Communication aid for non-vocal people using corpus-based concatenative speech synthesis

Akemi Iida [*1,*2] *, Yosuke Sakurada* [*3], *Nick Campbell* [*2,*4], *Michiaki Yasumura* [*3],

[*1] Keio Research Institute at SFC, Keio Univ., Kanagawa, Japan
[*2] Japan Science and Technology Corporation, CREST
[*3] Graduate school of Media and Governance, Keio Univ, Kanagawa, Japan
[*4] Information Sciences Division, ATR International, Kyoto, Japan
akeiida@sfc.keio.ac.jp, nick@slt.atr.co.jp

Abstract

This paper reports on the development of Chatako-AID, a communication aid for non-vocal people using corpus-based cocatenative speech synthesis by creating a speech corpus especially designed for such use. The concept of Chatako-AID; synthesis with the user's voice, which makes use of precomposed texts, is highly appreciated by the target user. This confirms that the recording of a minimum set of phonetically balanced sentences is insufficient for speech synthesis in the proposed method and that a combination of the above recording and a recording of well-read continuous-text material produces more natural sounded synthesised speech.

1. Introduction

Speech is one of the most important and basic means of communication. Some people affected by Motor Neurone Disease(MND) [1] or cerebrovascular diseases have lost their phonatory function although their intellect remains unaffected. We hear from them and their family that it is difficult to communicate with each other after they lost their voice [2].

Boards with words and icons have long been serving as their communication tool but in the past decade, PC-based communication assistive systems some incorporating a text-to-speech (TTS) synthesis have been developed and becoming popular. Examples of such systems are prototype of Murray's [3] and products of Eyegaze [4], Camereon [5], and Dennoshin [6]. The speech synthesisers used in these systems are either rule-based formant system or rule-based concatenative system using small segment units. The quality of synthesised speech generated by these systems has been improving greatly and there is no problem in intelligibility (understanding what is being said). However, when naturalness or voice quality is required, improvement is still desired from the users.

The work reported here uses the ATR CHATR as a concatenative speech synthesiser [7]. Unlike most commercial concatenative synthesisers, CHATR employs a natural speech corpus as a source of units for synthesis. Taking this as an advantage, the authors have developed a communication aid using an ALS (Amyotropic Lateral Sclerosis) patient's voice who is anticipating loss of his voice in the future. The notable characteristics of this study is our trials of testing several different speech corpora with our synthesis method. In this paper, we describe how the read materials are constructed to obtain good synthesised speech which matches the situation of use. We then explain about the system configuration of Chatako-AID, the TTS communication aid programmed with a list of precomposed words and short sentences.

2. Target user and the speaker

ALS is one of MND affecting the motor neurones in the brain and spinal cord, which leads to weakness and wasting of muscles. In ALS, the respiratory muscles also weaken and patients need to undergo a tracheotomy which results in losing the ability to speak in most cases.

Mr. Shinnichi Yamaguchi, age 62, is an ALS patient resides in Fukuoka, Japan. His occupation was an electric engineer and he has taught computer science in college. He was diagnosed as ALS five years ago. At the same time, spontaneous respiration became difficult for him and since then, he has been wearing a nasal pressure support ventilator 24 hours a day. He is aware of the possibility of losing his voice in the future. He views that speech synthesised by commercial system sounds less natural than human voice and has been hoping to use more human-like, expressive speech and if possible, his own voice [8]. He showed a keen interest in the authors' research of employing CHATR to synthesise emotional speech [9] and so the work of using his speech as a speech database for synthesis has began a year ago.

The precise population for the MND patients are uncertain but the prevalence is thought to be 7 per 100,000 people. Since MND is a progressive disease, it is possible to make speech database before the patients lose their voice. This is exactly the case for Mr. Yamaguchi.

The recording of Mr. Yamaguchi (the speaker)'s voice took place in a barrier-free sound-treated room with his caretaker and volunteer recording staffs. The speaker's nasal pressure support ventilator gave high pressure at the speaker's aspiration and low pressure at expiration and it made motor noises. To reduce its affect, a blanket was used to cover the system unit and also the speaker was asked try not to speak while the ventilator is in high pressure. Recording was conducted in two days paying attention to the speaker's health conditions and taking plentiful rests when needed.

