Dysarthric Speech Modification Using Parallel Utterance Based on Non-negative Temporal Decomposition

Ryo Aihara, Tetsuya Takiguchi, and Yasuo Ariki

Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, Japan

aihara@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

Abstract

We present in this paper a speech modification method for a person with dysarthria resulting from athetoid cerebral palsy. The movements of such speakers are limited by their athetoid symptoms, and their consonants are often unstable or unclear, which makes it difficult for them to communicate. In this paper, duration and spectral modification using Non-negative Temporal Decomposition (NTD) is applied to a dysarthric voice. F0 is also modified by using linear-transformation. In order to confirm the effectiveness of our method, objective and subjective tests were conducted, and we also investigated the relationship between the intelligibility and individuality of dysarthric speech.

Index Terms: speech modification, dysarthria, Non-negative Temporal Decomposition

1. Introduction

Dysarthria refers to a kind of speech disorder resulting from disturbances in the form or function of the speech mechanism. Some nervous system diseases, such as Parkinson's disease or amyotrophic lateral sclerosis (ALS), produce motor paralysis which results in dysarthric speech.

In this paper, we focused on a person with dysarthria resulting from the athetoid type of cerebral palsy. Cerebral palsy is a non-progressive disorder of movement, and most people with cerebral palsy are born with the athetoid type. About two babies in 1,000 are born with cerebral palsy [1]. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. Three general times are given for the onset of the disorder: before birth, at the time of delivery, and after birth. Cerebral palsy is classified into the following types: 1) spastic, 2) athetoid, 3) ataxic, 4) atonic, 5) rigid, and a mixture of these types [2].

Athetoid symptoms develop in about 10-15% of people with cerebral palsy [1]. In the case of a person with this type of dysarthria, his/her movements are sometimes more unstable than usual. That means their utterances (especially their consonants) are often unstable or unclear due to the athetoid symptoms. Athetoid symptoms also restrict the movement of their arms and legs. Most people with athetoid cerebral palsy cannot communicate by sign language or writing, so there is great need for voice systems for them.

In [3], we proposed robust feature extraction based on principal component analysis (PCA), which has more stable utterance data, instead of DCT. In [4], we used multiple acoustic frames (MAF) as an acoustic dynamic feature to improve the recognition rate of a person with dysarthria, particularly for speech recognition using dynamic features only. In spite of these efforts, the recognition rate of dysarthric speech. is still lower than that of non-dysarthric speech. The recognition rate using a speaker-independent model, which is trained by non-dysarthric speech, is 3.5% [3]. This recognition rate suggests that for people who have not communicated with a person with athetoid cerebral palsy, it will be very hard for them to understand what that person is trying to say.

Text-to-speech synthesis (TTS) has been applied to a person with dysarthria in recent years. Veaux *et al.* [5] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from ALS. Yamagishi *et al.* [6] proposed a project which is named "Voice Banking and reconstruction". In that project, various types of voices are collected, and they proposed TTS for ALS using that database. We also proposed TTS for a person with dysarthria using HMM-based speech synthesis [7]. However, in general, TTS systems need a large amount of training data. In [7], we used more than 500 sentences of dysarthric speech to construct a speaker-dependent model.

Voice conversion (VC) has also been applied to dysarthric speech. The difference between TTS and VC is that TTS needs text input to synthesize speech, whereas VC does not need text input. In [8], we proposed VC system for dysarthric speech and improved the intelligibility of dysarthric words. The amount of the training data for a VC system is less than that for a TTS system; however, more than 200 words, or 50 sentences, are used to construct dysarthric speech model. A large amount of the training data is a high hurdle for practical use of, especially for people with athetoid cerebral palsy.

Speech modification systems for dysarthric speech, that are different from TTS or VC have also been proposed. In this paper, speech modification refers to a kind of voice transformation, which transforms an input labeled speech signal by performing a detailed speech analysis. Kain *et al*. [9] proposed speech modification for the vowel portion of dysarthric speech. Rudzicz [10] proposed a speech modification method for people with dysarthria based on the observations from the database. In general, speech modification needs less training data than TTS. Moreover, with a speech modification system, it is easier to preserve the speaker individuality of dysarthric speech than VC with a system.

