A database design for a concatenative speech synthesis system for the disabled

Akemi Iida [*1,*2] , *Nick Campbell* [*2,*3]

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

Abstract

This paper reports on our research on designing a Japanese speech database for a concatenative speech synthesis system which are to be used for a specific purpose, For this work, the purpose was set to assisting communication of non-vocal people. Four kinds of speech database were developed by combining different speech corpora spoken by an Amyotropic Lateral Sclerosis (ALS) patient who was anticipating the imminent loss of his voice. This work confirmed that the recording of a minimum set of phonetically balanced sentences (129 sentences) was insufficient for concatenative speech synthesis and that a combination of these and a recording of well-read continuous-text material produced more natural sounding synthesised speech. A communication aid was developed using a concatenated speech synthesis with the database created in this work.

1. Introduction

Since the introduction of waveform processing techniques like PSOLA, concatenated speech synthesis using small concatenation units is becoming popular. To improve naturalness of the synthesised voice, corpus-based approaches have been introduced to concatenated speech synthesis by Sagisaka [1], and Campbell and Black [2]. Sagisaka proposed a scheme for the selection of non-uniform units to be excised at synthesis time from a database of 5000 recorded words from a frequency-based wordlist and used LMA resynthesis from cepstra to modify speech signals to the desired intonation. Campbell introduced the use of prosody as a selection criterion, resulting in a phone-based natural speech re-sequencing synthesis system which employs a large speech corpus of read sentences as a source of units for synthesis. By taking advantage of the natural variety of the source units both segmentally and prosodically, the amount of subsequent signal processing was reduced and yet natrualness of the original voice was maintained.

The ideal size and balance (segmental and prosodical) of the source corpus is yet to be determined. A larger corpus with richer variety produces better quality synthesised speech. However, when recording is taken into consideration, the efficient design of a speech corpus by selection of reading materials is an important task. This is especially important when the speaker has some disability and the need to reduce the amount of reading material is essential.

In this study, four kinds of speech database were created by combining four speech corpora spoken by an ALS patient who was anticipating loss of his voice. The work reported here used ATR CHATR speech synthesis system which employed a natural speech corpus as a source of units for synthesis. The first part of this paper describes how the read materials were constructed to obtain good synthesised speech which matches the situations of intended use. The latter part reports on the evaluation of the synthesised speech generated with four speech database and on the result of a perceptual experiments. Employing the database with the best result, the TTS communication aid, Chatako-AID (based on CHATR code) was developed. The system configuration of Chatako-AID is described in a paper to be presented at the Eurospeech 2001 [3].

2. Target user and the speaker

ALS is one of Motor Neurone Disease (MND) [4] 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 [5]. He showed a keen interest in the authors' research of employing CHATR to synthesise emotional

speech [6] 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. Since ALS patients have physical disabilities more or less, the reading materials are shorter the better. Although Mr. Yamaguchi was already wearing a nasal pressure support ventilator, he was willing to corporate in reading materials as much as possible for this study.

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 natural speech corpus to select units to create new utterances using an algorithm that optimises the unit selection [2]. 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 [7] 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. Prosody prediction is made based on prosody knowledge base and generates a phonetic transcription with accent and intonation marks. 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 the text corpora design in this study is to construct a minimum but sufficient speech corpus for concatenative TTS reflecting the speaker's natural voice. For practical use, it is also important that by using the created corpus, synthesised words and sentences that frequently appear in patients' daily lives are natural and understandable.

Maintaining phonetic balance is also important. There has been much concern, however, 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 with phonetic balance close to that of the standard read material (503 sentences) for CHATR [6].

The following is the read materials for the recording. The second and third materials can be prepared individually according to the individual user.

4.1. 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.

4.2. Familiar texts for the speaker (348 sentences)

The speaker's talk manuscript was selected as a main text set [5]. 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 unfamiliar texts. This text set contains 348 sentences with 385 biphone variation.

4.3. Frequently used words and sentences 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.

4.4. 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 could 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.

To sum up, the following speech corpora were created by having the speaker read the above materials.

