Estimation of Resonant Characteristics Based on AR-HMM Modeling and Spectral Envelope Conversion of Vowel Sounds

Nobuyuki Nishizawa*, *Keikichi Hirose***, *Nobuaki Minematsu****

*Graduate School of Engineering, University of Tokyo
**Graduate School of Frontier Sciences, University of Tokyo
***Graduate School of Information Science and Technology, University of Tokyo
{nishi, hirose, mine}@gavo.t.u-tokyo.ac.jp

Abstract

A new method was developed for accurately separating source and articulation filter characteristics of speech. This method is based on the AR-HMM modeling, where the residual waveform is expressed as the output sequence from an HMM. To realize an accurate analysis, a scheme of dividing HMM state was newly introduced. Using the AR-filter parameter values obtained through the analysis, we can construct a vocoder-type formant synthesizer, where the residual waveform is used as the excitation source. Through the listening test on the vowel sounds synthesized using AR-filter from a vowel and excitation waveform from another vowel, it was shown that a “flexible” synthesis with a high controllability on the acoustic parameters were possible by our formant synthesis configuration.

1. Introduction

Research works for speech synthesis have two major goals: high quality of synthetic speech and high controllability of acoustic features. Although a rather high quality has already been realized by the progress in waveform concatenation method, flexible control of acoustic features is still difficult. Flexible control is possible by analysis-synthesis methods, but obtainable speech quality is rather low. Recently, a good quality was realized even after a rather large change in F0 based on non-parametric analysis-synthesis methods, such as sinusoidal modeling[1] and STRAIGHT[2]. However, these methods tried to represent the envelope of speech spectrum as precisely as possible without considerations on the speech production process, and a large part of source features is wrongly included in the articulation filter parameters. Therefore, manipulation of filter parameters may easily lead to serious quality degradation. In contrast, parametric analysis-synthesis methods, such as those based on AR modeling with mathematical representation of voiced source waveform (ARX modeling)[3][4], could offer acoustic features (such as vocal tract transfer functions) with a good correspondence to the physical aspect of speech production. The major reason of their low quality is that the methods assume speech production modeling, which includes some mismatches with the real speech production process. Because of the complexity of speech production process, its full modeling is impossible. Therefore, we need a scheme to include such unseen features of speech production in a non-parametric expression. These considerations lead us to a semi-parametric modeling, where parametric representation of articulation filter and non-parametric representation of source waveform were combined.

The modeling is to represent articulation filter by an AR filter and source waveform by the outputs from an HMM. It was proposed by Sasoh *et al.*[5] and called AR-HMM modeling. Since the source HMM can be a good representation of distribution of voiced source waveform amplitudes, the obtained AR-filter coefficients can be a good estimation of articulation features.

The major problem of the AR-HMM modeling is that the data for training the HMM parameters are quite limited; the available data are those included in the analysis frame. We cannot solve this problem by reducing HMM state numbers, because the HMM’s ability of representing a source waveform is tightly related to the state numbers. To solve this situation, in the current paper, we newly introduced a recursive process to gradually increase the number of HMM states along the time axis.

After the analysis based on the AR-HMM modeling, we can only have AR-filter estimation, but cannot have source waveform estimation; the HMM output with the best likelihood will not be the counterpart of the estimated AR-filter. The source waveform should be calculated as the residual of the AR-filter. Therefore, in the framework of AR-HMM modeling, the speech synthesis is similar to the LP vocoder, where AR-filters are excited by the residual waveform. The major point of the method is that the estimated AR-filters is the better representation of the articulation, and can be modified with less degradation in the synthesized speech.

The rest of the paper is constructed as follows: In section 2, a vocoder type formant synthesizer is introduced as a possible configuration under the proposed AR-HMM modeling. Also, a hybrid configuration with waveform concatenative synthesis is proposed for synthesizing consonants, for which realization of good quality by formant synthesis is still difficult. Section 3 explains the AR-HMM modeling and the process of estimating its parameters. In section 4, an effective training method of the HMM part is proposed and is evaluated through analysis of synthetic speech of Japanese vowels. The method is based on recursively dividing HMM states. Section 5 shows the result of listening test for vowel sounds synthesized using source waveforms for different vowels, and section 6 concludes the paper.