This paper proposes a dysarthric speech modification system using parallel utterances in order to improve the intelligibility of dysarthric utterances. Non-negative Temporal Decomposition (NTD) [11] has been proposed in the field of speech coding and it is applied to the rhythm conversion of non-native English. We applied NTD to dysarthric speech. Duration (rhythm) of dysarthric speech is transformed into that of parallel non-dysarthric speech. The consonants of dysarthric speech are also

SLPAT 2016, 7th Workshop on Speech and Language Processing for Assistive Technologies, pages 75–79,
San Francisco, USA, 13 September, 2016.

replaced with the consonants of non-dysarthric speech based on NTD. F0 is also modified by using linear-transformation. The effectiveness of our method is evaluated by using mean opinion score (MOS) [12] test, and we investigated the relationship between the intelligibility and individuality of dysarthric speech.

The rest of this paper is organized as follows: In Section 2, the summary of the NTD algorithm is described. In Section 3, our proposed method is explained. In Section 4, the experimental data are evaluated, and the final section is devoted to our conclusions.

2. Non-negative Temporal Decomposition

In NTD, the i-th dimensional spectrum, $v_i(t)$ of time t is decomposed into spectral event basis $w_{i,l}$ and activity $h_l(t)$. The problem of NTD is defined as follows:

$$\begin{aligned} min \quad & \sum_{l=2}^{L}\sum_{t=t_{l-1}}^{t_l}\sum_{i=1}^{I}(v_i(t)-w_{i,l}h_l(t)-w_{i,l-1}h_{l-1}(t))^2 \\ s.t. \quad & h_l(t)\geq 0, h_{l-1}(t)\geq 0 \\ & w_{i,l}\geq 0, w_{i,l-1}\geq 0 \\ & h_l(t)+h_{l-1}(t)=1 \quad for \quad \forall t,i,l \end{aligned} \tag{1}$$

where l denotes the number of bases and t_l denotes the event timing of l-th basis. By applying the last constraint, activities are restricted to the range $[0, 1]$.

(1) is rewritten into the cost function as follows:

$$\begin{aligned} & d(\mathbf{V}, \mathbf{WH}) \\ = & \sum_{l=2}^{L}\sum_{t=t_{l-1}}^{t_l}\sum_{i=1}^{I}(v_i(t)-w_{i,l}h_l(t)-w_{i,l-1}h_{l-1}(t))^2 \\ & +\alpha\sum_{l=2}^{L}\sum_{t=t_{l-1}}^{t_l}(h_l(t)+h_{l-1}(t)-1)^2 \end{aligned} \tag{2}$$

where $1 = t_1 < t_2 < \cdots < t_l = T$. The second term of (2) is a penalty term to satisfy $h_l(t) + h_{l-1}(t) = 1$ with α as is its weight.

(2) is minimized by iteratively updating (3) - (5), which is shown at the top of the next page. These updating rules are derived in [11]. In [11], line spectral pair (LSP) is used as a spectral feature; however, we use a magnitude spectrum to estimate the event basis and activity more precisely. Moreover in [11], each event basis corresponds to a single phoneme. In order to estimate the event basis and activity more precisely, 3 event bases are extracted from a single phone.

3. Modification of Dysarthric Speech

3.1. Flow of our proposed method

Fig. 1 shows the flow of our speech modification process. First, a dysarthric utterance and a non-dysarthric utterance, which is parallel to the dysarthric utterance, are labeled by using HMM-based forced alignment. Here, parallel means that the utterances are spoken by different speakers, but the text is the same. Then we extract spectral features, F0, and aperiodic features from the parallel utterances by using STRAIGHT analysis [13]. The duration and extracted spectral features are modified by using NTD. The extracted F0 is also modified using linear conversion. The modified spectra and F0, and the aperiodic features of the dysarthric speech are synthesized using STRAIGHT synthesis [13].

