EVALUATION OF A SYNTHETIC SPEECH SYSTEM WHICH SIMULATES VOCAL EMOTION BY RULE

I R Murray & J L Arnott

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

ABSTRACT

The aevelopment of a system which can add vocal emotion effects by rule to synthetic speech is described. Perception experiments conducted to evaluate the realism of the simulated emotions are also described. The results of these experiments have indicated that the emotions are generally recognised by naïve listeners, and that the different emotions have different levels of recognition.

The system simulates six discrete emotions (anger, happiness, sadness, fear, disgust and grief) and operates on a microcomputer with a commercial high quality speech synthesiser. The emotion is selected by name, and the appropriate effects are generated by rule and form an integral part of the speech output; input text is unrestricted. Expansion of the system to include a three- dimensional model of emotions, allowing a wide range of emotions of different strengths to be simulated, is described. The proposed use of the system as part of a communication prosthesis for the non-vocal is described, and other applications of synthetic speech with emotion are discussed.

1. INTRODUCTION

Although the facial expression of emotions has been studied by numerous researchers (eg. Schlosberg [1], Ekman and Friesen [2]), the study of vocal emotion has been much more limited (eg. Kramer [3], Davitz [4], van Bezooijen et al. [5]), and the studies more diverse in nature (see Murray [6] for a detailed bibliography of the field). Despite the wide variety of techniques used by researchers, the nature of many of the vocal parameters analysed has been broadly consistent, allowing a patchy but coherent picture of the human voice under various emotion states to be built up. The literature has also indicated that emotion affected three elements of the voice; pitch contour, timing and voice quality. Summaries of some vocal correlates of emotion were given by Murray [6] and Scherer [7].

The intelligibility of some recent commercial speech synthesisers is approaching that of human speech (Greene et al. [8]), and some devices also offer the user easy control over a large number of voice parameters, including both voice quality and prosodic features. This project sought to determine if a commercial synthesiser could be made to produce emotion effects in its speech by systematically varying the three required vocal elements within the voice of the synthesiser.

The goal of the project was to incorporate the emotional voice into communication systems for the nonvocal, such as the CHAT conversation system (Alm et al. [9]). To facilitate simple integration into commercial communication systems, the program was to operate on an IBM-PC (including portable versions) and use a commercially available speech synthesiser. The best synthesiser available for use on the project was the Digital Equipment Corporation's DECtalk V2.0, offering considerable control flexibility and excellent synthetic speech intelligibility.

2. THE DEVELOPMENT OF THE SPEECH-WITH-EMOTION SYSTEM

The synthetic speech-with-emotion system was given the name HAMLET (Helpful Automatic Machine For Language and Emotional Talk), and was designed to add vocal emotion effects to the standard speech output of the DECtalk (Murray et al. [10]). For the prototype version developed in 1987-1989, six discrete emotions were selected for the system: anger, happiness, sadness, fear, disgust and grief. These were chosen because these were the most thoroughly studied in the literature, and hence more information was available about their acoustic features, and also because the first five are the emotions most commonly believed to be "basic" (all other emotions being altered or mixed forms of these five - see Ortony and Turner [11] for a review of basic emotion theory).

The rulebase around which HAMLET was developed comprised a series of synthesiser control rules based on the vocal emotion knowledge in the literature. These comprised fixed-value changes to the various voice parameters of the synthesiser, plus a series of eleven emotion-dependent prosodic rules for altering features of the pitch and timing of the utterance. The emotion effects are added on top of any existing prosodic or voice quality effects in the synthetic speech, normally the emotionally neutral default male voice of the synthesiser; other voices can be selected if required. The DECtalk default neutral intonation pattern is lost when using the device in phoneme mode, and the neutral intonation is recreated by the HAMLET program using rules based on Allen et al. [12].

Normally, the required phrase (unrestricted English text) is typed into HAMLET, and it is converted to phonemes and stored. Any of the six emotions can then be selected from a menu; the corresponding rules then operate on the stored phonemes, and the modified voice quality settings and phonemes are sent to the DECtalk. For a more detailed description of the HAMLET system, see Murray et al. [6 and 10].

3. EVALUATION OF THE EMOTIONAL SPEECH

In order to evaluate the recognisability and realism of the vocal emotions produced by the HAMLET system, a listening experiment was conducted in which naïve listeners were played a series of phrases generated by HAMLET and asked to comment on various aspects of the emotion they perceived in the voice.

As no previous synthetic speech-with-emotion system had been reported in the literature, there was no standard procedure for evaluating such a system, and the procedure used for the current project was based on techniques reported for evaluation of synthetic speech and human speech affect. The three main techniques used were paired opposite adjective scales (eg. Uldall [13], Rosson and Cecala [14]) where subjects indicated emotion ratings on linear scales between two opposite adjectives, free response (eg. Johnson et al. [15]) where subjects could say anything about the stimulus utterance, and forced response (eg. Johnson et al. [15], Fairbanks and Hoaglin [16]) where the subjects were forced to pick from a list of emotions (usually those under examination plus distractors).

