A Speech Synthesis System with Emotion for Assisting Communication

Akemi Iida, [*1, *2] *Nick Campbell,* [*2, *3] *Soichiro Iga,* [*4] *Fumito Higuchi,* [*5] *Michiaki Yasumura*[*5]

[*1] Keio Research Institute at SFC, Keio University [*2] CREST, JST (Japan Science and Technology),
[*3] ATR Information Sciences Division [*4] Ricoh Co. Ltd.
[*5] Grad. School of Media & Governance, Keio University

E-mail: akeiida@sfc.keio.ac.jp nick@slt.atr.co.jp

1. ABSTRACT

The authors have proposed a method of generating synthetic speech with emotion by creating three copora of emotional speech for use with CHATR [1][2], the concatenative speech synthesis system developed at ATR [3][4]. The corpora express joy, anger and sadness. For the previous trial, speech corpora were made with female voice. Having added speech corpora of male voice, the authors developed a prototype of a speech synthesis system with emotion, "CHATAKO," which works on a notebook PC in Japanese language environment. The ultimate goal of this system is to assist people with communication problems. In this paper, first, design principles are stated, followed by system configuration and the evaluation on the synthetic speech with emotion generated by our method.

2. CHATAKO'S DESIGN PRINCIPLES

The purpose of the implementation of this prototype system is to evaluate its validity as a communication assistive system. The target users are mainly people with communication problems, such as patients suffering ALS (Amyotropic Lateral Sclerosis) who have lost the ability to speak through tracheotomy operations.

Although the graphical user interface (GUI) of this prototype is made as simple as possible, Each indivisual needs to prepare additional input device, which compensates his or her disabilities in order to use this prototype.

For our research, 'joy,' 'anger' and 'sadness' were selected as a first trial from the four emotion types (joy, anger, sadness and surprise) appeared on the top layer of the hierarchical cluster analysis of emotion by Shaver [5].

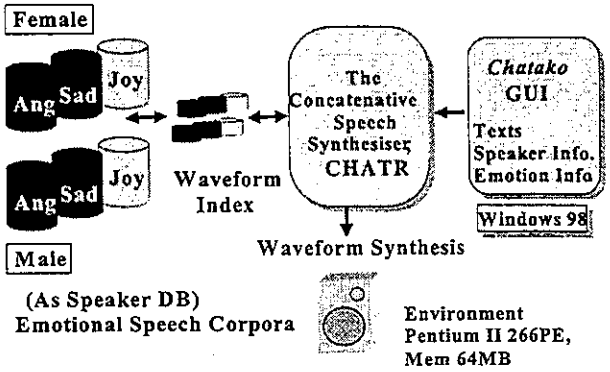

Figure 1: The system configuration

3. PROTOTYPE SYSTEM, CHATAKO

In this section, CHATAKO's system configuration and components are explained.

3.1. System Configuration

This system runs under Microsoft Windows 95/98 and consists of 1) speech synthesis system with emotion and 2) the GUI. CHATAKO's GUI offers text input windows, speaker selective icons and emotion selective icons along with command buttons. Figure 1 shows the system configuration of CHATAKO.

3.2. Speech Synthesis System with Emotion

CHATR for Microsoft Windows developed by ATR is used as a speech synthesis system. Six speech corpora (angry, joy and sadness with a male and a female voice) are equipped for the use for CHATR.

CHATR, the Concatenative Speech Synthesiser

Being a re-sequencing speech synthesiser, CHATR produces an index for random-access waveform sequences of an externally stored corpus to select target units to create new utterances [3][4]. Figure 2 shows the speech synthesis method of CHATR.