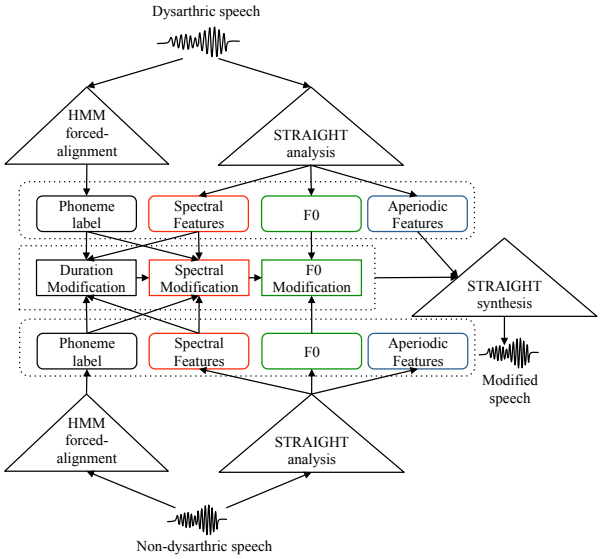

Figure 1: Flow of dysarthric speech modification

3.2. Duration modification

The duration of dysarthric speech tends to be longer than non-dysarthric speech [10]. In [7], we investigated that average duration per mora in 50 dysarthric sentences is 1.3 times slower than that of non-dysarthric speech. We modified the duration of dysarthric speech to that of non-dysarthric speech by using NTD.

First, parallel utterances of dysarthric and non-dysarthric speech are decomposed into the dictionary and activity. We refer to the basis set as the dictionary. Fig. 2 shows the flow of the decomposition. The dysarthric spectrum $\mathbf{V}^s \in \mathbb{R}^{(I\times J)}$ is decomposed into the source dictionary $\mathbf{W}^s \in \mathbb{R}^{(I\times L)}$ and its activity $\mathbf{H}^s \in \mathbb{R}^{(L\times J)}$ using NTD.

$$\mathbf{V}^s \approx \mathbf{W}^s\mathbf{H}^s \tag{6}$$

The non-dysarthric spectrum $\mathbf{V}^t \in \mathbb{R}^{(I\times K)}$ is decomposed into the target dictionary $\mathbf{W}^t \in \mathbb{R}^{(I\times K)}$, and its activity $\mathbf{H}^t \in \mathbb{R}^{(K\times J)}$ is the same way the dysarthric spectra is.

$$\mathbf{V}^t \approx \mathbf{W}^t\mathbf{H}^t \tag{7}$$

In NTD, the l-th event timing t_l is defined with the center frame of l-th phoneme label. Therefore, $\mathbf{W}^s$ and $\mathbf{W}^t$ will be parallel.

The durations of dysarthric spectra is modified as shown in Fig. 3.

$$\mathbf{V}^{s\to t} = \mathbf{W}^s\mathbf{H}^t \tag{8}$$

Because we use the source dictionary for duration modification, only the duration is modified.

3.3. Spectral modification

In general, the vowels voiced by a speaker strongly indicate the speaker's individuality. On the other hand, the consonants of people with dysarthria are often unstable. In [8], in order to improve the intelligibility of dysarthric utterances, we converted dysarthric consonants into non-dysarthric ones. Based on the same idea, we use a "combined-dictionary" that consists of the

$$w_{i,l} \leftarrow \frac{\sum_{t=t_{l-1}}^{t_{l+1}} v_i(t)h_l(t)}{\sum_{t=t_{l-1}}^{t_l-1}(w_{i,l-1}h_{l-1}(t)h_l(t)+w_{i,l}h_l^2(t))+\sum_{t=t_l}^{t_{l+1}}(w_{i,l+1}h_l(t)h_{l+1}(t)+w_{i,l}h_l^2(t))} w_{i,l} \quad (3)$$

$$h_l(t) \leftarrow \frac{\sum_{i=1}^{I} w_{i,l}v_i(t)+\alpha}{\sum_{i=1}^{I}(w_{i,l-1}w_{i,l}h_l(t)+w_{i,l}^2h_l(t))+\alpha(h_{l-1}(t)+h_l(t))} h_l(t) \quad (4)$$

$$h_{l-1}(t) \leftarrow \frac{\sum_{i=1}^{I} w_{i,l-1}v_i(t)+\alpha}{\sum_{i=1}^{I}(w_{i,l-1}w_{i,l}h_l(t)+w_{i,l-1}^2h_{l-1}(t))+\alpha(h_{l-1}(t)+h_l(t))} h_l(t) \quad (5)$$

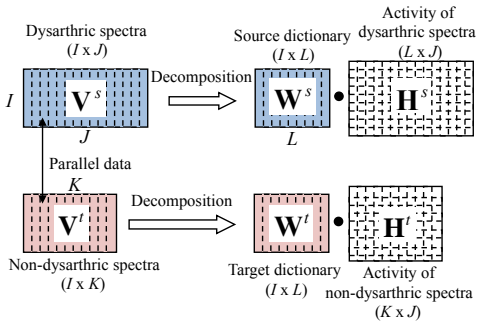

Figure 2: Decomposition using NTD

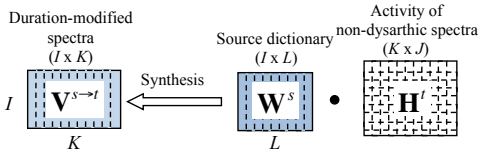

Figure 3: Duration modification

bases of dysarthric vowels from the source dictionary and bases of non-dysarthric consonants from the target dictionary.

