Proceedings of the Institute of Acoustics

METHODOLOGICAL ASPECTS OF THE IMPLEMENTATION OF EMOTIONAL CHARACTERISTICS IN SYNTHESISED SPEECH

E Abadjieva, I R Murray & J L Arnott

The MicroCentre, Dept. of Mathematics and Computer Science,
The University, Dundee, DD1 4HN, Scotland

1. INTRODUCTION

One significant disadvantage of synthesised speech is its monotony and lack of expressiveness to reflect the semantic content of the text and the emotional state of the 'speaker'. When used as a communication prosthesis, speech synthesisers need to be enhanced with additional features to convey more information about the personality, mood and affect of their users. The solution of this problem is restricted by the available control parameters of the synthesiser used and by the limited available knowledge on how the emotional features of speech can be distinguished and described using the formalism of conventional speech signal processing methods. Though the first type of restrictions are only technological and a suitable choice of a synthesiser with a wide range of accessible control parameters forms part of the solution, the problem of how to set up and combine these parameters to add emotional characteristics to the speech remains. These parameters are not orthogonal and a suitable emotion-dependent combination can only be achieved by trial and error [1]. It is only recently that results from research work providing descriptions of the basic emotions in terms of speech technology have been published [1,2]. In order to facilitate the process of enhancing the synthesised speech emotion effects, two main tasks can be defined:

- To verify and augment, by detailed analysis, the parametric descriptions given in the literature, of emotions with particular regard to their future simulation in synthesised speech;
- To deduce a strategy for the progressive implementation of emotional features at different stages of the process of speech synthesis making maximal use of the existing voice design and control parameters of a high quality synthesiser.

2. BASIC STRUCTURE OF A SYSTEM FOR EMOTIONAL SPEECH SYNTHESIS

The process of speech synthesis implemented by a synthesiser consists of two main stages: analysis of the input text string and production of the synthesised speech waveform. Each of these phases is in turn broken down into a cascaded set of functional modules which operate on well-defined input and output data structures. A detailed description of each module is given in [3]. In this paper the structure of the formant synthesiser will be discussed at its highest level with emphasis on the stages where relevant emotional characteristics can be introduced.

Figure 1 illustrates the main functional blocks of such a system. The input text is preprocessed to convert special symbols into text suitable for linguistic analysis, and the phonetic transcription for each word is computed by rules based on a dictionary of morphs. The result is the phonetic notation of the input text with the lexical stress and syntactic information about the type of the sentence and the clause boundaries. The DECtalk synthesiser which was used in the project is based on this model and allows the user to have access to the phonetic string for changes and alterations. The text

EMOTIONAL SPEECH IMPLEMENTATION

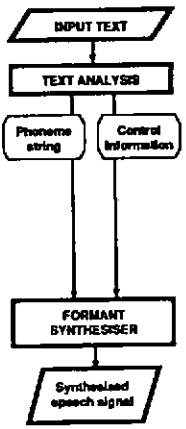

FIGURE 1

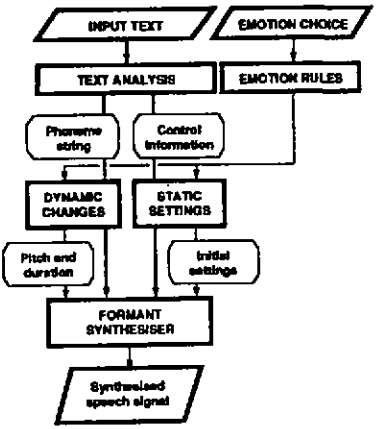

FIGURE 2

analysis block also includes modules for parsing and semantic analysis which provides extra control information based on the prosodic features and the semantic effects on the pitch contour; these parameters are used internally in the speech synthesis phase, but are not accessible to the user and thus can not be used to introduce emotional features. The determination of the fundamental frequency and the time duration of the generated speech waveform segments are the basic modules of the next stage. These two parameters play a significant role in conveying an emotional effect to the synthesised sentence and when the synthesiser is accessed in phoneme mode they can be modified according to the implemented emotion.

