Individuality-Preserving Voice Reconstruction for Articulation Disorders Using Text-to-Speech Synthesis

Reina Ueda, Tetsuya Takiguchi, Yasuo Ariki
Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe, 6578501, Japan
reina_1102@me.cs.scitec.kobe-u.ac.jp, takigu@kobe-u.ac.jp, ariki@kobe-u.ac.jp

ABSTRACT

This paper presents a speech synthesis method for people with articulation disorders. Because the movements of such speakers are limited by their athetoid symptoms, their prosody is often unstable and their speech rate differs from that of a physically unimpaired person, 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 a Hidden Markov Model (HMM)-based text-to-speech synthesis approach that preserves the individuality of a person with an articulation disorder and aids them in their communication. In our method, a duration model of a physically unimpaired person is used for the HMM synthesis system and an F0 model in the system is trained using the F0 patterns of the physically unimpaired person, with the average F0 being converted to the target F0 in advance. In order to preserve the target speaker's individuality, a spectral model is built from target spectra. Through experimental evaluations, we have confirmed that the proposed method successfully synthesizes intelligible speech while maintaining the target speaker's individuality.

Keywords

articulation disorders, speech synthesis system, hidden Markov Model, assistive Technologies

Categories and Subject Descriptors

H.5.5 [**INFORMATION INTERFACES AND PRESENTATION**]: Sound and Music Computing—*modeling, signal analysis, synthesis, and processing*
; K.4.2 [**COMPUTERS AND SOCIETY**]: Social Issues—*Assistive technologies for persons with disabilities*

1. INTRODUCTION

In this study, we focused on a person with an articulation disorder resulting from the athetoid type of cerebral palsy. Cerebral palsy results from damage to the central nervous system, and the damage causes movement disorders. It is classified into the following types: 1)spastic, 2)athetoid, 3)ataxic, 4)atonic, 5)rigid, and a mixture of these types [2]. In the case of a person with an articulation disorder resulting from the athetoid type of cerebral palsy, 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. This is why there is great need for speech synthesis systems to aid them in their communication. An HMM-based speech synthesis system [13] is kind of text-to-speech (TTS) system that can generate any signals from input text data, where the spectrum, F0 and duration are modeled simultaneously in a unified framework. Spectral features are modeled by continuous density HMMs, F0 patterns are modeled by a hidden Markov model based on multi-space probability distribution (MSD-HMM [7]), and state duration densities are modeled by single Gaussian distributions [12].

In the field of assistive technology, Veaux *et al.* [10] used HMM-based speech synthesis to reconstruct the voice of individuals with degenerative speech disorders resulting from Amyotrophic Lateral Sclerosis (ALS). They have proposed a reconstruction method for degenerative speech disorders using an HMM sound synthesis system. In this method, the subject's utterances are used to adapt an average voice model pre-trained on many speakers. Creer *et al.* [3] also adapt the average voice model of multiple speakers to the severe dysarthria data. And Khan *et al.* [1] uses such adaption method to the laryngectomy patient's data. Yamagishi *et al.* [11] proposed a project called "Voice Banking and Reconstruction".

In this paper, we propose an HMM-based speech synthesis method for articulation disorders. Because F0 patterns of articulation disorders are unstable and their duration becomes slow compared to physically unimpaired persons, the output synthesized signals to be indiscernible. To deal with these problems, it is necessary to develop a speech synthesis system in which the output signals are more intelligible and include the patient's individuality. To generate the intelligible voice while preserving the speaker's individuality, our training data include the voice of a physically unimpaired person. Because both the duration and pitch of a person with an articulation disorder are especially different from those of a physically unimpaired person, we use the state duration densities of a physically unimpaired person for the duration model, and the F0 model is trained from F0 patterns of a physically unimpaired person, where the average

ICMI 2015, November 9–13, 2015, Seattle, WA, USA.

DOI: http://dx.doi.org/10.1145/2818346.2820770.

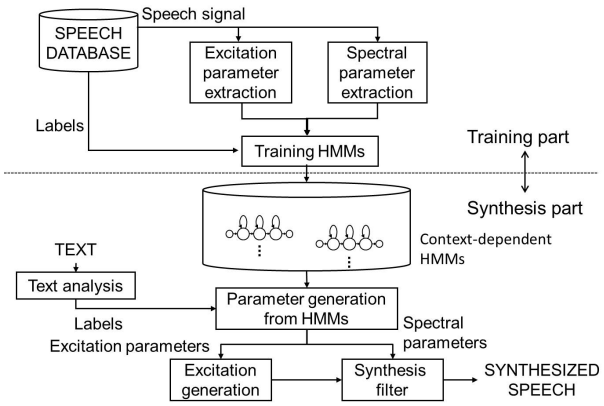


Figure 1: HMM-based speech synthesis system

F0 is converted to F0 of the person with an articulation disorder. To preserve the individuality of the person with an articulation disorder, their spectrum patterns are used for building the spectrum model in the HMM-based speech synthesis.

