Highly Intelligible Speech Synthesis for Spinal Muscular Atrophy Patients Based on Model Adaptation

Takuma Yoshimoto[1], *Ryoichi Takashima*[1], *Chiho Sasaki*[2], *Tetsuya Takiguchi*[1]

[1]Kobe University
[2]Kumamoto Health Science University

yoshimoto_t@stu.kobe-u.ac.jp, rtakashima@port.kobe-u.ac.jp

Abstract

Recently, speech signal processing technology has been used to assist people with disabilities, and the demand for such technology is increasing. In this study, we focus on spinal muscular atrophy (SMA) patients. SMA is a neuromuscular disease. Those with this disease have speech that is unclear compared to that of normal subjects which results from the use of a ventilator after a tracheotomy and from the atrophy of muscles that move the mouth, etc. Therefore, it is difficult to understand what they are saying, making communication difficult. In this paper, we analyze the speech of people with SMA and propose a text-to-speech (TTS) system to aid in communication. The proposed system uses an approach that adapts a TTS model pre-trained using normal speech to speech of a person with SMA. This system can synthesize speech having both of intelligibility derived from normal speech and individuality derived from the speech of the target subject with SMA.

Index Terms: speech synthesis, speech recognition, model adaptation, spinal muscular atrophy

1. Introduction

According to a survey conducted by the Cabinet Office [1], in Japan there are 4,360,000 people with physical disabilities, 1,094,000 with intellectual disabilities, and 4,193,000 with mental disabilities. If those with multiple disabilities are not double counted, this means that approximately 7.6% of the population has some form of disability. Among those with physical disabilities living at home, 341,000 are estimated to have hearing or speech disabilities [2]. These disabilities tend to be a major barrier to communication, and support to enable them to be able to communicate smoothly is essential.

This study focuses on dysarthria caused by spinal muscular atrophy (SMA). SMA is a type of lower motor neuron disease caused by lesions of motor nerve cells in the spinal cord [3, 4]. Most people with SMA are unable to move their bodies freely, and their voice is one of the most important means of communication for them. However, the speech of dysarthria patients, including those with SMA, differs from that of normal people in speech style, resulting in slurred speech and difficulty in their being understood [5, 6]. In recent years, text-to-speech (TTS) applications using smartphones and tablets have been developed and used to aid in the communication of patients with dysarthria. However, the voice produced by current TTS applications is based on the human voice that was used to train the model used in the application, resulting in a synthesized voice that is very different from the user's own. Users have a need to "speak in their own voice," and for this purpose, it is conceivable that the TTS model can be trained using only their own voice. However, to achieve this, a large amount of data is required, and the long recording time is very physically demanding for the user. Moreover, this will cause the synthesized voice to be as unclear as the user's original unclear voice.

There have been previous works on speech synthesis for motor neural disease (MND) and amyotrophic lateral sclerosis (ALS) (e.g. Voice Banking Project [7, 8]). There have also been several studies on speech recognition for people with dysarthria [9, 10]. On the other hand, to the best of our knowledge, there have been no studies addressing speech synthesis or recognition for SMA patients. Therefore, as the first step of our work, we measure the speech recognition accuracy of the speech of an SMA patient in order to quantitatively evaluate its intelligibility. Next, we propose a speech synthesis system to generate intelligible speech while maintaining individuality of the target SMA patient. The proposed method is based on a model adaptation approach that adapts a speech synthesis model of a normal person with intelligible speech to the speech of the target SMA patient.

2. The Voice of a Person with Spinal Muscular Atrophy

In this study, before examining the speech synthesis system, we conducted a simple automatic speech recognition (ASR) experiment to investigate how difficult it is for a person with SMA to understand speech. The spectrograms were compared with those of normal subjects, and the features of the speech of persons with SMA were analyzed.

2.1. Automatic speech recognition (ASR) experiment

2.1.1. Experimental conditions

In this experiment, we conducted a speaker-dependent isolated word-recognition experiment on one person with spinal muscular atrophy (SMA) and one healthy subject. Models were trained and evaluated using the Hidden Markov Model Toolkit (HTK) [11]. Speech recordings of one female SMA patient (label: DYS) uttering 216 phonetically-balanced words in the ATR digital speech database [12], repeated five times per word, were used for the speech recordings of a person with SMA. However, because some words were not recorded, 210 words were actually uttered five times, five words were uttered only four times, and one word was not uttered at all. The speech data of the normal subjects were obtained from 216 phonetically-balanced words spoken by one woman (label: FTK) in the corpus, and five variations per word (including the original) were created by manipulating the speech speed and pitch. The one of five utterances were used as evaluation data, and the remaining four were used as training data. For example, when recognizing the first utterance, the second to fifth utterances were used as training data. This was done for each of the five utterances, and the

Table 1: *Phoneme list*

I	N	U	a	b	by
ch	cl	d	e	f	g
gy	h	hy	i	j	k
ky	m	my	n	ny	o
p	py	r	ry	s	sh
t	ts	u	w	y	z

Table 2: *ASR results*

model	DYS	FTK
word	73.88%	100%
phoneme	15.41%	100%

recognition results were calculated as the average of the recognition rate.

