Individuality-Preserving Speech Synthesis System for Hearing Loss Using Deep Neural Networks

Tsuyoshi Kitamura[1], *Tetsuya Takiguchi*[1], *Yasuo Ariki*[1], *Kiyohiro Omori*[2]

[1]Graduate School of System Informatics, Kobe University, Japan
[2]Hyogo Institute of Assistive Technology, Kobe, Japan

kitamura@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

Abstract

Statistic parametric speech synthesis (SPSS) systems [1] are being widely researched in the field of speech processing. We present in this paper a speech synthesis method for people with hearing loss. Because of their disorders, their prosody is often unstable and their speech rate, pitch, and spectrum differ from those of physically unimpaired persons, which causes their speech to be less intelligible and, consequently, makes communication with physically unimpaired persons difficult. In order to deal with these problems, this paper describes an approach that makes use of a novel combination of deep neural networks (DNN)-based text-to-speech synthesis using the DNNs of a physically unimpaired person and a person with hearing loss, while preserving the individuality of a person with hearing loss. Through experimental evaluations, we have confirmed that the proposed method successfully synthesizes an intelligible speech signal from a hard-to-understand signal while preserving the target speaker's individuality.

Index Terms: hearing disorders, speech synthesis system, deep neural networks, assistive technologies

1. Introduction

In this paper, we focus on, as one assistive technology for a person with hearing loss, a speech synthesis system that assists persons in their speech communication. Their speech style may be different from those of persons without hearing loss and the utterances may be less intelligible due to hearing loss. It sometimes makes verbal communication with other people difficult.

A DNN-based speech synthesis system [2], [3], [4] is a text-to-speech (TTS) system that can generate signals from input text data. A TTS system may be useful for persons with hearing loss because the synthesized speech signal may become more intelligible by adjusting the utterance duration, pitch, and spectrum.

In this paper, we propose a DNN-based speech synthesis method for a person with hearing loss. To generate an intelligible synthesized speech signal while preserving the speaker's individuality, we use speech data from both a person with hearing loss and a physically unimpaired person. Because the speech rate of a person with hearing loss may be unstable, the duration model of a person with hearing loss is modified using the DNNs of a physically unimpaired person to stabilize the speech rate. In addition, the F_0 patterns of a person with hearing loss are often unstable. To solve this problem, in the synthesis step, the F_0 features predicted from the networks of a physically unimpaired person are used as the input of the networks of a person with hearing loss after being converted to the average F_0 of the hearing loss person using a linear transformation.

As for the spectral problem associated with a person with hearing loss, the consonant parts of utterances are sometimes unclear or unstable. To resolve the consonant problem, we generate the spectrum for some consonants from the acoustic model of a physically unimpaired person and the vowel spectrum from the acoustic model of a person with hearing loss in order to preserve the speaker's individuality.

The rest of this paper is organized as follows; In Section 2, an introduction to related work about assistive technology is presented. In Section 3, a speech synthesis system using deep neural networks is presented. Section 4 presents the proposed speech synthesis system for a person with a hearing disorder. In Section 5, in order to confirm the effectiveness of our method, the experimental data are evaluated. Finally, the conclusions are drawn in Section 6.

2. Related Works

To assist people with articulation disorders, a number of assistive technologies using information processing have been proposed. As one of the techniques used for statistic parametric speech synthesis, the Hidden Markov model (HMM)-based TTS approach [5], has been studied for a long time and a number of assistive technologies using a HMM-based TTS system have been proposed; for example, Veaux used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting form Amyotrophic Lateral Sclerosis (ALS) [6]. They have proposed a reconstruction method for degenerative speech disorders using an HMM sound synthesis system. In this method, the subject's utterances were used to adapt an average voice model pre-trained on many speakers. Creer also adapted the average voice model of multiple speakers to severe dysarthria data [7], and Khan used such an adaptation method on a laryngectomy patient's data [8]. The authors of this paper also proposed a HMM-based TTS system for people with articulation disorders [9].

Recently, deep learning has had success in speech synthesis in regard to naturalness and sound quality compared with hidden Markov models [1]. Deep neural networks contain many layers of nonlinear hidden units and represent a mapping function from linguistic features to acoustic features. In the field of speech processing technology, speech recognition (lip reading using deep learning) has also had success [10].