3. Speech synthesis system for this study

CHATR is a natural speech re-sequencing synthesis system that incorporates naturally read continuous-text materials as a source database for concatenative unit selection. Unlike the widely used concatenative synthesisers which produce synthesised speech with pre-recorded small units, CHATR produces an index for random-access retrieval of an externally stored nat-

ural speech corpus to select units to create new utterances using an algorithm that optimises the unit selection [7]. The quality of the resulting synthesis depends to a large extent on the phonetic and prosodic balance in the speech corpus as well as on the recording quality. ATR phonetically balanced text corpus of 503 sentences which can be read in about an hour [10] has served as a read material for CHATR synthesis. When the balance requirement is met, units can be selected so that signal processing, which often produces distortion, will be unnecessary.

For CHATR synthesis, a source database consists of a digitised waveform sequence without disfluency and redundancy, and its index file in text format must be prepared from a recorded speech corpus. The procedure to create index file is in three steps: 1) Converting an orthographic transcription of the speech corpus to a phonetic representation, 2) aligning the phones to the waveform to provide a key to the prosodic feature extraction, and 3) producing feature vectors for each phone (label, f0, duration, power etc.,) which are to be written in the index file. CHATR determines optimal weight vectors of each feature per phone that is used at unit selection. The weight vectors are also written in the index file. Each phone index holds the information of the current, the previous and the following phones. When texts are typed in, three steps are conducted before synthesising the re-sequenced waveform; a text processing, a prosody processing, and a unit selection. Units are selected by way of maximising continuity and minimising the distance from phonetic and prosodic targets.

CHATR runs equally on UNIX, Linux and Mircosoft Windows 95/98/2000. For this research, CHATR98 for MS Windows was used and no signal processing was applied.

4. Text corpora design

The objective of text corpora design is to construct a speech corpus which represents the speaker's natural daily speech, and which is strong in synthesising words and sentences that frequently appear in patients' daily lives. We also paid attention to maintain phonetic balance by adding a minimum set of phonetically balanced sentences. However, there has been much concern that phonetically balanced sentence are not easy to read and are remote from daily conversation. Our previous work of creating corpora of emotional speech showed that using read materials that were easy to read and contains appropriate words and sentences for the target speaking style and situation could result in close-to-desired synthesised speech [9].

The following is the read materials for the recording. The first two materials can be prepared individually according to each user.

4.1. Familiar texts for the speaker (348 sentences)

We set the speaker's talk manuscript [8] as a main text set. The speaker has been actively giving a talk about the usefulness of computer to disabled people. The manuscript for this talk is well digested by him and more natural phonation and prosody can be expected than having him read texts which is unfamiliar to him. This text set contains 348 sentences with 385 biphone variation.

4.2. Words and sentences frequently used among patients (459 words, 91 sentences)

One of the important requirements for a communication aid is to accurately produce words and lists that are frequently used by the users. The corpus-based concatenative synthesis, the method applied here can meet this requirement by modifying the synthesis unit from a phone to a word or even a longer sequence. The speaker database incorporates the waveforms for this set.

For this trial, words and short sentences are prepared based on the speaker's word and sentence list. They are categorised as follows; sentences for requests to the caretaker, for conversations with caretaker/with friends/ or on the phone, and words essential to his daily conversation (parts of the body, symptoms, directions and proper nouns). This set contains 495 words, 71 short sentences and 20 sentences and is formed into a list for Chatako-AID.

4.3. Phonetically balanced sentences (129 sentences)

107 sentences extracted from the ATR 503 sentences using the criteria of "biphone + with/without accent" are used as phonetically balanced sentences with 22 supplemental sentences. The purpose of making this subset was to see whether the recording of this set could serve as a minimum satisfactory source database for CHATR. It was also our aim to reduce the speaker's workload when recording. The biphone variation is 465.

4.4. Question sets (78 sentences)

This set was aimed to produce the question intonation pattern since the other text materials are less likely carry it. 78 short sentences with final syllables in all mora variation are included.

