SYNTHESIZING EMOTIONS IN SPEECH: IS IT TIME TO GET EXCITED?

Iain R. Murray and John L. Arnott

The MicroCentre, Applied Computer Studies Division,
University of Dundee, Dundee DD1 4HN, U.K.

ABSTRACT

Modern speech synthesis systems with very high intelligibility are readily available in a number of languages. However, the output from all present systems is still readily identifiable as being machine-generated - the output does not sound "natural". One aspect of naturalness is the variability introduced by the emotional state of the speaker, and related pragmatic effects; no current commercial systems include such variation. Comparatively little work has been done to investigate how a speaker's emotional state creates variation in the speech signal, and this work has traditionally been performed by psychologists and has remained distinct from mainstream speech science. Current research suggests that there will be considerable effort involved in producing any accurate description of pragmatic variations in speech, but there has recently been increasing interest in this area due to potential applications in many branches of speech technology.

This paper describes a prototype system which has been constructed to simulate emotion in speech synthesized by rule. The system is based on emotion information from the literature, and it simulates a range of emotions using a commercial synthesiser. The use of emotion models and their applicability in the area of speech technology is discussed.

The limitations of our current knowledge in the area of vocal emotion are discussed, and suggestions are presented for future research in this area.

1. VARIABILITY IN SPEECH

Speech is the principal mode of communication between humans, both for transfer of information and for social interaction. Consequently, learning the mechanisms of speech have been of interest to scientific research, leading to a wealth of knowledge about the production of human speech, and thence to technological system to simulate and to recognize speech electronically.

One problem encountered in dealing with speech which has proven to be a major challenge to researchers is that of variability. Different speakers say things in different ways at both a verbal and vocal level; there is also considerable variability within the speech of a single speaker. A speaker will not use the same words to say the same thing twice (either consciously or unconsciously), and even different instances of the same word will not be acoustically identical. There are a number of reasons for this variability:

Speaking style: Speakers alter their way of speaking in response to a number of conditions related to their environment and their status relative to those to whom they are speaking. Such attitudinal conditions include consciously increasing intelligibility (a speaker will alter their speech for a non-native listener, or due to increased background noise), familiarity (a speaker will speak more carefully to a listener with whom they are not familiar) and social status (a speaker will speak to a child differently from the way they would speak to a peer, and would speak in a different way again to a listener in a socially dominant position relative to the speaker). A discussion of the various factors related to speaking styles is presented in [1].

Emotion and mood: Different emotional states will affect the speech production mechanism of a speaker in different ways, and lead to acoustical changes in their speech; these changes can be perceived as being due to emotion by listeners. Generally, *emotion* refers to short-term states, with *mood* being longer-term, and *personality* can be regarded as the underlying state of an individual, although the terms overlap somewhat; mood and emotion are occasionally used synonymously.

Stress: A number of other factors relating to physiological arousal also contribute to changes in speech, and these are commonly labeled as *stress* (emotion is often included under this heading also). Such factors include fatigue, illness, and the effects of drugs and workload. Physical stress due to vibration or acceleration may also produce acoustic changes in speech due to direct action on the vocal apparatus itself. A taxonomy of stressors is given in [2], which also discusses the difficulties involved in dealing with stress in speech-based systems.

All three of these factors are essentially independent within the speaker, but all are present to a greater or lesser extent in all speech. The changes they produce within the speech signal are of a similar nature, and thus they are often considered together by researchers; indeed a listener might not be able to correctly attribute a specific change in speech to one of these factors. For the most part, the speech variabilities produced are generated unconsciously, and even where a speaking style is adopted consciously by a speaker, the actual vocal changes are made at an unconscious level. It is thus hard to quantify the changes which occur, and to produce a robust description of how they are produced. Studies in this area have traditionally been in the

domain of psychologists, and there is very little overlap between psychology and the speech sciences.

Most speech scientists have been interested in dealing with "normal speech", that is speech which does not display any of these variabilities, which have traditionally been viewed as unnecessary complications by speech scientists. Successful work on analysis, synthesis and recognition of speech has been achieved under this constraint, but the results break down when natural variability is present. The result is that present synthesis systems do not exhibit these variabilities, producing bland "neutral" speech, and the performance of speech recognition and verification systems falls dramatically when there is variability in the incoming speech signal. Variability, then, is present in all natural human speech, and thus if we wish to simulate natural sounding speech, we need to incorporate variability in some way.

2. NATURALNESS ASPECTS OF SYNTHETIC SPEECH

Modern speech synthesis systems are highly intelligible - that is, the words spoken can be clearly understood. However, most synthetic speech is readily identified as being "machine generated" - it sounds *unnatural*. High intelligibility has been the prime consideration of developers, and consequently an unnatural neutral voice output has been acceptable. However, for broad acceptability of systems which incorporate speech output, a more natural voice is required, and there is thus increasing research interest in improving the naturalness of the synthetic speech.

