Speech Synthesis based on Articulatory-Movement HMMs with Voice-source Codebooks

Tsuneo Nitta [1], *Takayuki Onoda* [1], *Masashi Kimura* [1], *Yurie Iribe* [2], *Kouichi Katsurada* [1]

[1] Graduate School of Engineering, Toyohashi University of Technology, JP
[2] Information and Media Center, Toyohashi University of Technology, JP

{nitta, katsurada}@cs.tut.ac.jp, Iribe@imc.tut.ac.jp

Abstract

Speech synthesis based on one-model of articulatory movement HMMs, that are commonly applied to both speech recognition (SR) and speech synthesis (SS), is described. In an SS module, speaker-invariant HMMs are applied to generate an articulatory feature (AF) sequence, and then, after converting AFs into vocal tract parameters by using a multi-layer neural network (MLN), a speech signal is synthesized through an LSP digital filter. The CELP coding technique is applied to improve voice-sources when generating these sources from embedded codes in the corresponding state of HMMs. The proposed SS module separates phonetic information and the individuality of a speaker. Therefore, the targeted speaker's voice can be synthesized with a small amount of speech data. In the experiments, we carried out listening tests for ten subjects and evaluated both of sound quality and individuality of synthesized speech. As a result, we confirmed that the proposed SS module could produce good quality speech of the targeted speaker even when the training was done with the data set of two-sentences.

Index Terms: speech synthesis, HMM-based speech synthesis, articulatory features, voice source codebook, LSP

1. Introduction

We have been developing one-model speech recognition (SR) and speech synthesis (SS) modules based on articulatory movement HMMs that are commonly applied to both SR and SS [1]. In general, SR and SS modules have been designed separately and current HMM-based SR modules use speaker-independent phone-models, while speaker-dependent models are embedded in HMM-based SS modules [2]. To unify two modules of SR and SS, the following three requirements should be satisfied. Namely, in one-model SR and SS modules, (i) speaker-invariant features such as articulatory features (AFs) should be extracted from speech signal, (ii) common HMMs for SR and SS should be designed using speaker-invariant features, and (iii) the feature parameters generated from HMMs should be converted into speaker-specific parameters such as vocal tract (VT) parameters.

In the SR module of one-model SR & SS, the introduction of speaker-invariant feature parameters to HMMs can realize an efficient SR engine. In our previous report [1], the experimental results in a speaker-independent phoneme recognition task showed that the articulatory movement HMM with a single mixture from a single speaker outperformed MFCC-based HMMs with 16 mixtures from 100 speakers. On the other hand, in the SS module, the articulatory movement HMM generates an AF sequence, like motor command in a human brain, and then the AF sequence is converted into a set of VT parameters by using multi-layer neural network (MLN). In the previous paper [1], residual signals of PARCOR analysis are applied to a PARCOR synthesizer [3].

In this paper, we focus on the improvement of voice source design, as well as the modification of synthetic digital filter, that is, the PACOR filter is changed to an LSP digital filter [4]. Moreover, CELP coding technique [5] is applied to improve voice sources when generating these sources from embedded codes in HMMs. Each state of the articulatory movement HMM has a code number corresponding to that in a voice-source codebook. The procedure of designing multiple codebooks for eight phoneme categories and allocating a code to each state is described in section 3 in detail.

The proposed speech synthesis system can separate phonetic information and the individuality of a speaker. That is, HMMs represent speaker-invariant phonetic information, and on the other hand, the MLN converter and the voice sources share the role of representing individuality like a speech production organ. Therefore, it is expected to synthesize targeted speaker's speech with a small amount of voice data. In the experiments, we carried out listening tests for ten subjects and evaluated both of sound quality (MOS test) and individuality (ABX test) of the synthesized speech.

This paper is organized as follows. Section 2 explains speech synthesis based on articulatory movement together with the outline of one-model SR and SS. Sections 3 describes voice source design using a closed loop of AbS. Section 4 then explains the results of evaluation for synthetic speech. Finally, Section 6 presents the conclusion and suggests future work.

2. Outline of One-model SR & SS using Articulatory Movement HMMs

Figure 1 shows an outline of the proposed one-model SR and SS based on articulatory movement HMMs. In the Figure, the upper block is a SR module and the lower, a SS module. Both modules use the same HMMs. The SR module has an AF extractor with three-stage MLNs that outputs an AF sequence to the articulatory movement HMMs [6], [7]. The HMMs represent probabilistic articulatory gestures in each mono-phone model.