The introduced functional changes resulting in emotional speech generation are shown in Figure 2. The chosen emotion activates a set of appropriate rules which produces changes in the emotional feature space. The rules operate on the control parameter in two chronologically distinct phases, one operating on fixed voice parameters, the other operating on dynamically changing features. The first phase consists of the *initial settings* of the voice design parameters to set up the relevant voice characteristics for the chosen emotion. For the DECtalk this is an *initial string of design* voice parameters which is sent out before the text to be synthesised. The next phase includes the changes done and sent together with the phonetic transcription of the text, which affect the speech signal characteristics (i.e. pitch contour and timing) at phoneme, word and clause levels.

3. DESCRIPTION OF THE VOCAL EMOTION EFFECTS FOR THE FIVE BASIC EMOTIONS

In order to introduce emotional changes into the synthesised speech a detailed description of each emotion in technical speech terms is needed. The different aspects of emotion effects in human speech have been discussed in the vocal emotion literature [2,4,5] but it has been almost totally separated from the main body of speech analysis literature. To this end a detailed analysis of recorded emotional speech by a male speaker has been carried out. The two emotionally free sentences: "This is not what I expected" and "You have asked me that question so many times" have been read within

an emotionally different background text which sets up the appropriate emotional state of the speaker. The recordings were made in a recording studio using a high quality microphone and a CD-quality DAT recorder; a PC based speech workstation was used for the analysis of the speech signal. Figure 3 presents some oscillograms and pitch contours obtained for the five basic emotions: anger, happiness, sadness, disgust and fear. The results of the analysis are summarised with emphasis on the features that can be modelled into the synthesised speech. The descriptions of the five basic emotions are grouped in three main categories: voice quality descriptions, pitch contour alterations, time domain changes.

3.1 ANGER

Voice quality:

- increased intensity;
- dynamic changes corresponding to the stressed content words;
- the voice is breathy, with tense articulation;
- greater high frequency energy over all the utterance;
- higher first formant frequencies.

Pitch contour:

- mid general level of the pitch contour;
- large dynamic range;
- the highest pitch values are on the first content word and for the longer sentence on the second last content word. In the second case the pitch rises twice and reaches the same high value;
- the highest part of the pitch contour corresponds to the highest intensity;
- strong downward inflection at the end of the sentence;
- sharp downward inflections at phoneme level and a few upward peaks on the stressed syllables of the content words.

Time characteristics:

- the speech rate is faster than neutral, but not as fast as for happiness.

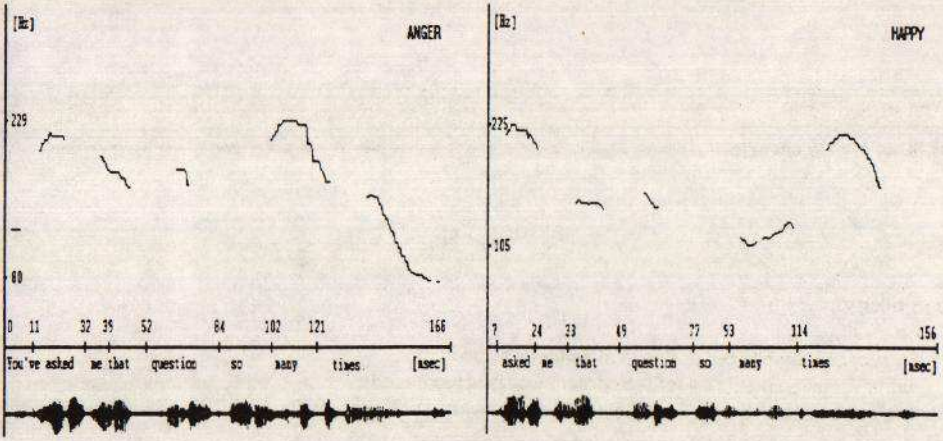

FIGURE 3(a)

