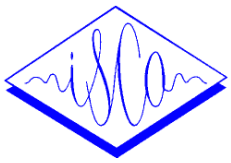

Speech and Language Technology
for Disabled Persons
Stockholm, Sweden
May 31 - June 2, 1993

ENHANCING A COMMUNICATION PROSTHESIS WITH VOCAL EMOTION EFFECTS

John L. Arnott, Norman Alm and Iain R. Murray
The MicroCentre, Dept. of Mathematics and Computer Science,
The University, Dundee DD1 4HN, Scotland, U.K.

ABSTRACT

Although computer-based vocal prostheses for non-speaking people are becoming increasingly common, they are very often slow in use, and they cannot convey the feelings of the user beyond the actual words spoken. This paper describes the integration of two systems designed to overcome these disadvantages of current systems.

INTRODUCTION

Non-speaking people often communicate using vocal prostheses incorporating synthesised speech. Many non-speakers are also motor impaired in some way, and the vocal prostheses often include some form of prediction system to help the user produce their utterances. Despite this, communication rate is often very low (2-10 words per minute) even with modern computer-based systems (Darragh and Witten (1992)). In addition, although the speech produced by these systems is often of high intelligibility, it does not convey any of the feelings or emotion of the user beyond the words spoken.

Two complementary systems for non-speaking disabled people have been developed at Dundee University MicroCentre. CHAT is a vocal prosthesis system for rapid prediction of appropriate conversational utterances for a non-speaking user, and HAMLET is a synthesiser control system for adding emotional effects into synthesised speech. This paper describes some of the problems to be overcome in combining the two systems into an integrated prosthesis, offering the user both verbal (textual) and nonverbal (pragmatic) control over their synthetic voice for the first time.

CHAT - SPEEDING UP COMMUNICATION

CHAT is a speech act prediction system designed to guide a non-speaking motor-impaired user through a simple "meeting in the corridor" conversation, and some simple discussion situations (Alm, Arnott and Newell (1992)). The system moves through a series of conversational stages from "greetings" to "farewells", and a user can take part in a simple conversation by repeatedly pressing the "Say It" button to have an utterance of the predicted type spoken by the system. CHAT's main advantage is that whole phrases appropriate to the stage in the conversation can be spoken very quickly, allowing the user to communicate at a faster rate - around 50 words per minute (Alm, Arnott and Murray (1992)). CHAT allows simple discussion to take place, and also acts as a framework for modules handling more detailed topic discussion, which is a complex task being addressed by other work (Alm, Arnott and Newell (1989), Waller, Broumley, Newell and Alm (1991), McKinlay(1991)).

The CHAT system uses text phrases prestored by the user; a set of ten phrases are stored for each speech act to allow some variation in the *actual* phrase spoken each time. The phrases can be displayed on the user's screen, a second screen positioned for the person talking to the user to read comfortably, or more commonly, the text can be sent to a text-to-speech synthesiser for audible output. To give the user the capability to express some emotion, one of four moods (polite, informal, humorous, angry) can be selected; different phrase sets are then used in each case.

HAMLET - ADDING VOCAL EMOTION

HAMLET is a rule-based system designed to add vocal emotion effects into the speech produced by a synthesiser (currently a DECtalk™); these rules were based on the literature on human vocal emotion research (reviewed by Murray and Arnott (1993a)). HAMLET alters the DECtalk's voice quality as well as the duration and pitch of the individual phonemes in the utterance to produce the required emotion effects in the output speech (Abadjieva, Murray and Arnott (1993)). Consequently, the system must convert the input text into a phonemic representation prior to processing; this was done using the DECtalk's own text-to-phoneme function. HAMLET operates on the currently set DECtalk voice, and so can use any of the synthesiser's built-in voices, or any others designed to suit the user (Murray and Arnott (1993b)).

INTEGRATING THE SYSTEMS

CHAT and HAMLET were both written in Pascal, and as both were written with eventual integration in mind, the process was accomplished with little difficulty. CHAT was modified to send the selected text phrase to HAMLET rather than directly to the synthesiser; the required emotion was also sent. HAMLET then converted the phrase to phonemes, added the appropriate vocal emotion effects according to its rules, and sent the modified utterance to the synthesiser.

When testing the integrated system, it was found that DECtalk's text-to-phoneme conversion process introduced a delay between the user's keystroke and the output of the emotive phrase. Although this delay was only a few seconds, this was unacceptable within CHAT where speed of response is a vital requirement. Consequently, ways of increasing the speed of the integrated system were explored.

The first solution was to store the CHAT phrases in their phonemic representation rather than as text. Although practical, this required considerable manual input during the conversion; conversion of large numbers of phrases (and of new phrases added to CHAT) would not be practical, and only a few phrases were converted for test purposes. The speed of the system was improved, but it was felt that the problems of generating and storing phrases as phonemes eliminated this as a long term solution.