3.1 Pilot Experiment

The evaluation was intended to determine how well HAMLET could produce recognisable emotions, and indicate which emotions were most realistic.

The three techniques named above were used in a pilot evaluation experiment; ten paired opposite adjective scales were selected from the literature for the current experiment, and the forced response test offered the six emotions under test plus seven distractors, plus "no emotion" and "other". Three groups of stimulus test phrases were used: those with text appropriate to the vocal emotion, those with semantically neutral text, and those with text not appropriate to the vocal emotion, a total of forty test utterances. The phrase utterances in each section were presented to each subject in a randomised order. The subjects were nineteen university students from various faculties who were not previously associated with the research; they were recruited by sign-up sheets and were paid for their participation.

Subjects were read a short verbal introduction by the experiment supervisor and invited to sit at a computer which conducted the experiment, sequencing the phrases heard by the subject and recording their responses. The subjects were fitted with a pair of headphones to enable them to hear the DECtalk while minimising the amount of background noise and distortion caused by the synthesiser's internal loudspeaker. The subjects were told that it was purely the voice from the synthesiser upon which the judgement was to be made, and not the actual words which were spoken.

After an on-screen introduction, including a demonstration of (normal) DECtalk speech and of the experimental procedure, the paired opposite test was presented to the subject using only the sixteen utterances with emotive text and corresponding vocal emotion; selection of responses on each of the ten paired-opposite scales was performed using a mouse. Following this section, a ten minute rest period was taken, followed by the free response section, in which the subjects typed in their comments about each stimulus phrase, and the forced response section where an emotion adjective selection was made from a list for each stimulus phrase. The entire procedure took about an hour for most subjects.

The results from the pilot experiment indicated that when the text was appropriate for the vocal emotion, identification was very good, but with neutral text it was poorer, and with text inappropriate to the vocal emotion, the vocal emotion was hardly recognised at all. In the latter case, subjects unexpectedly tended not to choose the text emotion either, suggesting a tertiary contradiction effect between the text and the voice giving rise to perception of a third emotion. The subjects found the experimental procedure itself quite straightforward.

The free response test was particularly valuable in locating test phrases which subjects had problems with, although some contradictions in responses between subjects (and even within responses) were noted.

3.2 Main Experiment

Thirty-five subjects took part in the main experiment. The stimulus test phrases were divided into two groups of forty (consisting of four semantically neutral phrases, eighteen semantically emotionless phrases and eighteen emotive phrases), one group having neutral vocal emotion, and the second (using identical texts) having appropriate vocal emotion effects. The eighty test utterances were presented to the subjects in a randomised order. The experimental procedure used for the pilot was repeated, although the paired-opposite test was discarded as it would have taken too much time with the larger number of test phrases, and the rest period was now taken after the free response test and before the forced response test.

3.3 Results

3.3.1 Emotion Recognition. To analyse the results, confusion matrices were drawn up for each of the four subgroups of stimulus phrases. The results from these and the subjects' free responses will now be discussed.

3.3.1.1 Neutral Text with Neutral Vocal Emotion. These were the phrase utterances in which there was no emotion expressed in either the voice or the text. These were, however, not perceived as emotionally neutral, but generally slightly sad, although in the free response section, terms such as "condescending", "disheartened", "frustrated" and "despondent" were used to describe the phrases.

3.3.1.2 Emotive Text with Neutral Vocal Emotion. These were the phrase utterances in which there was a particular emotion expressed in the text, but the phrase was spoken without vocal emotion. Unexpectedly, more subjects attributed "no emotion" to these phrase utterances than to the semantically neutral phrases spoken neutrally, and there was some indication that subjects were inclined to select the textual emotion.

3.3.1.3 Neutral Text with Vocal Emotion. These were the phrase utterances in which there was no emotion expressed in the text, but the phrase was spoken with a particular vocal emotion. Analysis of the confusion matrix for these results indicated that sadness and anger were correctly recognised by a large number of subjects, but the other emotions were poorly recognised.

3.3.1.4 Emotive Text with Vocal Emotion. These were the phrase utterances in which there was a particular emotion expressed in the text, and the phrase was spoken with the corresponding vocal emotion. Analysis of the confusion matrix of these results indicated that more than half of the phrases were identified correctly by more than half of the subjects, anger, sadness and grief being recognised most reliably.

3.3.2 The Effect of Adding Vocal Emotion. By subtracting the confusion matrix for the forty vocally neutral utterances from the corresponding vocally emotional utterances, a difference matrix was obtained; this showed the net effect of the vocal emotion on the phrase. These differences were analysed using McNemar's test (Sprent[17]).

