One-Model Speech Recognition and Synthesis Based on Articulatory Movement HMMs

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

[1] Graduate School of Engineering, [2] Information and Media Center
Toyohashi University of Technology
{nitta, katsurada}@cs.tut.ac.jp

10.21437/Interspeech.2010-31

Abstract

One-model speech recognition (SR) and speech synthesis (SS) based on a common articulatory movement model are described herein. The SR engine has an articulatory feature (AF) extractor and an HMM based classifier that models articulatory gestures. Experimental results of a phoneme recognition task show that the AF outperforms MFCC even if the training data are limited to a single speaker. In the SS engine, the same speaker-invariant HMM is applied to generate an AF sequence, and then, after converting AFs into vocal tract parameters, a speech signal is synthesized by a PARCOR filter, together with a residual signal. Phoneme-to-phoneme speech conversion, using AF exchange, is also described.

Index Terms: speech synthesis, speech recognition, articulatory features, phoneme-to-phoneme speech conversion

1. Introduction

Current HMM-based speech recognition (SR) systems have achieved acceptable performance in certain limited applications. However, because most of these systems use spectrum origin features that are often distorted by various factors, such as speakers, phoneme contexts, ambient noise or distant speech, the development of accurate SR system requires a large speech corpus and mixture components.

On the other hand, a human infant can acquire a speaker-independent phone system [1] by listening only to his/her parents' voices. To explain the mechanism that enables such linguistic acquisition, a theory of articulatory gestures, in which a human perceives a voice by referring to articulatory movement, has been presented [2]. In recent SR studies, various methods have been proposed to achieve articulatory feature (AF) extraction [3],[4],[5],[6],[7], and well-designed articulatory-based HMMs have been shown to outperform MFCC-based HMMs. In this paper, we present experimental results in which the high performance of phoneme recognition is achieved using articulatory-based HMMs, or articulatory movement (AM) models, even when the training data are limited to a single speaker and a mixture number of one.

The argument as to whether human speech production and perception run on a single system or independently, or two systems, has been discussed for quite some time [8], however, recent studies in brain science seem to lean toward the single system belief [9]. In this paper, we propose a novel method that realizes both a SR engine and a speech synthesis (SS) engine with a common articulatory movement model, that is a one-model SR and SS. A typical HMM-based speech synthesizer models a single speaker's voice using spectrum-origin features and thus cannot be applied to speaker-independent SR[10]. Our proposed method represents speaker-invariant articulatory gestures in an HMM that can generate an AF sequence, and, after converting the AF sequence into vocal tract parameters with a multi-layer neural network (MLN), a speech signal is synthesized by a PARCOR synthesizer together with a residual signal [11].

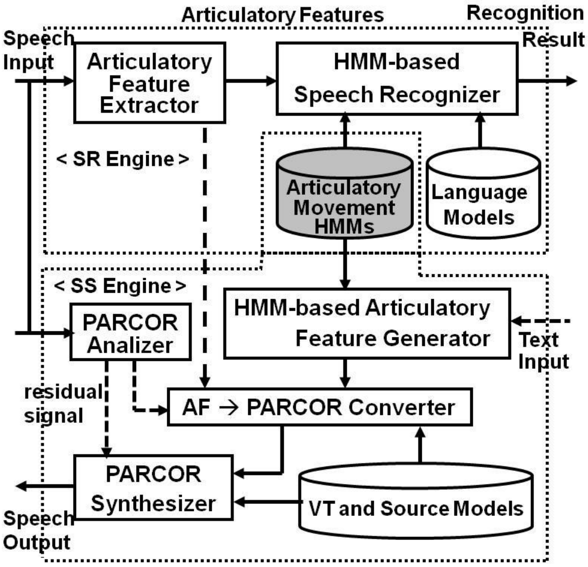

Figure 1: *One-model SR and SS.*

A one-model SR and SS framework enables other functionality as well, such as phoneme-to-phoneme speech conversion. In this paper, we investigate this functionality by exchanging the AF of a phoneme A (B) in a word with another AF corresponding to a phoneme B (A). We test them using an ABX listening test.

This paper is organized as follows. Section 2 explains the outline of the one-model SR and SS. Sections 3 and 4 describe SR and SS using the common articulatory movement HMMs and their experimental results, respectively. Section 5 then presents phoneme-to-phoneme speech conversion and the result of the ABX test. Finally, Section 6 presents the conclusion and suggests future work.