The second solution was to circumvent the phoneme conversion delay in the DECtalk by including text-to-phoneme and lexical stress rules within HAMLET itself; these rules were based on previously published rules (Elovitz, Johnson, McHugh and Shore (1976) and Allen, Hunnicutt and Klatt (1987)) with some additions, modifications and anglicisations. Although the operation of the rules appeared to be sufficiently accurate

and fast on normal utterances, it was found that they did not work well with the phrases used in CHAT as many were short phrases composed of short words which the rules did not stress correctly. They would, however, work satisfactorily with the longer topic-related phrases which would be needed by the user during extended discussion with a topic management module (op. cit.).

A third option was to implement only the voice quality changes required by HAMLET, leaving the prosodic effects unaltered. The control signals corresponding to the required emotions were sent to the DECtalk by CHAT, followed by the utterance text itself. This meant that the slow text-to-phoneme stage could be avoided, while retaining some of the vocal emotion effects.

In consequence from these trials, the integrated CHAT/HAMLET system currently uses the compromise solution of a core vocabulary of phrases stored as phonemes in order to maintain the speed of the system with the HAMLET voice quality effects only. This limits the ease of altering the CHAT phrases, however, and a better long term solution will be to use the rule-based modules once their performance has been improved.

USER CONTROL OF THE EMOTIONS

A further problem to be overcome was how to give the user control over the vocal emotion. Initially this was done automatically, with the CHAT anger mood (i.e. angry text phrases) selecting the HAMLET angry voice (i.e. angry vocal effects), humorous mood selecting the happy voice, and so on. However, it was desirable to have independent control over the textual mood and vocal emotion, so a control was added to CHAT to cycle through the six available vocal emotions (in the same way as the four textual moods are selected). This meant that the user could select numerous cross-matched combinations of textual and vocal emotion to convey different affect. The HAMLET emotions can also be selected using an input of a three-dimensional vector, giving a wide range of possible vocal emotions; this system may also be incorporated into CHAT if an efficient selection mechanism suitable for a disabled user can be found.

CONCLUSION

The integration of a phrase prediction system and a synthetic speech-with-emotion system to produce a prosthesis system for non-speakers capable of conveying vocal and textual emotion has been described. Work is continuing to find the most appropriate solutions to the problems encountered during the integration process. The system can then be evaluated with disabled users.

ACKNOWLEDGEMENT

This work was carried out under SERC/MoD Grant number GR/F 63862, and the authors also gratefully acknowledge the donation of researcher support and equipment by the Digital Equipment Corporation.

REFERENCES

Abadjieva, E., Murray, I. R. & Arnott, J. L. (1993), "An Enhanced Development System for Emotional Speech Synthesis for use in Vocal Prostheses", *Proceedings of the 2nd European Conference on the Advancement of Rehabilitation Technology*, Stockholm, Sweden, 26-28 May 1993.

Allen, J., Hunnicutt, M. S. and Klatt, D. H. (1987), *"From Text to Speech: The MITALK System"*, Cambridge University Press, Cambridge.

Alm, N., Arnott., J. L., Murray, I. R. (1992), "Bypassing communicational difficulties to allow satisfying conversational participation by a non-speaking person", *Proceedings of the Institute of Acoustics*, **14(6)**, pp. 637-644.

Alm, N., Arnott., J. L., Newell, A. F. (1989), "Database design for storing and accessing personal conversation material", *Proceedings of the 12th Annual Conference of the Rehabilitation Engineers Society of North America*, New Orleans, LA, USA, pp. 147-148.

Alm, N., Arnott., J. L., Newell, A. F. (1992), "Prediction and conversational momentum in an augmentative communication system", *Communications of the ACM*, **35(5)**, pp. 46-57.

Darragh, J., Witten, I. (1992), *"The Reactive Keyboard"*, Cambridge University Press, Cambridge.

Elovitz, H. S., Johnson, R., McHugh, A., Shore, J. E. (1976), "Letter-to-sound rules for automatic translation of English text to phonetics", *IEEE Transactions on Acoustics Speech and Signal Processing*, **ASSP-24(6)**, pp. 446-459.

McKinlay, A. (1991), "Using a social approach in the development of a communication aid to achieve perceived communicative competence", *Proceedings of the 14th Annual Conference of the Rehabilitation Engineers Society of North America*, Kansas City, MO, USA, pp. 204-206.

Murray, I. R. and Arnott, J. L. (1993a), "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *Journal of the Acoustical Society of America*, **93(2)**, pp. 1097-1108.

Murray, I. R. and Arnott, J. L. (1993b), "A tool for the rapid development of new synthetic voice personalities", *this volume.*

Waller, A., Broumley, L., Newell, A. F. and Alm, N. (1991), "Predictive retrieval of conversational narratives in an augmentative communication system", *Proceedings of the 14th Annual Conference of the Rehabilitation Engineers Society of North America*, Kansas City, MO, USA, pp. 107-108.