The dysarthric spectra $\mathbf{V}^{s\to t}$ are modified as shown in Fig. 4 where $\hat{\mathbf{W}}^{st}$ denotes the combined-dictionary.

$$\hat{\mathbf{V}}^{s\to t} = \hat{\mathbf{W}}^{st}\mathbf{H}^t \quad (9)$$

By using the combined-dictionary, only consonants are modified, and we can preserve the speaker's individuality. Moreover, by using target activity, the duration of dysarthric speech is also modified.

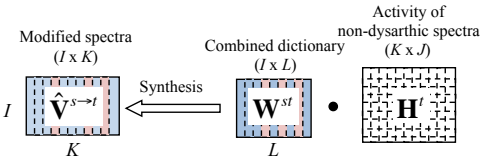

Figure 4: Spectral modification

3.4. F0 modification

Fig. 5 shows an example of non-dysarthric F0. Fig. 6 shows an example of dysarthric F0, which is a parallel utterance of Fig. 5. Although these utterances are parallel, F0 trajectories are different between the two utterances. In a TTS system [7] the F0 model is trained from non-dysarthric speech in order to synthesize an intelligible voice.

In the proposed F0 modification, we use non-dysarthric F0, which is linearly transformed in order to preserve the source speaker's individuality as follows:

$$f0^{conv}(t) = \frac{\sigma^{(s)}}{\sigma^{(t)}}(f0^t(t)-\mu^{(t)})+\mu^{(s)}, \quad (10)$$

where $f0^s(t)$, $f0^t(t)$, and $f0^{conv}(t)$ denote log-scaled F0 of dysarthric speech, non-dysarthric speech, and modified speech at frame t, respectively. $\mu^{(s)}$ and $\sigma^{(s)}$ denote the mean and standard deviation of the log-scaled F0, as calculated from dysarthric speech. $\mu^{(t)}$ and $\sigma^{(t)}$ are the mean and standard deviation of non-dysarthric speech.

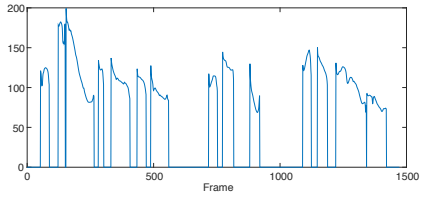

Figure 5: Example of non-dysarthric F0

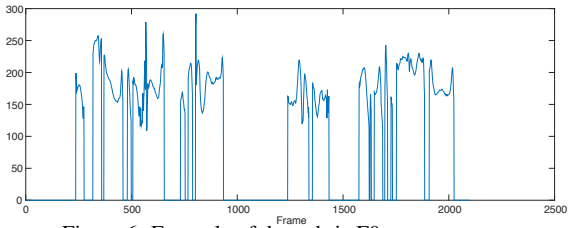

Figure 6: Example of dysarthric F0

4. Experimental Results

4.1. Experimental Conditions

The proposed method was evaluated on sentence-based speech modification for one Japanese male with dysarthric speech. We recorded 50 sentences, which are found in the ATR Japanese database [14]. The speech signals were sampled at 12 kHz, and the frame shift was 5 ms.

Label data were obtained by HMM-based forced alignment using HTK. In the case of dysarthric speech, it is difficult to

obtain precise label data using HMM because some phonemes in dysarthric speech fluctuated due to the speaker's inability to speak clearly. Moreover, dysarthric speech includes the unexpected sound of breath. Therefore, some labels are replaced.

Acoustic and prosodic features were extracted using STRAIGHT. Duration and spectra are modified using NTF. The dictionary is initialized with the spectra of the center frame of each phoneme. The activity of the l-th event area (between t_{l-1} and t_{l+1}) is initialized with positive random values. The number of dimensions of STRAIGHT spectra is 513. α is set at 100.