Recently, to develop sound quality and naturalness, the architectures of the DNN have been improved; for example, using long-short-term-memory to take the continuity of speech into account [11], and using i-vectors to adapt the average voice model of multiple speakers [12]. In the adaptation task, a small amount of speech data is required to create synthesized speech because it is difficult for a person with an articulation disorder to say many sentences.

In this paper, we employ trajectory training [3] to train DNN. Trajectory training is regarded as a successful method and has been widely used in recent years for various tasks.

3. DNN-based Speech Synthesis

Fig. 1 shows the overview of the basic approach to text-to-speech (TTS) synthesis based on deep neural networks. The figure shows the synthesis parts of a DNN-based TTS system. In the training part of networks, inputs, the linguistic features extracted from an input text-by-text analysis, are mapped to the output acoustic features (spectral, F0, and aperiodicity) using back propagation. In the synthesis part, the linguistic features are mapped to the output acoustic features using forward propagation. In the parameter generation part, the output acoustic parameters including static, delta, and acceleration parameters are generated into smooth parameter trajectories using the speech parameter generation algorithm [13]. In the waveform synthesis part, a vocoder, such as STRAIGHT [14] or WORLD [15], is used to synthesize speech signals from the smooth parameter trajectories. State duration densities are modeled by the method used in HMM-based speech synthesis systems [16] to control rhythm and tempo, where state durations of each phoneme HMM are modeled by a multi-dimensional Gaussian distribution.

DNN-based speech synthesis comprises the training part and the synthesis part. Acoustic features consist of D-dimensional static features $\mathbf{c}_t = [c_t(1), c_t(2), ..., c_t(D)]$ and corresponding dynamic features $\Delta\mathbf{c}_t$ and $\Delta^2\mathbf{c}_t$, written as

$$\mathbf{o}_t = [\mathbf{c}_t^{\mathrm{T}}, \Delta\mathbf{c}_t^{\mathrm{T}}, \Delta^2\mathbf{c}_t^{\mathrm{T}}]^{\mathrm{T}} \tag{1}$$

Dynamic features are computed from the sequence of static features. The sequence of acoustic features $\mathbf{o} = [\mathbf{o}_1^{\mathrm{T}}, \mathbf{o}_2^{\mathrm{T}}, ..., \mathbf{o}_T^{\mathrm{T}}]$ can be caluculated from the sequence of static features $\mathbf{c} = [\mathbf{c}_1^{\mathrm{T}}, \mathbf{c}_2^{\mathrm{T}}, ..., \mathbf{c}_T^{\mathrm{T}}]$ by

$$\mathbf{o} = \mathbf{W}\mathbf{c} \tag{2}$$

where T is the number of frames included in the sequence and $\mathbf{W}$ is a matrix used to extend static features $\mathbf{c}$ to acoustic features $\mathbf{o}$ [13].

In the training part, the input text is analyzed and transformed into labels, which contain linguistic contexts. The networks learn the complex mapping function from linguistic features $\mathbf{x}_t$ to acoustic features $\mathbf{o}_t$, where the frame-level mean square errors between the predicted acoustic features $\hat{\mathbf{o}}_t$ and the observed acoustic features $\mathbf{o}_t$ are minimized using the back-propagation algorithm.

In the synthesis part, output features include static, delta, and acceleration features. To generate the smooth parameter trajectories, the maximum likelihood parameter generation (MLPG) algorithm [17] is used to take the dynamic features as constraints. The smooth parameter trajectory $\hat{\mathbf{c}}$ is given by

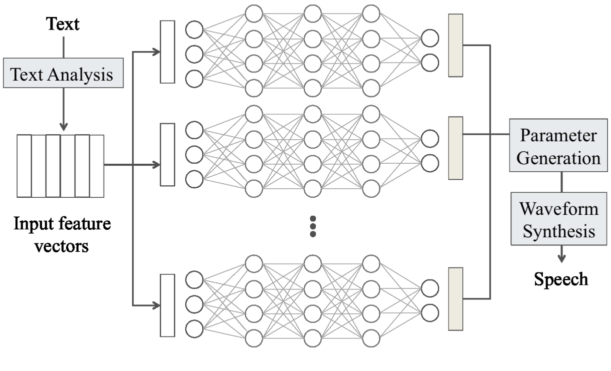


Figure 1: *A flow of speech synthesis using deep neural networks.*

$$\hat{\mathbf{c}} = \arg\max_{\mathbf{c}} P(\mathbf{o}|\boldsymbol{\lambda}) = \arg\max_{\mathbf{c}} \mathcal{N}(\mathbf{W}\mathbf{c}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \bar{\mathbf{c}} \tag{3}$$