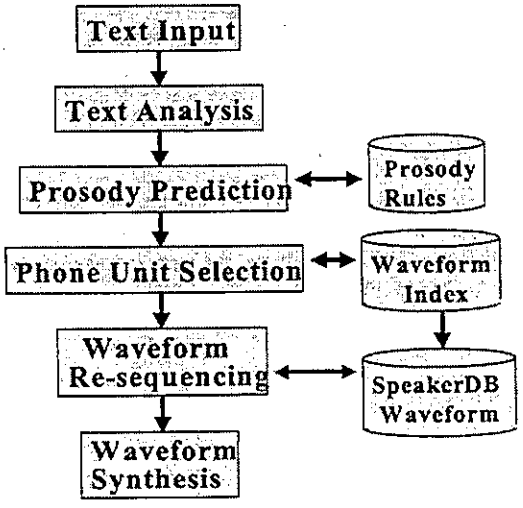

Figure 2: Speech synthesis method of CHATR

Since CHATR selects phone units from naturally spoken human speech, the quality of the synthetic speech it generates is one of the closest to human voice.

Designing Corpora of Emotional Speech

First, a text corpus for each emotion was created. The main text-level requirement was to be able to induce natural emotion in the speaker when read. Monologue texts were chosen so that a particular emotion could be sustained for a relatively long period of time. From newspapers, the WWW and self-published autobiographies of disabled people, essays and columns expressing joy, anger and sadness were collected. Some expressions typical to each emotion were inserted in appropriate places in order to enhance the expression of each target emotion [2]. Table 1 shows the size of each text corpus and Figure 3 shows an example from joy corpus.

	Texts	Sentences	Moras	Phonemes
Joy	12	461	21676	40916
Anger	15	495	21085	39171
Sadness	9	426	16189	31840

Table 1: Sizes of text corpora

Joy Text No. 5
Mattaku teashi no ugokanai watashi nimo jibun de yareru kotoga dekita no desu. "Sugoizo, sugoizo! Oi, konna koto mo dekiruzo! Miteroyo, iika? Hora, mou ichido yatte mirukarana! Iya, gokigendayo, kore!"
(Even I, whose body is completely paralysed, could do it. "It's great! Just great! Hey, I can do things like this, too! Look at me, are you ready? See, I'll do it again. Oh, it's absolutely fantastic!")

Figure 3: An example text from joy corpus

When texts were collected, the phonetic balance of each corpus was not considered due to giving priority to the naturalness of the texts. Instead, more texts were gathered than previously created corpora for CHATR for other purposes. Each of our corpora, however, reached equivalent balance with others. The average number of biphone of our corpus is 357 and that of nine ATR corpora is 354. (The number of biphones for the ATR phonetically balanced text corpus is 403.)

Non-professionals were selected for both female and male speakers in order to avoid exaggerated expression. All texts were read in a sound treated room and the speech was digitised at a 16kHz, 16bit sampling rate. Hence, total of six corpora (joy, anger, and sadness for both female and male) were created.

Acoustic Characteristics of the Corpora

For each corpus, means and standard deviations (SD) of fundamental frequency (f0), power (RMS), and duration were measured per phone. Means comparisons by ANOVA showed that f0 means of three emotions were significantly different from one another at a significance level of $p < 0.05$ for both male and female voice. Table 2 shows f0 means for each corpus. This result supports the reports from previous studies [6][7].

	Joy	Anger	Sadness
Male	149.8	154.2	124.1
Female	256.6	262.4	242.9

Table 2: f0 means for each corpus (kHz)

On the contrary, there was no common tendency in mean duration or mean RMS between the two speakers. The following is means for duration for each in order. Male: sadness < joy < anger, Female: anger < joy < sadness. Means for RMS: Male: anger < joy < sadness, Female: anger = joy = sadness.

3.3. Graphical Interface for CHATAKO

The GUI was implemented in Tcl/Tk 8.1 (a script programming language). In this section, system window layout and each component is explained [8].

CHATAKO's Window Layout

The window layout of CHATAKO is shown in Figure 4. The text window is located in the centre, speaker icons (male and female) and emotion icon (joy, anger and sadness) is located to the right of the text window. Beneath the text windows are command buttons for users to select the operation and synthesis method of his or her preference. In this prototype, commands implemented are 'open,' 'whole synthesis,' 'part synthesis' 'save speech,' 'save sentence,' 'window clear,' and 'exit,' from left to right in order. In this way, all the selection can be seen explicitly.