2. Inverse-filtered waveform excited formant synthesis

If we select complex conjugate pole pairs for the AR-filter, the speech synthesizer comes a formant synthesizer with inverse-filtered waveform excitation. Since the formants have a good correspondence with the articulation filter features, a high flexibility for the parameter change can be realized. A problem for the synthesis method is the lack of controllability on source waveform features. Figure 1 shows a possible configuration of the speech synthesizer to cope with this problem. In the synthesizer, manipulation of voiced source waveform features, such as F0, is realized by the TD-PSOLA (Time-Domain Pitch-Synchronous OverLap and Add[6]). Also, the synthesizer has a hybrid configuration of formant and waveform concatenative synthesizers. Natural speech segments are utilized as consonants whose source-filter representation is difficult. Through an experiment, we already have shown that the hybrid of natural speech and synthesized speech will not degrade the final synthetic speech quality[7].

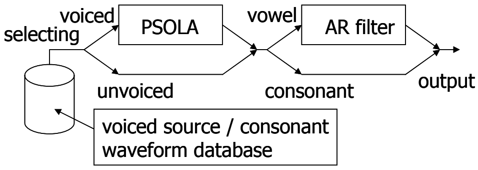


Figure 1: *Configuration of the hybrid synthesizer. The formant synthesizer is excited by the pre-stored source waveforms (obtained by the inverse-filtering) after the TD-PSOLA.*

3. AR-HMM modeling

Precision in the AR-filter estimation is very important to realize high quality in synthetic speech after parameter change. The exact formant information cannot be obtained by the well-known LP method, because it assumes a simple Gaussian process for the source waveform. Also analysis based on the ARX modeling cannot offer the exact information, because the mathematical expression of source waveforms cannot cover all the waveform movements. The AR-HMM modeling has an ability of precisely representing complicated source waveform features.

3.1. Structure of an AR-HMM model

Figure 2 shows structure of the AR-HMM model. In the model, a source waveform is represented as outputs from an HMM. Different from the case of well-known HMM used in speech recognition, it has a recursive structure with a returning path from the last state to the initial state. This structure corresponds to the periodicity of the voiced source waveform. Output probability of each state is modeled with a single Gaussian distribution. Irregularities in source waveform, which are difficult to be represented by mathematical formulae, can be automatically included as output probabilities through the HMM training. Thus, better separation of source and filter characteristics is expected than LP analysis and other related methods.

3.2. Estimation of model parameters

In the following model parameter estimation process, N, p and y_n respectively denote analysis frame length, order of AR-filter and output from AR-HMM model at time n. Also, $\tilde{\mu}_s$ and $\tilde{\sigma}_s$ indicate estimated mean and variance of Gaussian distribution for the HMM state s, respectively. Sequence $(S_p, S_{p+1}, \cdots, S_{N-1})$ is the state transition corresponding to the model output with maximum likelihood.

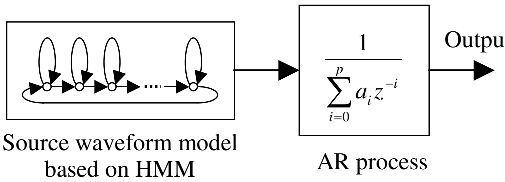


Figure 2: *Structure of an AR-HMM model.*

According to the proposal by Sasoh *et al.*, the AR-HMM model parameters are estimated through the following steps:

Step 1: Initialize source HMM and maximum likelihood state sequence $\{S_n\}_{n=p}^{N-1}$.

Step 2: Calculate AR parameters and residual with maximum likelihood. The estimated filter coefficient of AR process $\hat{\theta}$ is given by

$$\hat{\theta} = -[\Omega^T \tilde{\tilde{\Sigma}}_p^{-1} \Omega]^{-1} \Omega^T \tilde{\tilde{\Sigma}}_p^{-1} (\mathbf{y}_p - \tilde{\mathbf{m}}_p) \qquad (1)$$

where

$$\begin{aligned} \tilde{\mathbf{m}}_p &= [\tilde{\mu}_{s_p}\ \tilde{\mu}_{s_{p+1}}\ \cdots\ \tilde{\mu}_{s_{N-1}}]^T \\ \tilde{\tilde{\Sigma}}_p &= diag(\tilde{\sigma}^2_{s_p}, \tilde{\sigma}^2_{s_{p+1}}, \cdots, \tilde{\sigma}^2_{s_{N-1}}) \\ \Omega &= [\mathbf{y}_{p-1}\ \mathbf{y}_{p-2}\ \cdots\ \mathbf{y}_0] \\ \mathbf{y}_p &= [y_p\ y_{p+1}\ \cdots\ y_{N-1}]^T \end{aligned}$$

Step 3: Stop the estimation when likelihood of residual for HMM converges. Else, go to the next step.