where $\boldsymbol{\lambda}$ is the model parameter and $\mathcal{N}(|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with mean vectors μ and covariance matrix $\boldsymbol{\Sigma}$. The smooth parameter trajectories calculated by the MLPG algorithm can be written by (4).

$$\hat{\mathbf{c}} = (\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{4}$$

In the synthesis part, $\boldsymbol{\mu}$ is the frame obtained by performing a forward propagation and $\boldsymbol{\Sigma}$ is computed from the training data. We can reconstruct the speech waveform from the smooth parameter trajectory $\hat{\mathbf{c}}$ by using a vocoder.

3.1. Trajectory training

To take the interaction between the static and dynamic features into account, the trajectory training minimizes the utterance-level trajectory error than the frame-level error [3]. This training criterion is called the minimum generation error (MGE). The Euclidean distance between the predicted trajectory $\hat{\mathbf{c}}$ (calculated by MLPG) and the observed static trajectory is called the trajectory error. The objective function is written as

$$L = (\hat{\mathbf{c}} - \mathbf{c})^{\mathrm{T}}(\hat{\mathbf{c}} - \mathbf{c}) = (\mathbf{R}\hat{\mathbf{o}} - \mathbf{c})^{\mathrm{T}}(\mathbf{R}\hat{\mathbf{o}} - \mathbf{c}) \tag{5}$$

where

$$\hat{\mathbf{R}} = (\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \tag{6}$$

Mean-variance normalization is performed to $\hat{\mathbf{c}}$ and $\mathbf{c}$ before calculated the trajectory error. The mean and variance values are calculated from the training data in advance. The parameters of DNN are trained by back-propagation using gradient, as is the case with conventional frame-level training.

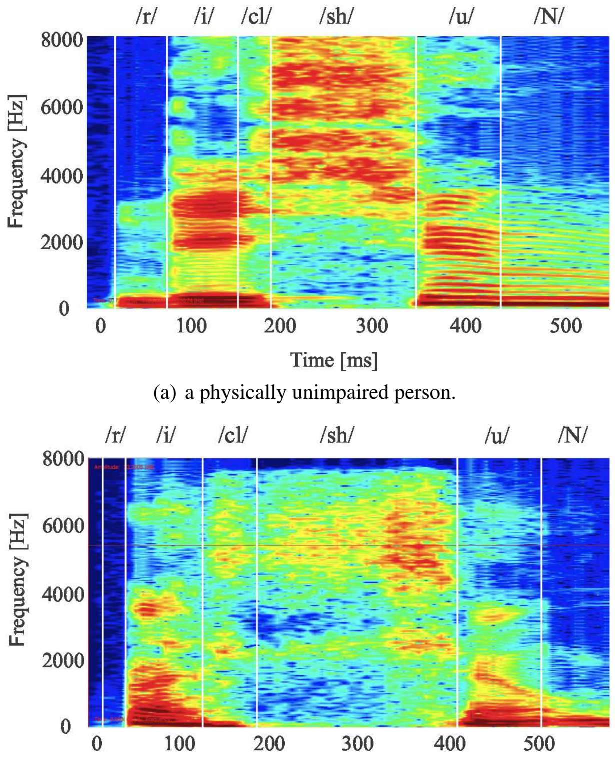
4. DNN-based Speech Synthesis for a Person with Hearing Loss

In our method, the voice of two people, a person with hearing loss and a physically unimpaired person, are used to generate a more intelligible synthesized speech signal that preserves the individuality of the person with hearing loss. Fig. 2 shows the original spectrograms for the word "/r/ /i/ /cl/ /sh/ /u/ /N/""/r/ /i/ /cl/ /sh/ /u/ /N/" of a physically unimpaired person and a person with hearing loss.

