



Building an English Speech Synthesis System from a Japanese ALS Patient's Voice

Akemi Iida†, Jun Ito†, Shimpei Kajima‡, Tsutomu Sugawara‡

†Tokyo University of Technology, ‡Sophia University, Tokyo, Japan

E-mail: take@media.teu.ac.jp

Abstract

This paper reports on the development of an English speech synthesis system for a Japanese amyotrophic lateral sclerosis patient as part of the project of developing a bilingual communication aid for this patient. The patient had a tracheotomy three years ago and anticipates the possibility of losing his phonatory function. His English speech database for Festival, a free speech synthesis system, was generated from his reading of a US diphone list. There were two problems with the recording. The first was the noise that the artificial ventilator made and the second was his difficulty in pronouncing English. Although the speaker's English database was successfully built by Festvox and the voice was recognized as his voice, the utterance was unintelligible. We therefore proposed reconstructing the patient's database by partially combining it with an English native speaker's database. Results showed that the proposed approach can be promising for those facing this problem.

1. Introduction

There are many groups of people who are unable to communicate either vocally or physically but have unimpaired cognitive abilities. Common causes of these communication disorders include muscular dystrophy, and motor neuron diseases (MND). Not only do these individuals not speak, but they also often suffer severe physical disorders, involving extreme difficulty in conveying their intention to others in any way as symptoms progress.

As technology improves, various kinds of voice output communication aids (VOCA) have been developed and become effective tools for the communication for vocally disabled people[1]. From the 1980s, text-to-speech (TTS) synthesis began to be used in VOCA where texts were read aloud by the synthesized voice. Many systems now offer several different voices, ranging from child to adult in both male and female registers[2]. However, while speech produced by these synthesizers is intelligible (i.e. understandable), it is true that the voice is completely different from the individual speech quality that users once had.

The intended beneficiary of this research is an ALS patient. ALS is a subtype of MND which is an all-embracing term used to cover a number of progressive illnesses that affect motor neurons in the brain and spinal cord [3]. The dominant symptom of these illnesses is the weakening and wasting of muscles, while the intellect and emotions remain unimpaired. Respiratory muscles also weaken eventually, and many

individuals need to undergo tracheotomies that make it difficult for them to phonate.

The participant in our study wishes to communicate with his own voice even after his tracheotomy. The following information is given with the participant's permission and at his request. Mr. Shinichi Yamaguchi, age 67, resides in Fukuoka, Japan and was diagnosed with ALS 10 years ago. At the time of diagnosis, he already had difficulty with spontaneous respiration, and he had a tracheotomy in 2003. He has been wearing an artificial ventilator 24 hours a day since then. Formerly he was an engineer and taught computer science at college. Since being diagnosed with ALS, the speaker has been active in giving talks to public audiences on the effectiveness of computers for people with disabilities. He can still talk but is aware of the possibility of losing his voice. He thinks that speech synthesized by current commercial systems sounds less natural than the human voice and he hopes to use more human-sounding, expressive speech generated from his own voice [4]. He has shown a keen interest in research in expressive speech and has been collaborating with the first author and her colleagues since the beginning of 2000.

A Japanese speech synthesis system with participant's voice was created using ATR CHATR, a concatenative speech synthesis system in 2000 to offer him the ability to phonate with his own natural speech quality [5]. CHATR was a corpus-based text-to-speech (TTS) synthesis system, and concatenation units were selected from a natural speech database [6]. The participant read his own speech manuscript, a list consisting of words and phrases frequently used in his daily conversation and some expressive phrases in addition to a phonetically balanced sentence set. Results showed that the proposed method successfully synthesized intelligible speech with his natural voice quality.

The patient's next wish is to be able to communicate in English using speech synthesis. This paper reports on our research in developing an English speech synthesis system for a Japanese ALS patient. The next section introduces the speech synthesis system used, Section 5 describes the initial database creation and synthesis experiment and Section 6 describes our proposed method of mixing the speech database to improve the intelligibility of the synthesized speech.

2. The speech synthesis system used

2.1. Festival and Festvox

Festival is free software developed at the University of Edinburgh, and it offers a multi-lingual speech synthesis workbench that runs on multiple-platforms [7].



The main factor for selecting Festival as the speech synthesis system for this study is largely due to the free use of Festvox, which offers environments to build new speech databases that work with Festival [8]. In addition, fully documented instructions on how to build new diphone databases are available and it comes with example speech databases for US and UK English. We also thought that reading a diphone list would be relatively easy for the non-native speakers (but we found out that it was not likely to be true after the experiment).