The acoustic features used in the experiment are the 12-dimensional mel frequency cepstrum coefficients (MFCC) and their first derivative. The sampling rate of the speech was 16 kHz, the hamming window length was 25 msec, and the frame shift was 10 msec.

The GMM-HMM word and phoneme models were used for the automatic speech recognition (ASR) model. The number of states for the HMM was set to 3 (excluding start/end states) and the number of mixtures for the GMM was set to 4. The dictionary was designed to include only 216 phonetically-balanced words. Although there are multiple definitions of the Japanese phonetic system, we used the 36 phoneme models shown in Table 1.[1]

2.1.2. Automatic speech recognition results

Table 2 shows the experimental results. The recognition rate of a healthy subject was 100% for both the word and phoneme models, regardless of which set of utterances was used for evaluation. On the other hand, there was a large difference in the recognition rate between the word model and the phoneme model for the speech of a person with SMA. The result of the word model showed that the recognition rate of the SMA subjects was not as high as that of normal subjects, but they were able to distinguish words fairly well. However, the extremely low result of the phoneme model indicate that the phoneme model was not learned well. The reason for this may be that the phonetic system of SMA patients does not match that of normal subjects.

2.2. Comparison of spectrograms

Actually, the speech of dysarthria patients is more likely to lack high-frequency components and to have phoneme spacing than the speech of those with normal speech. As an example, Figure 1 shows the spectrogram and phoneme alignment of "ikioi /i k i o i/" uttered by a normal person and a person with SMA. Compared to the speech of normal subjects the speech of SMA patients ...

- the power of the high-frequency component is weaker than that of the low-frequency component,
- the duration of each phoneme is not constant (in the figure, the second phoneme /i/ is prolonged),

[1]In addition, we also provided 'pau,' which represents the part of speech that is not speech.

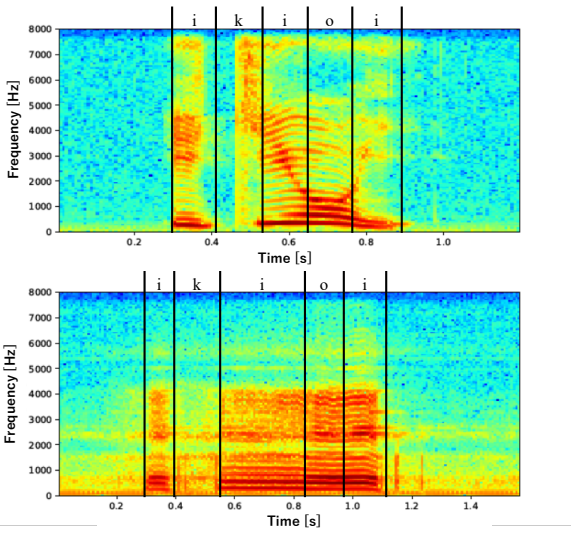

Figure 1: *Sample spectrograms of a physically unimpaired person (top) and a person with SMA (bottom)*

- changes in the vowel not being clear (in the figure, the change from the phoneme /o/ to the phoneme /i/ cannot be judged from the spectrogram alone), etc.

As seen in the spectrograms of normal subjects, consonants contain many high-frequency components, consonants are particularly difficult to understand in the speech of people with SMA, whose high-frequency components are weak.

3. Speech synthesis system using speaker-adaptation

Figure 2 shows an overview of the system proposed in this study. The arrows in both directions in the figure represent the loss function in learning, and in this study, the mean squared error (MSE) is used for both. The labels here refer to full-context labels [13] as shown in Figure 3.

For training, as in conventional text-to-speech (TTS) speech synthesis systems, the first step is to train two models using a large amount of normal human speech data and corresponding labels — a duration model that estimates phoneme duration and an acoustic model that estimates acoustic features. Both models consist of three layers of bidirectional LSTM (long short-term memory) [14, 15] having 1,024 cells in each layer. Next, of the two learned models, only the acoustic model is replicated. Finally, speaker adaptation is carried out by retraining the replicated acoustic model using a small amount of speech data from a person with SMA and the corresponding labels.

