SEPARATION OF VOICED SOURCE CHARACTERISTICS AND VOCAL TRACT TRANSFER FUNCTION CHARACTERISTICS FOR SPEECH SOUNDS BY ITERATIVE ANALYSIS BASED ON AR-HMM MODEL

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

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

ABSTRACT

A new method was developed for the separation of source and transfer function characteristics of speech sounds, with an aim of utilizing it to "flexible" speech synthesis. The method is based on representing source waveform by an HMM, and transfer function by the AR process (AR-HMM model). As compared to methods based on ARX model, where a parametric representation is assumed for source waveform, a better separation is possible. By introducing a process of recursively deleting real poles of AR filters, which represent source waveform features, and including them into HMM source waveform, the resulting AR filters may correctly represent transfer function features. Experiments were conducted for Japanese vowel sounds in continuous speech, and the results were compared with those by conventional LP analysis and AR-HMM model analysis without recursive process. After representing obtained source and transfer function features respectively as DFT cepstrum and LPC cepstrum, variations of cepstrum parameters for each vowel sound were compared for the three analysis methods. The smallest variations were obtained by the proposed method, indicating that the proposed method can separate source and transfer function features well, and, thus, has potential ability of generating good quality of speech when applied to "flexible" speech synthesis.

1. INTRODUCTION

In speech synthesis, methods based on the source-filter modeling, such as LP vocoder, are widely used, because control of fundamental frequency is easily realized. However, a good separation of speech sounds into source and articulation filter characteristics is not an easy problem to solve, causing certain limitations in the quality of synthetic speech sounds especially when their acoustic features (such as fundamental frequency) are changed a lot. Recently, several methods, like sinusoidal modeling[1], STRAIGHT[2] and so on, were developed for the good separation and applied to corpus-based speech synthesis. It was shown that, after a rather large change in fundamental frequency, a rather high-quality is maintained in the speech sounds. However, these methods are trying to represent spectral envelope as precisely as possible, and, thus, including source waveform features into filter parameters. This leads to the difficulty in concatenating filter parameters of stored segments in speech synthesis, and reduces the "flexibleness" in speech synthesis.

If the complete separation into source and filter characteristics is realized, we can rather freely change them to modify the speech quality without degradation. (Surely, because of interaction between source and filter in the speech production process, "truly complete" separation is impossible.) Since, in linear predictive analysis, the AR process is assumed, the obtained LP parameters mostly correspond to formants, and rather good information on filter characteristics is obtainable. However, the LP parameters usually include source waveform characteristics as real poles and others of the spectral envelope. Decision of appropriate LP orders is rather crucial.

In order to realize a good separation, several methods are tried to incorporate vocal source waveform models in the AR process[3][4]. As for the vocal source waveform models, they are mostly dividing the waveform of one fundamental period into several portions and representing each of them by a mathematical formula. The methods are called ARX model analysis ones, and a good separation is realized if the source waveform model can simulates the actual waveforms. However, this is not the case. Moreover, an iterative process is required because of non-linearity in the ARX model, making a reliable and automatic analysis difficult.

Based on these considerations, we tried to model the source waveform using HMM scheme so that a flexible representation of source waveform is possible, and, then include the model in AR process. This AR-HMM model was proposed already by Sasoh *et al.*[5], but it was without considerations on the good separation. We have newly incorporated an iterative process in which vocal source waveform features included in the AR filter are moved to the HMM source waveform by removing real poles.

The rest of the paper is constructed as follows: Section 2 gives an explanation on the AR-HMM model, section 3 proposes the analysis method developed for the good separation of source and vocal tract filter characteristics, and section 4 shows that the proposed method is better than the conventional LP analysis for the purpose of extracting parameters realizing flexible speech synthesis. Section 5 concludes the paper.

2. AR-HMM MODEL

Reliable analysis of speech sounds comes difficult when their F0 is high. AR-HMM model was first adopted for the speech analysis for such a case by Sasoh *et al.*, with results better than those by conventional LP analysis[5].

Figure 1 schematically shows the AR-HMM model. In the model, source waveform was represented as outputs from an HMM. Different from the case of well-known HMM used in speech recognition, it has a ring structure with a path from the last state to the initial state. This structure corresponds to the periodicity of the source waveform. Output probability of each state is modeled in a single Gaussian distribution. Irregularities in source waveform, which is difficult to be represented by mathematical formulae, can be automatically included as output probability through the HMM training. Thus, better separation of source and filter characteristics is expected than LP analysis and other related methods.

