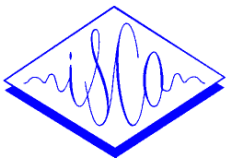

2nd International Conference on
Spoken Language Processing (ICSLP 92)
Banff, Alberta, Canada
October 12-16, 1992

ADDING EMOTION TO SYNTHETIC SPEECH DIALOGUE SYSTEMS

Katherine Morton

Linguistics Dept., Essex University, Colchester UK–CO4 3SQ

ABSTRACT

Current synthetic speech systems produce voice output which is unacceptable for general public use. Listeners report it sounds mechanical, machine-like and is not pleasant to listen to. A model is being developed which adds variability to the waveform; this variability conveys emotion or attitude and produces more natural sounding synthetic speech.

I. INTRODUCTION

Synthetic speech is still rarely used as a means of communicating between machines and humans. However, pre-recorded speech used in limited domains is becoming widely used. So there is a clear need for voice output replacing a human speaker. The reason for not using synthetic speech is simply that it is not yet good enough quality.

Telephone inquiry systems have been introduced where the subscriber is connected directly to the company's database and the required number is spoken without human intervention. What is heard, though, is a sequence of concatenated coded and stored recordings of a human being speaking the digits and a few connecting words. Even though good quality speech is heard, it is not equivalent to *fluent* human speech.

Users report the systems lack 'naturalness' although the speech is intelligible. The difference between *intelligible* and *natural* was obscured by problems in early systems such as mispronunciation of words, small dictionary storage for uncommon words, etc. These problems have been resolved in current systems and correctly pronounced intelligible speech is being produced. But the problem of naturalness remains. Until this is solved, text to speech can only be used where there is no alternative, and users will have to tolerate the monotonous, irritating, machine-like quality.

Commercial need is growing for good voice output in two areas:

1. interactive dialog systems, essential in some public areas such as telebanking, telephone enquiry systems, and

2. multimedia, a rapidly developing field involving both basic research and application of technology where full voice input and output systems will be needed.

These two developments create a real need for researchers to understand and model naturalness in such a way that it leads directly to acceptable synthesis systems.

There are two questions to address in building an applications model:

1. What is it in human speech which constitutes its humanness or naturalness that we cannot yet adequately model?

2. How can a model of naturalness be incorporated into a synthesis system?

These two questions are related: how the simulation is constrained affects the way in which the phenomenon is modelled. The objective is to model naturalness *for the purposes* of building a good simulation; this does not necessarily mean that all aspects of naturalness need to be understood before attempting to incorporate generalizations within the working simulation.

II. THE SYNTHESIS MODEL

The work described in this paper assumes synthetic speech produced by parallel formant synthesis: a common well-tried method [1]. The model is based on summing a number of parameters of continuously varying values. Those parameters include the center frequency and bandwidths of four or more of the lowest formants and their amplitudes. Fundamental frequency is treated as a separate parameter. A typical parameter file is shown in Fig.1.

ms	F1	A1	F2	A2	F3	A3	AHF	S	f0
10	150	33	1600	26	2650	28	35	36	31
20	125	33	1600	26	2650	28	35	36	30
30	125	33	1600	26	2600	28	35	36	30
40	175	33	1600	26	2550	28	35	36	29
50	225	33	1650	26	2550	28	35	36	29
60	250	49	1700	38	2450	45	43	63	29
70	350	49	1700	38	2400	45	43	63	29
80	350	49	1850	38	2350	45	43	63	29
90	350	49	1900	38	2350	45	43	63	29
100	350	49	1950	38	2350	45	43	63	29
110	300	49	1950	38	2400	45	43	63	29
120	275	49	1950	38	2400	45	43	63	29

Fig.1 Synthesizer parameter file. The first five rows represent *th*, the next seven rows represent a transition between *e* and *ea* in 'the early...'. (For the sake of clarity the nasal frequency and the low frequency amplitude parameters have been omitted.)

If parameter values are determined by hand analysis of spectrograms using samples every 10ms and these values are delivered to the synthesizer at a rate of 100 samples/s, speech output is virtually indistinguishable from real speech. The speech output is not only intelligible, but also natural sounding. Therefore, the parallel formant model is adequate and a synthesizer built to this specification can deliver good speech. The problem lies with determining the values to be delivered to the synthesizer.