1. A minimum set of phonetically balanced sentences (Balance corpus)
2. Familiar texts for the speaker (Speaker corpus)
3. Words and phrases used daily (Daily corpus)
4. Question corpus
5. Emotionally-coloured corpora (containing Angry, Joy and Sad corpus)

5. Source database creation

Following four kinds of source database were created as main database (default database)with all speech corpora except corpora of emotional speech which is designed to be loaded separately (Details are described in a paper to be presented at the Eurospeech 2001 [3]. Speech was synthesised with CHATR using each database.

1. Balanced corpus
2. Speaker corpus
3. The combination of 1) and 2)
4. The combination of 1), 2), Daily corpus and Question corpus.

5.1. Analysis on generated synthesised speech

As for Database 1 and 2, no need for explanation from which corpus the units are selected since they are consisted of a single corpus. For Database 3, most units were selected from the speaker corpus and several units were selected from the balanced corpus. Same coul be said for Database 4, but further, sevelral were selected from the daily corpus and the question corpus. Following is the selected unit using Database 3 and 4 for "hajimetenanode totemo fuanndesu (I am worried since it is my first time)."

- Database 3

Speaker corpus345 ; h
Speaker corpus037 ; a j i m e
Speaker corpus009 ; t
Speaker corpus079 ; e
Balanced corpus010 ; n
Speaker corpus055 ; a n o d
Speaker corpus188 ; e #
Speaker corpus147 ; t o t
Speaker corpus345 ; e m
Balanced corpus027 ; o #
Balanced corpus026 ; f
Speaker corpus345 ; u a
Speaker corpus259 ; N d
Speaker corpus145 ; e s U

- Database 4

Balanced corpus047 ; h
Speaker corpus037 ; a j i m e
Speaker corpus228 ; t
Question corpus058 ; e n
Speaker corpus055 ; a n o d
Speaker corpus238 ; e
Speaker corpus205 ; # t
Speaker corpus298 ; o t
Speaker corpus165 ; e m o
Speaker corpus347 ; #

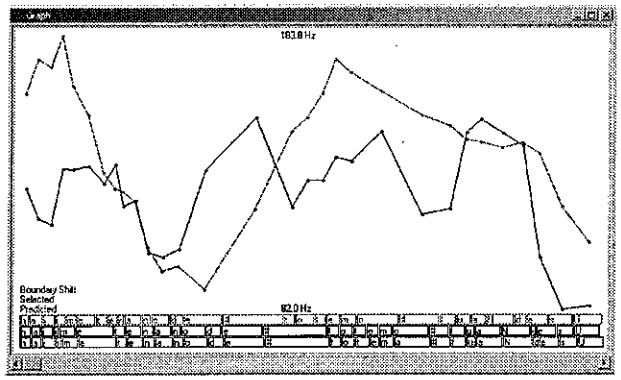

Figure 1: The predicted and selected units for Database 1

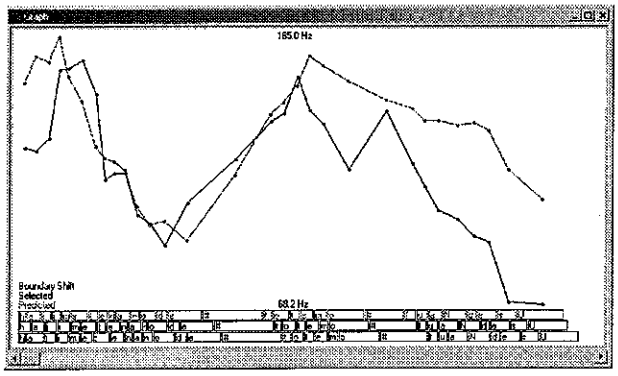

Figure 2: The predicted and selected units for Database 2

Daily corpus004 ; f

Speaker corpus345 ; u a

Speaker corpus259 ; N d

Speaker corpus047 ; e s

Speaker corpus044 ; U

Fig.1 to 4 shows both the selected and predicted units for the same sentence. In current Chatr98, the most weighted feature for unit selection is f0 at midpoint in each phone. From the figures, we can see that when Database 1 is used, (Balanced corpus), the selected units are less closer to the predicted prosody than when other three database are used.