We conducted the objective evaluation to evaluate the precision of the decomposition of NTF. We used log spectrum distance (LSD) as a measurement.

$$LSD[dB] = \sqrt{\frac{1}{I}\sum_{i}^{I}(20\log_{10}\frac{v_i^s(t)}{v_i^{conv}(t)})^2} \qquad (11)$$

We compared 3 methods: 1) duration modification, 2) duration and spectral modification, and 3) duration, F0, and spectral modification. We conducted subjective evaluations using a 5-scale MOS test. A total of 10 Japanese speakers took part in the listening test using headphones. We evaluated both the aspect of listening intelligibility and the aspect of speaker similarity. For listening intelligibility, dysarthric speech and non-dysarthric speech are presented as reference voices, and the opinion score was set as follows: (5: very intelligible, just like non-dysarthric speech, 4: intelligible, like non-dysarthric speech, 3: fair, 2: not so intelligible, like dysarthric speech, 1: unintelligible, just like dysarthric speech). For speaker similarity, dysarthric speech and non-dysarthric speech are also presented as the references, and the opinion score was set as follows: (5: very similar to a person with dysarthria, 4: similar to a person with dysarthria, 3: fair, 2: similar to a physically unimpaired person, 1: very similar to a physically unimpaired person).

4.2. Results and Discussion

We evaluated log spectrum distance (LSD) using the different number of bases in the dictionary, and the results are shown in Table 1. We obtained a better result when we used three bases for one phoneme than when the number of the bases is the same as that of the phoneme (default setting as [11]). The LSD of dysarthric speech is worse than that of non-dysarthric speech. We assume that this is because dysarthric speech fluctuates.

Table 1: LSD of using different dictinoaries

#basis of phoneme	Dysarthric [dB]	Non-dysarthric [dB]
1	2.33	2.17
3	1.93	1.52

Fig. 7 and Fig. 8 show an example of non-dysarthric spectra and dysarthric spectra, respectively. Comparing Fig. 7 to Fig. 8, the duration of dysarthric speech tends to be long and dysarthric spectra have weak energy. Fig. 9 shows an example of duration-modified spectra. Fig. 10 shows an example of duration and spectrum-modified spectra.

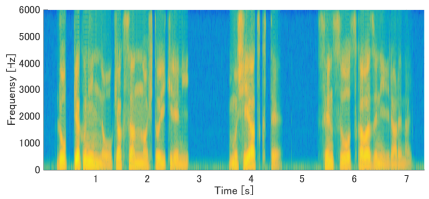

Figure 7: Example of non-dysarthric spectra

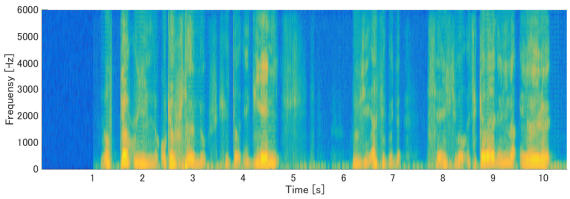

Figure 8: Example of dysarthric spectra

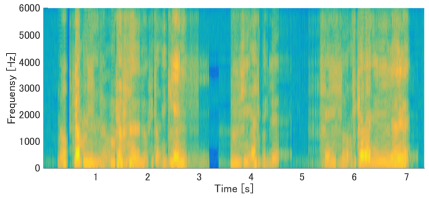

Figure 9: Example of duration-modified spectra

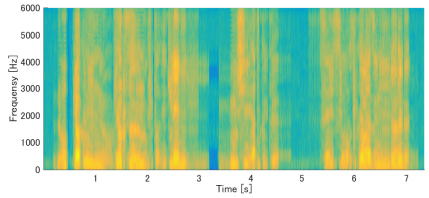

Figure 10: Example of duration and spectrum-modified spectra

Fig. 11 and Fig. 12 show the results of the subjective evaluation on intelligibility and similarity to a person with dysarthria, respectively. Error bars show 95% confidence area, and the results are confirmed with the p-value test result of 0.05. Fig. 11 shows that duration, spectrum, and F0 modification significantly improve the intelligibility of dysarthric speech. Duration- and spectrum-modified speech (without F0 modification) is slightly improved the intelligibility of dysarthric speech. Fig. 12 implies, that because we focus on consonants in spectrum modification, duration and spectrum modification preserve speaker individuality. Considering the results shown in Fig. 11 and Fig. 12, F0 is important for improving intelligibility, and speaker similarity is also impacted greatly by it.

