AN AUTO-REGRESSIVE, NON-STATIONARY EXCITED SIGNAL PARAMETER ESTIMATION METHOD AND AN EVALUATION OF A SINGING-VOICE RECOGNITION

Akira Sasou[1], *Masataka Goto*[1], *Satoru Hayamizu*[2], *Kazuyo Tanaka*[3]

[1]National Institute of Advanced Industrial Science and Technology (AIST)
{a-sasou,m.goto}@aist.go.jp
[2]Faculty of Engineering, Gifu University
hayamizu@cc.gifu-u.ac.jp
[3]Institute of Library and Information Science, University of Tsukuba
ktanaka@ulis.ac.jp

ABSTRACT

We have previously described an Auto-Regressive Hidden Markov Model (AR-HMM) and an accompanying parameter estimation method. The AR-HMM was obtained by combining an AR process with an HMM introduced as a non-stationary excitation model. We demonstrated that the AR-HMM can accurately estimate the characteristics of both articulatory systems and excitation signals from high-pitched speech. In this paper, we apply the AR-HMM to feature extraction from singing voices and evaluate the recognition accuracy of the AR-HMM-based approach.

1. INTRODUCTION

The linear prediction (LP) method is widely used for the analysis of speech signals [1, 2]. However, several problems remain to be resolved. For example, (1) local peaks of LP spectral estimates are strongly biased toward harmonics, especially for high-pitched speech [3], and (2) the addition of white noise to the Auto-Regressive (AR) process markedly alters the spectral estimate [4]. These phenomena result in deterioration of the perceived quality of re-synthesized speech and can also cause speech recognition errors.

LP methods assume that the excitation signal conforms to an Identically Independent Distributed (IID) Gaussian. However, actual excitation signals exhibit non-stationary properties, especially in the case of a high fundamental frequency. As a result, local peaks in the LP spectral envelope estimated from high-pitched speech are strongly biased toward harmonics. To correct this, we have proposed an Auto-Regressive Hidden Markov Model (AR-HMM) and an accompanying parameter estimation method [5] in which the HMM is introduced as a non-stationary excitation model. We have also demonstrated that the proposed method can accurately estimate the characteristics of both articulatory systems and excitation signals from high-pitched speech.

Ozaki et al.[6] demonstrated that the recognition accuracy for a singing voice drastically deteriorates in comparison with that for normal speech, and that the deterioration is caused by high-pitched sounds and prolonged sounds. In [6], the Mel-Frequency Cepstral Coefficient(MFCC) was adopted as a speech feature. In the frequency domain, it is difficult for the MFCC to retain the formant information for each sound as is the case with the LPC because harmonic components of high-pitched speech become sparse. In order to overcome the difficulties of singing-voice recognition, we applied the speech analysis method based on the AR-HMM. Using the AR-HMM-based method, we extracted the features of the singing voices and evaluated the recognition accuracy.

2. AUTO-REGRESSIVE HIDDEN MARKOV MODEL

Previously, we proposed an AR-HMM that was obtained by combining an AR process with an HMM introduced as a non-stationary excitation model. Figure 1 illustrates an example of the AR-HMM. The output probability distribution of each node in the excitation HMM is assumed to be a single Gaussian. The nodes in the figure are concatenated in a ring state, so the state transitions occur in order. Therefore, this type of AR-HMM can be used to represent periodically excited signals. The AR-HMM can represent various types of signals through appropriate design of the network topology. The number of nodes and the prediction order are determined according to the signal.

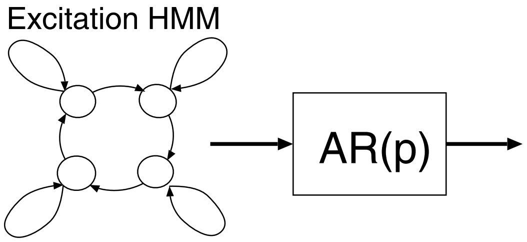

Fig. 1. Example of AR-HMM.