5.2. Perceptual evaluation of the synthesised speech

Intelligibility and subjective impression were evaluated by conucting a perceptual experiment. Four short sentences were synthesised with CHATR using each database. All samples were saved as 16kHz, 16 bit wav-format and were presented to 20 listeners. For listeners, all four sentences are synthesised with different database and total of 5 responses were obtained for each sentences. To maintain fairness, words and sentences in daily corpus were not used in the evaluated samples.

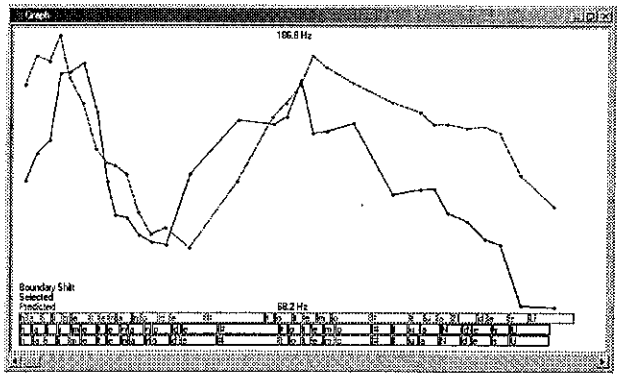

Figure 3: The predicted and selected units for Database 3

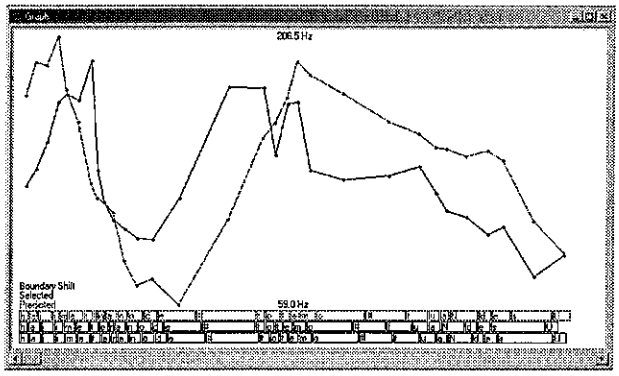

Figure 4: The predicted and selected units for Database 4

For intelligibility evaluation, listeners 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". As shown in Fig.5, overall mean intelligibility score and standard deviation (SD)(written in parenthesis) for each database was 70% (0.33) for Database 1, 77% (0.30) for Database 2, 92% (0.15) for Database 3 and 95%(0.11) for Database 4. Under means comparisons of p 0.05, no significant difference were recognized among Database 2 to 4 but a significant difference was noticed between Database 1 and the rest. No significant difference was noticed between Database 1 and 2.

For subjective impression, "understandability" and "overall preference" were evaluated. Listeners were asked to rate each speech sample using 5-point scales (5 = excellent, 1 = very poor) for both items. For "understandability," MOS (mean opinion score) and standard deviation from Database 1 through 4 was 2.4 (1.4), 2.2 (1.0), 3.4 (1.3) and 3.2 (1.0). For "overall preference," MOS and SD was 2.8 (1.3), 2.5 (1.0), 3.3 (0.9) and 3.4 (0.9).