3.3.2.1 The Effect of Adding Vocal Emotion to Phrases with Neutral Text. Of the eighteen phrase pairs, improvements in emotion perception caused by adding vocal emotion which are significant at the 5% level occurred for one of the anger phrases, two of the sadness phrases, and one of the grief phrases. For the other phrases, minor differences in emotion recognition occurred, suggesting that significant changes occurred only for specific phrases.

3.3.2.2 The Effect of Adding Vocal Emotion to Phrases with Emotive Text. Of the eighteen phrase pairs, improvements in emotion perception caused by adding vocal emotion which are significant at the 5% level occur for nine of the phrases, with seven of these also significant at the 1% level. Recognition of all three grief utterances was significantly improved at the 1% level, indicating a high reliance on the context for this emotion. Other emotions showed high variability between phrases as in the neutral text phrases. Ranking recognition of the emotions tested (with emotive texts), anger was the most recognisable, followed by grief, sadness, happiness, fear and disgust, similar to the results of van Bezooijen et al. [5] and Johnson et al. [15] for human speech. The order for neutral text with vocal emotion was similar, except that sadness was the most recognisable.

3.4 Conclusion

For the six emotions studied in the experiment, all were perceived best when used with semantically appropriate text, although anger and sadness were often also recognised out of context. It was found that, in general, adding vocal emotion effects to neutral text did not increase subjects' recognition of the intended vocal emotion (significant improvements occurring for only a fifth of the test phrases), but it was found that adding such effects to phrases with emotive text generally did improve recognition of the intended vocal emotion in these phrases (significant improvements occurring for half of the test phrases). The experiment also showed that the emotions produced were not all equally recognisable, although this effect is also apparent in human emotional speech (van Bezooijen et al. [5]) and is thus not due entirely to differences in realism of the simulated emotion effects.

4. FURTHER WORK

The HAMLET prototype system simulated effects for six discrete emotions. However, emotions can also be regarded as forming a continuum (discrete emotions being at particular points within this continuum) and a number of models have been proposed to model the relationships between emotions, usually with two or three "emotion dimensions" (eg. Schlosberg [1], Davitz [4], and Scherer [18]). The 3-dimensional Schlosberg model appears to have been accepted most widely, and was chosen as the basis of a modified version of the HAMLET system, capable of producing a range of emotions. The three dimensions were labelled as Pleasantness, Attention and Tension (see descriptions op. cit. for details of these dimensions), and in order to select an emotion in the new system , only the P.A.T. co-ordinates are specified.

For the three dimensional version of HAMLET, new rules were written for each of the voice quality parameters (each a function of P, A and T) to replace the fixed values used in the prototype. The eleven prosodic rules were modified so that the magnitude of their effects was also determined by the P, A and T values. The new rules were developed heuristically from the original rules, as the emotion literature indicates that it is unclear how most voice features depend on either the state of the subject's nervous system or relate to the three emotion dimensions (although some parameters such as fundamental frequency, loudness and rate are believed to be related to the activity dimension (eg. Davitz [4])). Work is continuing to improve these rules and relate them more coherently to possible emotion models.

At the same time as the current project, the MIT Media Laboratory has been developing an "Affect Editor" system similar to HAMLET, reported by Cahn [19 and 20], conceived as a "tool for exploring what is needed in an affect generating system". This system uses as input an "annotated utterance" and "implements a transfer function from an acoustical description of emotional speech to synthesized speech" involving "both one-to-many and many-to-one mappings from the acoustic parameters to the synthesizer settings". The Affect Editor used a DECtalk V3, an improved version of the device used for the HAMLET system, and an experiment described briefly [20] has indicated that the system also produces "recognizable affect".

5. APPLICATIONS

The HAMLET system was designed for incorporation into an unlimited vocabulary communication prosthesis for the nonvocal, giving a new capacity to such systems, and consequently offering the nonvocal an improved communication medium. Incorporated into a CHAT-type system (Alm et al. [9]), the HAMLET rules would automatically add to the output speech the effects for the current emotion required by the user. The ability of HAMLET to use any existing voice settings means that if any customised voice has been implemented to suit the personality of the user, this is not lost when the emotion is added.

However, as affect forms a part of all human speech, it's incorporation into synthetic speech could be used to advantage in (theoretically) any situation where synthetic speech is used, verbal warning systems being perhaps the most obvious. It would be possible to include an emotion module into existing text-to-speech systems largely as an addition to the intonation module, as the prosodic effects produced by emotion can be superimposed on the neutral affect prosody of an utterance. Control of the emotion selection within such a system could be by discrete emotion selection or by parameterised selection if a suitable emotion model was incorporated into the system.