4.5. Emotionally-coloured texts (464 sentences)

Three types of materials were prepared, reflecting happy, angry and sad emotions respectively. The authors' previous work showed that emotional speech can be synthesised by preparing a corpus of emotional speech. We have reduced the size of the text set for each emotion from the ones used in the previous studies by about 1/4 due to the same reason as reducing that of balanced sentence set. Anger texts were taken from the speaker's writings with permission. The number of sentences of each set is as follows; happiness, 185, anger, 138 and sadness, 141. Due to time limitations, only 1/3 of the text materials were recorded.

4.6. Speech corpus to be used

Four kinds of source database were created from the following speech corpus: The balanced corpus (the recordings of 4.3), the speaker-familiar corpus (4.1), a combination of the two, and a combination of all except the corpora of emotional speech. Speech was synthesised with CHATR using each database and the evaluation comparing the distance between predicted and selected units was conducted along with a perceptual evaluation. When synthesising arbitrary sentences, the combination of 4.3 and 4.1 showed the best results. However, when synthesising the words and sentences in the precomposed list, the combination of all corpora (except the corpora of emotional speech) is better to be incorporated since in CHATR, although phone-based synthesis is employed, a recorded word as a whole is likely to be selected by the CHATR's unit selection algorithm if the predicted intonation variation matches with that of the word. The details of the evaluation is described in a paper to be presented at the Eurospeech satellite 4th Speech Synthesis workshop [11]. Hence for Chatako-AID, the source database created from "a combination of all corpora except the corpora of emotional speech" is employed as a main database.

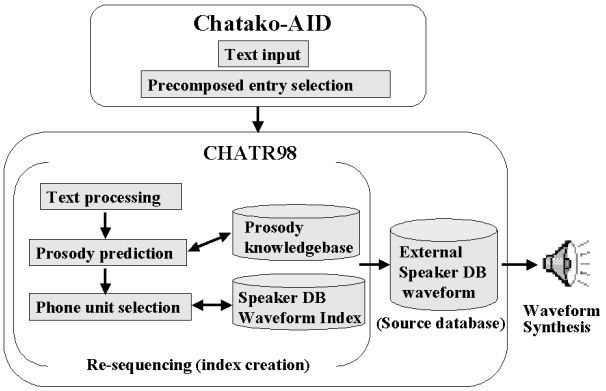

Figure 1: *Chatako-AID's system configuration.*

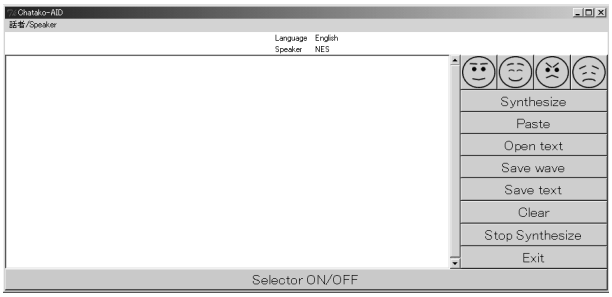

Figure 2: *Main window.*

5. Communication Aid, Chatako-AID

Using CHATR and the main database, a communication aid, Chatako-AID was developed. The main features of this system are 1) speech synthesis with the user's own voice, 2) programmed with precomposed lists of words and short sentences, 3) designed to incorporate speech segments from precomposed lists to TTS, and 4) designed for multilingual use. In this section, the system configuration and main features are introduced.

5.1. System configuration

Chatako-AID runs on Microsoft Windows using CHATR98 as a speech synthesis engine. Fig. 1 shows the system configuration of Chatako-AID. The Graphical User Interface (GUI) for Chatako-AID is implemented in Tck/TK 8.3 (a script programming language) with a text window, speech database selection menu, command buttons and a selectable list of precomposed words and sentences.

5.2. Main window

Fig. 2 shows the English-mode main window at the initial stage. On the top left is a pull-down menu for source database selection. Since, with CHATR, synthesis in any language is possible by loading a source database of the target language, Chatako-AID can also be used multilingually by database selection. The current system can be used in English and Japanese. The language on the screen will change automatically according to the selected source database.