Step 4: Update HMM parameters by Baum-Welch algorithm.

Step 5: Update maximum likelihood state sequence $\{S_n\}_{n=p}^{N-1}$ by Viterbi algorithm.

Step 6: Go to step 2.

4. HMM estimation with recursive state division process

To realize an accurate estimation of AR-filter coefficients using the AR-HMM modeling, its HMM part should have an ability of representing complex source waveform features. This requires an enough number of HMM states. However, since available data for HMM training are limited to the speech samples in the analysis frame, increase of the state numbers may result in the inaccurate modeling. Moreover, if the training data were distributed unevenly to each state, the variation of the output distribution of the state with smaller training data is tend to be estimated smaller. Since likelihood of the outputs from such state is calculated higher, likelihood of the outputs from other states is over-lightened. This situation will further degrade the analysis results. To cope with these problems, we have developed a method of recursively dividing HMM states to obtain a better modeling from the fixed number of data.

4.1. Recursive state division of the HMM

The method starts from a small number of states and increase the number by splitting the state with the largest training data along the time axis. This process ends when the number of states reaches a fixed value. In the case of male vowel sounds in

16 kHz sampling frequency, the start number and end number were set to 4 and 16, respectively, based on the result of a preliminary experiment. The state division process is done through the following steps:

1. Divide a pitch period into 4 segments with the same length.

2. Assign a segment to each of 4 states, and set their initial distributions to match with the LP residual.

3. Train the HMM by the Baum Welch algorithm and reassign the training data to each state by the Viterbi alignment.

4. Divide the state with largest number of training data. Division is done just adding the state with the same transition probabilities and output distribution.

5. Go back to the 3rd step.

The process ends at the 3rd step when the state number is 16.

4.2. Evaluation of formant extraction for synthetic vowel sounds of Japanese

In order to evaluate the proposed method, an experiment was conducted using synthetic speech of Japanese vowels (/a/, /i/, /u/, /e/, /o/), which was generated by the formant synthesizer with cascade connection of filters representing formants[8]. The merit of using synthetic speech for the analysis is that the correct answer is known and, therefore, a fair evaluation is possible. The FL (Fujisaki-Ljungqvist) glottal waveform model[3] was adopted to generate the excitation source of the synthesizer. Taking the sampling frequency (16 kHz) into consideration, the number of the formants was set to 6, which corresponded to 12 orders in the AR-filter (6 complex conjugate pole pairs). For each vowel, 9 samples with different F0 were prepared, since the method performance may change by the pitch period. The F0's were selected from 100.0 Hz to 251.2 Hz with equal separation in log frequency. The parameters of formant filters were fixed during speech synthesis for each vowel.

The method was compared with 2 baseline methods: 16-order LP analysis and AR-HMM without state division process. Since the AR-filters of the LP analysis may include real poles and additional complex poles to compensate over-simplified representation of source characteristics, the order of LP analysis was fixed to 16. The baseline AR-HMM had 16 states HMM from the first point. The first assignment of speech samples to each state was done by evenly dividing a pitch period into 16 segments. Henceforth, the AR-HMM method without state division and the proposed method shall be called ARHMM1 and ARHMM2, respectively. All samples were pre-emphasized by $1 - 0.97z^{-1}$ before analyses. For each method's results, root mean errors in formant frequencies and bandwidths were obtained. The errors were represented in logarithmic scale.

Figure 3 shows the errors for each method and for each vowel as the averages over 9 samples. In most of the cases, it is apparent from the figure that the errors are smaller for ARHMM2 than those for ARHMM1 and LP analysis. Notable improvements were observed in the bandwidth estimation, which was rather difficult by the LP analysis. Analysis based on the AR-HMM modeling occasionally causes large estimation errors when the HMM training process was done poorly. Such errors can be avoided by introducing the state dividing process as clearly shown in the figure (see the result for /i/ sound).