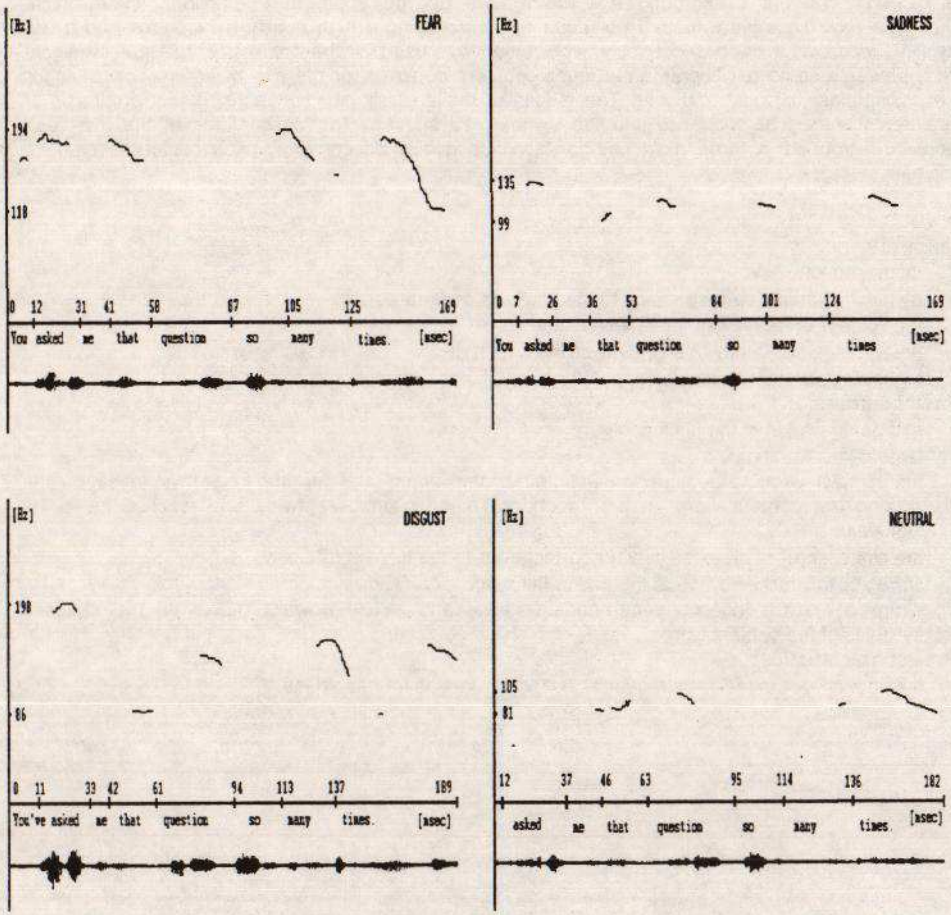

FIGURE 3(b)

3.2 HAPPINESS

Voice quality:

- increased intensity but not as loud as for anger;
- the articulation precision for the content words is increased;
- the voice sounds breathy and light without tension.

Pitch contour:

- the average pitch contour is higher with large dynamic range and the highest peak values (compared to the other emotions);
- the general form of the pitch contour has a second upward part towards the end of the sentence which falls down afterwards;
- the line of the pitch contour is not smooth. It has sharp small oscillations at the primary stressed syllables and local downward pitch changes which seem to be rhythmic.

Time characteristics:

- the fastest speech rate;
- the rhythmic structure introduced corresponds to the stressed phonemes occurring at regular intervals.

3.3 FEAR

Voice quality:

- low intensity with no dynamic changes over the whole utterance;
- vowels and consonants are more precisely articulated;
- disturbed respiratory pattern leading to voice irregularity;
- relatively little energy in the lower frequencies.

Pitch contour:

- the average pitch contour is slightly higher than for neutral speech;
- the dynamic range of the pitch contour is wider but not as wide as for anger and happiness;
- the content words exhibit changes in pitch contour with upward inflections at the beginning of each segment and downward at the end; the downward inflection is well pronounced; the upward parts reach approximately the same peak values;
- at phoneme level the line of the pitch contour has very small fluctuations up and down from the main base line and irregular downward inflections.

