Proceedings of the Institute of Acoustics

MODELLING VOCAL EMOTION EFFECTS IN SYNTHETIC SPEECH TO IMPROVE AUGMENTED COMMUNICATION FOR NON-VOCAL PEOPLE: A REVIEW AND FUTURE OBJECTIVES

Iain R. Murray, John L. Arnott and Elissaveta A. Rohwer

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

1. INTRODUCTION

Synthetic speech systems are now common in a number of applications, but have been most widely accepted as part of communication systems for non-vocal people. In such a situation, the synthesiser is acting as the person's voice, and consequently gives the disabled individual access to a powerful means of communication otherwise denied them. As there is so much information about the speaker, the speaker's state and the speaker's intentions in the speech signal (often this is actually more important than the words spoken) that it is highly desirable to offer this capability to people using synthesised speech as their voice; non-vocal users are often perceived as dull and uncommunicative due to the unexpressive nature of their synthetic speech systems. Thus any improvements in speech synthesis technology will be of great interest and benefit to non-speakers who use such systems.

There are a variety of techniques employed for machine production of speech, ranging from speech recordings (analogue or digital), through numerous coded speech methodologies, to text generated entirely by rule (commonly known as text-to-speech (T-T-S)). Most techniques have been used in the various applications of speech technology, though within a communication system, the T-T-S methodology is most appropriate as the user is not constrained by a set of pre-stored words or phrases, but can speak as freely as required. The output stage of a T-T-S system can use concatenation of pre-stored units (e.g. diphones), or constructive (e.g. formant) synthesis where the speech signal is generated in real time. Most current communication prosthesis systems use stored phrases (selected by the user or via some prediction mechanism) which are output using a text-to-speech system (often an off-the-shelf component), so any improvement to the latter could be of immediate benefit to non-vocal users.

There are three features of synthetic speech systems which can be used (formally or informally) to measure their performance:

Variability: this is the capability of a synthesiser to change the voice with which it speaks. This may be in the form of variable speech rate and similar parameters, but is often extended into voice quality features which allow the "personality" of the voice to be changed. Many synthesisers have a range of "standard" voices, and some also allow the user to edit the personality of the voice (though usually not the accent) to produce a new voice to their own requirements. This process is very important for non-vocal users, who have in the past declined to use synthetic speech systems because it does not sound like them; as voice conveys so much about one's character and personality, they would rather do without a voice prosthesis at all. Even in the case where the synthesiser's voice is user alterable, the process is often quite complicated, though attempts have been made to simplify and speed up this process [1].

Intelligibility: The most rigorously applied of synthetic speech measurements, intelligibility essentially measures how well the synthesiser can produce different speech sounds in a recognisable way. Many synthetic speech systems can score very highly in intelligibility tests, often scoring almost as well as natural speech. Intelligibility ratings are also affected by the synthesiser's lexical stress capability, as

incorrect stress placement makes words harder to identify for the listener.

Naturalness: This parameter is not rigorously defined, but is intended to measure how human the synthesiser sounds; thus a highly natural voice may be often mistaken for a human voice, while an unnatural voice is clearly machine-generated. The parameter is also often used to describe how pleasing or how "easy to listen to" the synthetic speech is. Although today's synthesisers are often highly intelligible, they are very often not very natural (though many are substantially better than the Daleks and other famous robots), and one goal of speech synthesis research has been to produce a synthetic voice which is highly natural. This is certainly a desirable feature for a communication prosthesis system, though may not be so desirable for other applications (e.g. dialogue systems) where it may be appropriate to remind the user that the speaker is a machine, due to differences between human-human and human-machine dialogue [2]. There is also some evidence [3] that factors affecting intelligibility correlate negatively with those affecting naturalness.

2. IMPROVING NATURALNESS

There are three ways in which to improve the naturalness of synthetic speech:

Voice quality: Improvements in the speech reproduction or synthesis process can lead to more a natural frequency distribution within the speech signal, leading to more natural-sounding speech. One major contributor to this process in constructive synthesis has been found to be simulation of the voice source itself, and improving the voicing model has been found to greatly enhance the output speech. Voice quality as related to naturalness is thus generally a feature of the synthesiser being used, and can generally not be improved by the user.

Speaking style: Research (summarised in [4]) has shown that humans speak in different ways depending on a number of factors related to their speaking environment, such as the type of material being read, intelligibility projection, the audience and the speaker's social standing relative to the audience. These changes are in timing, pitch contour, and stress placement (at both word and utterance level).