2. HMM-BASED SOUND SYNTHESIS

2.1 Basic approach

Fig. 1 shows the overview of the basic approach for TTS based on HMMs. This figure shows the training and synthesis parts of the HMM-based TTS system. In the training part, parameters (spectral, F0, and aperiodicity) are extracted as feature vectors. These features are modeled by context-dependent HMMs. Also, by installing the duration model, it is possible to model both each parameter and duration in a unified framework.

In the synthesis part, a context-dependent label sequence is obtained from an input text by text analysis. A sentence HMM is constructed by concatenating context-dependent HMMs according to the context-dependent label sequence. Then, HMM state sequences $q = [q_1, \cdots, q_T]$ are decided from the duration model as follows:

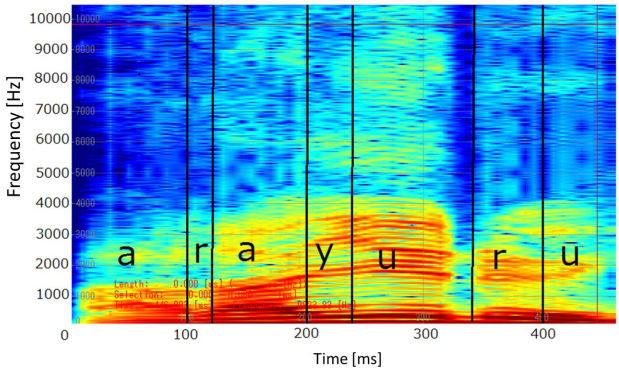
$$\hat{q} = \arg\max_{q} P(q|\lambda) \tag{1}$$

where T, q_t, and λ represent the number of frames, index of the HMM-state of the t-th input frame, and the parameter sets of HMM, respectively. The explicit constraint between static and dynamic features, and signal parameter sets are generated with maximizing HMM likelihood [8] as follows:

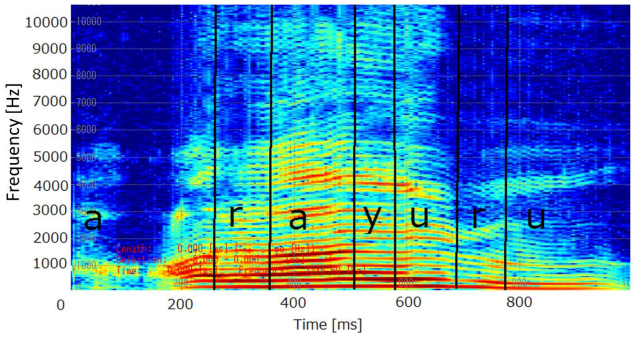
$$c = \arg\max_{c} P(\mathbf{W}c|\hat{q}, \lambda) \tag{2}$$

where $c = [{c_1}^\mathsf{T}, \cdots, c_t^\mathsf{T}, \cdots, c_T^\mathsf{T}]^\mathsf{T}$ represents signal parameter sequences, $c_t = [c(1), \cdots, c(D)]^\mathsf{T}$ represents a signal parameter vector of the t-th frame, and $\mathbf{W}$ represents the matrix constructed from weights which are used for calculating dynamic features [14].

Finally, by using MLSA (Mel-Log Spectrum Approximation) filter [4], speech is synthesized from the generated parameters.



(a) a physically unimpaired person



(b) a person with articulation disorders

Figure 2: Examples of spectrogram /a r a y u r u/

2.2 HMM-based speech synthesis for articulation disorders

If each feature parameter in the HMM-based speech synthesis system is trained using acoustic features obtained from a person with an articulation disorder, the synthesized speech becomes indiscernible. Fig. 2 shows the original spectrograms for the word “a ra yu ru” (“all” in English) of a physically unimpaired person and a person with an articulation disorder. The duration of a person with an articulation disorder is longer than 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 that preserves speaker individuality is generated by using features of both a person with an articulation disorder and a physically unimpaired person. Fig. 3 shows the overview of our proposed method. In this method, the voices of a physically unimpaired person and a person with an articulation disorder are prepared for the training data. First, we extract these two person’s speaking voices into three acoustic parameters (F0 contour, spectral envelope, and aperiodicity index (AP)) estimated by using STRAIGHT analysis [6]. For making the F0 feature characteristics close to those of a person with an articulation disorder, both the F0 mean and variance are estimated from training data. Then, the F0 features of a physically unimpaired person are converted to those of a person with an articulation disorder by using the linear transformation as