Time characteristics:

- the speech rate is slower than for anger and happiness but still faster than for neutral speech;
- the speech is interrupted by pauses between words forming about one quarter of the speaking time.

3.4 SADNESS

Voice quality:

- low intensity with no dynamic changes;
- the articulation precision is decreased.

Pitch contour:

- the average pitch contour is within the same range as for neutral speech;
- the pitch range is narrow, similar to neutral;
- there are almost no dynamic changes (or very slight, smooth over all the utterance) at word level;
- small downward inflections at phoneme level.

Time characteristics:

- slow speech rate;
- rhythm with regular pauses.

3.5 DISGUST

Voice quality:

- low intensity with no dynamic changes;
- increased articulation precision especially with emphasis at stressed content words;
- the voice is not breathy.

Pitch contour:

- the average pitch contour is lower than neutral with very wide downward inflections at the phrase endings;
- noticeable downward pitch inflections at word level matching the words ends;
- for stressed content words the pitch contour goes up at the beginning of the word;
- smooth general line.

Time characteristics:

- slow speech rate;
- large number of introduced pauses;
- increased phonation time;
- lengthening of the stressed syllables well noticeable at the stressed content words.

4. IMPLEMENTATION OF THE EMOTIONAL CHARACTERISTICS IN SYNTHESISED SPEECH

4.1 Static Settings

As was mentioned above, the introduction of emotional features to the synthesised speech can be done in two phases: initial settings of the voice design parameters and dynamic changes integrated with the phoneme string. From the descriptions of the five basic emotions it can be concluded that for each emotion there is a set of parameters that have specific content and remain static within all the analysed utterances of the group; they convey a significant part of the emotional effect. These static features can be implemented by the initial string of the design voice parameters. The description of each emotion in this case includes the following DECtalk parameters:

4.1.1 Voice quality: An appropriate voice quality can be modelled by subtle changes in the head size, laryngealization, brilliance, smoothness and richness of the DECtalk design voice parameters.

The head size is a parameter that strongly changes the quality of the synthesised voice. It affects the voice characteristics in a global way. By changing only the head size parameter and keeping the other parameters constant one can obtain a significantly different voice. The volume of the synthesised speech changes as well and may cause an internal overload in the amplifiers, which has to be corrected by changes in the other gain parameters. As a result a careful use of this parameter can produce a suitable voice quality for the chosen emotion. For instance reducing the default voice head size parameter for Paul's neutral voice to a specific value can produce a metallic, unpleasant, louder voice, which conveys anger.

Smoothness is obtained by a decrease in the voicing energy at higher frequencies opposite to brilliance which results from an increase. Breathiness is the third parameter to be adjusted together with them to obtain suitable voice quality.

The overall intensity can be adjusted using the two groups of gain parameters: resonator gains - connected with the structure of the synthesiser and a group phonetically affecting the articulation which includes: frication gain, aspiration gain, voicing gain and nasal gain. Changes in both groups affect the voice quality as a whole.

4.1.2 Pitch Contour. The general form of the pitch contour is affected by two emotion dependent consistent factors - the average pitch and the pitch range, which are to be included in the initial string.

4.1.3 Time Characteristics. The speech rate parameter sets up the initial tempo according to the implemented emotion.

When choosing the values for the initial string of the voice design parameters it is important to limit the amount of change introduced by the modelled emotion. If the changes are too large, the result may convey a different voice "personality" rather than introducing emotional features. Another property of the initial voice design parameters is that most of them are within a range of relative values. The descriptions of the five basic emotions are also presented in relative terms with references to each other and based on emotionally neutral speech which is relevant to the DECtalk unchanged synthetic speech. A maximal use of the synthesiser's feature space can be achieved if the combinations for the initial settings are designed well apart from each other according to the number of emotions implemented. For instance if the synthesiser has to model only three basic emotions: anger, happiness and sadness, their initial settings may be slightly different from the case when they form part of a set of five or more emotions which are to be distinguished.