Emotion and mood: Numerous internal factors within the speaker, commonly referred to as emotion and mood (for longer duration phenomena) can also lead to changes in speech produced (summarised in [5]). There are also other physical factors (i.e. which are not commonly thought of as emotions) which can lead to changes in the pragmatics of speech, such as tiredness, ill health (such as a blocked nose caused by a cold), or other pathological speech disorders. Unlike speaking style, emotion and related pragmatic factors are not normally under the conscious control of the speaker, though they can contribute at least as much to the speech signal (a speaker can, of course, take conscious control of some features for deliberate affect). These features also alter the pitch and timing of the spoken utterance, though they do not generally affect stress placement, and can be considered as additional to any stress-related effects such as inflections and vowel lengthenings.

Thus to achieve a truly natural-sounding synthetic voice, it must have a good underlying voice quality, speak with features appropriate to the style of dialogue and speaking context, and have the capability to express emotion and other pragmatics effects within the speech. The principal problem for researchers endeavouring to improve the naturalness of synthetic speech is the limited knowledge of the effect of speaking styles and emotion upon the speech signal. Some fragmented studies have been carried out in these areas, but it is only now that a direct application of this knowledge exists that any sustained research is being performed in these areas. It is apparent, however, that both style and emotion cause

changes in the same ways within the speech signal, so for the purposes of speech synthesis, it would be appropriate for both to be considered together during the synthesis process.

3. EMOTION KNOWLEDGE

Our knowledge of the ways in which emotion effects our bodies is very limited indeed [6], and the ways in which emotion leads to outwardly perceptible changes is even less well known, speech being even less well researched than facial expression. This is despite our constant experience of emotions during our lives, and the ubiquitous use of emotion-related terminology; however, these are by their very nature subjective, and hence often contradictory. Despite this limited knowledge, several prototype synthetic speech-with-emotion systems have been developed and demonstrated (e.g. the HAMLET system [7], the Affect Editor [8], [9]) and interest in this area of research is increasing as the number of potential applications grows.

Emotion modeling research has been conducted in two main directions, namely attempting to model the internal patterns of the emotional process (from stimulus through physiological changes to outwardly perceivable changes in the subject) and modeling inter-relationships between emotions themselves. Scherer [10] has been very active in the former area, and proposes a theory of Stimulus Evaluation Checks (SECs) which define a hierarchy of effects leading to a subject's response to a stimulus, depending on various factors.

Models of the inter-relationships between emotions are of particular interest for application to speech synthesis, and have been divided between two theories. "Basic" emotion theories define a series of discrete emotions as variations and combinations of a closed set of basic emotions, and dimensional theories define emotions as an area within some form of dimensional space. These theories are not clearly defined by psychologists, and both have a number of variations; the "basic" emotion set varies in size between two and eighteen ([11]), and the dimensional theory exists in two-, three- and even four-dimensional form. However, it is generally the three-dimensional form (originally proposed by [12]) which has received most support, and discrete emotions can of course be considered as points within a dimensional model. Both types of model appear to have some validity, though the dimensional type are perhaps of greater potential application.

It is known that some physiological changes correlate with parameters of speech ([6, 10]), and these also correlate with some of the proposed emotional dimensions (e.g. the "tension" dimension in the Schlosberg model correlates strongly with both speech rate and heart rate).

4. USING A 3-DIMENSIONAL MODEL TO CONTROL SYNTHETIC SPEECH

The prototype HAMLET system [7] simulated six discrete emotions, these being the five most commonly accepted "basic" emotions (anger, happiness, sadness, fear and disgust) plus grief. Rules within the system selected pre-defined voice quality parameters and introduced preset changes to the pitch contour and timing of the utterance being spoken, depending on the emotion selected by the user.

To enhance the prototype system, two approaches were taken. The first was to perform voice analysis on emotional human speech and use the knowledge gained to improve the existing HAMLET rules; a summary of this process is given in [14]. The second approach was to extend the range of emotions simulated by HAMLET by incorporating an emotion model. To convert the system to accept three-dimensional co-ordinate input, the pitch contour and timing rules were rewritten to utilise the co-ordinates selected for activation, and also to have some degree of variability of the effect produced

depending on the co-ordinates selected. To obtain variations in voice quality, the use of preset parameters was superseded by a series of rules (one for each of the voice quality parameters affected), the output of the rules again depending on the selected co-ordinates. For some parameters, the correlation between the emotion dimensions and the magnitude of the parameter was known from the literature; for others, the rule has developed heuristically. For user input within the enhanced HAMLET system, a display of the three emotion co-ordinates was produced, with cursor keys used to alter the co-ordinate values. The dimensional model was able to reproduce a range of emotions by varying the three co-ordinates (including the previous six discrete emotions by appropriate selection of the co-ordinates).