2. One-model SR & SS based on articulatory movement models

Figure 1 shows an outline of the proposed one-model SR and SS based on AM models. In the Figure, the upper block is a SR engine and the lower, a SS engine. Both engines use the same AM HMMs. The SR engine has an AF extractor with three-stage multi-layer neural networks (MLNs), described in section 3, that outputs an AF sequence to the AM HMMs [12],

[13]. The HMMs represent probabilistic articulatory gestures in each mono-phone model.

In the SS engine, the same speaker-invariant HMMs generate an AF sequence by concatenating mono-phone models, and then converting them into vocal tract parameters, or PARCOR parameters, using a speaker-specific model. A speech signal is synthesized by a PARCOR filter together with a residual signal. The proposed one-model SR and SS can also output the speech input directly by adding the AF extractor output into the AF→PARCOR converter as shown in Figure 1. Such functionality is useful for talk-back services in spoken dialogue systems, especially when an out-of-vocabulary (OOV) word is detected. In Section 5, the same output from the AF extractor is modified and then input to the AF→PARCOR converter for corrective training of pronunciation.

3. Speech Recognition Based on Articulatory Movement HMMs

3.1 Articulatory Feature Extraction

The proposed SR engine is divided into two parts: an AF extractor that converts input speech into AFs, and an articulatory movement HMM classifier. Figure 2 shows the AF extractor. Input speech is sampled at 16 kHz and a 512-point FFT of the 25 ms Hamming-windowed speech segment is applied every 10 ms. The resultant FFT power spectrum is then integrated into 24-ch BPFs output with mel-scaled center frequencies. At the acoustic feature extraction stage, the BPF-outputs are first converted to local features (LFs) by applying three-point linear regression (LR) along the time and frequency axes [14], [15], [16], [13]. LFs represent variation in a spectrum pattern along two axes. After compressing these two LFs, from 24 dimensions into LFs in 12 dimensions, using discrete cosine transform (DCT), a 25-dimensional (12 Δt, 12 Δf, and ΔP, where P stands for the log power of a raw speech signal) feature vector, called a LF, is extracted. Our previous work shows that a LF is superior to an MFCC when used as input to MLNs for the extraction of AFs, or distinctive phonetic features (DPFs) [7].

The LFs are then entered into a three-stage AF extractor [13]. The first stage extracts 45-dimensional AF vectors from the LFs of the speech input, using two MLNs. The first MLN maps acoustic features, or LFs, onto discrete AFs and the second MLN reduces misclassification at the phoneme boundaries by constraining the AF context. Figure 3 shows an example AF sequence for the utterance /jiNkoese (artificial satellite)/. In the figures, "Solid thin line" and "Solid bold line" represent "ideal segmentation" and extracted AF sequences at the first stage, respectively. The second stage incorporates Inhibition/ Enhancement (In/En) functionalities to obtain modified AF patterns [13]. The third stage decorrelates three context vectors of AFs using the Gram-Scmidt (GS) orthogonalization procedure [15] before connecting with the HMM-based classifier.

3.2 Evaluation of SR

The proposed AF-based HMMs are compared with MFCC-based HMMs. A standard MFCC feature set consists of a vector in 38 dimensions (12 MFCC, 12 Δ and 12 ΔΔ coefficients of MFCC, Δ and ΔΔ coefficients of the log energy of the speech signal). On the other hand, the AF extractor outputs an AF-vector in 45 dimensions (15 preceding context AF patterns, 15 current frame AF patterns, and 15 following context AF patterns) for each input frame.