3. ITERATIVE AR-HMM PARAMETER ESTIMATION METHOD

The AR-HMM parameters are the AR coefficients and the parameters of the HMM. Previously, we presented an algorithm that iteratively estimates these parameters from a signal $x(t), t = 0, \cdots, T-1$ [5]. In the following, P denotes the prediction order of the AR process. Let $\mathbf{a}^{(i)} = [a^{(i)}(1), \cdots, a^{(i)}(P)]^T$ represent the ith estimate of the AR coefficients. The ith estimate of the excitation signal $e^{(i)}(t), t = P, \cdots, T-1$ is given by:

$$\mathbf{e}_P^{(i)} = \mathbf{x}_P - \Omega \mathbf{a}^{(i)} \tag{1}$$

where

$$\mathbf{e}_P^{(i)} = [e^{(i)}(P)\; e^{(i)}(P+1)\; \cdots\; e^{(i)}(T-1)]^T \in R^{T-P},$$

$$\mathbf{x}_t = [x(t)\; x(t+1)\; \cdots\; x(t+T-P-1)]^T \in R^{T-P},$$

$$\Omega = [\mathbf{x}_{P-1}\; \mathbf{x}_{P-2}\; \cdots\; \mathbf{x}_0] \in R^{(T-P)\times P}$$

We allocate a unique number from $S = \{1, \cdots, N\}$ to each node of the excitation HMM to distinguish them from other nodes, where N is the number of nodes. Let $\mu_n^{(i)}, \sigma_n^{2\,(i)}, n \in S$ represent the ith estimates of the output distribution population parameters in each node. Given a state-transition sequence $s(t) \in S, t = P, \cdots, T-1$, the population parameters of an excitation signal at time t are given by $m^{(i)}(t) = \mu_{s(t)}^{(i)}, v^{(i)}(t) = \sigma_{s(t)}^{2\,(i)}$. Hence, the expectation vector of the excitation signal vector is represented by:

$$\mathbf{m}_P^{(i)} = [m^{(i)}(P)\; m^{(i)}(P+1)\; \cdots\; m^{(i)}(T-1)]^T \tag{2}$$

Based on the assumption that the samples of the excitation signal at different instants are mutually independent, the covariance matrix of the excitation signal vector is defined as a diagonal matrix given by:

$$\Sigma_P^{(i)} = \text{diag}(v^{(i)}(P), v^{(i)}(P+1), \cdots, v^{(i)}(T-1)) \tag{3}$$

The algorithm for parameter estimation consists of the following processes.

1. The initial population parameters of the excitation signal are prepared as $\mathbf{m}_P^{(0)} = \mathbf{0}$, $\Sigma_P^{(0)} = \mathbf{I}$. Repeat the following processes from $i = 0$.

2. The AR coefficients $\mathbf{a}^{(i+1)}$ and the excitation signal $\mathbf{e}_P^{(i+1)}$ are estimated by maximizing the likelihood given by $L(\mathbf{e}_P^{(i+1)}; \mathbf{m}_P^{(i)}, \Sigma_P^{(i)})$.

3. The population parameters $\mathbf{m}_p^{(i+1)}$, $\Sigma_p^{(i+1)}$ of the excitation signal vector are estimated by maximizing the likelihood given by $L(\mathbf{e}_P^{(i+1)}; \mathbf{m}_P^{(i+1)}, \Sigma_p^{(i+1)})$.

4. If the likelihood has converged, the algorithm stops. Otherwise, repeat the above processes for $i \leftarrow i+1$ from step 2.

By repeating the above processes, the likelihood increases almost monotonically in practical situations and converges to the optimum or to a local optimum value.

The details of each step are as follows. In step 2, the AR coefficient vector can be obtained by

$$\mathbf{a}^{(i+1)} = [\Omega^T(\Sigma_P^{(i)})^{-1}\Omega]^{-1}\Omega^T(\Sigma_P^{(i)})^{-1}(\mathbf{x}_P - \mathbf{m}_P^{(i)}). \tag{4}$$

The excitation signal vector $\mathbf{e}_P^{(i+1)}$ is derived from (1).

In step 3, the population parameters of the excitation signal vector are estimated according to the following processes.

3.1 The Baum-Welch algorithm estimates the population parameters $\mu_m^{(i+1)}, \sigma_m^{2\,(i+1)}, m \in S$ of each output distribution using $\mathbf{e}_P^{(i+1)}$.

3.2 The Viterbi algorithm estimates a state transition sequence $s(t), t = P, P+1, \cdots, T-1$.

3.3 The expectation vector $\mathbf{m}_P^{(i+1)}$ and the diagonal covariance matrix $\Sigma_P^{(i+1)}$ of the excitation signal vector are estimated using (2) and (3).

4. TRANSFORMATION OF AR-HMM PARAMETERS TO MFCC

The transformation of the AR coefficient to an LPC cepstrum can be efficiently computed using a simple recursive formula.

$$c(n) = -a(n) - \frac{1}{n}\sum_{i=1}^{n-1}(n-i)a(i)c(n-i). \tag{5}$$

Also, we can obtain an LPC Mel-Cepstrum converting the frequency axis to a Mel-frequency axis in a simple recursive way.