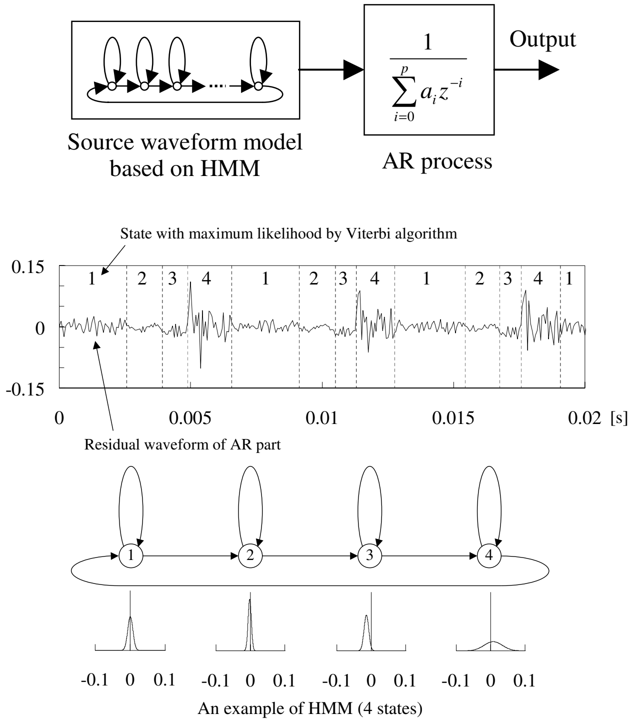
Output
$\frac{1}{\sum_{i=0}^{p} a_i z^{-i}}$
Source waveform model based on HMM
AR process
State with maximum likelihood by Viterbi algorithm
0.15
0
-0.15
1 2 3 4 1 2 3 4 1 2 3 4 1
0
0.005
0.01
0.015
0.02 [s]
Residual waveform of AR part
1
2
3
4
-0.1 0 0.1 -0.1 0 0.1 -0.1 0 0.1 -0.1 0 0.1
An example of HMM (4 states)

Fig. 1. Schematic illustration for AR-HMM model.

2.1. 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 state s of HMM, respectively. Sequence $(S_p, S_{p+1}, \cdots, S_{N-1})$ is the state transition corresponding to the model output with maximum likelihood.

According to proposal by Sasoh *et al.*[5], parameters of AR-HMM model is estimated as follows:

Step 1: Initialize source HMM and state sequence with maximum likelihood.

Step 2: Calculate AR parameters and predictive error sequence 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: End the estimation when likelihood of error sequence for HMM converges. Else, go to the next step.

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

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

Step 6: Go to step 2.

2.2. Comparison with LP Analysis

When analyzing speech by LP analysis method, its order is empirically decided as the double of estimated formant number plus a small number around 2 to 4. This is intended by the fact that a formant corresponds to a pair of conjugate complex poles, and spectral tilt originated by the source waveform features corresponds to a real pole. The extra poles may originate from the source waveform features or from the non-linearity in the speech generation process. Since actual source waveform shows rather complex features, the above order is not enough for the precise analysis. However, if we simply increase the order of LP analysis, the correspondence between poles and formants comes not clear. Moreover, by doing so, the source waveform characteristics come to be included in the LP parameters, and the LP residual may only represent the periodicity of the source waveform. This situation is far from what we planned; a good separation of source and transfer function characteristics.

Contrarily to the case of LP analysis, in AR-HMM model analysis, all the source waveform features can be included in the HMM representation. By increasing HMM states, we can easily increase the degree of freedom to represent complicated waveforms. However, this also reduces the stability of the analysis. In addition, since the HMM training is done by Baum-Welch algorithm, which is the process of finding a local-optimum, the initial assignment of HMM parameters largely affects the final result. The algorithm in section 2.1 trains HMM starting from LP residual-like waveform, and, therefore, the final result may converge to one close to the LP analysis result.

3. ITERATIVE ANALYSIS BASED ON AR-HMM MODEL

As mentioned already, the AR-HMM model can handle a complex shape in source waveform. However, AR part of AR-HMM model still includes components corresponding to the source waveform. Automatic separation of these components is rather difficult and the result may fluctuate a lot.

To cope with this problem, we developed an iterative analysis method based on the AR-HMM model. By this method, poles corresponding to source waveform features are removed from the AR part and their characteristics are included into the HMM part. At the result, AR parameters directly representing transfer functions and an HMM expressing source waveform can be obtained. The method is based on the assumption that the vocal tract transfer functions can be fully represented by the resonance poles corresponding to the formants.

3.1. Procedure

Analysis is done in the following steps:

Step 1: Analyze speech waveform using AR-HMM model as indicated in section 2.1.