Speaker-Emotion Selection and Text Input

The user should first select speaker and then emotion. When user select the speaker by clicking either a male or a female icon, then emotion icons appear beneath the speaker icon area for the user to select. After selecting both items, the user types in the text into the text window. When the user wishes to

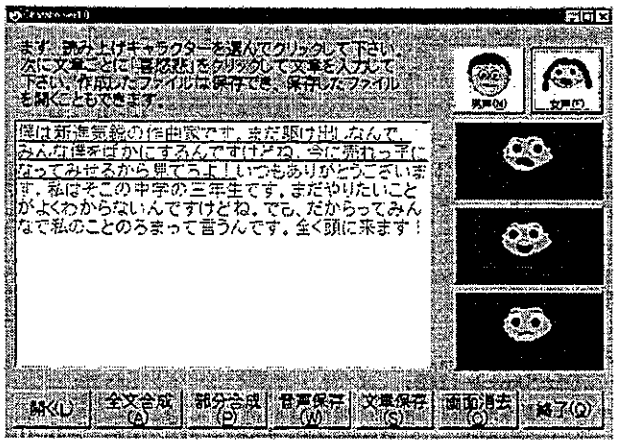

Figure 4: The graphical user interface of CHATAKO

change speaker or emotion types while typing, he or she can do so by re-selecting the speaker and emotion icons and restart typing.

Tagging and Storing the Speaker-Emotion Information

Following the above procedure, the user automatically generates tags which maps speaker-emotion information to texts so that the synthetic speech will be generated with phone units selected from a corresponding corpus when synthesise command is executed. Speaker-emotion information can be stored with the text when 'save text' command is elected. Figure 5 shows the example of the saves tagged texts.

TAGMale:happy boku wa shinnshinn kiei no sakkyokuka desu. TAGMale:sad mada kakedashi nandesukedone TAGFemale:angry mattaku attamani kimasu.

Figure 5: An example of the stored tagged texts

Displaying the Speaker-Emotion Information

When the male speaker is selected, the texts are underlined, while plain text is used for the female speaker. Emotion types are distinguished by font colours: Red for anger, yellowishgreen for joy and blue for sadness.

Changing the Speaker-Emotion Information

Speaker-emotion information once tagged to texts can be changed by partially marking the portion of the text the user wishes to change and reselect the speaker and emotion.

Generating and Storing the Synthetic Speech with Emotion

Three types of speech generation are offered:

- Synthesis of the entire message ('whole synthesis').
- Synthesis of a part of the message ('part synthesis).
- Synthesis a line at a time (triggered by carriage return)

The synthetic speech can be saved with 'save speech' command button as a 16kHz, 16bit wav-format file.

4. PERCEPTUAL EVALUATIONS

Perceptual experiments were conducted to evaluate the quality of synthetic speech with emotion generated with our method. All speech samples used were Japanese and all listeners were Japanese natives.

4.1. Identification of Emotion Types

The purpose of this experiment is to verify whether listeners could correctly recognise the system user's intended emotion from the synthetic speech. Since the date of completion for male and female corpora differs, the experiment for each speaker's corpora was conducted separately.

For each speaker corpora, subjects were 18 university students. Semantically neutral five texts were chosen for the speech sample. An example sentence is shown in Figure 6. Texts were then synthesised by CHATR with corpora of emotional speech and total of 15 speech samples were generated (5 sentences * joy, anger and sad corpus). Speech samples were then stored in a notebook PC (Toshiba Dynabook SS3300) as 16kHz, 16bit wav-format files. Each subject was asked to put on a headset (SONY MDR-NC20) and then was given a forced-choice selection of joy, anger and sadness for each sample.

Chataa wa iroiro na koe de shaberu kotono dekiru atarashii onsei gosei no shisutemu desu. (CHATR is a new speech synthesiser that can speak in various voices).

Figure 6: An example text for synthesis

Previously reported emotion identifying rates for female speech were joy: 51%, anger: 60% and sadness: 82% [2]. For male speech, joy: 52%, anger: 51% and sadness: 74%. The identifying rate for female speech is shown in Figure 7 and for male speech, Figure 8. It can be concluded that emotion types have been correctly recognised at a significance level of $p < 0.01$.