2.2. US diphone list

The participant read all 1369 entries in the US diphone list generated by Festvox. Each entry was a three-syllable nonsense word. The diphones in the parentheses in Figure 1 were generated from these nonsense words

1	(us_0001 "pautaababaapau" ("b-aa" "aa-b"))
2	(us_0002 "pautapapaapau" ("p-aa" "aa-p"))
⋮	
1368	(us_1368 "pautaaraxapau" ("r-dx"))
1369	(us_1369 "pautataapapau" ("pau-pau"))

Figure 1 Diphone list.

3. Initial development

3.1. Preliminary experiment

In order to become familiar with the procedure, we went through the recording and building process, during which the first author, who had lived in Australia and the U.S. for a total of nine years, served as the speaker. Difficult pronunciations were marked and recording settings for the participant were planned after the recording. Following the instructions given in [9], her speech database (hereafter "teu_us_ai") was successfully built. We managed to synthesize intelligible utterances with her voice quality.

3.2. Recording of the participant's voice

The recording of the participant's voice was conducted in August, 2005. The recording took place in his house located in a quiet neighborhood since it was difficult for the participant to move. Figure 2 shows how we recorded his voice.

After tracheotomy, the speaker usually cannot phonate. He communicates by articulating sounds without voice and his family read what he wants to say by watching his mouth movement. However, he can still speak by using a cuffed tracheostomy tube [10] as shown in Figure 3 and that was how he was able to speak when he was giving talks. The cuff expands during inspiration and all the air from the ventilator is sent to the lungs. During expiration, the cuff deflates and some air passes out around the deflated cuff and discharges through the glottis, allowing sufficient ventilation and also enabling the vocal folds to vibrate. Using this air, patients can speak by training themselves. At the time of recording, Mr. Yamaguchi's lung function decreased compared to what it was several months before. Therefore, he used an oxygen tank to double the amount of air that he inspired, enabling sufficient air to be discharged to the larynx. The recording was conducted from 9 a.m. to 7 p.m.

allowing the speaker plentiful rest when needed. The speaker pronounced each diphone into the microphone (Sony ECM-330) after listening to a prompt through headphones (Sony MDR-Z900). Prompts were synthesized using a US English native speaker's voice ("voice_kal_diphone" - hereafter "kal") offered by Festival. Speech was recorded directly into a laptop PC (Panasonic Let's note W2) using Festvox running on Linux, Vine 3.1. The default interval from one prompt to another was 2 seconds, but the speaker could not phonate within that duration so the interval was set to 5 seconds. The speaker found it difficult to pronounce diphones that involved phonemes unfamiliar to Japanese. In order to minimize mispronunciation, phonemes appearing in entry were shown on the PC display from about a halfway point.

There were two problems with the recording. The first was the noise that the artificial ventilator made and the second related to the English pronunciation. The latter issue was more of a general problem that occurs to non-native speakers of English. Vowels are especially difficult for them since the Festvox US diphone list has 14 vowels while Japanese has only five. For several mispronounced entries, the speaker was asked to re-pronounce them but due to time constraints not all mispronounced entries were re-pronounced.



Figure 2 Recording scene.

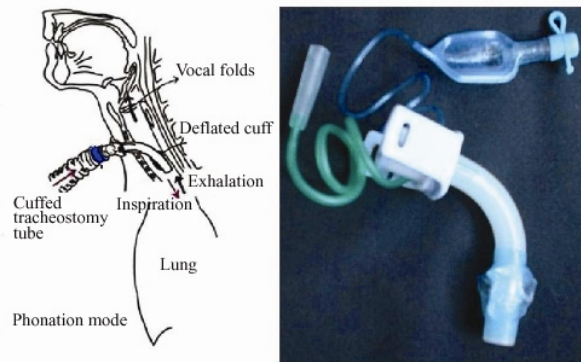


Figure 3 Cuffed tracheostomy tube



3.3. Building the contributor’s database

Each entry was automatically saved as a 16kHz, 16bit wav-format file. First, since we extended the recording time for each entry, it was necessary to cut some period of time prior to and after the phonation. Following the Festvox instructions, the speaker’s English database (“teu_us_ys”) was built. The results of the initial synthesis test were examined by listening to several sentences. Table 1 shows some examples. Five male college students judged that the voice was recognizable as the speaker’s but they could not understand what the synthesizer was saying except Example 3. There was also a noise similar to a beep occurring at the end of each sentence. One possible cause of this noise could be as follows: In festival, a “pau” (which refers to a pause) should be silence but in this recording, the periodic noise of the ventilator was active and thus “pau” might have been identified as a voiced sound for this database. As a result, the source pulse might have been generated for the diphone including a “pau”. More thorough investigation is needed to determine this, however. Figure 4 shows the spectrogram of “I am Festival” synthesized with “teu_us_ys”.

