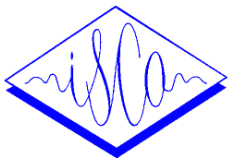

ESCA-NATO Workshop on
Speech under Stress
Lisbon, Portugal
September 14-15, 1995

The Strains of Emotional Stress in Synthetic Speech

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

The MicroCentre, Dept. of Mathematics & Computer Science,
The University, Dundee DD1 4HN, U.K.

ABSTRACT

Text-to-speech and other speech output technologies are becoming increasingly widespread, especially as input technologies improve and facilitate applications with both speech input and output. Although speech output systems now generally have very high intelligibility, most are still easily identified as artificial voices and no commercial systems yet allow prosodic variation due to emotion and related factors. This is largely due to the complexity of incorporating such naturalness factors, and our very limited knowledge of what voice changes actually occur due to the speaker's emotion. However, prosodic content in synthetic speech is seen as increasingly important as interactive computer systems become more common, and there is presently renewed interest in the investigation of human vocal emotion and the expansion of synthesis models to allow greater prosodic variation.

This paper will review progress to date in the investigation of human vocal emotions and their simulation in synthetic speech, and requirements for future research which is required to develop this area will be presented.

1. INTRODUCTION

Emotion, mood, personality and other pragmatic information about the state of the speaker are present in *every* spoken utterance. However, ***none*** of these effects are available in commercial text-to-speech systems. This is partly due to the limited complexity of such systems to incorporate a range of pragmatics effects, but mostly due to our very scant knowledge of how physiological emotion contributes to acoustical changes in speech.

2. WHY IS EMOTION SO DIFFICULT TO DEAL WITH?

The first problem encountered in any study of emotion is one of terminology. We are all constantly experiencing emotion and a vast vocabulary of emotion-related terms is available, but as emotion and its various forms are so intangible and, like taste or colour, different to every individual, these terms are open to totally subjective interpretation. Thus despite often common usage, few of these terms have common meaning and even fewer have any formal definition, making it very difficult to produce any rigorous descriptions of emotion and its associated features, and this has led to a wide range of emotions being described and analysed by a wide range of means. A study of emotion terminology is described in [1].

Our knowledge of the ways in which stimuli lead to emotion changes within our bodies is exceptionally limited [2], and the ways in which the somatic emotion is carried through to outwardly perceptible changes is even less well known; in this respect, speech is less well researched than expression via the face [3] and posture.

Another problem of marrying emotion effects to speech technology (both for input and output) is the distinct nature of the two research disciplines. Speech scientists have almost entirely avoided the emotion area (though there are exceptions, e.g. [4]); for speech recognition, emotion information is redundant, and for speech synthesis it has been of lower importance than intelligibility. Vocal emotion studies have also tended to be isolated in nature, whereas speech technology has been in development by major research groups over long periods of time.

We need:

- formal definitions of emotion terminology and their corresponding effects;
- greater integration between psychologists and speech scientists.

3. WHAT DO WE KNOW ABOUT EMOTION IN SPEECH?

From the human vocal emotion analyses reported in the literature, it is known that emotion causes changes in three groups of speech parameters:

- **voice quality:** these effects define the "character" of the voice, and typically include hoarseness, whispering, croakiness and similar effects.
- **pitch contour:** the intonation of an utterance, especially the range of pitch and the nature of pitch inflections, contributes greatly to the emotion content.
- **timing:** the overall speed of an utterance, together

with changes in duration of certain parts of the speech, also conveys some emotion affect.

It should be noted that these parameters are also used to convey other features during normal (that is, unemotional) speech; pitch and timing changes, for example, are used to indicate stressed words in an utterance. Any emotion-related changes to these parameters must therefore be considered *in addition* to the underlying changes present.

It is also worth considering a number of other pragmatic features alongside emotions (some occasionally are considered *as* emotions), as they affect the same sorts of voice parameters in often similar ways to emotions. Such features would include "pseudo-emotions" such as tiredness, non-emotive bodily changes such as having a cold, and the more conscious-level effects of "speaking style" which a speaker uses to impart particular intention into their speech.

Thus any useful model of emotions would be able to separately define all of these features, and combine them correctly into one speech signal.