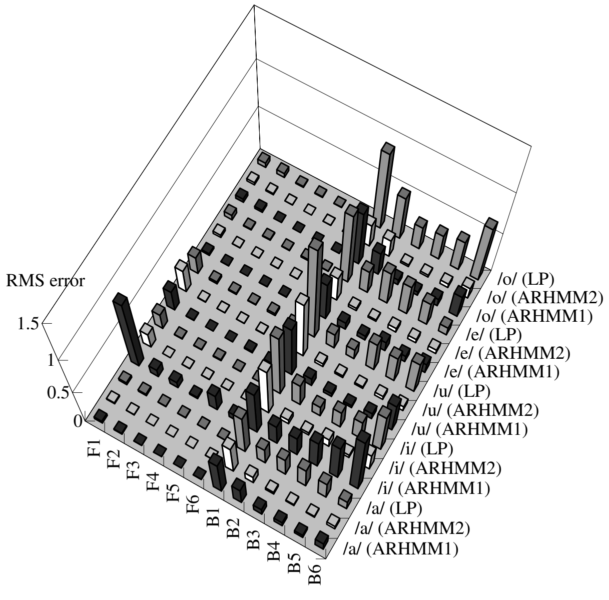


Figure 3: *Root mean square (RMS) errors of estimated formant frequency and bandwidth. Fx and Bx denote x-th formant frequency and x-th formant bandwidth respectively. ARHMM1 and ARHMM2 mean AR-HMM-based analyses without and with the proposed state division method, respectively. LP means LP analysis. Errors were evaluated in log frequency scale.*

5. Evaluation of vowel sound conversion

If the separation between source and filter is complete, quality of synthetic speech will not be degraded so much after a large change in the AR-filter parameters. In the extreme, even the source waveform was substituted to that obtained from another vowel, the synthesized speech may sound as the original vowel. Surely, due to the interaction between source and filter characteristics, source waveform for a vowel sound should be different from that for another vowel sound, and this difference may cause certain degradation. However, if the source-filter separation is complete, the degradation will be small, especially between vowels with similar articulation, such as /a/ and /o/, or /i/ and /e/ of Japanese.

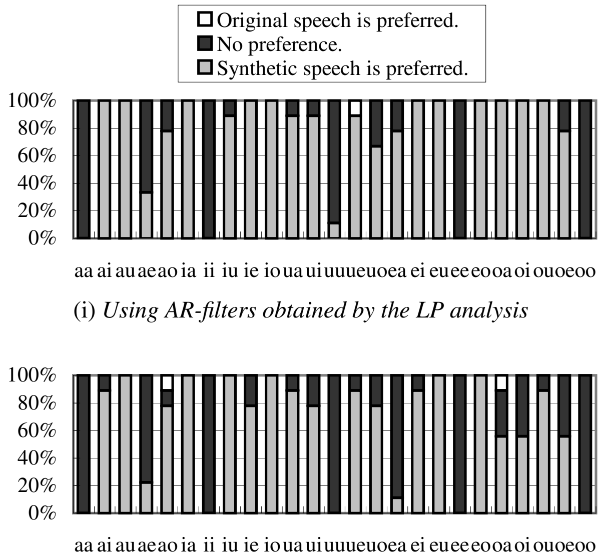
To prove this prospect, a listening test was conducted for the vowel sounds synthesized by changing the source waveforms to those obtained from the other vowel sounds. Vowel speech samples (one sample for each vowel) by male speaker MHT were selected form the ATR speech corpus[9]. Their F0's distributed around 110 to 120 Hz. They were first down-sampled from 20 kHz to 16 kHz for the further process. Pre-emphasis of $1 - 0.97z^{-1}$ was applied before the analysis. The frame length and shift were 336 samples and 160 samples, respectively.

The vowel sounds were first analysed using the proposed AR-HMM model-based method to obtain 6 complex conjugate pole pairs (corresponding to formants). The order of AR part was set to 16 for the analysis. When more than 6 complex conjugate pole pairs were obtained, 6 pairs with sharper resonance were selected. Then, the source waveform was calculated by applying the inverse filtering of the complex conjugate pole pairs thus obtained to the original speech waveform. This means that the real poles (and complex conjugate pole pairs with broad resonance) of the AR-filter were moved to the source waveform features.