Table 1. *Intelligibility of the speech synthesized with original database, “teu_us_ys”.*

	Sentences	Intelligibility
1	This hotel is nice.	NG
2	I like baseball.	NG
3	I am festival.	OK
4	How are you?	NG

(OK: Intelligible, NG: Unintelligible)

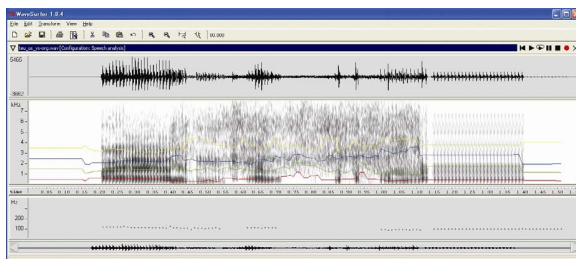


Figure 4 *Spectrogram of “I am Festival” synthesized with “teu_us_ys”*

4. Reconstruction: Combining voices

4.1. Preliminary experiment

As an experiment we combined the first author’s (“teu_us_ai”) entries with those of prompts synthesized using a US native speaker’s voice (“kal”) with a ratio of 9:1. The breakdown was as follows: “teu_us_ai” was used for diphone list No.11-1233, “kal” was used for diphone list No.1-10, 1234-1369. No specific criteria were set for this composition. A synthesis test was given and all five listeners judged that the voice was recognized as “teu_us_ai” and the contents were intelligible. Needless to say, a synthesis test using the “kal” database (all entries were

generated with “kal”) was also conducted and the results showed that all sentences were intelligible.

4.2. Building the database by combining voices

By changing the composition, we made the five sets of database listed below. As mentioned earlier in 3.2, English vowels are more difficult for Japanese to pronounce, and this was also noticeable with the patient’s pronunciation. Therefore, we started off by making a combined set (“teu_us_ys2”) while avoiding vowels from the patient’s database (“teu_us_ys”) as much as possible. However, since the vowels are more sonorant than consonants and the voice quality of the voice synthesized was not recognizable as the patient’s voice without vowels, we gradually added entries with vowels.

- teu_us_ys2: “teu_us_ys”: 545 (entries containing consonant-consonant diphones), “kal”: 824 (the rest including vowel-vowel (VV), vowel-consonant (VC), consonant-vowel (CV) diphones).
- teu_us_ys3: “teu_us_ys”: 807 (adding entries containing VC, CV diphones which were pronounced well to “teu_us_ys2”), “kal”: 562 (the rest).
- teu_us_ys4: “teu_us_ys”: 960 (adding entries containing VC, CV diphones pronounced with understandable pronunciation to “teu_us_ys3”), “kal”: 409 (the rest).
- teu_us_ys5: “teu_us_ys”: 1007 (adding entries containing VC, CV diphones pronounced with acceptable pronunciation to “teu_us_ys4”), “kal”: 362 (the rest).
- teu_us_ys the best: “teu_us_ys”: 933 (all entries judged as acceptable by listening one by one, each compared with the equivalent “kal”). The total number of entries was smaller than “teu_us_ys5” since some entries containing consonant-consonant diphones were judged as defected, “kal”: 436 (the rest).

4.3. Result and evaluation

Table 2 shows the results of a listening test with the same sentences used in 6.1. The speech synthesized with “teu_us_ys2” had a clear pronunciation but was not recognizable as having the participant’s voice quality. With “teu_us_ys3”, the participant’s voice quality was recognized only for “How are you?”. As we proceeded with the test with “teu_us_ys4”, “teu_us_ys5”, we felt that the voice quality of the generated speech got closer to the participant’s voice quality but with beeping sounds appearing at the end and also at the

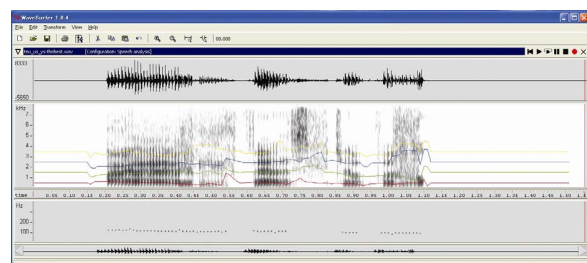


Figure 5 *Spectrogram of “I am Festival” synthesized with “teu_us_ys the best”*

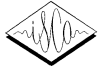


Table 2. *Intelligibility and similarity of each database.*
 (Int: Intelligibility, Sim: Similarity, A: Very Good, B: Good, C: Fair, NG: Poor)

	Sentences	ys2		ys3		ys4		ys5		thebest	
		Int	Sim	Int	Sim	Int	Sim	Int	Sim	Int	Sim
1	This hotel is nice.	A	NG	A	NG	B	NG	B	NG	B	C
2	I like baseball.	A	NG	A	NG	B	NG	B	NG	B	NG
3	I am festival.	A	NG	A	NG	B	C	B	C	B	C
4	How are you?	A	NG	C	C	NG	C	NG	C	C	B

beginning of the sentence “How are you?”. The best result was obtained with “teu_us_ys_thebest”. Figure 5 shows the spectrogram for “I am Festival” synthesized with “teu_us_ys_thebest”. However, again, “How are you?” had the same beeping sounds at the beginning and at the end of the sentence for both “teu_us_ys5” and “teu_us_ys_thebest”.