In the SS module, the same speaker-invariant HMMs generate an AF sequence by concatenating mono-phone models, and then converting them into vocal tract parameters, or LSP parameters, using a speaker-specific model. A speech signal is synthesized by a LSP digital filter together with a voice-source signal. The voice-source signal is read from the same HMMs and is modified along pitch contour by using PSOLA technique [8]. The proposed one-model SR and SS can also output the speech input directly by adding the AF extractor output into the MLN of an AF-LSP converter as shown in Figure 1. Such functionality is useful for talk-back

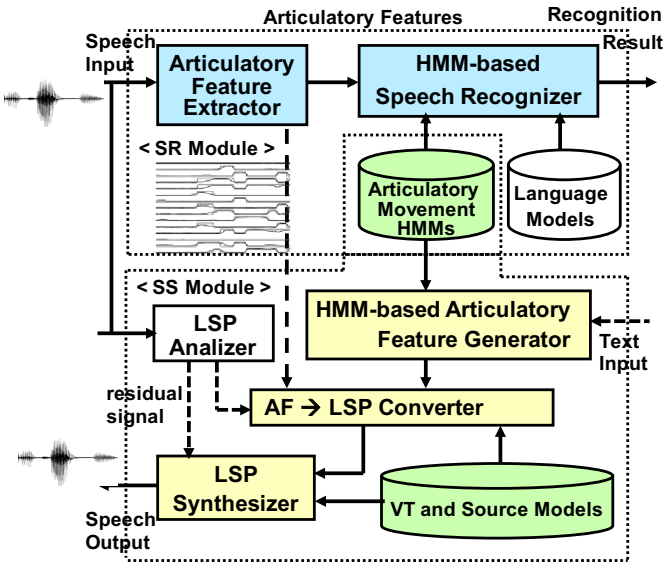

Figure 1: One-model SR and SS.

services in spoken dialogue systems, especially when an out-of-vocabulary (OOV) word is detected.

3. Speech Synthesis using Articulatory Movement HMMs

3.1 Outline of speech synthesis using articulatory movement HMMs

A typical HMM-based speech synthesizer models a single speaker's voice using features originated in spectrum [9]. The proposed SS module shown in Figure 2 introduces speaker-invariant aticulatory movement models to HMMs that are commonly used for a SR module. HMMs generate AF sequences by concatenating mono-phone models, and then feeding the average AF vectors in each state into an AF-LSP converter. The current frame data of the inputs of the converter, AF(m, t), m=1,2,…15, are combined with the other two frames, which are three points prior to and following the current frame (AF(m, t-3), AF(m, t+3)) to form articulatory movement.

3.2 Conversion from AFs to LSP parameters

Because the articulatory movement HMMs are speaker-invariant models, speaker-specific acoustic features are represented in the AF-LSP converter and/or voice-source signals. In the AF-LSP converter, the AF sequence is converted into a set of vocal tract parameters, or LSP parameters [4], which are line spectrum pairs in an LPC vocoder, known as the most effective speech compression algorithm. The LSP parameters are orthogonalized with respect to each other. The AF-LSP converter is designed with a three-layer neural network (MLN). The MLN has 45 input units (15-AFs × 3-frames) corresponding to a set of context-dependent AF vectors (a preceding context, AF(m,t-3), a current context, AF(m,t), and a subsequent context, AF(m,t+3)), each in 15 dimensions. The MLN has 42 output units (14-LSPs × 3-frames) corresponding to a set of context-dependent LSP parameters. The hidden layer of the MLN has 450 units.

Speaker adaptation, or training, is executed at the AF-LSP conversion stage and the voice-source codebook design stage

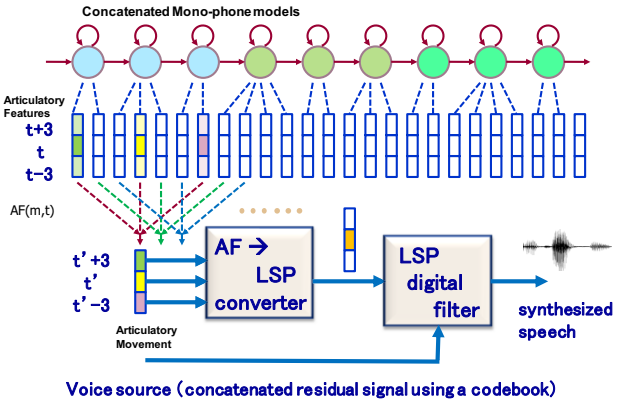

Figure 2: HMM-based speech synthesis using articulatory movement HMMs.