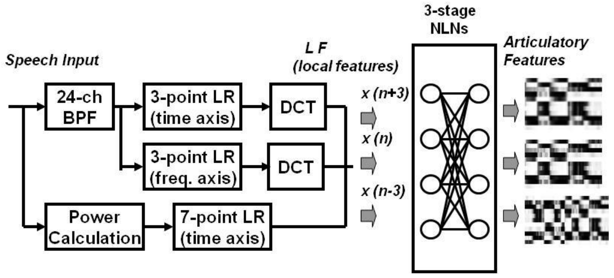

Figure 2: *Articulatory feature extraction.*

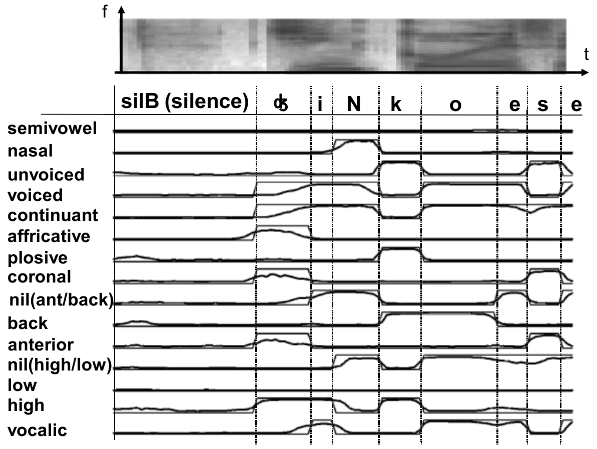

Figure 3: *Articulatory feature sequence.*
: /jiNkoese/ (artificial satellite)

- ***Speech Data:***

D1: Training data set-1 for MLNs
A subset of the Acoustic Society of Japan (ASJ) Continuous Speech Database comprising 4,503 sentences uttered by 30 male speakers (16 kHz, 16 bit) is used [17].

D2: Training data set-2 for HMMs
This training data set comprises 5,000 JNAS [18] sentences uttered by 33 male speakers (16 kHz, 16 bit).

D3: Test data set
The test data set comprises 2,719 JNAS sentences uttered by 17 male speakers (16 kHz, 16 bit).

- ***Experimental Setup:***

The Japanese phoneme correct rate for the D3 data set is evaluated using an HMM-based classifier. In the experiments, The D2 data set is used to design 38 Japanese monophone HMMs with five states, three loops, and left-to-right models. Input features for the classifier are MFCC features or AFs. In the HMMs, the output probabilities are represented in the form of Gaussian mixtures using diagonal matrices. The number of mixture components in the HMM is varied between 1, 2, 4, 8, and 16. In this experiment, we do not implement language models, because we focus our research on the design of an accurate phonetic typewriter.

- ***Experimental Results and Discussion:***

Figure 4 shows the experimental results of the AFs by comparison with MFCCs in a phoneme recognition task. The proposed AFs exhibit an almost equivalent level of performance with the different numbers of speakers used for training, as well as the different number of mixtures in HMMs. This result suggests that the extracted features are speaker invariant parameters.

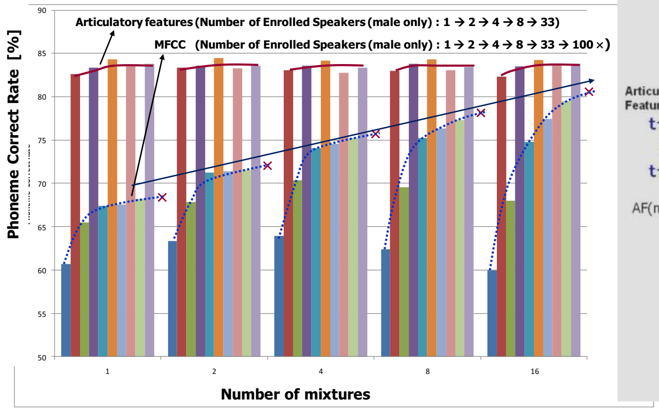

Figure 4: *Phoneme correct rate vs. number of mixtures and enrolled speakers.*

4. Speech Synthesis Based on Articulatory Movement HMMs

4.1 HMM-based speech synthesis using articulatory movement models

A typical HMM-based speech synthesizer models a single speaker's voice using features originated in spectrum [10]. The proposed SS engine shown in Figure 5 introduces speaker-invariant aticulatory movement models to HMMs that are commonly used for a SR engine. HMMs generate AF sequences by concatenating mono-phone models, and then feeding the average AF vectors in each state into an AF-PARCOR 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.

4.2 Conversion from articulatory fatures (AFs) to PARCOR parameters

In the AF-PARCOR converter, the AF sequence is converted into a set of vocal tract parameters, or PARCOR parameters [11], which are k-parameters in an LPC vocoder, related to the reflection coefficients of a vocal tract. The PARCOR parameters are orthogonalized with respect to each other. The AF-PARCOR 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 39 output units (13-PARCORs × 3-frames) corresponding to a set of context-dependent PARCOR parameters. The hidden layer of the MLN has 450 units.