Current speech systems generate speech in one of two ways: using recorded speech segments which are resynthesized in some way, or using rule-based speech generation. As the speech is generated entirely by the computer in the latter systems, the speech is often of low quality compared to recorded speech. Despite this, rule-based synthesis is more attractive for many applications, since the vocabulary is effectively unlimited as the rules can cope with any new words which are encountered; a number of packages based on this principle are now available, either as plug-in devices or increasingly as software packages which run on standard computer hardware without the need for additional specialized devices. Most importantly, as the speech is generated entirely by rules, there is the greatest scope for the inclusion of variability in the speech, assuming appropriate rules can be derived.

There are three main contributors to the naturalness of synthetic speech: intonation , voice quality and variability.

2.1 Intonation

One of the main factors in intelligibility of speech has been identified as the intonation contour of the utterance, and in particular the placement of word- and utterance-level accents; hence this is also of major importance for naturalness, as an incorrect intonation contour immediately suggests an unnatural voice. Much work has been done on the description of intonation contours, and rules produced for assigning contours to synthetic speech based on parsing their verbal content. While current systems are very good in this respect, the limitations of the word parsing and intonation rules mean that no system can correctly assign the correct contour for *every* possible utterance.

2.2 Voice quality

The underlying "personality" of the voice is also a major contributor to naturalness. Systems based on recorded speech perform well in this respect, as the voice quality from the speaker comes through in the resynthesized speech (depending on the coding used, this may be high fidelity, or the quality may be reduced somewhat). In rule-based synthesis, the voice quality is generated synthetically, and so does not have all the vocal "brilliance" of human speech; however, improving synthesis techniques (especially in the accurate synthesis of the glottal pulse) mean that the performance of these systems is increasing.

2.3 Variability

As discussed above, there are various variability elements in speech, but these are not included in current synthetic speech systems - all speech is "neutral". However, if we wish synthetic speech to be natural, we must include some variability into the speech output.

Information about speaking style is largely at a level above the speech itself, and this data would have to be available to the synthesis system; even then, although some acoustic correlates of speaking style are known [1], it is not wholly clear how different speaking situations correlate to acoustic changes. Similarly, changes due to stress are very poorly defined, and several stress factors are often present simultaneously. Vocal emotion factors, while still not rigidly defined, have perhaps been examined more than speaking style and stress, and offer the first opportunity to add variability to synthetic speech.

3. ADDING EMOTION TO SYNTHETIC SPEECH

Vocal emotion research has generally been separate from mainstream speech research, and has been characterized by a small number of isolated studies rather than lengthy research effort. However, a number of results have been presented regarding the acoustic correlates of vocal emotion; these are summarized in [3]. These studies indicate that emotion affects the pitch, timing and voice quality of utterances.

Based on this literature and a study of actor-generated emotions [4], a system was developed by the current authors with the aim of demonstrating vocal emotion within synthetic speech generated by rule - HAMLET (the Helpful Automatic Machine for Language and Emotional Talk). It was decided to use a commercially available speech synthesiser, and the DECtalk device was selected; this has since become the *de facto* industry standard for rule-based speech synthesis.

The HAMLET system is based around a series of rules which systematically alter the voice of the synthesiser in ways

appropriate to the emotion being simulated. The parameters controlled by the system were the underlying voice quality of the synthesiser used (at the utterance level), together with the pitch and timing of individual phonemes within the utterance to maintain detailed control over the intonation contour. These alterations are applied above the basic intonation contour of the utterance (to retain accent information). The prototype HAMLET system simulated six emotions (anger, happiness, sadness, anger, fear, disgust and grief) - these were selected explicitly by the user in the prototype, and the emotion effects applied to the input of unrestricted text. The output of HAMLET was tested in a perception experiment with naïve listeners, and it was found that the vocal emotion produced by the system was identifiable in most utterances (with statistically significant recognition for half of the test utterances). A detailed description of the prototype HAMLET system and the perception experiment is given in [5]. Enhancements subsequently made to the system [4] included the addition of a 3-D emotion model (based on [6]) which allowed selection of emotion via three co-ordinates; the rules were modified to alter the various synthetic speech parameters based on these co-ordinates. Use of the emotion model allowed simulation of a range of emotions, including combinations and different depths of expression. Samples of the HAMLET test utterances used are given in the Appendix.