4.2 Dynamic Changes

The dynamic changes include the emotional characteristics introduced together with the phonetic transcription of the text to be synthesised.

4.2.1 Voice quality. Once set up according to the emotion, the voice quality does not vary significantly during the utterance. The only parameter that needs to be adjusted in order to augment the emotional effect is the intensity which for ANGER and HAPPINESS increases with the first content word. This can be done by inserting a string with the intensity adjustments at the appropriate place in the phonetic string.

4.2.2 Pitch Contour. The dynamic changes in the pitch contour play an important role in conveying emotional effects. They affect the general form for the whole sentence and the fluctuations at word and phonemic levels (see Figure 3). The pitch contour generated by the synthesiser is first sketched using syntactic information about the sentence type, clause contour, phrase contour and individual word contour. At the lower level it is further augmented by considering the effect of individual segments. Two global level "tunes" are assigned depending upon the sentence type: declarative or yes/no question. The number of global level tunes can be augmented to include emotion - dependent variations. They are further modelled according to the clause and phrase contours by adding initial F0 rise and final F0 fall or continuation rise according to the content words and the phrase type. Emotion dependent adjustments can be introduced at this point too. The individual content words within a phrase have most of the F0 fluctuation. Emphasising the content words helps understanding the utterance because they are less predictable. From the analysis it was noticed that the content words are also used by the speaker to enhance and convey emotional effects. The F0 fluctuations in each word depend upon its rank and the number of its syllables, they also can be made emotion dependent.

The low level of the pitch contour generator is controlled by a set of "prosodic indicators" and reflects the effects of phonemics, lexical stress, and the number of syllables of the words in the utterance. The algorithm first sets the peaks on the lexically stressed syllables. Falls and rises are then assigned around these peaks. Continuation rises are added to the last syllable of most non-

sentence-final phrases, and sentence-final words are given rises or falls depending upon their tune. The F0 contour is finally completed by specifying the amount of fall on other nonstressed syllables. The fact that this algorithm is rule based can be used by changing some of the existing rules and adding some new ones according to the implemented emotional features.

4.2.3 Time Characteristics. The dynamic changes in the time characteristics include: introducing pauses between words and clauses appropriate to the modelled emotion, modelling the increased phonation time in some cases and adjusting the rhythm characteristics by changing the phoneme length. Content words with higher accent number for some emotions, DISGUST for instance, need lengthening to emphasise the emotional effect.

5. CONCLUSION

This paper has summarised the descriptions of the five basic emotions as observed from a detailed analysis of human emotional speech. These findings have been presented in a form suitable for implementation by a synthesiser, and a strategy for their gradual integration using the DECtalk control parameters has been presented.

Further research would obtain a more detailed phonetic description of emotional speech, such as formant structure changes. Improvements in the quality of emotional synthetic speech would also be obtained by a better, low level integration of the emotional changes with the rule-based structure of the synthesiser.

This work was funded by SERC/MoD Research Grant No. GR/F 63862.

6. REFERENCES

[1] J E CAHN, "Generating Expression in Synthesised Speech", Technical Report, Media Laboratory, M.I.T. (1990).
[2] I R MURRAY, "SIMULATING EMOTION IN SYNTHETIC SPEECH", PhD Thesis, University of Dundee (1989).
[3] J ALLEN, M S HUNNICUTT & D KLATT, "FROM TEXT TO SPEECH: THE MITALK SYSTEM", Cambridge University Press (1987).
[4] G FAIRBANKS & W PRONOVOST, "An Experimental Study of the Emotion", Speech Monograph (1939).
[5] I FONAGY, "Emotions, Voice and Music", in J SUNDBERG (Ed), "RESEARCH ASPECTS ON SINGING", Royal Swedish Academy of Music No.33, pp. 51-79 (1981).