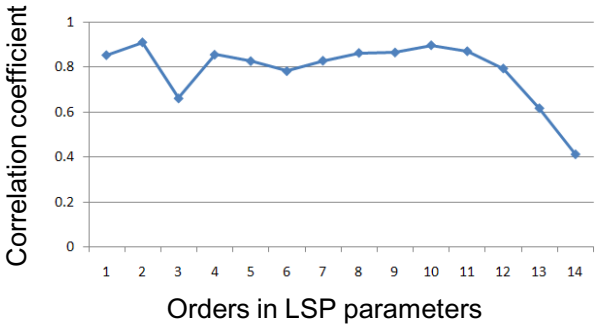

Figure 3: Correlation coefficients between original speech LSPs and converted LSPs from articulatory features (AFs).

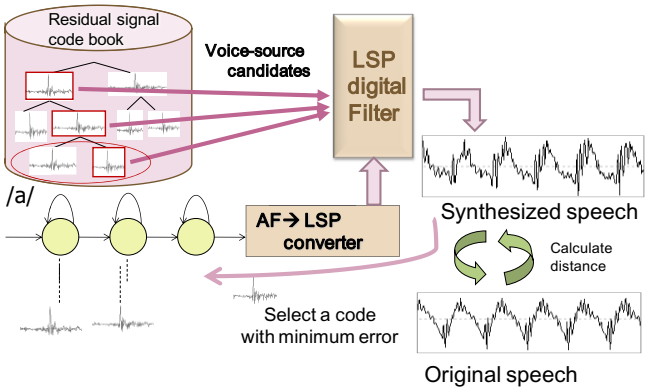

Figure 4: Voice-source codebook design and allocation of a code using CELP coding technique.

described in 3.3. When training the MLN, the initial set of weighting coefficients is trained with AF data, uttered by multiple speakers or a referenced single speaker but with a large number of speech data. These are then adapted to a specific user using small number of his/her speech samples. A speech signal is finally synthesized using a LSP synthesizer. Figure 3 shows correlation coefficients between LSPs of the original speech and the converted LSPs from articulatory

features (AFs). Speech samples are two short sentences. Two LSPs have comparatively high correlation.

3.3 Voice-source design by using CELP technique

Figure 4 shows a procedure of voice-source codebook design and allocation of a code using CELP coding technique.

3.3.1 Designing initial residual segment codebooks

The large number of residual segments in a codebook gives high quality speech, however, the computation cost should also be considered. Firstly, we stored a large number of segments in the form of binary trees using a clustering algorithm of LBG [10]. The following is the four steps to design residual codebooks.

Step1:
Extracts residual signals using LPC analysis.
Step2:
Cuts each segment by centering the pitch mark point and windowing the segment with double length of pitch interval.
Step3:
Analyzes each segment by using FFT.
Step4:
Applies LBG hierarchical clustering to all the segments (binary trees).

Sep 4 is repeated until all the residual segments are enrolled.

The multiple codebooks are finally designed for eight phoneme categories each. The eight groups are five vowels /a, i, u, e, o/, one independent Japanese nasal sound /N/, a voiced consonant group /Cvoiced/, and an unvoiced consonant group /Cunvoiced/.

3.3.2 Allocating a code to each state of HMMs

Figure 4 shows the procedure to allocate residual codes to all the states in HMMs by using CELP technique, that is called Analysis-by-Synthesis (AbS), where the encoding (analysis) is performed by optimizing the decoding (synthesis) in a closed loop.

The following is the three steps to allocate a code to the corresponding state.

Step 1:
Estimates LSP parameters from an averaged AF vector in the HMM state.
Step 2:
Synthesizes speech by combining the estimated LSP parameters and a residual signal in the codebook by adjusting pitch mark point.
* This step is started from a root node and ended at a leaf node. At each node the residual code with minimum error is selected.
Step 3:
Allocates the code with minimum error to the HMM state.