Step 2: If no real pole is included in the estimated AR part, end the process. Else, proceed to the next step.

Step 3: Remove real poles form the AR part to obtain a new one.

Step 4: Re-calculate the source waveform by the inverse filtering using the new AR filter.

Step 5: Re-train the HMM and re-estimate AR-HMM parameters.

Step 6: Back to step 2.

To guarantee the procedure working correctly, HMM should have enough states to represent source waveforms. Currently this is decided empirically.

4. EXPERIMENTS AND EVALUATION

In order to evaluate the method's ability to separate source and transfer function features, experiments were conducted for Japanese vowel sounds (/a/, /i/, /u/, /e/, /o/) in ATR continuous speech corpus. These sounds were analyzed by the 3 methods: the conventional LP method, the method based on the original AR-HMM model and the proposed method. Comparison of the results was conducted for LPC cepstrum and DFT cepstrum, obtained respectively from vocal tract transfer functions and source waveforms.

For comparison, real poles obtained by the two reference methods are moved from transfer function features to source waveform features. The vowel sounds for analyses were those found in 50 sentences by male speaker MHT. Table 1 summarizes the sample number for each vowel. Although the original corpus is recorded with 20 kHz sampling and 16 bit accuracy, it is downsampled to 16 kHz for the current analyses. Pre-emphasis of $1 - 0.97z^{-1}$ was applied before the analyses. For each vowel included in the speech samples, a period of 528 samples points was segmented from the vowel center and was used for analysis. Through preliminary experiments, the order of LP analysis was fixed to 16, which was also the same with the order of AR filter of the original AR-HMM model and the initial AR filter order of the proposed method. The number of HMM states was set to 8.

Estimated vocal tract transfer function was represented as LPC cepstrum of 32nd order, while source waveform was windowed by Blackman window and represented as DFT cepstrum of 32nd order. Then, for each vowel, LPC and DFT cepstrum centers were calculated. The Euclidian distance from the center point in cepstrum space (not including 0th coefficient) was calculated for each sample. The mean squared distance may serve as an index for the consistency of the analysis; smaller value indicates a better separation of source and transfer function characteristics.

Table 1. Number of vowel samples used for the experiment.

vowel	a	i	u	e	o
number	325	223	197	185	279

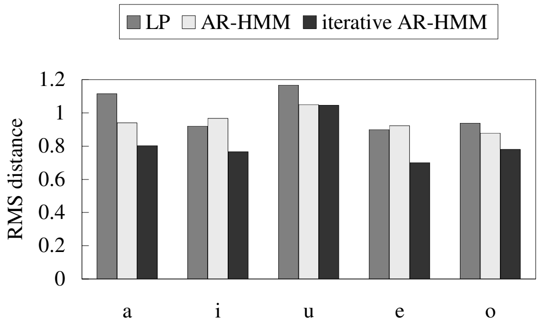

Fig. 2. Variance of estimated vocal tract characteristics of each Japanese vowel in the LPC cepstral space. Indicated as root mean square (RMS) distance between each sample point and the center.

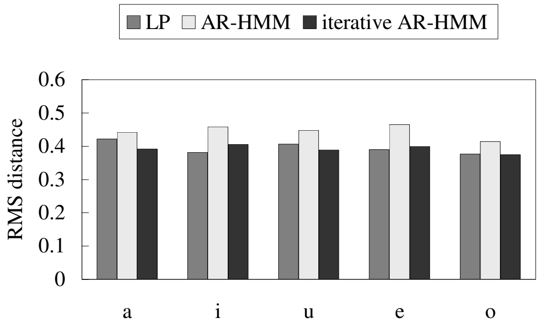

Fig. 3. Variance of estimated source characteristics of each Japanese vowel in the DFT cepstral space. Indicated as root mean square (RMS) distance between each sample point and the center.

Figures 2 and 3 show root mean square distances for LPC cepstrum (transfer function) and DFT cepstrum (source waveform), respectively. Smallest value in LPC cepstral distance is clearly obtained by the proposed method for every vowel. While, value in DFT cepstral distance by the proposed method is similar to that by the LP analysis. The larger values in DFT cepstrum by the original AR-HMM are due to the increased freedom in the source waveform modeling by HMM. The results implies that the acoustic parameters extracted by the proposed method represent the source and transfer function features well, and can be used for the "flexible" speech synthesis.