5. USER CONSIDERATIONS FOR A COMMUNICATION PROSTHESIS

Addition of emotion capability to a communication prosthesis potentially increases the load on the user, as an emotion is required for each spoken output, as well as selection of the output text itself. However, by relying on the fact that emotion tends to be relatively constant over a short period of time, and tends to return to a "neutral" (i.e. non-emotional) state over a longer period, this load increase could be made almost negligible by appropriate design of the system. Whilst one or more explicit user inputs are always required to select and output a phrase from a typical communication system, once an emotion was initially selected, this selection could be maintained throughout following utterances without further user input being required. If an emotion model such as that described above was incorporated into the system, the emotion co-ordinates could be programmed to tend back towards neutral over a number of utterances in order to mimic the natural calming effect.

An additional possibility for controlling the emotion produced by such a system, which is made possible within a communication prosthesis system because it has a user whose voice the system is providing, is the unconscious input of emotion information directly from the user. By using suitable sensors to monitor physiological parameters such as heart rate, blood pressure, and skin resistance, and correlating these to the speech parameters (possibly via the dimensional model), the speech could be altered appropriately without any explicit user input at all. Work by the current authors is continuing to investigate possible ways of acquiring useful physiological data for input to the speech-with-emotion system. It remains to be seen if users will object to being attached to a direct-input system, as many physiological changes (particularly those not related to emotion) may produce undesired alterations in the emotional voice, and also the user may at times wish to hide his true feelings while communicating. At any rate, such direct input systems and their sensors would have to be easily portable and extremely discrete in order to be acceptable, even though many disabled individuals are already used to relying on technology systems for movement and communication.

The dimensional version of HAMLET has been tested informally within a communication prosthesis which predicts conversational phrases for the user [15] with promising results, although the emotion selection process was simplified for the user in this prototype.

6. THE FUTURE

Earlier (e.g. [16]) and current ([14, 17]) emotion research using voice analysis techniques has attempted to correlate different emotional states directly with specific changes within speech. The "generative theory of affect" proposed by Cahn [8] attempts to describe emotional changes (and other features of speech, such as intonation) directly from the mental and physical emotional state of the speaker, and such a model would be a valuable tool for automating the addition of emotion effects to synthetic speech.

MODELLING VOCAL EMOTION EFFECTS

To fully describe the emotion process, we must be aware of user stimuli, the way these stimuli affect the user's physiology, and thence affect his vocal apparatus and speech (see Figure 1). However, the exact way in which physiology is affected by external stimuli is not known, nor is the exact way in which this in turn affects the vocal tract. Conversion of vocal tract dynamics at this level to speech sounds can only be achieved using articulatory synthesis, and while this technique has been in existence for a long time [18], it is comparatively little used in speech research at present due to its computational complexity compared to other methods of speech synthesis. While this model seems an attractive one upon which to base a speech system, particularly a vocal prosthesis where a user is available, there are great practical difficulties in its implementation, and it is unlikely to be realised within a decade. In particular, our knowledge of the acoustic variables related to details of phonology, speaking style, emotion and other related factors and the way they influence perception of speech is still very poorly understood.

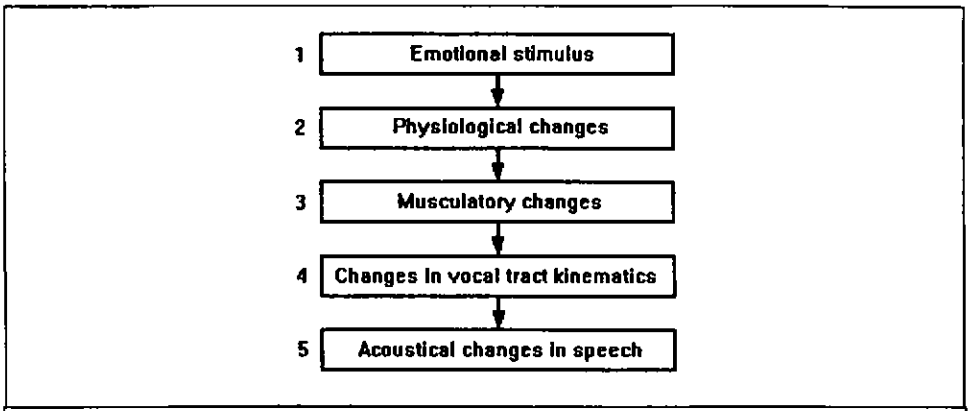

FIGURE 1 - The Processes Leading to Emotive Speech

A further benefit to future synthetic speech-with-emotion systems would be from a recognised strategy for the testing of such systems. There has been wide acceptance of the need for a standard testing strategy for speech recognition systems, and also for measuring the intelligibility of speech synthesis systems, but there is not as yet a standard strategy for such features as emotion and the broader naturalness. As in speech recognition, future developments and comparison of results of synthetic speech systems would be more easily measured if there were a common test procedure covering the various aspects of the speech signal (intelligibility, naturalness, etc.).