Below the menu, a text window is located on the left, and command buttons on the right. The following commands are implemented: "Synthesise," "paste," "open text file," "save wave," "save text," "clear screen," "stop synthesis" and "exit."

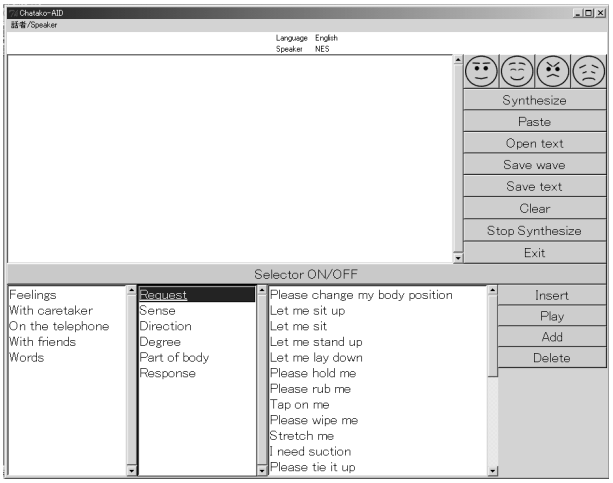

Figure 3: *List window*

Above the command buttons are emotion selection keys. These keys are designed to synthesise emotional speech of the user's choice while typing in the text in the text window. The default is set to the main database which is "normal speech," the leftmost face icon. Then "happy," "angry" and "sad" from left to right. When the user wishes to change emotion types while typing, he or she can do so by re-selecting emotion icons. Texts in the text windows are distinguished by font colours representing emotion categories. For this trial, since recording of the emotional speech database is still incomplete, this function was not activated. Details of emotional speech synthesis using concatenative speech synthesis is reported in [9].

5.3. List window

Below the text window and command buttons is a selector bar. By clicking it, a list of precomposed words and short sentences appears as shown in Fig. 3. The list and its format can be tailored to each user's preference. As a prototype, we composed a three-layered structure, the top layer is classified based on who the user wants to talk to, the second layer is classified based on the content (request, salutation, etc.) and the third layer is the words and short sentences to be synthesised.

Command buttons are located on the right. "Play" button will synthesise or play the speech of the selected entry. "Insert" will copy the text entry to the text window for CHATR to synthesise. The "add" button will create a pop-up window for adding words and short sentences to the list, and the "delete" window removes the entry from the list. Adding and deleting of category labels can be implemented by the same proccedure. Fig. 4 shows the pop-up add window.

5.4. User evaluation

5.4.1. Evaluation on intelligibility of the synthesised speech

Three Japanese speech samples were synthesised using CHATR and as a comparison, three speech samples synthesised with a commercial system which is broadly used for speech synthesis LSI chips in communication aids. Intonation, speed, pitch for this system were set to the most natural sound level and sound volume was adjusted to equivalent level for all 6 sentences. All samples were saved as 16kHz, 16 bit wav-format and were presented to Mr. Yamaguchi (the user) and his wife via internet.

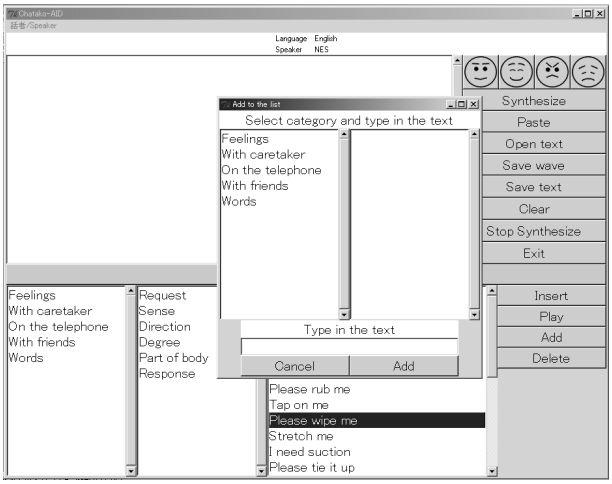

Figure 4: *Add window.*

They were asked to type in the exact words they heard, and the intelligibility for each sentence was calculated by summing up the correct number of "bunsetsu," a notion for Japanese language equivalent to "phrase". To maintain fairness, words and sentences in the precomposed list were not used in the evaluated samples.