5. Discussion on noise reduction

In order to reduce the noise from the ventilator, noise reduction was attempted as described in [11] and also using “Water Glass”, a shareware [12] as suggested by a U.S. Experiments with a few entries proved to be effective. Due to time constraints, we could not attempt to apply this to all entries and the synthesis experiments reported in this paper were performed without noise reduction. We will continue to apply this method to all the entries.

6. Conclusion

This paper reported on developing an English speech synthesis system for a Japanese ALS patient. The patient had a tracheotomy three years ago and anticipating the possibility of losing his phonatory function. His English speech database for Festival was generated from his reading of a Festvox US diphone list. Although the speaker’s English database was successfully built and the voice was recognized as his voice, the utterance was unintelligible. Our proposed method of reconstructing the contributor’s database by partially combining the English native speaker’s database showed some achievements and we consider that it was a good start as an initial attempt. Further trial and investigation with knowledge about Festival and Festvox in depth are needed. We consider that more careful look not only at manuals but also shell scripts and intermediate files would help us achieve a promising outcome. Further, finer evaluation should be performed. If proposed method can be used practically, it can contribute not only to people with speaking disorders but also for general use for non-native speakers of English who wish to communicate in English with near-native pronunciation.

7. Acknowledgements

This research was conducted under the Grant-in-Aid for Scientific Research (A-2, 16203041) of the Japan Society of Promotion of Science. The authors would like to appreciate Mr.

Shinichi Yamaguchi and his family for participating in the research. We also would like to thank all the members of the project for their cooperation. Further appreciation goes to Professors Kiyooki Aikawa and Sumio Ohno of Tokyo Univ. of Tech., Messrs. Wataru Imatake and Takashi Sato of Animo Limited, and Mr. John Kominek of CMU for their advice and assistance. Lastly we would like to thank Professor Paul Brocklebank for proofreading this paper.

8. References

- [1] Possum Controls Ltd. Retrieved Feb 25, 2006 from <http://www.possum.co.uk/>
- [2] VOCA. Arcadia’s Voice Aid. Retrieved Feb 25, 2006 from <http://www.arcadia.co.jp/VOCA/> (In Japanese).
- [3] Motor Neurone Disease Association (n.d.) What is MND? Retrieved Feb., 25, 2005 from <http://www.mndassociation.org/full-site/what/index.htm>
- [4] Yamaguchi, S. Pasokon wo Tsukaikonasou [Let’s Use PC]. <http://www.ne.jp/asahi/laconic/ikiru/> (Click on “Essay”) (in Japanese).
- [5] Iida, A. and Campbell, N. Speech database design for a concatenative text-to-speech synthesis system for nonspeaking individuals. *International Journal of Speech Technology*, Vol. 6, Issue 4, pp.379-392, 2003.
- [6] Black A. and Campbell, N. Optimising Selection of Units from Speech Databases for Concatenative Synthesis. *Proceedings of Eurospeech 95*, Madrid, Spain, pp. 581-584, 1995.
- [7] The Festival Speech Synthesis System Homepage. Retrieved Feb., 25, 2006 from <http://www.cstr.ed.ac.uk/projects/festival/>
- [8] Festvox: Home. Retrieved Feb.,25, 2006 from <http://festvox.org/index.html>
- [9] Building Synthetic Voices. Chapter 19. US/UK English Diphone Synthesizer. Retrieved Feb.,25, 2006 from <http://festvox.org/bsv/bsv-usukdiphone-ch.html>
- [10] Hiroaki. N. Tracheostomy Tube Enabling Speech During Mechanical Ventilation. Retrieved Feb., 25, 2006 from <http://www.chestjournal.org/cgi/content/full/125/3/1046>
- [11] Kajima, S., Takeshita, O., Yasu, K. and Arai, T. Study on noise reduction of ventilator noise from speech signals, to appear in this technical report, pp. 49-53, 2006.
- [12] Clone ensemble. Retrieved Feb., 25, 2006 from <http://www.cloneensemble.com> (Click on “Psycho Toolkit Bundle”).