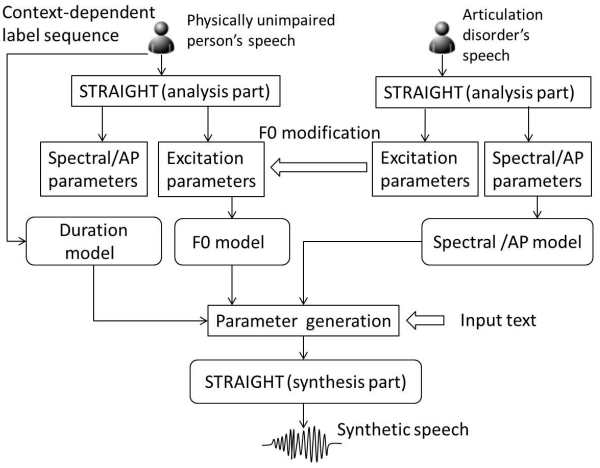


Figure 3: Diagram of HMM-based speech synthesis method for articulation disorders

follows:

$$\hat{x}_t = \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) + \mu_y \quad (3)$$

where x_t represents a log-scaled F0 of the physically unimpaired person at the frame t, μ_x and σ_x represent the mean and standard deviation of x_t, respectively. μ_y and σ_y represent the mean and standard deviation of a person with an articulation disorder's log-scaled F0, respectively. An F0 model in the HMM-based speech synthesis, which is trained using the F0 features from Eq. (3), generates the F0 sequences so that they include the individuality of a person with an articulation disorder.

The duration model of the physically unimpaired person is used in the HMM-based TTS system, where it is trained using the context-dependent label sequences of the physically unimpaired person because the utterance duration of a person with an articulation disorder is unstable. To preserve the articulation disorder's individuality, the spectral model and the AP model are trained using only spectral and AP sequences of the voice of the person with an articulation disorder. In the synthesis part, after the text to be synthesized is converted to a context-dependent label sequence, parameter sequences are generated from the context label and these models. Parameter sequences are converted to the features (spectral envelope, F0 contour, aperiodicity index), which can be handled in the STRAIGHT. Finally, the output signal is synthesized from these features by using the synthesis part of the STRAIGHT.

3. EXPERIMENTS

3.1 Experimental conditions

We prepared the training data for two men. One is a physically unimpaired person, and the other is a person with an articulation disorder. We used 513 sentences in the ATR Japanese database for a physically unimpaired person, and recorded 429 sentences in the same database uttered by a person with an articulation disorder. (The other 84 sentences were mis-recorded). The speech signals were digitized 16bit/48 kHz. The frame shift was 5 ms. Acoustic and prosodic features were extracted by using STRAIGHT. Mel-cepstrum coefficients (their dynamic, acceleration coefficients) were used as spectral parameters. Log-F0 and 5 band-filtered aperiodicity measures [5] (their dynamic and acceleration coefficients) were used as excitation parameters. Context-dependent phoneme HMMs with five states were used in the speech synthesis system [13].

In order to confirm the effectiveness of our proposed method, we evaluated the aspects of both listening intelligibility and speaker similarity by having subjects listen to voices synthesized under five conditions listed in Table 1.

Table 1: Voices compared in the evaluation tests

Type	Duration Model	F0 Model	AP/Spectral Model
ADM	AD	AD	AD
Ref1	PU	AD	AD
Prop	PU	convPU	AD
Ref2	PU	PU	AD
PUM	PU	PU	PU

ADM: Model of a person with an articulation disorder
Prop: Proposed method
PUM: Model of a physically unimpaired person
AD: Articulation Disordered
PU: Physically Unimpaired
convPU: Creating the model from a physically unimpaired person's parameter sequences which are converted to those of a person with an articulation disorder by Eq. (3)

Ten sentences included in ATR Japanese database were synthesized under those five conditions. A total of 8 Japanese speakers took part in the test using headphones. For the speaker similarity, we performed a MOS (Mean Opinion Score) test [9]. In the MOS test, an opinion score was set to a 5-point scale (5: Identical, 4: Very Similar, 3: Quite Similar, 2: Dissimilar, 1: Very Dissimilar). For the listening intelligibility, a paired comparison test was carried out, where each subject listened to pairs of speech converted by two methods and selected which sample sounded more intelligible.

3.2 Results and discussion

We calculated the average synthesized duration per mora of synthesized signals for 50 sentences. The average duration of ADM is 219.768 [$ms/mora$] and PUM is 179.69 [$ms/mora$]. As compared to the duration of PUM, that of ADM is quite slower, which leads to the unintelligibility of the synthesized speech.