As shown in Fig. 2, the high-frequency spectral power of a person with hearing loss is weaker compared to that of a physically unimpaired person. In addition, the duration of a person with hearing loss is unstable that some phones (ex: "/cl" and "/sh/") are too long compared to other phones although the speech length is almost the same as that of a physically unimpaired person. This may be one of the reasons behind the unintelligibility. Therefore, in our method, a more intelligible synthesized speech signal that preserves the speaker's individuality is generated by using the features of both a person with hearing loss and a physically unimpaired person.

4.1. F_0 model modification

Fig. 3 shows the overview of the approach to F_0 modification. As the F_0 patterns of a person with hearing loss are often unstable, we modify it using the F_0 features of a physically unimpaired person.



(a) a physically unimpaired person.

(b) a person with hearing loss.

Figure 2: *Sample spectrograms of /r i cl sh u N/.*

In the training part, first, DNNs of a physically unimpaired person and a person with hearing loss are trained independent of each other. For the DNNs of a physically unimpaired person, the input is the linguistic features and the output is the spectral, aperiodicity, and F_0 features of a physically unimpaired person. For the DNNs of a person with hearing loss, the input is the linguistic features and the F_0 features (static, delta, and acceleration) and the output is spectral features and aperiodicity features of a person with hearing loss.

As shown in Fig. 3, in the synthesis part, first, linguistic features are mapped to the spectral, aperiodicity, and F_0 features using the DNNs of a physically unimpaired person. The output F_0 features of a physically unimpaired person are converted to those of a person with hearing loss by using the linear transformation in Eq. (7) and then, they are used as the input for networks of a person with hearing loss.

$$\hat{w}_t = \frac{\sigma_x}{\sigma_w}(w_t - \mu_w^{(F_0)}) + \mu_x^{(F_0)} \tag{7}$$

where w_t represents a log-scaled F_0 of a physically unimpaired person at the frame t, $\mu_w^{(F_0)}$ and σ_t represent the mean and standard deviation of w_t, respectively. $\mu_x^{(F_0)}$ and σ_x represent the mean and standard deviation of log-scaled F_0 of a person with hearing loss, respectively.

4.2. Duration model modification

The speech rhythm and tempo of a person with hearing loss differ from those of physically unimpaired persons, and this causes their speech to be less intelligible. To solve this problem, the speech rhythm and tempo of a physically unimpaired person are used in the synthesis part. However, as the average speech rate contains rich speaker individuality, the average speech rate of the synthesized speech signal is fit to that of a person with hearing loss. To implement these ideas, the duration model is modified as follows:

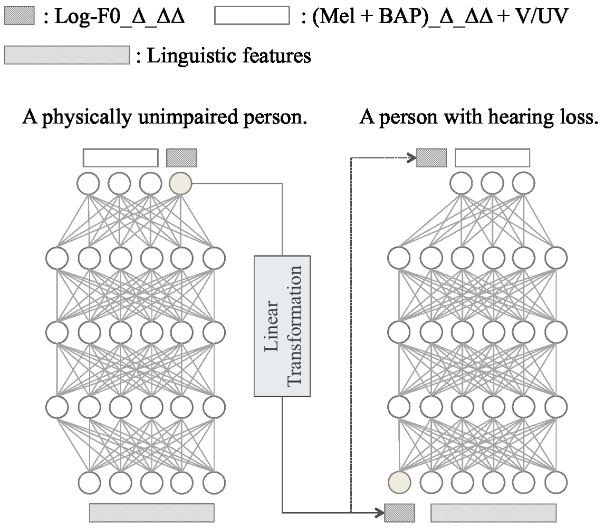

▨ : Log-F0_Δ_ΔΔ
: (Mel + BAP)_Δ_ΔΔ + V/UV
: Linguistic features
A physically unimpaired person.
A person with hearing loss.
Linear Transformation