When training the MLN, the initial set of weighting coefficients is initially trained with AF data, uttered by multiple speakers. These are then adapted to a specific user using his/her voice. A speech signal is finally synthesized using a PARCOR synthesizer.

4.3 Evaluation of Synthetic speech

To investigate voice quality of the SS engine, the following three data points are compared. The initial MLN of the AF-PARCOR converter is trained with 5,000 JNAS sentences, which constitutes the D2 data described in section 3. It is then adapted to a new male speaker.

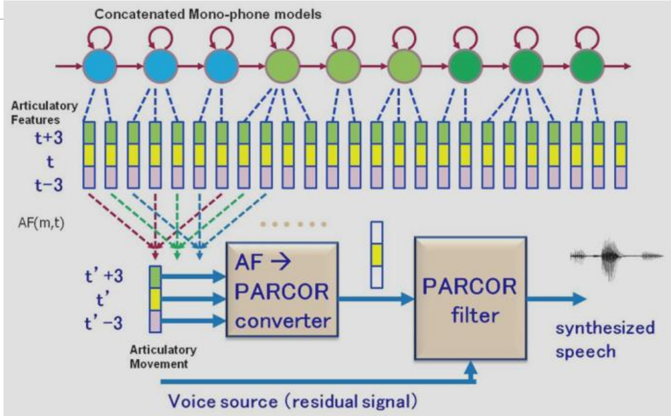

Figure 5: *HMM-based speech synthesis using articulatory movement model.*

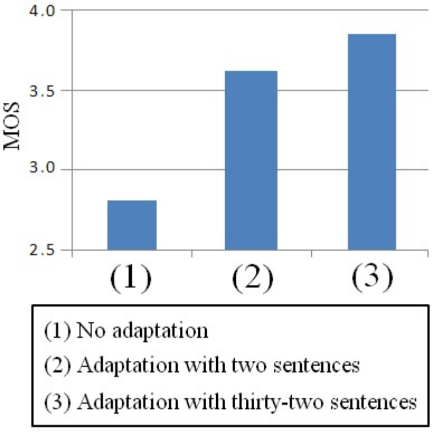

Figure 6: *MOS of Synthetic speech generated with articulatory movement HMMs. (voice source is residual signal)*

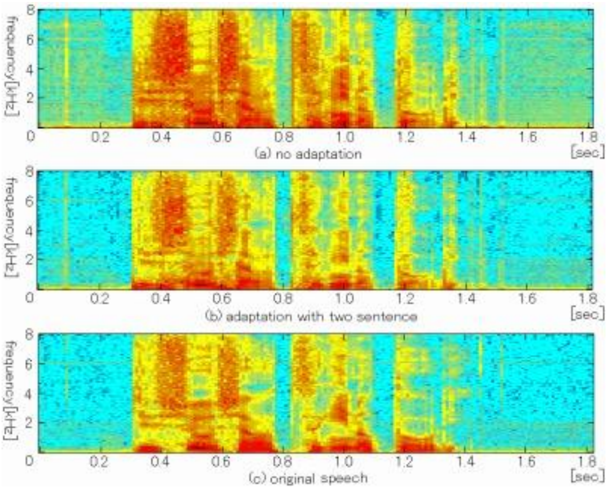

Figure 7: *Sound spectrogram of synthetic speech.*

Figure 6 shows MOS of synthetic speech with:
(1) no adaptation,
(2) two sentences adaptation,
(3) thirty two sentences adaptation.

In the MOS test, eleven subjects heard the original speech of the male speaker before testing took place. Residual signals of the PARCOR analysis are used for the voice source in this test.

Figure 6 shows the result of a MOS test. Adaptation with short sentences in the AF-PARCOR converter is found to be effective, however further improvement is needed. Figure 7 shows the sound spectrograms that result from (a) no adaptation, (b) adaptation with two sentences, and (3) original speech (target speaker).

5. Phoneme-to-phoneme speech conversion

The one-model SR and SS framework enables easy phoneme-to-phoneme speech conversion by exchanging the AFs of phoneme A to the AFs of phoneme B. Modified AF sequences are input to the AF-PARCOR converter, and speech sound is then output through the PARCOR filter, descrived in Figure 5.