III. TEXT-TO-SPEECH SYNTHESIS

Text-to-speech synthesis has one advantage over prerecorded speech: flexibility. A waveform can be produced for utterances which can't be predicted in advance. However, for use in interactive systems, we need a generalized model of speech production that will reproduce the naturalness required.

10.21437/ICSLP.1992-226

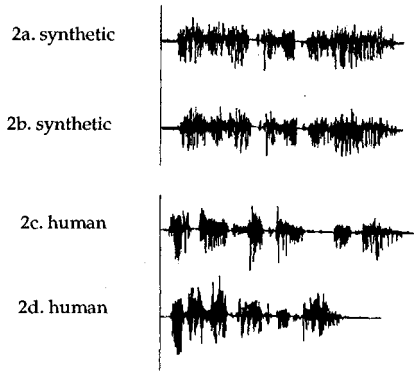

Fig.2 Waveforms of human and synthetic speech. 2a, 2b — synthetic ouput and 2c, 2d — human repetitions of the same sentence: 'There'll be a high tide in two hours.'

Fig.2 shows the waveforms of a single sentence spoken twice by a human being and twice by a modern text-to-speech synthesis system, SPRUCE, described in the proceedings of this Conference [2]. Repeated output from a synthesizer produces an identical waveform each time. But when a human speaker repeats a sentence, variability is introduced. This variability, seen in spectrograms and waveforms of human speech, is not encoded in the parameter files of synthetic speech systems.

Speech researchers have observed that some types of variability are perceived by the listener as linguistically significant. Furthermore, within a sequence of consecutive utterances, such as in dialogue, variability appears to be associated with response to previous utterances. Simulating this variability is not currently addressed in synthetic speech systems.

III. TYPES OF VARIABILITY AND NATURALNESS

Variability can be classified into two types: *intended* and *unintended* [3]. Intended variability is linguistically significant e.g. in conveying changes of mood, unintended variability is the result of coarticulatory and myodynamic process in the human speech system.

In this paper, I am dealing with intended variability, the limits within which it occurs, and what parameters are affected. It can be modelled as a predictable event, and is treated as overlaid on unintended variability. Intended variability gives speech its emotive content.

The speech waveform produced is perceived as meaningful. In the case of a sentence, the meaning can be regarded as primarily a function of the meaning of the individual words within the sentence and the relationship which holds between the words. Moreover, in dialogue a sentence can be considered an *object* in relationship with other sentences. The plain message can be regarded as a baseline, and the relationships among sentences defined by the type of variability associated with the plain message of each sentence.

In this model, overlaid on the baseline meaning of a sentence, there is an additional element which listeners perceive as the mood of the speaker, his/her attitude to what is being said, attitude toward the listener, etc. This added effect, a pragmatic effect, can be described as an *emotive* overlay on what a speaker says.

Thus, there are two types of meaning encoded in the speech waveform: the plain message and the pragmatic effects which give rise to the listener's perception of the speaker's emotion and attitude.

Synthesis systems do not add this meaningful variability in speech to their output. But since it is usually there in spoken language, adding pragmatic effects to synthesis should increase the perception of naturalness.

IV. MODELLING EMOTIVE VARIATION

Emotive variation is modelled as an overlay on a neutrally spoken sentence. Thus emotions or attitudes of 'happiness', 'gloom', 'contrast', 'question', 'emphasis', are overlays on a neutral sentence which itself conveys none of these effects. The parameter file of the neutral sentence is modified to create the perceived effect [4].

In the model presented here those modifications are made to duration and the fundamental frequency contour of the sentence. The neutral values are generated by the phonological and prosodic components of the synthesis system. In the system I use, phonology and prosody refer to a pronouncing dictionary and rely on a simple syntactic parse of the input text and on assignment of abstract Highs and Lows based on Pierrehumbert [6]. For a neutral sentence, the *parameter file* is delivered to the formant synthesizer in the last stage of synthesis. The procedure for adding an emotive overlay intercepts the delivery and transforms the parameter file before it is sent on for synthesis.

Therefore, adding emotion to synthetic speech relies on two inputs:

1. the neutral sentence generated by the synthesis system and

2. the second input associated with a pragmatic procedure, and triggered by a pragmatic marker for emotion, attitude, which can be

- placed on the input text by hand, or
- generated automatically in a concept input system, or
- derived within the pragmatic procedure and applied to a plain text input.