The above results show that listeners can identify the system user's intended emotion from the synthetic speech generated with both male and female corpora.

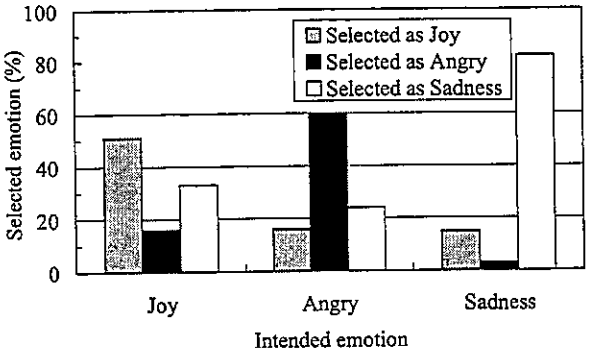

Figure 7: Selected emotion for female synthetic speech

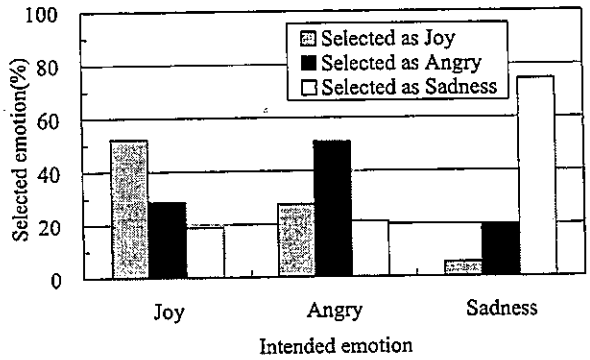

Figure 8: Selected emotion for male synthetic speech

4.2. The Phrase Intelligibility per Sentence

Prior to the subjective evaluation, the intelligibility, the prerequisite for speech, was examined [9]. Subjects were 12 university students. Speech samples were generated with our method and as a comparison, with a speech synthesis system works on MS Windows 95/98 in commercial use. This speech synthesiser is commonly used for a speech synthesis LSI chip in communication aid. Setting levels as intonation, speech, pitch, and volume for that system were set to the most natural level in the experimenter's perceptual impression. Speech samples were all saved as 16kHz, 16bit wav-format files and stored in a notebook PC. Six short sentences (each sentence includes either 1) numbers, 2) uncommon collocations or 3) emotional phrases) were synthesised by the two systems. Subjects were told to write down the exact words and the intelligibility for each sentence was calculated by summing up the correct number of phrases ('bunsetsu' in Japanese) in the sentence.

As figure 9 shows, The score for the intelligibility per sentence was 92.1% for synthetic speech of our method while 81.9% for the synthesiser in comparison. The result of t-test with significance level of $p < 0.05$ showed that there was no significant difference within the two synthesisers. This indicates that the synthetic speech with emotion generated by our system maintains an equivalent or higher level of the intelligibility with the synthesiser in comparison currently in use.

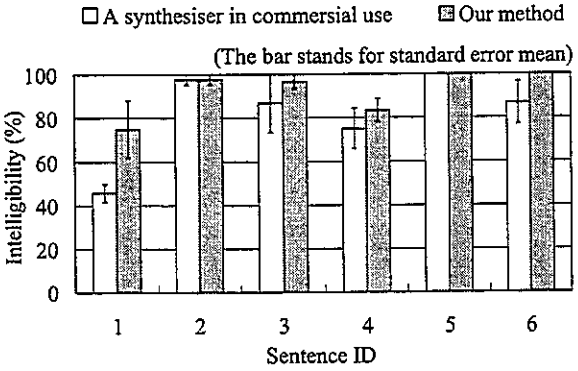

Figure 9: The phrase intelligibility per sentence

4.3. Subjective Evaluation

In this experiment, listeners' overall preference and the subjective degree of expressed emotion (i.e. not the identification rate of the emotion type) were evaluated. The same speech samples were used for three groups of subjects: 1) university students, 2) ALS patients and 3) visually impaired.