5. Conclusions

We proposed speech modification for dysarthric speech resulting from athetoid cerebral palsy. Input dysarthric speech is labeled by HMM-based forced alignment. Using the label data and parallel non-dysarthric speech, the duration, spectra, and F0

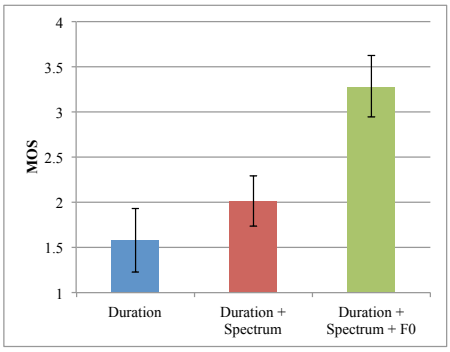

Figure 11: MOS test on intelligibility

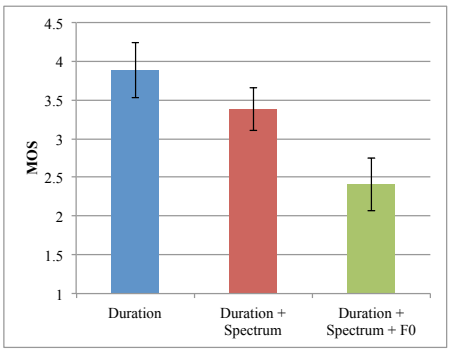

Figure 12: MOS test on similarity

are modified to improve the intelligibility of dysarthric speech. We applied NTF for dysarthric duration and spectrum modification and linear-conversion for dysarthric F0.

Using a subjective testing approach, we investigated the relationship between modified features, intelligibility and similarity to a person with dysarthria. Experimental results show that our speech modification effectively improved the intelligibility of dysarthric speech. However, it was also confirmed that speaker similarity is quite sensitive to F0. Therefore, intelligibility-preserving F0 modification will be the subject of future work. In this paper, there was only one test subject, so in future experiments, we will increase the number of test subjects and further examine the effectiveness of our method. Future work will also include efforts to study the co-articulation effect between phonemes.

6. References

[1] M. V. Hollegaard, K. Skogstrand, P. Thorsen, B. Norgaard-Pedersen, D. M. Hougaard, and J. Grove, "Joint analysis of SNPs and proteins identifies regulatory IL18 gene variations decreasing the chance of spastic cerebral palsy," *Human Mutation, Vol. 34*, pp. 143–148, 2013.

[2] S. T. Canale and W. C. Campbell, "Campbell's operative orthopaedics," Mosby-Year Book, Tech. Rep., 2002.

[3] H. Matsumasa, T. Takiguchi, Y. Ariki, I. Li, and T. Nakabayachi, "Integration of metamodel and acoustic model for dysarthric speech recognition," *Journal of Multimedia*, vol. 4, no. 4, pp. 254–261, 2009.

[4] C. Miyamoto, Y. Komai, T. Takiguchi, Y. Ariki, and I. Li, "Multimodal speech recognition of a person with articulation disorders using AAM and MAF," *in Proc. IEEE International Workshop on Multimedia Signal Processing (MMSP)*, pp. 517–520, 2010.

[5] C. Veaux, J. Yamagishi, and S. King, "Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders," *in Proc. Interspeech*, 2012.

[6] 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.

[7] R. Ueda, R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving spectrum modification for articulation disorders using phone selective synthesis," *in Proc. Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2015.

[8] R. Aihara, T. Takiguchi, and Y. Ariki, "Individuality-preserving voice conversion for articulation disorders using phoneme-categorized exemplars," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 4, pp. 13:1–13:17, 2015.

[9] A. B. Kain, J. Hosom, X. Niua, J. Santen, M. Fried-Oken, and J. Staehely, "Improving the intelligibility of dysarthric speech," *Speech Communication*, vol. 49, no. 9, pp. 43–759, 2007.

[10] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech and Language*, vol. 27, no. 6, pp. 1163–1177, 2014.

[11] S. Hiroya, "Non-negative temporal decomposition of speech parameters by multiplicative update rules," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2108–2117, 2013.

[12] INTERNATIONAL TELECOMMUNICATION UNION, "Methods for objective and subjective assessment of quality," *ITU-T Recommendation P.800*, 2003.

[13] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequencybased F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[14] 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, pp. 357–363, 1990.