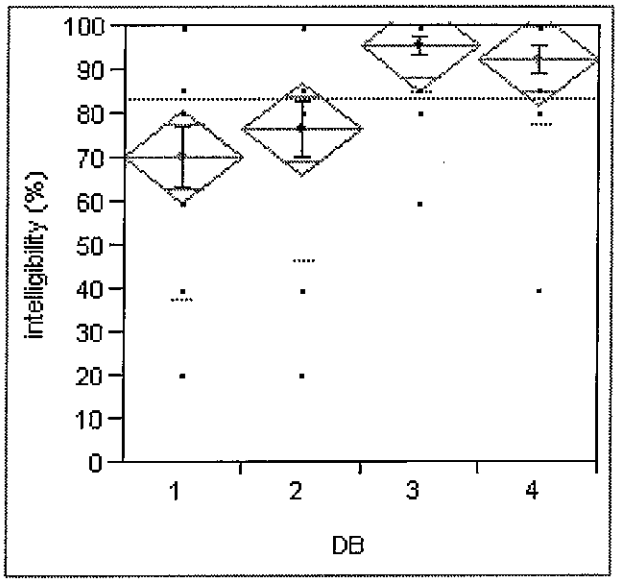

Figure 5: *Mean intelligibility score and SD per DB*

6. Source database for implementation

As shown in 5.2, when synthesising arbitrary sentences, the combination of Database 3 the combination of balanced corpus and speaker corpus scored slightly higher in perceptual evaluation than Database 4, the combination of all corpora but there was no significant difference in mean comparison.

For implementation of communication aid, we selected Database 4 since it includes Daily corpus 4.5 which contains necessary words and phrases for ALS patients. This is to take an advantage of CHATR's characteristics of selecting phones in ontinuum even with the current phone-based unit selection algorithm. Fig. 6 shows the pitch contour of naturally uttered speech of the speaker and the synthesised speech using Database 3 and 4 of "Kyuunyuu shite kudasai (Please suck my sputa)." Along with the unit selection below, "kyuuin" as a whole is selected from Daily corpus which enables natural intonation in synthesised speech.

- Database 3

  Balanced corpus037 ; k
  Speaker corpus028 ; y u:
  Balanced corpus027 ; i N sh
  Speaker corpus065 ; I t e k
  Speaker corpus039 ; u d a s a i

- Database 4

  Daily corpus011 ; ky u: i N sh
  Speaker corpus125 ; I t
  Speaker corpus055 ; e k

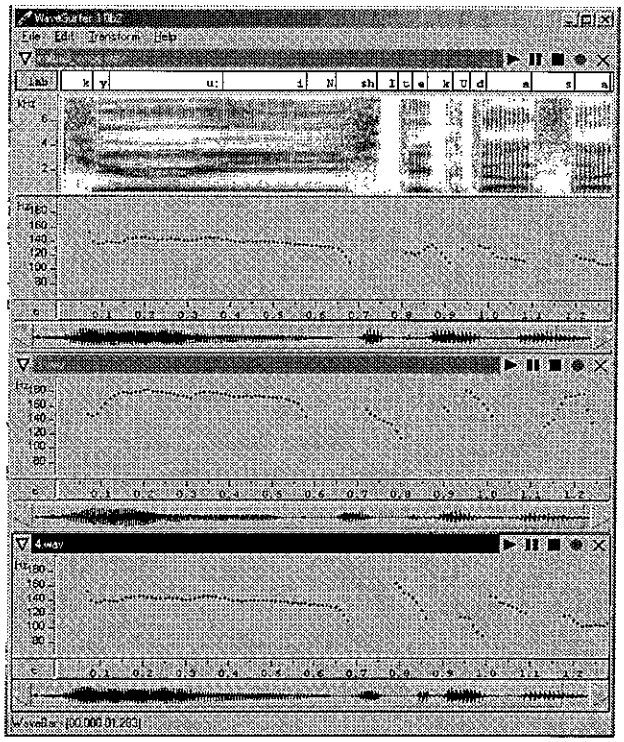

Figure 6: *pitch contour of natural speech, Database 3 and 4*

Daily corpus019 ; u d a s
Daily corpus009 ; a i

6.1. Perceptual evaluation on practicality level

To evaluate the feasibility for practical use, the same perceptual experiment as 5.2 was conduced with Database 4 and as a comparison, with a commercial system which is broadly used for speech synthesis LSI chips incorporated in communication aids.

Six short Japanese sentences were synthesised by two methods. For the latter, intonation, speed, pitch for this system were set to the most natural sound level and sound volume was adjusted to equivalent level for all six sentences. All samples were saved as 16kHz, 16 bit wav-format. 20 listeners were divided into two groups of 10 people. The first group were asked to evaluate three synthesised speech generated by our method and other three, generated by the counterpart. Switching the sentence and the synthesised method combination, the latter group were asked to do the same task. Consequently total of 10 responses were obtained for each speech samples. None of the listeners participated in 5.2 and this experiment.