Other research synthesis systems have been developed which include a capability for emotion and similar variability. The Affect Editor [7] uses a DECtalk synthesiser controlled by rules to process manually tagged input text, and the SPRUCE system [8] includes pragmatic variability within its synthesis model. Parameters of coded speech systems can also be manipulated to simulate emotion affect, but this has generally been done for the investigation of emotion rather than for synthesis purposes [9].

4. THE WAY FORWARD

Current efforts to explore the fields of stress and emotion have been plagued by a lack of commonality of reference [2]. This is due to the ephemeral nature of these concepts which denies easy description which in turn prevents robust definition of related terms. Language translation problems add a further facet to the problem, despite the largely cross-cultural nature of emotion. Thus any progress in this field would be aided by the production of a reference framework which could add structure and definition to a topic based on transient and largely subjective data.

Another contributing factor to the poorly defined terminology is the various angles from which the field of vocal emotion has been approached. However, it is increasingly recognized that this area of research is by its nature interdisciplinary [10], and continued attempts to involve all relevant disciplines in the formulation of a research and assessment framework will be of practical value.

Several models for the inter-relationships between emotions are available (e.g. [6]); models for the processes of emotional response within an organism are also available (e.g. [11]). However, these models are currently somewhat simplistic given the complexity of what they are trying to represent, and these need to be improved and expanded in light of our improving knowledge. We do not even have a clear picture of how particular stimulus events correlate to specific acoustical changes in speech (either for an individual or for the general population); one possible option to deal with this is suggested in [2], but this is a very complex problem.

The various published studies of vocal emotion have generally independently selected the emotions for study and analysis techniques, making for a somewhat fragmented literature. Any framework thus also needs to include guidelines on how to conduct vocal analyses, and how to present results in a broadly understandable format. This also applies to experiments including synthetic speech output, where a common test strategy covering all aspects of the voice is required; current intelligibility ratings are widely accepted, but do not include any measurement of variability or naturalness factors including emotion.

One area of research which somewhat surprisingly has not so far been united under the banner of vocal emotion is that of music. It is self-evident that musical patterns produce emotional responses in listeners. In speech, intonational patterns convey part of the emotive content of the speech, and it seems attractive to suggest that the patterns in speech and music are in some way related. Some brief studies have investigated this (e.g. [12]), and although no direct link between the pitch of speech and music has been found, this seems to be a potentially rich avenue of exploration. A review of emotion and music studies is presented in [13].

5. CONCLUSION

The first steps in synthesizing speech which includes variability in the form of emotional effects have been taken, and there is an increasing interest in further improving the naturalness of synthesized speech. However, in order to make further progress in this area, a robust descriptive framework for speech, which includes variability factors is required. Production of this framework will require an interdisciplinary approach; greater co-operation between disciplines related to speech science will be of great benefit to the research communities, and improvement in our basic knowledge of variability, its causes and effects, will be of benefit in many other speech disciplines and technologies as well as in speech synthesis.

6. REFERENCES

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

2. Murray, I.R., Baber, C. and South, A. "Towards a definition and working model of stress and its effects on speech", paper in preparation.

3. Murray, I.R. and Arnott, J.L. "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.

4. Abadjieva, E., Murray, I.R. and Arnott, J.L. "Applying analysis of human emotional speech to enhance synthetic speech", *Proc. Eurospeech '93*, Berlin, Germany, pp. 909-912, 1993.

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

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

7. Cahn, J.E. "The generation of affect in synthesized speech", *J. American Voice Input/Output Society*, 8, pp. 1-19, 1990.

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

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

10. *Proc. ESCA/NATO Workshop on Speech under Stress*, Lisbon, Portugal, 1995.

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

12. Fonàgy, I. and Magdics, K. "Emotional Patterns in Intonation and Music" Z. Phonetik Sprachwissenschaft Und Kommunicationsforschung, 16, pp. 293-326, 1963.

13. Scherer, K.R. "Expression of emotion in voice and music", *J. Voice*, 9(3), pp. 235-248, 1995.

APPENDIX - HAMLET Sound Samples

Two sets of HAMLET demonstration utterances are included.

1 - "This is not what I expected"

Unemotional	[SOUND A998S01.WAV]
Anger	[SOUND A998S02.WAV]
Happiness	[SOUND A998S03.WAV]
Sadness	[SOUND A998S04.WAV]
Fear	[SOUND A998S05.WAV]
Disgust	[SOUND A998S06.WAV]

2 - "You have asked me that question so many times"

Unemotional	[SOUND A998S07.WAV]
Anger	[SOUND A998S08.WAV]
Happiness	[SOUND A998S09.WAV]
Sadness	[SOUND A998S10.WAV]
Fear	[SOUND A998S11.WAV]
Disgust	[SOUND A998S12.WAV]