The user's score was 93% for the commercial system and 92% for the current method and his wife's was 89% for the former, 93% for the latter. The five-scale rating, where 5 is the highest score, was also used for testing overall preference and both gave higher scores for the speech synthesised by proposed method. The equivalent perceptual experiment was conducted with 16 listeners and the result showed the similar tendency [11].

5.4.2. Evaluation on GUI

The user has been using the system with a special input device consisting of a joystick and several buttons in Japanese envornment. He was asked to evaluated the system's concept and usability. He was asked to rate subjectively, using five-point scale. As shown in Table 1, evaluation items for concept were rated high. The evaluation items for GUI were evaluated in comparison with commercial editors and rated with severe criteria which resulted in 3 and 4.

6. Conclusion

A communication aid for use by non-vocal people was developed. For this work, an ALS patient who anticipating the loss of his voice participated as both the speaker and the user of the system. Several materials were read to produce speech corpus aimed for specific situation. From this work, we confirmed that a minimum set of balanced sentences is insufficient for synthesising natural sounding speech and an effective approach is to have the speaker read the material which he/she knows well. Adding words and short sentences frequently used in the target situation is also effective. As shown in the user's evaluation and the authors' earlier evaluation results from a group of target users [9], the concept of Chatako-AID, synthesis with the user's voice and with a precomposed list is highly appreciable. The evaluation by the user shows that the system is usable in practical use.

Table 1: *Usability Evaluation*

Evaluation Items	Rating
Items on Concept	
You can synthesise with your own voice	5
Daily used words and short sentences can be played from a precomposed list	5
Words and short sentences can be inserted from the list into the desired position in the text you are creating in the text window	5
Items on GUI	
Window layout is easy to understand	3
Window layout is appropriate for your process routine	3
Sufficient functions are implemented	4
Label for command buttons are easy to understand	4
List selector button is easy to understand	4
The word selection operation in list window is easy to understand	4
The add window operation is easy to understand	4
The delete window operation is easy to understand	4
Prompting and warning messages are appropriate	4

7. Acknowledgement

Authors would like to express their sincere appreciation to Mr. Shinnichi Yamaguchi of Fukuoka-pref., Japan for his participation to this work as a speaker and a user. Authors also would like to thank Mr. Eiji Mitsuya and Mr. Masahiro Nishimura of ATR, Mr. Fumito Higuchi of Keio University for their kind cooperation.

8. References

[1] http://www.mndassociation.org/yindex.htm

[2] Ohira, Y., "Watashirashiku ningenrashiku (Autobiography in Japanese)", Kanagawa Chuo-shuppansha, 1995.

[3] Muray, I.R., Alm, N., Newell, A.F., "A communication system for the disable with emotional synthesised speech produced by rule", Proc. Eurospeech '91, pp.311-314.

[4] http://www.eyegaze.com

[5] http://www.cameleon-we.com

[6] http://www.hitachi.co.jp.

[7] Campbell, W. N., "CHATR: A High-Definition Speech Re-Sequencing System", Proc. 3rd ASA/ASJ Joint Meeting, 1996, pp.1223-1228.

[8] Yamaguchi, S., "Pasokon wo tsukaikonasou", http:// www.isn.ne.jp/ kamata/ftp/jals.html (in Japanese), 2000.

[9] Iida, A., Iga, S., Higuchi, F., Campbell, N. and Yasumura, M., "A Speech Synthesis System with Emotion for Assisting Communication", Proc. ISCA Workshop on Speech and Emotion, pp.167-172, 2000.

[10] Abe, M., Sagisaka, Y., Umeda, T., Kuwabara, H., "Speech Database User's Manual", ATR Technical Report TR-I-0166, 1990.

[11] Iida, A., Campbell, W. N., "A corpus Design for a Concatenative Speech Synthesis System for the Disabled", submit to 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.