During synthesis, the duration of each phoneme in the input text is first estimated using a duration model trained on a healthy subject's data. Next, a speaker-adapted acoustic model is used to estimate acoustic features from linguistic features at the frame level based on the previously estimated phoneme duration. Finally, synthesized speech is created based on the estimated acoustic features.

The acoustic features are estimated using a speaker-adapted acoustic model, whereas the duration of each phoneme is estimated using the duration model learned on healthy subject's data. This is done in order to deal with the problem that for those with SMA, the duration of each phoneme is not con-

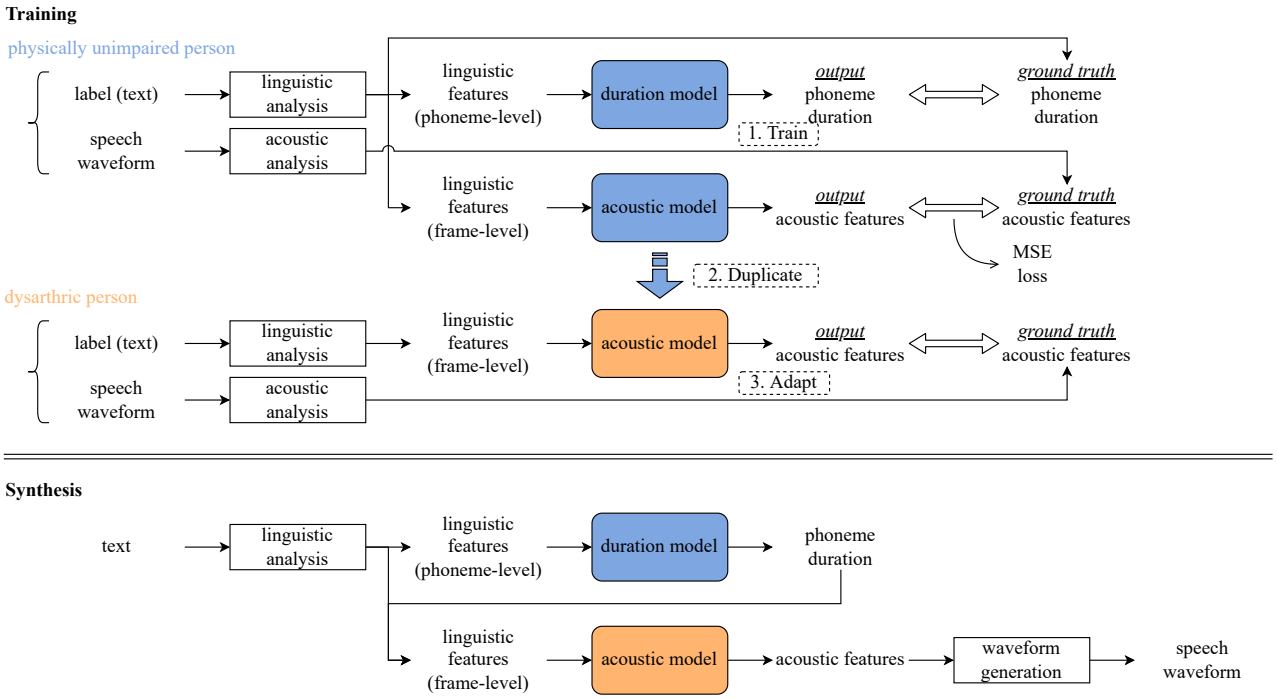

Figure 2: *Overview of the proposed TTS system*

```
3900000 4850000 xx^sil-a+r=a/A:-2+1+4/B:xx-xx_xx/C:07_xx+xx/D:02+xx_xx/E:xx_xx!xx_xx-xx/F:4_3#0_xx@1_2|1_9/G:5_5%0_xx_1/H:xx_xx/I:2-9@1+2&1-5|1+26/J:3_17/K:2+5-26
4850000 5100000 sil^a-r+a=y/A:-1+2+3/B:xx-xx_xx/C:07_xx+xx/D:02+xx_xx/E:xx_xx!xx_xx-xx/F:4_3#0_xx@1_2|1_9/G:5_5%0_xx_1/H:xx_xx/I:2-9@1+2&1-5|1+26/J:3_17/K:2+5-26
```

Figure 3: *An example of labels*

stant, which is one of the characteristics of their speech. However, if this method is used without modification, the individuality of the speaker, such as speech speed, is lost. Therefore, the proposed method utilizes the average duration of phoneme durations of a target SMA patient to denormalize the normalized phoneme duration output from the duration model. The phoneme duration $d^{(syn)}$ used to create frame-level linguistic features can be expressed as

$$d^{(syn)} = d^{(norm)} \times s_{un} + \bar{d}_{dys} \quad (1)$$

where $d^{(norm)}$ is the normalized phoneme duration, s_{un} is the standard deviation of the phoneme duration of the normal subject, and $\bar{d}_{dys}$ is the mean phoneme duration of the speaker with SMA.