For further evaluation, standard deviations for /a/ sound by the three methods are summarized for each cepstral coefficient. The results are shown in Fig. 4 for LPC cepstrum and in Fig. 5 for DFT cepstrum. It is clear the standard deviations of the first order rep-

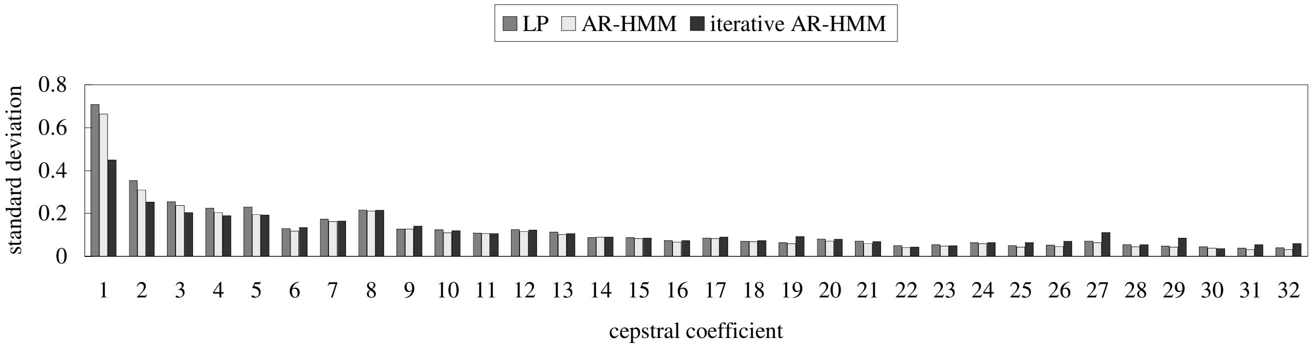

Fig. 4. Standard deviations of LPC cepstral coefficients for estimated vocal tract characteristics of vowel /a/. Numbers on the horizontal axis indicate cepstral coefficient order.

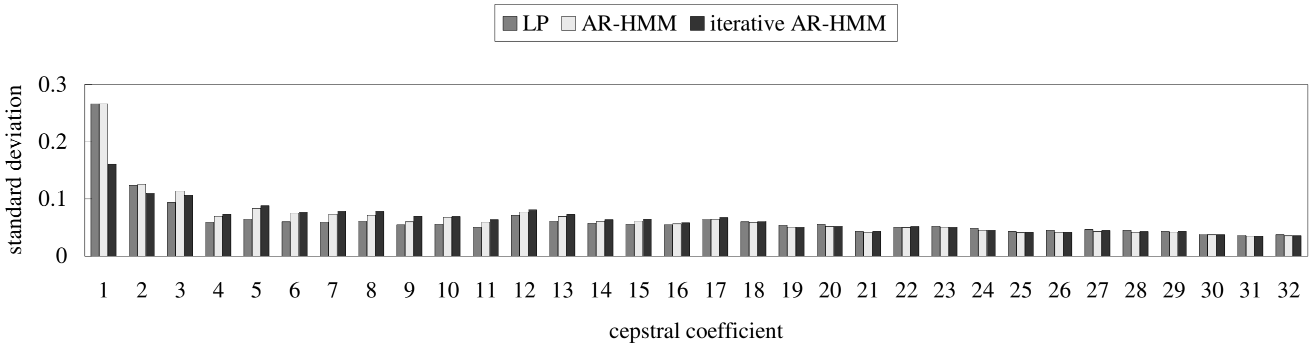

Fig. 5. Standard deviations of DFT cepstral coefficients for estimated source characteristics of vowel /a/.

resenting the spectral tilt are largely reduced for both cepstra. The larger deviations for the LP analysis method and the original AR-HMM model method are partly due to the fact that the AR filter order is fixed in these methods. When speech sounds are viewed within a fixed bandwidth, number of formant frequency varies and AR filter order should be changed accordingly. The results indicate that the recursive process introduced in the proposed method can cope with this problem.

5. CONCLUSION

With the final goal of realizing “flexible” speech synthesis, a source-filter analysis method was developed enabling a good separation of source and transfer function characteristics of speech sounds. The method is based on recursively finding the transfer function represented by a set of complex poles under the framework of AR-HMM model. Validity of the method was experimentally showed through analyses of Japanese vowel sounds in continuous utterances. Further evaluation experiments are going on for consonants. Also, hearing tests are planned for synthetic speech, whose source parameters (such as F0) are varied from the original values during the analysis-re-synthesis process by the proposed method.

6. REFERENCES

[1] McAulay, R. J., and Quatieri, T. F., “Speech Analysis/Synthesis Based On a Sinusoidal Representation,” IEEE Trans. on ASSP, ASSP-34, 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, 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.