drawing attention to the particular adjective (perhaps in contrast with another adjective used earlier in connection with the particular noun). Apart from detecting this subtle variation of meaning on the part of the human speaker, the dialogue system would adjust its response appropriately. This might require a particular wording for the next sentence, but perhaps just a simple tonal adjustment would satisfy the dialogue.

Recorded illustrations accompany this paper, and demonstrate the application of rules to neutral speech produced by the SPRUCE text-to-speech synthesis system [11]. The speech altered by rule sounds more natural and more appropriate for a dialogue system for lay users.

FUTURE WORK

We plan more experiments

1. to determine how the acoustic parameters of speech are manipulated to cue pragmatic effects;
2. to establish how these natural variations are best incorporated into synthetic speech;
3. to determine appropriate global tones of voice for specific dialogue systems.

REFERENCES

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

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

[3] Morton, K. (1992) Adding emotion to synthetic speech dialogue systems. Proceedings of the International Conference on Spoken Language Processing, Banff, Canada

[4] Morton, K. (1991) Pragmatic phonetics. In Advances in Speech, Hearing and Language Processing (ed. W.A. Ainsworth). London: JAI Press

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

[6] ‘t Hart, J. and Cohen, A. (1973) Intonation by rule: a perceptual quest. Journal of Phonetics, Vol. 1

[7] Ladd, D.R. (1984) Declination: a review and some hypotheses. Phonology Yearbook 1

[8] O’Shaughnessy, D. (1977) Fundamental frequency by rule for a text-to-speech system. Proceedings IEEE International Conference ASSP

[9] Tatham, M.A.A. (1986) Cognitive phonetics – some of the theory. In In Honor of Ilse Lehiste (eds. R. Channon and L. Shockey). Dordrecht: Foris

[10] Morton, K. (1986) Cognitive phonetics – some of the evidence. In In Honor of Ilse Lehiste (eds. R. Channon and L. Shockey). Dordrecht: Foris

11. Lewis, E. and Tatham, M.A.A. (1991) SPRUCE – a new text-to-speech synthesis system. Proceedings of Eurospeech ‘91. Genova: ESCA