3.3.3 Speaker adaptation in voice-source

To adapt the voice-source to a new speaker with a small number of speech data, the training segments are limited to five vowel parts and an independent nasal sound /N/.

4. Evaluation of Synthetic Speech

4.1 Speech Data

D1*: Training data set-1 for HMM training.*
This training data set comprises 5,000 JNAS [11] sentences uttered by 38 male speakers (16 kHz, 16 bit).

D2-1*: Training data set-2 for AF-LSP converter training.*
This second training data set comprises 503 ATR phonetically balanced sentences [12]. A male speaker "A" is set to reference and 503 sentences uttered by "A" are trained for AF-LSP converter.

D2-2*: Testing data set-1 for AF-LSP converter adaptation.*
The same data set as ***D2-1***, but uttered by a new male speaker "B" is set to a targeted speaker.
Two sentences are used to adapt the AF-LSP converter to the speaker "B".

D3-1*: Training data set-3 for voice-source codebook training.*
The same data set as ***D2-1***. The male speaker "A" is set to reference and 150 sentences are used to design codebook.

D3-2*: Testing data set-2 for voice-source codebook adaptation.*
The same data set as ***D2-1***, but uttered by a new male speaker "B", is set to a targeted speaker and two sentences are used to adapt the codebook to speaker "B".

4.2 Experimental Setup

The *D1* data set is used to design 38 Japanese monophone HMMs with seven states, five loops, and left-to-right models.

4.3 Evaluation of voice quality

To investigate voice quality of the SS module, the following six data points are compared. The initial MLN of the AF-LSP converter, or VT converter, is trained with a ***D2-1*** data set of speaker "A". It is then adapted to a new male speaker "B" with a ***D2-2*** data set. The initial codebook is trained with a ***D3-1*** data set of speaker "A". It is then adapted to a new male speaker "B" with a ***D3-2*** data set. In the listening test, nine sentences without two sentences used in the adaptation are evaluated.

(1) Original speech of speaker "B".
(2) VT converter of speaker "A" ***(D2-1)*** driven with pulse and noise signal.
(3) VT converter of speaker "A" ***(D2-1)*** driven with codebook of speaker "A" ***(D3-1)***.
(4) VT converter of speaker "A" ***(D2-1)*** driven with codebook of speaker "B" ***(D3-2)***.
(5) VT converter of speaker "B" ***(D2-2)*** driven with codebook of speaker "A" ***(D3-1)***.
(6) VT converter of speaker "B" ***(D2-2)*** driven with codebook of speaker "B" ***(D3-2)***.

Figure 5 shows the result of MOS test of synthetic speech. In the MOS test, twelve subjects heard the original speech (1) and pulse and noise excited synthetic speech (2) before listening test. Adaptation with short sentences both in the VT converter and in voice-source codebook is found to be effective, however further improvement is needed.

4.4 Evaluation of individuality

Figure 6 shows the result of ABX test of Synthetic speech with speaker adaptation by VT and/or codebook. The targeted speaker is B. The results show that both adaptation in VT conversion and in voice-source codebook contribute to individuality verification scores. The result also shows that the proposed speech synthesis method based on articulatory move-

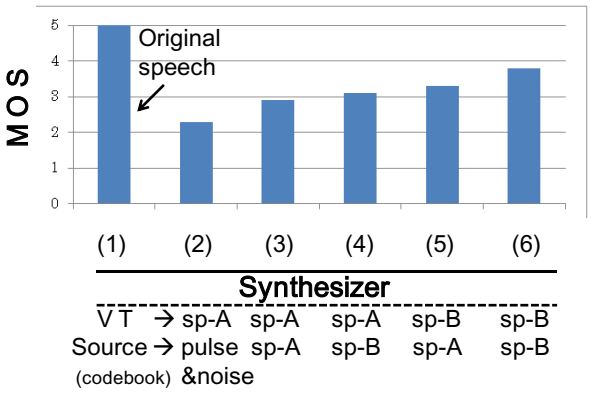

Figure 5: MOS of Synthetic speech generated with articulatory movement HMMs.