Subject Group 1, University Students

Subjects were 143 university students (Male 69, Female 74). Text samples used were one sentence each from angry, joy and sadness corpus and a combination of sentences consisted of a sentence from all three corpora. Again as a comparison, all sentences were synthesised with our method and a speech synthesis system in commercial use. This time, the speech output of a communication aid which synthesise speech with the speech synthesis a LSI chip manufactured by the same company as the comparison system used in 4.2. Setting levels were set to the equivalent level as the speech synthesiser used in 4.2 by the experimenter's perceptual impression. Synthesised samples made by both systems were saved as 16kHz, 16bit wav-format files and stored in a notebook PC. The speech samples were presented to the listeners by speakers (SONY SRS-38) in a large classroom and students were told to fill in the questionnaires.

For each sentence, synthetic speech generated by both systems were randomly presented to listeners who were asked to rate each speech sample using 5-point scales (5 = excellent, 1 = very poor) for the overall preference and the subjective degree of expressed emotion. MOS (mean opinion score) and standard error of the mean was obtained. As shown in Figure 8 and 9, for both items, MOS was higher for our method compared to the synthesiser in comparison currently in use.

From the above result, we can verify that the synthetic speech generated by our method can express emotions with favourable impression.

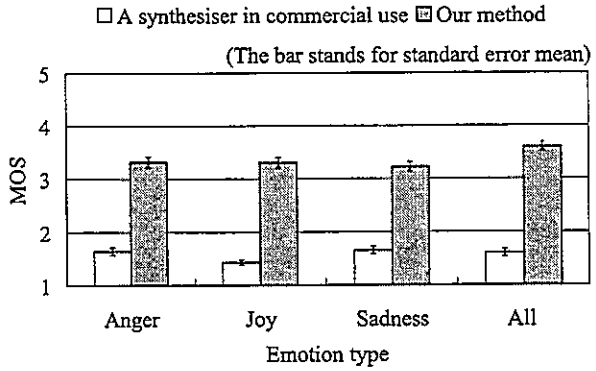

Figure 10: The overall preference of Group 1

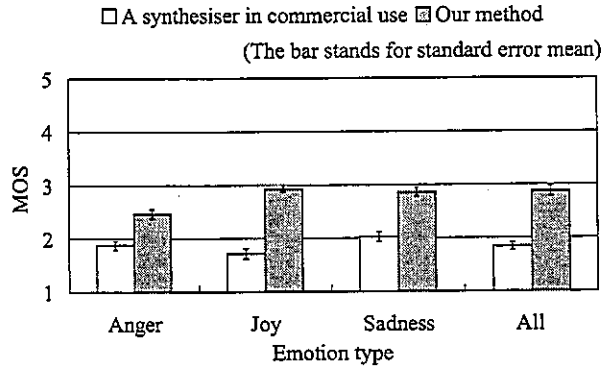

Figure 11: Subjective degree of expressed emotion of Group 1

Subject Group 2, Target Users – ALS patients

Subjects were five ALS patients of age 46 through 61 (3 males and 2 females). Speech samples and the audio equipment were the same as the experiment conducted to Subject group 1. Each experiment was conducted as individual interview and the experimenter transcribed the subject's answer. For each speech

sample, the subjects were asked to rate by marking 5-point scales for the overall preference and the subjective degree of expressed emotion. Subjects were also given a forced-choice selection of joy, anger and sadness for each speech sample and were asked to indicate the words they could not understand by pointing to the words printed out on the paper after listening to the synthetic speech.

The identifying rates for both male and female emotional synthetic samples were joy: 66.6%, anger: 93.3%, sadness: 86.6%. It can be concluded that emotion types have been correctly recognised at a significance level of $p < 0.01$. As for the intelligibility, for our method was 91.7% and for the synthesiser in comparison was 85.2% and no significant difference under t-test of $p < 0.05$. Figure 12 and 13 shows the overall preference and subjective degree of expressed emotion. For both items, MOS were higher for our method than the synthesiser in comparison.

Subject Group 3, Visually Impaired

We have asked people who were visually impaired to participate in the experiment as frequent users of speech synthesis. Subjects were five visually impaired (3 males, 2 females), age ranged from 20 through 40. Speech samples were the same with those used in the experiment conducted to subject group 1.