Figure 3: *A flow of the F_0 modification method.*

$$y_i = t_i - \mu_w^{(Dur)} + \mu_x^{(Dur)} \tag{8}$$

$$\mu_w^{(Dur)} = \frac{\sum_{i=1}^{I} \mu_{ti}}{I} \tag{9}$$

$$\mu_x^{(Dur)} = \frac{\sum_{i=1}^{I} \mu_{xi}}{I} \tag{10}$$

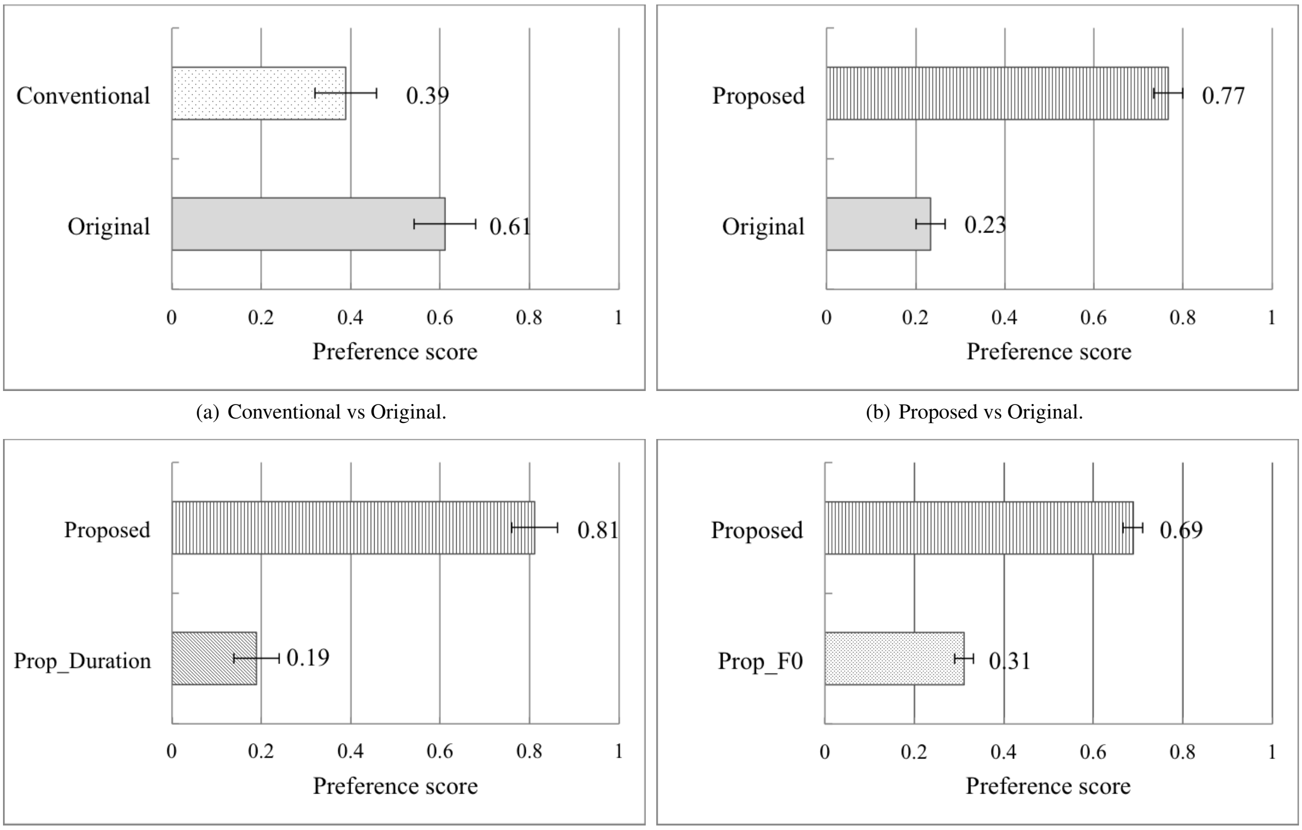
In Eq. (8), t_i represents the value of the i-th node in the duration model of a physically unimpaired person. In Eqs. (9) and (10), I represents the total number of nodes in the model, u_{ti} represents the mean the value of the i-th node in the model of a physically unimpaired person, and u_{xi} represents the mean the value of the i-th node in the model of a person with hearing loss.