The various emotion analyses reported in the literature have rarely investigated the same set of emotions, and differing techniques have been used. New studies using state-of-the-art equipment and a clearly defined study procedure are required; these may be facilitated by the increasing availability of speech libraries now becoming available (though emotion analysis has not been a deliberate goal of these recordings).

We need:

- detailed descriptions of how different emotions affect voice quality, pitch contour and timing of utterances;
- new emotion analyses using modern equipment and rigorous analysis procedures;
- cross-cultural analyses to determine the applicability of the results.

4. MODELLING EMOTION

Emotion modelling research has been conducted in two main directions:

- modelling the internal patterns of the emotional process (from stimulus through physiological changes to outwardly perceivable changes in the subject), and;
- modelling inter-relationships between emotions, to define which emotions are similar to others, and determine sub-groups or families of emotions.

4.1 How Do We Model Emotion Processes?

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. However, the exact way in which physiology is affected by external stimuli is not well understood, nor is the exact way in which this in turn affects the vocal tract. Stimulus Evaluation Checks (SECs) are one model (proposed in [3]), using a hierarchy of effects leading from a stimulus through physiological changes to the subject's response, depending on various internal and external factors, such as the magnitude of the stimulus and the perceived threat from it.

Older (reviewed in [5]) and more recent (e.g. [6, 7]) 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. Such a model would be a vital part of any unified theory of speech [9].

Conversion of vocal tract dynamics at this level to speech sounds can only be achieved using articulatory synthesis, and while this technique is not new, 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, there are great practical difficulties in its implementation. 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. However, this is potentially a very important area, as a good understanding of the emotion processes, and particularly how they affect the speech apparatus and thus the audible speech could be used as a convenient way to further develop an articulatory speech model to produce emotional speech; the system would be given a series of emotional stimuli and the emotional content of the speech would change accordingly. The entire process is enormously complex, and a simplified model is presented in Figure 1.

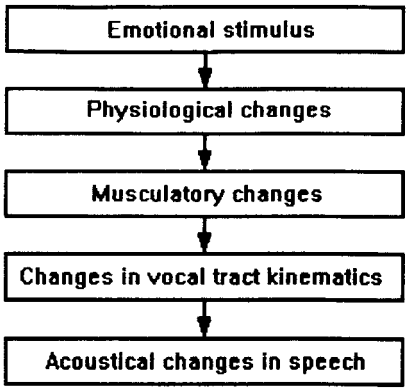

Figure 1 - How emotion causes changes in speech

4.2 How Do We Model Inter-Relations Between Emotions?

Models of the inter-relationships between emotions have been divided between two main theories - "basic" (or "palette") emotion theories define a closed set of basic discrete emotions and define all others as variations and combinations of these, while "dimensional" theories define emotions as points 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 (reviewed by [10]) varies in size between two and eighteen, and the dimensional theory exists in two-, three- and even four-dimensional form. However, it is generally the three-dimensional form (originally proposed by [11]) which has received most support (and the theories are not mutually exclusive - discrete emotions can 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 application, as they offer the potential to easily create a wide range of emotions from a simple set of input parameters.

It is known that some physiological changes correlate with parameters of speech ([2, 3]), and these also correlate with some of the proposed emotional dimensions (e.g. the "tension" dimension in the Schlosberg model [11] correlates strongly with both speech rate and heart rate).

We need:

- accepted models for both emotion processes and the inter-relationships between emotions;
- acoustic and other correlates to the emotion models.

5. HOW DO WE MODEL EMOTION FEATURES IN SYNTHETIC SPEECH SYSTEMS?

Despite our limited knowledge, several prototype synthetic speech-with-emotion systems have been developed and demonstrated. The HAMLET system [12] uses rules to alter voice, pitch and timing in speech via a commercial synthesiser, and is similar in concept to the Affect Editor system [8]. The SPRUCE text-to-speech system also includes a capability for including pragmatic effects [13].

While articulatory synthesis in theory gives wholly definable synthetic speech at a computational price, text-to-speech systems using diphone or formant synthesis are potentially the most useful form of synthesiser, though more complex than systems using recorded or coded speech. However, coded speech can have certain parameters manipulated, and emotion effects can be added in this way. Whilst commercial systems have not exploited this potential, manipulation of emotion parameters in coded speech has been used in many experiments investigating emotion (e.g. [4, 7]).