4. Speech synthesis experiment

4.1. Experimental conditions

In this experiment, the recorded speech of a person with SMA was the same as the recorded speech used in the automatic speech recognition (ASR) experiment (Section 2.1.). For the normal subject's speech, 503 phoneme-balanced sentences from the ATR digital speech database were used. The sampling rate of the speech was 16 kHz, and the frame shift was 5ms. All phoneme segmentation (mapping phonemes to their start and end times) was done manually for the SMA subject. The full-context labels for the normal subject and the SMA subject were created using the front-end of Open JTalk [16]. We implemented our proposed method by modifying a parametric speech synthesis toolkit [17]. As the vocoder, which converts the synthesized acoustic features into the speech waveform, we used WORLD toolkit [18, 19].

The acoustic features used in this experiment consist of a 60-dimensional melcepstrum, a band aperiodicity parameter (BAP), a logarithmic fundamental frequency (F0), and a voiced/unvoiced flag. In addition to static features, dynamic features up to the second order are included except for the voiced/unvoiced flag. The acoustic features were normalized (standardized) to have a mean of 0 and variance of 1 for each dimension during training. The number of dimensions for the linguistic features was 975 (979 for the frame-level features, which include additional frame features), and min-max normalization was performed so that the minimum was 0 and the maximum was 1 for each dimension.

The synthesized speech obtained in the experiment was evaluated using a subjective evaluation experiment, in which intelligibility and individuality were evaluated. The intelligibility was evaluated by comparing the raw speech of a person with SMA and the synthesized speech by AB evaluation. ABX evaluation was used for individuality evaluation to determine whether the synthesized speech was more similar to the raw speech of a person with SMA or to the speech of a person with normal speech.

4.2. Experimental results

4.2.1. Spectrogram changes

As an example, Figure 4 compares the spectrogram of the synthesized speech of the word "zenshu" with the original recorded

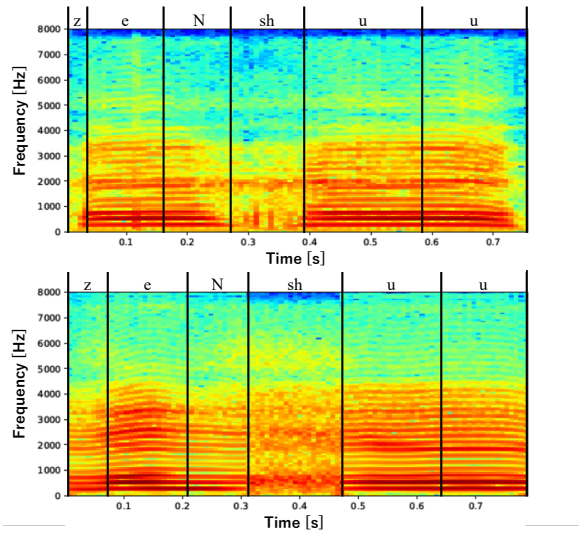

Figure 4: *Sample spectrograms of recorded speech (top) and synthesized speech (bottom)*

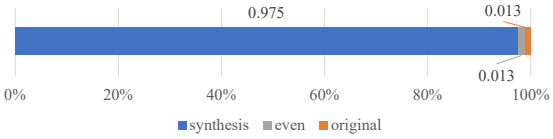

Figure 5: *Subjective evaluation of intelligibility*

speech of a person with SMA. Focus on the phoneme /sh/ in the figure. Since /sh/ is a fricative sound, the high-frequency component is large in the speech of normal subjects. However, in the recorded speech of a person with SMA, the high-frequency component of the part corresponding to /sh/ is much smaller. In contrast, in the synthesized speech, the high-frequency component, especially at 5,000 to 6,000 Hz, is larger than in the recorded speech, suggesting an improvement in intelligibility.