Fig. 4 shows the results of MOS test on speaker similarity, where the error bars shows a 95% confidence score. As shown in Fig. 4, the synthesized voice from ADM, which trained using only the voice of a person with an articulation disorder, is the most similar to the original voice of the person with an articulation disorder. Also, it is confirmed that the more we use features of a physically unimpaired person for modeling, the more the speaker individuality (of a person with an articulation disorder) will be lost.

Fig. 5 shows the preference score on the listening intelligibility, where the error bar shows a 95% confidence score. As shown in Fig. 5, our proposed method obtained a higher score than Ref1 and ADM. In comparison to Ref2 in regard to intelligibility, there may not be so much influence on intelligibility because only the average F0 is different between our method and Ref2. These results imply that intelligi-

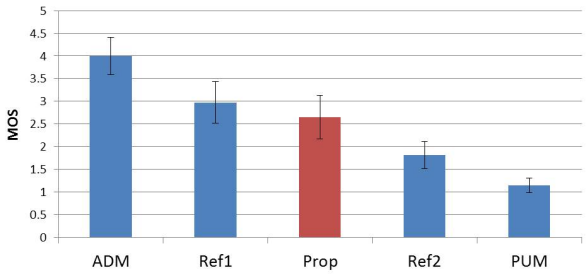


Figure 4: Speaker similarity to the articulation disorder

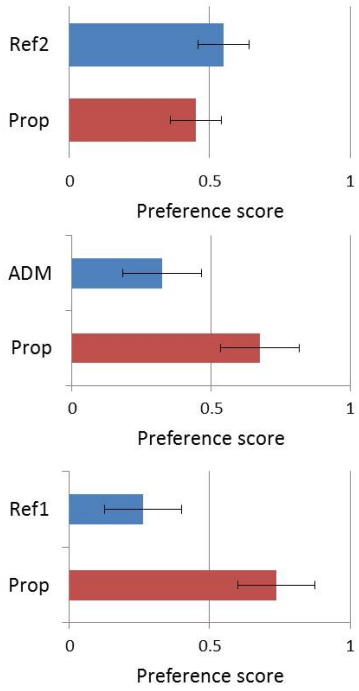


Figure 5: Preference scores for the listening intelligibility

bility can be improved by replacing the duration model of the person with an articulation disorder with the physically unimpaired person's duration model and using the F0 patterns from the physically unimpaired person (with the average F0 being converted to the target (articulation disorder's person) F0 in advance). Therefore, from Figs. 4 and 5, it is confirmed that our proposed method generates synthesized signals that are fairly intelligible and include the individuality of a person with an articulation disorder.

4. CONCLUSION

We proposed a text-to-speech synthesis method based on HMMs for people with articulation disorders. In our method, to generate more intelligible synthesized sounds, the duration model of a physically unimpaired person is used, and the F0 model is trained using F0 features of a physically unimpaired person, where the average F0 is converted to the F0 of the person with an articulation disorder by using the linear transformation. In order to preserve the target individuality, the spectral and AP models are trained using only features of the person with an articulation disorder. The experimental results showed our proposed method greatly improves the listening intelligibility of speech of a person with an articulation disorder. In our future research, we will modify spectral and AP models as well as duration and F0 model to improve the listening intelligibility even more.

5. REFERENCES

[1] Z. Ahmad Khan, P. Green, S. Creer, and S. Cunningham. Reconstructing the voice of an individual following laryngectomy. *Augmentative and Alternative Communication*, 27(1):61–66, 2011.

[2] T. Canale and W. C. Campbell. *Campbell's operative orthopaedics*, volume 12. Technical report, Mosby Year Book, June 2002.

[3] S. Creer, S. Cunningham, P. Green, and J. Yamagishi. Building personalised synthetic voices for individuals with severe speech impairment. *Computer Speech & Language*, 27(6):1178–1193, 2013.

[4] S. Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (MLSA) filter for speech synthesis. *Electronics and Communications in Japan (Part I: Communications)*, 66:10–18, 1983.

[5] 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 *Proc. of MAVEBA*, pages 59–64, 2001.

[6] H. Kawahara, I. Masuda-Katsuse, and A. D. Cheveigné. 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*, 27:187–207, 1999.

[7] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi. Hidden Markov models based on multi-space probability distribution for pitch pattern modeling. In *Proc. of ICASSP*, pages 229–232, 1999.

[8] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of ICASSP*, pages 1315–1318, 2000.

[9] I. T. Union. ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology. Technical report, International Telecommunication Union, July 2006.

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

[11] 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*, 33(1):1–5, 2012.

[12] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Duration Modeling in HMM-based Speech Synthesis System. In *Proc. of ICSLP*, pages 29–32, 1998.

[13] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Proc. of Eurospeech*, pages 2347–2350, 1999.

[14] H. Zen, K. Tokuda, and A. W. Black. Statistical parametric speech synthesis. *Speech Communication*, 51:1039–1064, 2009.