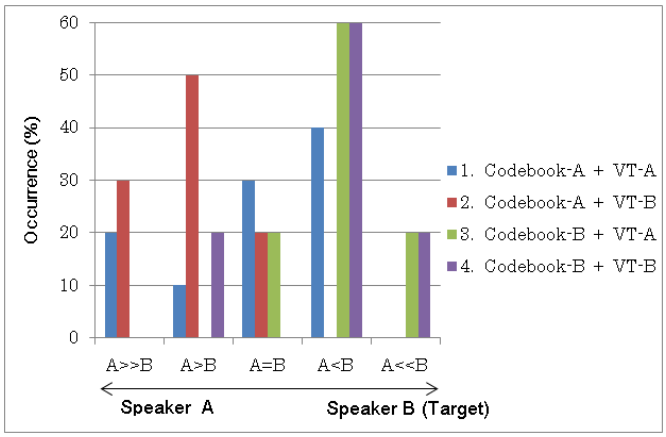

Figure 6: ABX test of Synthetic speech with speaker adaptation by VT and/or codebook. (targeted speaker is B)

ment HMMs can achieve individuality change with a small amount of speech data by two sentences.

5. Conclusion

Speech synthesis based on one-model of articulatory movement HMMs that are commonly applied to both speech recognition (SR) and speech synthesis (SS) is proposed. The proposed speech synthesis system separate phonetic information and speaker individuality and a target speaker's voice can be easily synthesized with a small amount of speech data. Experimental results on the evaluation of voice quality showed that adaptation with short sentences both in the VT converter and in voice-source codebook is effective, however further improvement is needed. The other evaluation results on individuality show that both adaptation in VT conversion and in voice-source codebook contribute to individuality verification scores. These results shows that the proposed speech synthesis method can achieve individuality change with a small amount of speech data by two sentences.
Future work will include further improvement of voice quality, as well as implementation of prosody control for Text-to-Speech.

6. References

[1] Nitta, T., Onoda, M., Kimura, M., Iribe, Y., and Katsurada, K., One-model speech recognition and synthesis based on articulatory movement HMMs, Proc. Of INTER-SPEECH2010, pp.2970-2973 (2010-9).

[2] Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T., and Kitamura, T., *Speech parameter generation algorithms for HMM-based speech synthesis,* Proc. of ICASSP, pp.1315-1318 (2000-6).

[3] Itakura, F. and Saito, S., Analysis synthesis telephony based on the maximum likelihood, Proc. of 6th ICA, C-5-5 (1968).

[4] Itakura, F., Line spectrum representation of linear predictor coefficients of speech signals, J. Acoust. Soc. Am. Vol. 57, Issue S1, pp.35-35 (1975).

[5] Schroeder, M. R., Atal,B. S., Code-excited linear prediction (CELP): high-quality speech at very low bit rates, Proc. of ICASSP'85, vol.10, pp.937–940 (1985).

[6] Huda, M.N., Katsurada, K. and Nitta, T., Phoneme recognition based on hybrid neural networks with inhibition/ enhancement of Distinctive Phonetic Feature (DPF) trajectories, Proc. of Interspeech'08, pp.1529-1532 (2008).

[7] Huda, M.N., Kawashima, H. and Nitta, T., Distinctive Phonetic Feature (DPF) extraction based on MLNs and Inhibition/ Enhancement Network, IEICE Trans. Inf. & Syst., Vol.E92-D, No. 4, pp.671-680 (2009).

[8] Charpentier, F. J., and Stella, M. G., Diphone synthesis using an overlap-add technique for speech waveforms concatenation, Proc. of ICASSP'86, pp. 2015-2018 (1986).

[9] Masuko, T., Tokuda, K., Kobayashi, T. and Imai, S., Speech synthesis from HMMs using dynamic features, Proc. of ICASSP1996, pp.389-392 (1996).

[10] Linde, T., Buzo, A., and Gray, R. M., An algorithm for vector quantizer design," IEEE Tran. Commun., COM-28, No. 1, pp.84-95 (1980).

[11] JNAS: Japanese Newspaper Article Sentences. http://www.milab.is.tsukuba.ac.jp/jnas/instruct.html

[12] Abe, M., Sagisaka,Y., Umeda, T. and Kuwabara, H., Speech Database User's Manual. *ATR Technical Report*, TR-I-0116 (1990). (in Japanese)