6. CONCLUSION

A rule-based system for producing synthetic speech with emotion effects has been produced and evaluated using naïve listeners. The results of this evaluation have indicated that the system is capable of producing recognisable vocal emotions. The system has now been expanded to produce a range of emotions based on a three-dimensional emotion model, and work on this system is continuing.

The majority of this work was funded by the SERC Postgraduate Research Studentship No. 8631679X. Work on the three dimensional version of HAMLET is continuing under SERC Research Grant No. GR/F 63862.

7. REFERENCES

[1] H SCHLOSBERG, "Three Dimensions of Emotion", PSYCHOLOGICAL REVIEW, 61 (2), pp. 81-88 (1954).
[2] P EKMAN & W V FRIESEN, "The Repertoire of Nonverbal Behaviour: Categories, Origins, Usage and Coding", SEMIOTICA, 1, pp. 49-98 (1969).
[3] E KRAMER, "Judgement of Personal Characteristics and Emotions from Nonverbal Properties of Speech", PSYCHOLOGICAL BULLETIN, 60 (4), pp. 408-420 (1963).
[4] J R DAVITZ, "A Review of Research Concerned with Facial and Vocal Expressions of Emotion", in J R DAVITZ (Ed), "THE COMMUNICATION OF EMOTIONAL MEANING", McGraw-Hill, New York, pp. 13-30 (1964).
[5] R VAN BEZOOIJEN, S A OTTO & T A HEENAN, "Recognition of Vocal Expressions of Emotion: A Three-Nation Study to Identify Universal Characteristics", JOURNAL OF CROSS-CULTURAL PSYCHOLOGY, 14(4), pp. 387-406 (1983).
[6] I R MURRAY, "SIMULATING EMOTION IN SYNTHETIC SPEECH", PhD Thesis, University of Dundee (1989).
[7] K R SCHERER, "Vocal Affect Expression: A Review and a Model for Future Research", PSYCHOLOGICAL BULLETIN, 99 (2), pp. 143-165 (1986).
[8] B G GREENE, J S LOGAN & D B PISONI, "Perception of Synthetic Speech Produced Automatically by Rule: Intelligibility of Eight Text-to-Speech Systems", BEHAVIOUR RESEARCH METHODS, INSTRUMENTS, AND COMPUTERS, 18 (2), pp. 100-107 (1986).
[9] N ALM, J L ARNOTT & A F NEWELL, "Discourse Analysis and Pragmatics in the Design of a Conversation Prosthesis", JOURNAL OF MEDICAL ENGINEERING AND TECHNOLOGY, 13(1/2), pp. 10-12 (1990).

[10] I R MURRAY, J L ARNOTT & A F NEWELL, "HAMLET - Simulating Emotion in Synthetic Speech", PROCEEDINGS OF SPEECH '88, THE 7th FASE SYMPOSIUM, Edinburgh, pp. 4:1217-1223 (1988).

[11] A ORTONY & T J TURNER, "What's Basic About Basic Emotions?", PSYCHOLOGICAL REVIEW, 97(3), pp. 315-331 (1990).

[12] J ALLEN, M S HUNNICUTT & D H KLATT, "FROM TEXT TO SPEECH: THE MITALK SYSTEM", Cambridge University Press, Cambridge (1987).

[13] E ULDALL, "Attitudinal Meanings Conveyed by Intonation Contours", LANGUAGE AND SPEECH, 3, pp. 223-234 (1960).

[14] M B ROSSON & A J CECALA, "Designing a Quality Voice: An Analysis of Listeners' Reactions to Synthetic Voices", PROCEEDINGS OF CHI '86, pp. 192-197 (1986).

[15] W F JOHNSON, R N EMDE, K R SCHERER & M D KLINNERT "Recognition of Emotion from Vocal Cues", ARCHIVES OF GENERAL PSYCHIATRY, 43, pp. 280-283 (1986).

[16] G FAIRBANKS & L W HOAGLIN, "An Experimental Study of the Durational Characteristics of the Voice During the Expression of Emotion", SPEECH MONOGRAPH, 8, pp. 85-91 (1941).

[17] P SPRENT, "APPLIED NONPARAMETRIC STATISTICAL METHODS", Chapman and Hall, London (1989).

[18] K R SCHERER, "Nonlinguistic Vocal Indicators of Emotion and Psychopathology", in C E IZARD (Ed), "EMOTIONS IN PERSONALITY AND PSYCHOPATHOLOGY", Plenum Press, New York, pp. 495-529 (1979).

[19] J E CAHN, "From Sad to Glad: Emotional Computer Voices", PROCEEDINGS OF SPEECH TECH '88, New York, pp. 35-36 (1988).

[20] J E CAHN, "Generation of Affect in Synthesized Speech", PROCEEDINGS OF AVIOS '89, MEETING OF THE AMERICAN VOICE INPUT / OUTPUT SOCIETY (1989).