7. CONCLUSION

Several prototype systems for producing emotion-by-rule in synthetic speech have been developed, and their application to vocal prosthesis systems for non-vocal people, and for improving the naturalness of synthetic speech for other applications is apparent. It appears that we are now on the threshold of producing natural-sounding synthetic speech systems, but our knowledge of emotions in general, and their affects on speech in particular, are neither clearly known nor well defined, and thus progress is likely to be slow beyond the type of speech systems already seen. Models for the emotion processes

and inter-relations between emotions would be particularly useful in this area to guide experimentation, but the field is so complex that such models would have to be very much simplified for the foreseeable future.

8. ACKNOWLEDGMENTS

This work was funded by SERC/MoD Research Grant No. GR/F 63862. Donation of equipment from the Digital Equipment Corporation is also gratefully acknowledged.

9. REFERENCES

[1] I R MURRAY & J L ARNOTT, "A tool for the rapid development of new synthetic voice personalities", *Proc. ESCA ETRW on Speech and Language Technology for Disabled Persons*, Stockholm, Sweden, pp. 111-14 (1993).
[2] R K MOORE & S R BROWNING, "Results of an exercise to collect 'genuine' spoken enquiries using WOZ techniques", *Proc. Institute of Acoustics*, **14(6)**, pp. 613-620 (1992).
[3] C K COWLEY & D M JONES, "Assessing the quality of synthetic speech", in C BABER & J M NOYES (Eds), *INTERACTIVE SPEECH TECHNOLOGY*, Taylor & Francis, London (1993).
[4] M ESKÉNAZI, "Trends in Speaking Styles Research", *Proc. Eurospeech '93*, Berlin, Germany, pp. 501-509 (1993).
[5] I R MURRAY & J L ARNOTT, "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *J. Acoustical Society of America*, **93(2)**, pp. 1097-1108 (1993).
[6] J R DAVITZ, *THE COMMUNICATION OF EMOTIONAL MEANING*, MacGraw-Hill, New York (1964).
[7] I R MURRAY, *SIMULATING EMOTION IN SYNTHETIC SPEECH*, PhD thesis, University of Dundee (1989).
[8] J E CAHN, *GENERATING EXPRESSION IN SYNTHESISED SPEECH*, MIT Media Laboratory Technical Report (1990).
[9] K MORTON, "Naturalness in synthetic speech", *Proc. Institute of Acoustics*, **12(10)**, pp. 125-132 (1990).
[10] K R SCHERER, D R LADD & K E A SILVERMAN, "Vocal affect expression - a review and a model for future research", *Psychological Bulletin*, **99(2)**, pp. 143-165 (1986).
[11] A ORTONY & T J TURNER, "What's basic about basic emotions?", *Psychological Review*, **97(3)**, pp. 315-331 (1990).
[12] H SCHLOSBERG, "Three dimensions of emotion", *Psychological Review*, **61(2)**, pp. 81-88 (1954).
[13] P EKMAN, R W LEVINSON & W V FRIESEN, "Autonomic nervous system activity distinguishes among emotions", *Science*, **221**, pp. 1208-1210 (1983).
[14] E ABADJIEVA, I R MURRAY & J L ARNOTT, "Applying analysis of human emotional speech to enhance synthetic speech", *Proc. Eurospeech '93*, Berlin, Germany, pp. 909-912 (1993).
[15] I R MURRAY, J L ARNOTT, J.L., N ALM & A F NEWELL, "A communication system for the disabled with emotional synthetic speech produced by rule", *Proc. Eurospeech '91*, Genova, Italy, pp. 311-314 (1991).
[16] C E WILLIAMS & K N STEVENS, "Emotions and speech: some acoustic correlates", *J. Acoustical Society of America*, **52(4(2))**, pp. 1238-1250 (1972).
[17] J VROOMEN, R COLLIER & S MOZZICONACCI, "Duration and intonation in emotional speech", *Proc. Eurospeech '93*, Berlin, Germany, pp. 577-580 (1993).
[18] C H COKER, "A model of articulatory dynamics and control", *Proc. IEEE*, **64**, pp. 452-460 (1976).