The speech synthesis was conducted for all the combinations of 5 AR-filter sets and 5 source waveforms. The combinations included those of AR-filter sets and source waveforms selected from the same vowel sounds. The original speech without analysis-synthesis process was used as the reference speech. For each vowel, 5 pairs were prepared. (Random order of original speech and synthesized speech.) Total of 25 pairs were randomized and were presented to 9 subjects (Japanese native speakers) through a headphone. The interval between two pairs was set to 1 second. During the interval, the subjects were asked to select one from the following choices: (a) The first one is with better quality, (b) The second one is with better quality, and (c) Two sounds are with the same quality.

The listening test was also conducted using synthetic speech obtained by the 16-order LP analysis instead. The source waveform was obtained through the same procedure; select 6 complex conjugate pole pairs with sharp resonance and apply the inverse filtering using these pairs.

Figure 4 shows the result of the test. Each bar indicates whether the original speech or the synthetic speech was preferred. The alphabet pair at the bottom of each bar mans the combination of the AR-filter and source waveform; for instance “ai” means the combination of AR-filter from /a/ and source waveform /i/. Naturally, for “aa, ii, uu, ee, and oo,” two sounds of the pair are judged as the same quality. The result shows that, in most cases, the speech by the AR-HMM-based method sounded more natural than that by the LP method. Clear tendency on the relationship between AR-filter and source waveform combination and speech quality was not observed. Experiments, such as direct comparison between synthetic speech by the proposed method and by the LP analysis, are left for the future research.



(i) *Using AR-filters obtained by the LP analysis*

(ii) *Using AR-filters obtained by the proposed AR-HMM analysis*

Figure 4: *Result of the listening test.* V_1V_2 *at the bottom of each bar means that the synthetic speech is generated using AR-filter obtained from* V_1 *and source waveform obtained from* V_2.

6. Conclusion

An accurate separation of articulation filter and source characteristics was realized for vowel sounds by the analysis method based on AR-HMM modeling. A process of automatically dividing HMM states was successfully introduced for the better results. The results of the listening test on the synthetic speech, where filter and source features were transplanted from different vowels, indicated the possibility of realizing a “flexible” synthesis by the vocoder-type formant synthesizer, where the pre-stored inverse-filtered waveform was used as the excitation source. For the future work, a detailed examination on how the combination of the filter and source features influence the quality of the synthetic speech. Also, total evaluation of the speech synthesis system shown in Figure 1 is planned.

This work was partly supported by Grant-in-Aid for Scientific Research (#12480079).

7. References

[1] McAulay, R. J., and Quatieri, T. F., “Speech analysis/synthesis based on a sinusoidal representation,” IEEE Trans. on ASSP, vol. 34, no. 4, pp. 744–754, 1986.

[2] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A., “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, pp. 187–207, 1999.

[3] Fujisaki, H., and Ljungqvist, M., “Proposal and evaluation of models for the glottal source waveform,” In. Proc. ICASSP, vol. 31, no. 2, pp. 1605–1608, 1986.

[4] Ding, W., Kasuya, H., and Adachi, S., “Simultaneous estimation of vocal tract and voice source parameters Based on an ARX Model,” IEICE Trans. Inf. & Syst., vol. E78-D, no. 6, pp. 738–743, 1995.

[5] Sasoh, A., and Tanaka, K., “Glottal excitation modeling using HMM with application to robust analysis of speech signal,” Proc. of ICSLP2000, pp. 704–707, 2000.

[6] Moulines, E., and Charpentier, F., “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” Speech Communication, vol. 9, no. 5/6, pp. 453–467, 1990.

[7] Nishizawa, N., Minematsu, N., and Hirose, K., “Formant speech synthesis partly using waveform concatenative synthesis – Experimental study on VCV sounds –,” IEICE Techical Report, SP2001-20, vol. 101, no. 87, pp. 35–42, 2001 (in Japanese).

[8] Hirose, K., and Fujisaki, H., “A system for the synthesis of high-quality speech from texts on general weather conditions,” IEICE Trans. Fundamentals, vol. E76-A, no. 11, pp. 1971–1980, 1993.

[9] Abe, M., Sagisaka, Y., Umeda, T., and Kuwabara, H., Speech Database User’s Manual, ATR Interpreting Telephony Research Laboratories, TR-I-0166, 1990 (in Japanese).