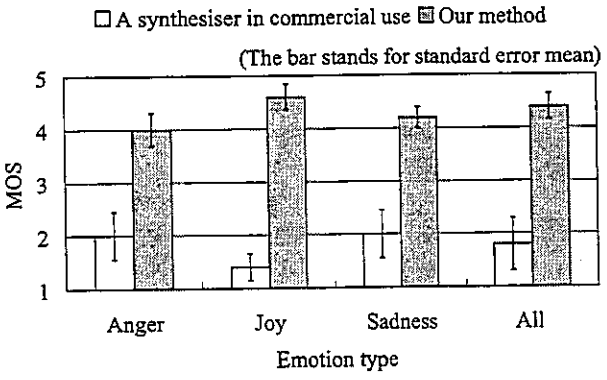

Figure 12: The overall preference of Group 2

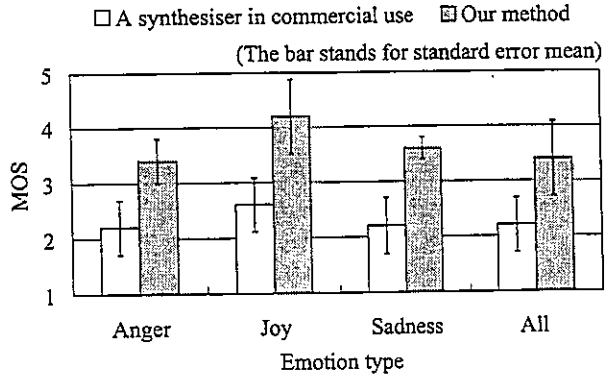

Figure 13: Subjective degree of expressed emotion of Group 2

Except for 1 male subject who was interviewed, subjects were presented speech samples via WWW and were sent questionnaires by e-mail. Subjects returned answers by e-mail and telephone interview was followed to ask them their impressions. Again, MOS for the overall preference and subjective degree of expressed emotion were higher to our method (Figure 12 and 13).

6. THE GUI EVALUATION

Target users (subject group 2) were asked to evaluate the GUI of CHATAKO, the prototype system. They were also asked to subjectively rate using 5-point scale (5 = the most important, 1 = the least important) for the important factors as a communication assistive system. Since the current prototype system is not equipped with input device for disabled, subjects observed the experimenter's system operation after given an explanation for each GUI component and commands.

As shown in Table 3, all the ratings for GUI items were 4 or higher. As for the important factors, the highest priority was given to the intelligibility of synthetic speech, but emotional expression was also considered to be more important than response time.

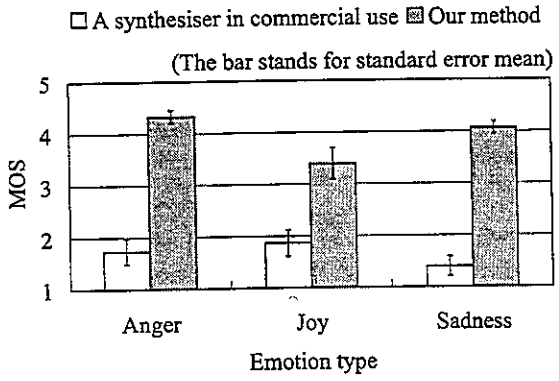

Figure 14: The overall preference of Group 3

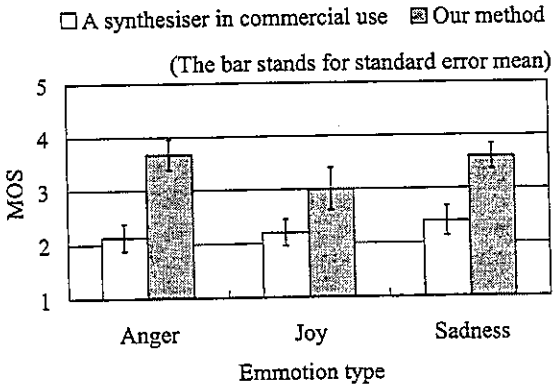

Figure 15: Subjective degree of expressed emotion of Group 3

Evaluation Items	MOS
Functions are sufficient	4.6
Window layout is easy is understand	4.8
GUI is preferable	4.2
Ability to express emotion is desirable	4.0
Speaker-emotion selection is easy to understand	4.6
Graphical expression speaker-emotion selection is easy to understand	4.6