4.2.2. *Subjective evaluation results*

First, the results of the intelligibility evaluation are shown in Figure 5. The legends "synthesis," "even," and "original" indicate the choice of "synthesized speech is more intelligible," "not distinguishable," and "original recorded speech is more intelligible," respectively.

In most of the evaluations, the synthesized speech was superior to the original recorded speech in intelligibility. Some evaluators said that the intelligibility of the "S" line sound showed a large difference in intelligibility. As mentioned in the previous section, the intelligibility of consonants was improved by reinforcing the high-frequency component, and the intelligibility was improved by suppressing the variation in the duration of each phoneme.

Figure 6 shows the results of the individuality evaluation. The legends DYS and FTK in the figure indicate the selection of "synthesized speech is close to the recorded speech of a person with SMA" and "synthesized speech is close to the speech of a normal person," respectively.

The synthesized speech was predominantly close to the speech of SMA patients, at around 60%. Although successful

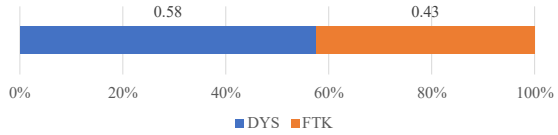

Figure 6: *Subjective evaluation of individuality*

adaptation of the model was able to make many speech sounds more similar to the speech of a person with SMA, it was necessary to reduce adaptation so that intelligibility was not compromised, and the effect of this adaptation may have affected the speaker's individuality.

5. Discussion and conclusion

In this study, we analyzed the speech of people with spinal muscular atrophy (SMA) using automatic speech recognition (ASR) system, and investigated speech synthesis to improve the cause of intelligibility. We showed that speech synthesis based on speaker adaptation of a normal human acoustic model can produce synthesized speech with improved intelligibility while maintaining individuality of speech. Speaker adaptation is a key factor in speech synthesis.

At this stage, in carrying out speaker adaptation, we have only considered the loss between calculated acoustic features by speech of a person with SMA and estimated features by the acoustic model. In order to prevent the loss of intelligibility due to too much adaptation, it may be necessary to consider the loss with the features of normal subjects during speaker adaptation. In our future works, we will explore more suitable model structure and the training methodology in order to synthesize speech having more individuality. In addition, we will evaluate our proposed method on more SMA subjects. We will also investigate objective evaluation criteria that can quantitatively evaluate the intelligibility and individuality of synthesized speech (e.g., speech recognition accuracy and speaker recognition accuracy).

6. References

[1] Cabinet Office, *Annual Report on Government Measures for Persons with Disabilities (Summary) 2020*. Government of Japan, 2020.

[2] Ministry of Health, Labour and Welfare, *Annual Health, Labour and Welfare Report 2020*. Government of Japan, 2020.

[3] G. Hamilton and T. H. Gillingwater, "Spinal muscular atrophy: going beyond the motor neuron," *Trends in molecular medicine*, vol. 19, no. 1, pp. 40–50, 2013.

[4] M. R. Lunn and C. H. Wang, "Spinal muscular atrophy," *The Lancet*, vol. 371, no. 9630, pp. 2120–2133, 2008.

[5] G. Zappa, A. LoMauro, G. Baranello, E. Cavallo, P. Corti, C. Mastella, and M. A. Costantino, "Intellectual abilities, language comprehension, speech, and motor function in children with spinal muscular atrophy type 1," *Journal of neurodevelopmental disorders*, vol. 13, no. 1, pp. 1–11, 2021.

[6] L. Pennington, N. K. Parker, H. Kelly, and N. Miller, "Speech therapy for children with dysarthria acquired before three years of age," *Cochrane Database of Systematic Reviews*, no. 7, 2016.

[7] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, 2012.

[8] C. Veaux, J. Yamagishi, and S. King, "Using hmm-based speech synthesis to reconstruct the voice of individuals with degenerative

speech disorders," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[9] R. Takashima, T. Takiguchi, and Y. Ariki, "Two-step acoustic model adaptation for dysarthric speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6104–6108.

[10] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition." in *Interspeech*, 2018, pp. 471–475.

[11] S. J. Young and S. Young, "The HTK hidden Markov model toolkit: Design and philosophy," 1993.

[12] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H. Kuwabara, and K. Shikano, "ATR Japanese speech database as a tool of speech recognition and synthesis," *Speech communication*, vol. 9, no. 4, pp. 357–363, 1990.

[13] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0." in *SSW*, 2007, pp. 294–299.

[14] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[16] "Open JTalk," http://open-jtalk.sourceforge.net/.

[17] https://github.com/r9y9/nnmnkwii.

[18] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[19] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.