These markers are general, sentence-wide, indicators of the mood with which the sentence is to be spoken. In the model so far built, modifications are applied to the fundamental frequency contour of the sentence and to the duration of voiced segments of nouns and verbs, or adding silence before or after nouns or verbs linked in a dialogue system.

V. PATTERNS EXPRESSING EMOTIVE CONTENT

It was not possible to determine patterns of intended variability from waveforms or spectrograms using normal measuring techniques. There was simply too much variability. F0 and duration changes could be seen, but it was not clear which changes could be correlated with a particular emotive content.

However, this is precisely the kind of association task a neural network is designed to do [5]. A neural network can discriminate between significant and nonsignificant features in an input if it has been trained to associate a range of inputs paired with a particular output. In this case the output was an emotive marker — for example, contrast — and the input was a number

of samples of a sentence judged by a human listener to communicate contrast. Once trained, the network can reversed to produce a typical or generalized contour when given a particular marker as input (Fig.3).

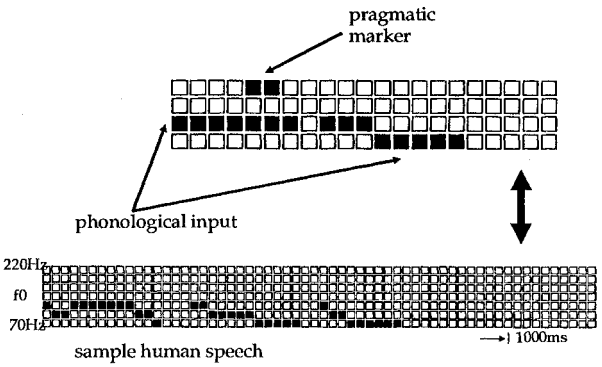

Fig.3 Sample training data for the neural network: pragmatic and phonological inputs associated with human speech — 'Five thirty'.

VI. AN ARTIFICIAL NEURAL NETWORK TO ADD EMOTIVE CONTENT

Using a neural network to identify significant features of the overlay associated with a particular emotive marker is straightforward, but is not suitable for the process of overlaying the *difference* between a neutral sentence and an emotive one.

A second network, a multi-layer perceptron with one hidden layer, was built and trained using a double input [4]. One part of the input consisted of the pragmatic marker appearing on the sentence to be processed, and the second part of the input consisted of the neutral intonation pattern, together with abstract timing information, as generated by the phonology and prosodic procedures within the synthesis system.

For the initial training phase this double input was paired with samples of the fundamental frequency contours and duration derived from multiple samples of human speech. When trained, the network settled on the significant features in human speech extracted from the sample utterances. The network learned to associate these features with the appropriate pragmatic marker.

Once trained the network can be incorporated in the synthesis system, intercepting the output of the phonology and prosodics, and, by reference to the pragmatic marker, generating the required transformed output contour and timing.

One shortcoming of this procedure is that unintended variability, a result of properties of the speaking process, is removed. If unintended variability contributes to perceived naturalness, then that contribution is lost. Variations of pronunciation of a neutral sentence to express emotion have displaced the variations associated with repetitions of the sentence. Listening experiments tend to show that listeners prefer emotive content to variability of sentences lacking such content [4], but I suspect that they would prefer sentences which had both types of variability.

VII. RULE-BASED EMOTIVE PROCEDURE

A rule-based system is being developed as an alternative to a neural network within the synthesis system, processing a suitable pragmatic overlay for emotive expression. One reason for this is that a neural network does not easily reveal how it makes the association between its input and output. In addition to producing good synthetic speech, as researchers we also want to contribute to a descriptive model of speech. Using the results of application of a neural network, data can be gathered, and generalized into a set of rules.

Another reason for wanting explicit rules is that the network could only handle sentences of the same syntactic type as it had been trained on. And it is not realistic to run training sessions on all possible syntactic types at the moment. Therefore, a rule system that would apply across sentence types would be useful.

The data for setting up a rule base was based on the patterns produced by the neural network. For a given phological and pragmatic input the network produced a particular output, expressed in terms of the two parameters, fundamental frequency and duration. The system developed by Pierrehumbert [6] was used to characterize abstract phonological intonation change. Once a neutral mapping is determined, changes corresponding to pragmatic effects are expressed as a percentage change within the range of a neutral utterance (Fig.3).