Interest in this area of research is increasing as the number of potential applications grows.

6. HOW DO WE MEASURE THE EMOTION CONVEYED?

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

- **intelligibility:** The most rigorously applied of synthetic speech measurements usually implemented by some form of rhyme test, intelligibility measures how well a 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.
- **variability:** this is the capability of a synthesiser to change the characteristics of the voice with which it speaks, from simple alterations of the speech rate to more involved voice quality alterations which allow the "personality" of the voice to be changed.
- **naturalness:** This parameter is not rigorously defined, but is intended to measure "how human" a 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, and one goal of speech synthesis research has been to produce a synthetic voice which is highly natural. There is also some evidence [14] that factors affecting intelligibility correlate negatively with those affecting naturalness.

Synthetic speech-with-emotion systems require a recognised testing strategy which includes measurement of all of these features. 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 using rhyme tests and similar techniques, 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 all aspects of the speech signal (intelligibility, naturalness, etc.).

We need:

- a formal testing procedure for speech synthesis systems including emotion and pragmatic effects;
- this procedure to be applied to existing and future systems.

7. WHAT DO WE NEED TO DO TO IMPROVE REALISM?

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.
- **speaking style:** Research (summarised in [15]) 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 or mood, and other pragmatic features, can also lead to changes in speech produced (summarised in [5]).

Thus, to ultimately 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, emotion and other pragmatics upon the speech signal.

We need:

- improved voice production models;
- better understanding of the various pragmatics components, and the way they are combined in natural speech.

8. CONCLUSION

With increasing demand for speech technology systems, there is increasing need for text-to-speech systems which sound natural with emotion and other pragmatic effects. Some research systems have been produced, but development is hindered by our poor knowledge of emotion processes and their correlates in speech. Co-ordinated research to address the various issues within this field is required to ensure continuing progress.

9. REFERENCES

[1] C. Storm & T. Storm, "A taxonomic study of the vocabulary of emotions", *J. Personality and Social Psychology*, **53(4)**, pp. 805-816, 1987.

[2] J.R. Davitz, *THE COMMUNICATION OF EMOTIONAL MEANING*, MacGraw-Hill, New York, 1964.

[3] K.R. Scherer, "Vocal affect expression - a review and a model for future research", *Psychological Bulletin*, **99(2)**, pp. 143-165, 1986.

[4] K.R. Scherer, D.R. Ladd & K.E.A. Silverman, "Vocal cues to speaker affect: testing two models", *J. Acoust. Soc. of America*, **76(5)**, pp. 1346-1356, 1984.

[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. Acoust. Soc. of America*, **93(2)**, pp. 1097-1108, 1993.

[6] 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.

[7] J. Vroomen, R. Collier & S. Mozziconacci, "Duration and intonation in emotional speech", *Proc. Eurospeech '93*, Berlin, Germany, pp. 577-580, 1993.

[8] J.E. Cahn, *GENERATING EXPRESSION IN SYNTHESISED SPEECH*, MIT Media Laboratory Technical Report, 1990.

[9] R.K. Moore, "Speech pattern processing: from 'blue sky' ideas to a unified theory?", *Proc. Inst. Acoustics*, **16(5)**, pp. 1-13, 1994.

[10] A. Ortony & T.J. Turner, "What's basic about basic emotions?", *Psychological Review*, **97(3)**, pp. 315-331, 1990.

[11] H. Schlosberg, "Three dimensions of emotion", *Psychological Review*, **61(2)**, pp. 81-8, 1954.

[12] I.R. Murray & J.L. Arnott, "Implementation and testing of a systems for producing emotion-by-rule in synthetic speech", *Speech Communication*, **16**, pp. 369-390, 1995.

[13] M.A.A. Tatham & E. Lewis, "Prosodic assignment in SPRUCE text-to-speech synthesis", *Proc. Inst. Acoustics*, **14(6)**, pp. 447-454, 1992.

[14] 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.

[15] M. Eskénazi, "Trends in speaking styles research", *Proc. Eurospeech '93*, Berlin, Germany, pp. 501-509, 1993.