For overall mean intelligibility score and SD (in parenthesis) was 92% (0.16) for the proposed method and 93% (0.13) for the system in comparison. Under t-test of p 0.05, no significant difference was recognised. For "understandability," MOS (mean opinion score) and stan-

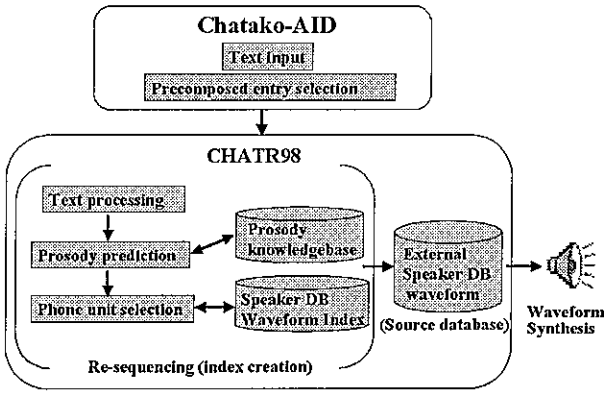

Figure 7: *Chatako-AID's system configuration.*

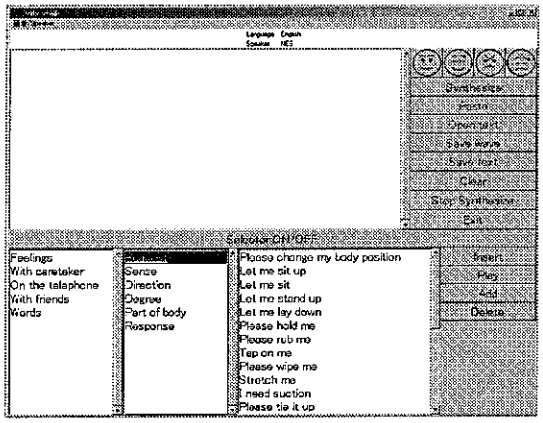

Figure 8: *GUI of Chatako-AID*

dard deviation for our method was 3.2 (1.1), for the compared, 3.0 (1.0) and for "overall preference," MOS and SD was 3.4 (0.9) and 2.9 (1.0). The result of this evaluation indicates that the proposed method is feasible for a practical use.

6.2. Communication Aid, Chatako-AID

Using CHATR98 and Database 4, a communication aid, Chatako-AID was developed. Fig.7 shows the system configuration of Chatako-AID and Fig.8 shows its GUI. This system is equipped with a text window, speech Database selection menu, command buttons and a selectable list of precomposed words and sentences. 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. The details of the system configuration and main features are described in a paper to be presented at the Eurospeech 2001 [3].

7. Conclusion

This paper reports on our research on designing a practical speech database for a concatenative speech synthesis system which are to be used for a specific purpose, For this work, the purpose is set to assisting communication of non-vocal people.

Four kinds of speech database were developed by combining different speech corpora spoken by an ALS patient who is anticipating the imminent loss of his voice. The perceptual evaluation confirmed that the recording of a minimum set of phonetically balanced sentences (129 sentences) was insufficient for concatenative speech synthesis and that a combination of these and a recording of well-read continuous-text material could produce more natural sounding synthesised speech. The perceptual evaluation by comparing with a synthesiser in commercial use showed the feasibility of the system in practical use A communication aid was created using concatenated speech synthesis based on the database created in this work.

8. Acknowledgement

Authors would like to express their sincere appreciation to Mr. Shinnichi Yamaguchi of Fukuoka-pref., Japan. Authors also would like to thank Mr. Eiji Mitsuya and Mr. Masahiro Nishimura of ATR for their kind cooperation.

9. References

[1] Sagisaka, Y., Kaiki, N., Iwahashi, N., Mimura, K., "ATR nyu-talk speech synthesis system", In Proc. ICSLP, pp483-486, 1992.

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

[3] Iida, A., Sakurada, Y., Campbell, W. N., Yasumura, M., " Communication aid for non-vocal people using corpus-based concatenative speech synthesis", submit to Eurospeech 2001.

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

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

[6] 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.

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