Table 3. Evaluation on graphical user interface

Important Factors	MOS
Intelligibility of synthetic speech	4.8
Emotional Expression	4.4
Response time	3.4

Table 4. Evaluation on factors to be considered

7. FURTHER DIRECTION

The acoustic analysis of the corpora indicated f0 as a powerful feature to cue emotion. The authors will further pursue research to identify features that are relevant for specifying a particular emotion in a parametric way so that those features could be included as selection criteria for CHATR. Our goal for corpus research is to combine the three corpora into one corpus, which results in smaller-sized database.

Also it is the authors' objectives to improve the ease-of-use of the prototype system, CHATAKO by including prepared responses and commonly used texts that can be accessed by pictographic icons.

Enhancing CHATAKO to communication assistive system in a remote environment is another research topic for the authors. By connecting to MAPI (Message Application Programming Interface), CHATAKO supports visually impaired as an e-mail reader with the sender's intended emotion.

8. CONCLUSION

This paper reported on a prototype of a speech synthesis system that can convey user's emotion (joy, anger and sadness) by using corpora of emotional speech. Target users of this system are people with speaking problems. Perceptual experiments were conducted using the synthetic speech generated and the results proved that emotion types of speech samples were significantly identifiable. Experiments were also carried out to evaluate the intelligibility, the overall preference and the subjective degree of expressed emotion of which results showed positive values for our method. Further directions for this research are 1) to identify features that are relevant for emotion, 2) to extend the prototype system for practical use and 3) to enhance it to a remote environment use.

9. ACKNOWLEDGEMENTS

Authors would like to express their sincere appreciation to Mr. Patrick Davin of ATR, Mr. Shinnichi Yamaguchi of Fukuoka-pref. Japan. Authors further appreciation goes to all the subjects participated in the experiments and to the labellers for their work on labelling the speech corpora.

10. REFERENCES

1. Iida, A., Iga, S., Higuchi, F., Campbell, N. and Yasumura, M. Acoustic Nature and Perceptual Testing of a Corpus of Emotional speech. *Proceedings of The 5th International Conference on Spoken Language Processing (ICSLP)* Vol. 4, 1559-1592, 1998.

2. Iida, A., Campbell, N. and Yasumura, M. Design and Evaluation of Synthesised Speech with Emotion. *Journal of Information Processing Society of Japan* Vol. 40, No. 2, 479-486, 1998.

3. Campbell, W. N. Processing a Speech Corpus for CHATR Synthesis. *Proceedings of The International Conference on Speech Processing* 183-186, 1997.

4. Campbell, W. N. and Black, A. W. Chatr: a multi-lingual speech re-sequencing synthesis system. *Technical Report of IEICE* SP96-7, 45-52, 1996.

5. Shaver, P., Schwartz, J., Kirson, D. and O'Connor, C. Emotion Knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology* Vol. 52, No. 1, 1061-1086, 1987.

6. Murray, I. R. and Arnott, J. L. Towards the simulation of emotion in synthetic speech: A review of the literature of human vocal emotion. *Journal of Acoustic Society of America* 93, No. 2, 1097-1108, 1993.

7. Ichikawa, A., Nakayama, T. and Nakata K. Experimental consideration on the Naturalness of the Synthetic Speech. *Proceedings of Acoustic Society of Japan Spring Meeting* 95-96, 1967.

8. Iida, A., Desirazu, S., Iga, S., Campbell, N. and Yasumura, M. A Prototype of Communication Aid using Speech Synthesis with emotion. *Technical Report of IEICE* WIT99-14, 83-88, 1999.

9. Iida, A., Iga, S., Higuchi, F., Campbell, N. and Yasumura, M. A Prototype of a Speech Synthesis System with Emotion for Assisting Communication. *Journal of Human Interface Society of Japan* Vol. 2, No. 2, 169-176, 2000.