Figure 8 shows the results of our experiments, where three phonemes, /b/, /p/, and /m/ are converted into /d/, /t/, and /n/, respectively [19]. In the ABX listening test, voice sources are applied with each residual signal of the original phoneme (b, p, m). Twelve subjects are then asked to judge the phoneme. The results show that the proposed method of using a combined, one-model SR and SS can deliver satisfactory phoneme-to-phoneme speech conversion, although further improvement is needed. Such functionality is useful for pronunciation training, where users' incorrect pronunciations can be compared with correct ones.

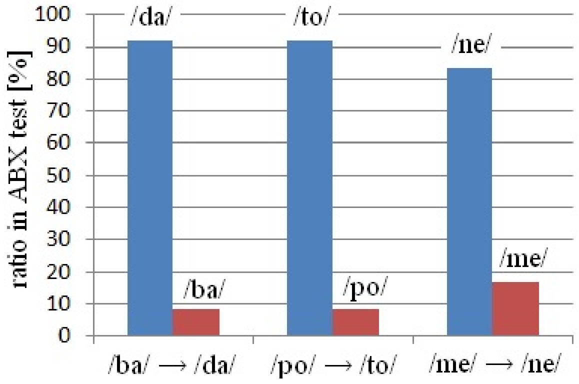

Figure 8: *HMM-based speech synthesis using articulatory movement model.*

6. Conclusion

One-model speech recognition and synthesis, based on common articulatory movement models were proposed herein. The articulatory movement HMMs showed high recognition performance, even when the training data are limited to a single mixture from a single speaker. In the SS engine, the same speaker-invariant HMMs generate AF sequences, and they are then converted into PARCOR parameters using a speaker-adapted MLN. Synthetic speech is realized by feeding the parameters to a PARCOR synthesizer.

Future work will include the improvement of voice quality in the SS engines, applying a CELP approach, as well as implementation of SR engines based on articulatory movement HMMs.

References

[1] Miller, J. L. and Eimas, P. D., Internal structure of voicing categories in early infancy, Percept. Psychophys., 58, 1157-1167 (1996).

[2] Liberman, A. M. and Mattingley, I. G.: The motor theory of speech perception revised, Cognition, 21, 1-36 (19845).

[3] King, S. and Taylor, P., Detection of phonological features in continuous speech using neural networks, Comput. Speech Lang., vol.14, no.4, pp.333-345 (2000).

[4] Eide, E, Distinctive features for use in an automatic speech recognition system, Proc. Eurospeech 2001, vol.III, pp.1613-1616 (2001).

[5] Kirchhoff, K., et al., Combining acoustic and articulatory feature information for robust speech recognition, Speech Commun., vol. 37, pp.303-319 (2002).

[6] Sivadas, S., and Hermansky, H., Hierarchical tandem feature extraction, ICASSP'02, vol.I, pp.809-812 (2002).

[7] Fukuda, T., Yamamoto, W. and Nitta, T., Distinctive phonetic feature extraction for robust speech recognition, Proc. ICASSP'03, vol.II, pp.25-28 (2003).

[8] Miller, G. A.: The science of word, Scientific American Library (1991).

[9] Wilson, S. M., Saygm, A.P., Sereno, M. I. and Iacoboni, M., Listening to speech activates motor areas involved in speech production, Nat. Neurosci., 7, 701-702 (2004).

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

[11] Itakura, F. and Saito, S., Analysis Synthesis Telephony based on the Maximum Likelihood, 6th ICA, C-5-5 (1968).

[12] 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. Interspeech'08, pp.1529-1532 (2008).

[13] 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).

[14] Nitta, T., Feature Extraction for Speech Recognition Based on Orthogonal Acoustic feature Planes and LDA, Proceedings of IEEE ICASSP'1999, pp.421-424 (1999).

[15] Fukuda, T. and Nitta, T., Noise-robust Automatic Speech Recognition Using Orthogonalized Distinctive Phonetic Feature Vectors, Proc. of Eurospeech 2003, Vol.III, pp.2189-2192 (2003).

[16] Fukuda, T. and Nitta, T., Orthogonalized Distinctive Phonetic Feature Extraction for Noise-robust Automatic Speech Recognition, IEICE Trans. Inf. & Sys, vol. E87-D, No. 5, pp.1110-1118 (2004).

[17] Kobayashi, T., Itahashi, S., Hayamizu, S. and Takezawa, T., "ASJ Continuous Speech Corpus for Research," Acoustic Society of Japan Trans. Vol.48, No.12, pp.888-893 (1992).

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

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