Syntactic categories are supplied by the parser and identifying labels are carried through the system to the phonetic level. Duration changes on flagged categories can be increased or decreased according to rule.

Sample dialogue:

- 1. *passenger:* What time does the first train leave tomorrow?
- 2. ***computer:*** The early train leaves at 5:30.
- 3. *passenger:* Did you say 9:30?
- 4. ***computer:*** No — 5:30. It leaves at 5:30.

From sentence 1: *time* and *train* are identified as nouns. Sentence 2 picks up the word *train* and states the time: *5:30*. Sentence 3: query on time. Sentence 4 has pragmatic marker on time: changes intonation pattern and duration on time *5:30* (Fig.4).

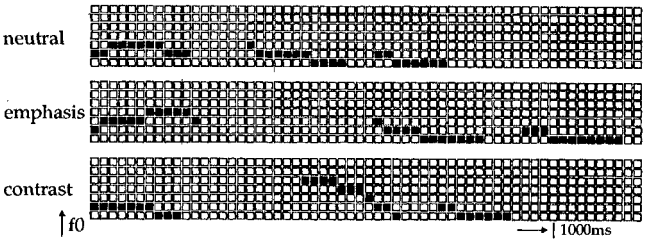

Fig.4 Sample neural network outputs for neutral speech, emphasis and contrast — f0 of 'Five thirty'.

For example: rules such as the following can be derived:

1. If a noun, then insert pause of xms before it.

2. If a noun identified as repeated from the previous utterance, then increase the duration on the vowel of the noun in the second utterance.

3. Using the system [6] of assigning Hs and Ls, increase the range of f0 for the effect of *emphasis* by x% (Fig.5).

Replicating patterns produced by the network produces better results than application of generalized rules. There is a tradeoff between greater generality and accuracy of detail; listener experiments will provide information about the breakover point in this tradeoff.

				H[
the	[1]	[1]	[det]	
early '	[2]	[1]	[adj]	H
train	[1]	[1]	[nou]	H*
leaves	[1]	[1]	[ver]	H
at	[1]	[1]	[pre]	
five	[1]	[1]	[nou]	H**
thirty	[2]	[1]	[nou]	H* L–
				L]

```
Focus assigned: five H**
column:   1 - sentence
          2 - number of syllables
          3 - stressed syllable number
          4 - syntactic category
          5/6 - intonation markers
```

Fig.5 Assigning abstract intonation pattern to a typical sentence following a parse.

VIII. CONCLUSION

Since potential users of synthetic speech are dissatisfied with current speech output from a variety of devices including interactive dialogue systems, an investigation was conducted into modelling what in human speech gives rise to the perception of naturalness. A large part of naturalness is assumed to be that speakers include emotive content in how they speak in addition to meaning conveyed by the syntax and semantics.

The model assumes that naturalness can be characterized in part by changes in the fundamental frequency contour, changes in duration of words and segments within words. The first step is to generate fundamental frequency and rhythm to produce neutral output: the second step consists of developing a procedure to overlay modifications on this neutral output which cause a listener to perceive the emotive content of the speech.

Data concerning fundamental frequency and duration of this aspect of naturalness were gathered using a neural network to extract significant features present in the waveform.

A model of pragmatically derived naturalness has been built based on the output of the neural network and from which generalized rules have been derived for application to similar types of sentence.

REFERENCES

[1] J.N. Holmes. *Speech Synthesis and Recognition.* Wokingham: Van Nostrand, 1988.

[2] M.A.A. Tatham and E. Lewis. "Prosodics in a Syllable-based Text-to-Speech Synthesis System." *ICSLP 92,* this volume, 1992.

[3] K. Morton. "Improving Naturalness in Speech Synthesis Using a Neural Network." *Neural Networks and Their Applications,* Neuro-Nîmes 91, pp.161-173, Nanterre: EC2, 1991.

[4] K. Morton. "Pragmatic Phonetics." *Advances in Speech, Hearing and Language Processing,* ed. W. A. Ainsworth, vol 2 pp. 17-55, London: JAI Press, 1992.

[5] *ANSim.* San Diego: Science Applications International Corporation, 1988.

[6] J.B. Pierrehumbert. "Synthesizing Intonation," *JASA,* vol. 70, pp. 989-995, 1981.