5. Experiments

5.1. Experimental conditions

We prepared the training data for two men. One is a physically unimpaired person, and the other is a person with hearing loss. We used 503 sentences from the ATR Japanese database B-set for a physically unimpaired person and we recorded 503 sentences uttered by a person with hearing loss. 450 and 53 utterances were used for training and development, respectively. In addition, we recorded 10 sentences uttered by a person with hearing loss for testing. The speech signal was sampled at 16kHz and the frame shift was 5 msec. Acoustic and prosodic features were extracted using WORLD [15]. As spectral parameters, the 0-th through the 49-th mel-cepstrum coefficients [18], and their dynamic and acceleration coefficients were used. As excitation parameters, log-F_0 and 25 band-filtered aperiodicity [19] were used, along with their dynamic and acceleration coefficients.

In order to confirm the effectiveness of our method, four systems were compared.



(a) Conventional vs Original.

(b) Proposed vs Original.

(c) Proposed vs Prop_Dur.

(d) Proposed vs Prop_F_0

Figure 4: *Preference scores for the listening intelligibility based on subjective evaluations.*

- **Conventional**: DNN-based speech synthesis system using trajectory training
- **Prop_Dur**: **Conventional** + "Duration model of a person with hearing loss was modified in Eq. (8)"
- **Prop_F_0**: **Conventional** + "F_0 modification based on section 4.1"
- **Proposed**: **Prop_Dur + Prop_F_0**

In **Conventional**, input features consisted of 395 features, which comprised 386 binary and 9 numeric features. Binary features were derived from categorical linguistic contexts including quinphone identities, accent type, position of phone, mora, word, and so on. Numeric features include frame position information. Output features consisted of 50 mel-cepstrum coefficients, log-F_0, and 25 band-filtered aperiodicity, their deltas and accelerations coefficients, and a voiced/unvoiced value (3 + (50 + 25 + 1) + 1 = 229). Input features were normalized to the range 0.0-1.0 based on min-max and output features were normalized to zero mean and unit variance. To reduce the training cost, silence frames were removed from the training data of a person with hearing loss. The architecture of the networks was 4-hidden-layers, with each hidden layer containing 700 units. The sigmoid activation function was used for hidden layers, and the linear activation function was used for the output layer. In order to complement some consonant parts of a person with hearing loss, the consonants /s, sh, k, t, ts, z, ch/ were replaced by those of a physically unimpaired person.

In order to evaluate the models, we evaluated the listening intelligibility and the speaker similarity by listening to voices recorded and synthesized under the five conditions (original speech, **Conventional**, **Prop_Dur**, **Prop_F_0**, **Proposed**). A total of 9 Japanese speakers took part in the listening test using headphones. For speaker similarity, a DMOS (Degradation Mean Opinion Score) test was performed. In the DMOS test [20], the original speech signal was used as the reference signal and the option score was set to a 5-point scale (5: degradation is inaudible, 4: degradation is audible but not annoying, 3: degradation is slightly annoying, 2: degradation is annoying, 1: degradation is very annoying). For the listening intelligibility, a paired comparison test was carried out, where each subject listened to pairs of speech signals converted by two methods, and then selected which sample was more intelligible.

5.2. Experimental results

Fig. 4 shows the preference score on the listening intelligibility, where the error bar shows a 95% confidence score. As shown in Fig. 4, our proposed method obtained a higher score than the original recorded speech signal, **Prop_F_0** and **Prop_Dur**. The synthesized speech of **Conventional** is less intelligible than the original recorded speech, but the synthesized speech of **Proposed** is more intelligible than the original speech signal and **Conventional**. Also, as shown in Fig. 4 (c) and (d), the modification of both the F_0 and duration model will result in synthesizing more intelligible speech signals.

Fig. 5 shows the results of the DMOS testing on speaker similarity, where the error bars show a 95% confidence score. As shown in Fig. 5, the synthesized voice from **Conventional** is

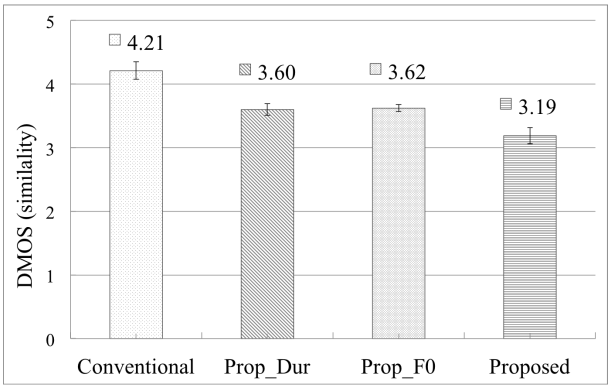


Figure 5: *Speaker similarity to the hearing loss person based on subjective evaluations.*

the most similar to the original voice of the person with hearing loss. Also, it was confirmed that the speaker individuality of a person with hearing loss was lost when using features of a physically unimpaired person. The DMOS score of the proposed method was 3.19 (4: degradation is audible but not annoying, 3: degradation is slightly annoying) and this means speaker individuality is slightly annoying but preserved.

Therefore, from Figs. 4 and 5, it is confirmed that our proposed method generates synthesized signals that are intelligible and include the individuality of a person with hearing loss.

6. Conclusions

In this paper we have proposed a text-to-speech synthesis method using deep neural networks for a person with hearing loss. In our method, to generate more intelligible synthesized sounds while preserving the individuality of a person with hearing loss, a novel combination approach of deep neural networks was employed. The F0 features of a person with hearing loss were modified using those of a physically unimpaired person. The duration model of a physically unimpaired person was used to modify the model of a person with hearing loss. In order to complement some consonant parts of a person with hearing loss, the consonant parts were replaced by those of a physically unimpaired person. The experimental results showed that our method was highly effective in improving the listening intelligibility of speech spoken by a person with hearing loss. In future research, we will complement the vowel parts of the spectral parameters in the training part.

7. Acknowledgements

This work was supported in part by PRESTO, JST (Grant No. JPMJPR15D2).

8. References

[1] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *IEEE ICASSP*, 2013, pp. 7962–7966.

[3] Z. Wu and S. King, “Improving trajectory modelling for DNN-based speech synthesis by using stacked bottleneck features and minimum generation error training,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1255–1265, 2016.

[4] T. Toda and S. Young, “Trajectory training considering global variance for HMM-based speech synthesis,” in *IEEE ICASSP*, 2009, pp. 4025–4028.

[5] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis,” in *Sixth European Conference on Speech Communication and Technology*, 1999.

[6] C. Veaux, J. Yamagishi, and S. King, “Using HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders,” in *Interspeech*, 2012, pp. 967–970.

[7] S. Creer, S. Cunningham, P. Green, and J. Yamagishi, “Building personalised synthetic voices for individuals with severe speech impairment,” *Computer Speech & Language*, vol. 27, no. 6, pp. 1178–1193, 2013.

[8] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham, “Reconstructing the voice of an individual following laryngectomy,” *Augmentative and Alternative Communication*, vol. 27, no. 1, pp. 61–66, 2011.

[9] R. Ueda, T. Takiguchi, and Y. Ariki, “Individuality-preserving voice reconstruction for articulation disorders using text-to-speech synthesis,” in *ACM ICMI*, 2015, pp. 343–346.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, and T. Sainath, “Deep neural networks for acoustic modeling in speech recognition,” *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[11] Y. Fan, Y. Qian, F.-L. Xie, and F. K. Soong, “TTS synthesis with bidirectional lstm based recurrent neural networks,” in *Interspeech*, 2014, pp. 1964–1968.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors.” in *IEEE ASRU*, 2013, pp. 55–59.

[13] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, “Speech parameter generation algorithms for HMM-based speech synthesis,” in *IEEE ICASSP*, vol. 3, 2000, pp. 1315–1318.

[14] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, “Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds,” *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[15] 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.

[16] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, “Duration modeling for hmm-based speech synthesis,” in *ICSLP*, vol. 98, 1998, pp. 29–32.

[17] K. Hashimoto, K. Oura, Y. Nankaku, and K. Tokuda, “The effect of neural networks in statistical parametric speech synthesis,” in *IEEE ICASSP*, 2015, pp. 4455–4459.

[18] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, “An adaptive algorithm for mel-cepstral analysis of speech,” in *IEEE ICASSP*, vol. 1, 1992, pp. 137–140.

[19] H. Kawahara, J. Estill, and O. Fujimura, “Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT,” in *MAVEBA*, 2001, pp. 59–64.

[20] “Methods for subjective determination of transmission quality.” p. 800.