Almost all recent Large-Vocabulary Continuous Speech Recognition (LVCSR) decoders have adopted the MFCC as a speech feature, so it would be useful if the AR-HMM-based features could be directly recognized by the MFCC-based decoders. This requires transforming the AR-HMM parameters into corresponding features of the MFCC. In order to do that, we processed the AR coefficients according to the following steps. The first step is an evaluation of the logarithmic spectral amplitude,

$$u(n) = -\log\left|1 - \sum_{i=1}^{P} a(i)\exp(-i2\pi n/N)\right|, \tag{6}$$

which corresponds to the logarithmic amplitude of the FFT in a typical MFCC calculation. As a second step, Mel-filter banks are constructed by summing the logarithmic spectral amplitudes, weighted by triangle windows. Finally, we can obtain the AR-HMM-based MFCC by calculating the DCT for the Mel-filter bank outputs.

Table 1. Fundamental frequencies [Hz]

Song No.	3	4	7	11	21	27	34	37	41	44	55	74
Avg.	315.3	279.4	416.9	261.5	409.5	275.1	342.2	305.1	234.1	225.0	365.5	216.3
Std.Dev.	84.38	89.69	76.54	54.99	72.04	60.01	99.26	48.39	44.80	61.12	105.6	30.21

5. EXPERIMENTS

5.1. Popular Music Database

For our experiments, we used 12 Japanese songs of the popular-music database *"RWC Music Database: Popular Music"* (RWC-MDB-P-2001 No. 3, 4, 7, 11, 21, 27, 34, 37, 41, 44, 55, 74) [7]. Instead of using the original audio signals after mixdown, we used the "Vocal-only Version" that records only the vocal part. The singers of those 12 songs consisted of 5 females and 7 males.

The audio signals of the 12 songs were manually segmented into several vocal phrases of time interval ranging from 2s to 6s. A total of 580 phrases were generated. Table 1 presents the average and standard deviation of fundamental frequencies of each song.

5.2. Acoustical Assessment of AR-HMM-Based Feature

In this section, we describe an acoustic assessment we performed of the AR-HMM-based MFCC. First, we extracted the AR-HMM parameters from all the vocal phrases in all 12 songs. The analysis frame size was set to 25ms. The frame shift was set to 10ms. The prediction order was set to 16, and the number of nodes in the excitation HMM was set to a range of 10 to 14. The evaluated AR coefficients were transformed into MFCC-compatible features using the method described in Section 4. In addition, a conventional MFCC was evaluated for comparison.

The number of nodes to be contained in the excitation HMM for each singer was determined as follows. First, we evaluated the forced-alignment acoustic scores for the AR-HMM-based MFCC. The number of nodes was then determined by selecting the highest acoustic score. For the acoustic score evaluations, we adopted a Phonetically Tied Mixture (PTM) model, which was trained by using conventional MFCCs extracted from the continuous speech corpus of Japanese Newspaper Article Sentences (JNAS) [8]. Trained PTM model was thus completely open for singing voices. The forced-alignment acoustic scores were evaluated using Julius, an LVCSR decoder [9].

Table 2. Number of Nodes Selected

Song No.	3	4	7	11	21	27	34	37	41	44	55	74
No.of Nodes	10	10	10	10	14	10	11	11	10	12	10	10

Table 3. Acoustic scores evaluated by forced alignment

Song No.	3	4	7	11	21	27	34	37	41	44	55	74	Avg.
ARHMM	-24.47	-24.77	-24.97	-24.41	-26.47	-24.33	-25.16	-24.69	-25.55	-23.97	-25.39	-23.98	-24.85
MFCC	-27.00	-27.37	-27.29	-26.35	-28.51	-26.55	-26.94	-26.90	-27.66	-26.19	-28.04	-26.68	-27.12

Table 2 presents the final node counts for the 12 songs. Table 3 shows the evaluated acoustic scores for the selected AR-HMM-based MFCC, as well as the acoustic scores for the conventional MFCC. The acoustic scores are normalized by the number of frames. All of the acoustic scores for the AR-HMM-based MFCC were better than those of conventional MFCC.

5.3. Recognition Results

Each vocal phrase was processed by Julius, the two-pass LVCSR decoder [9]. A language model and a dictionary were prepared for each song, which were generated from the lyrics using Chasen, a Japanese morphological analysis system [10]. Each coefficient of back-off smoothing was set to an extremely small value. The acoustic models used were the same as those used for the acoustic assessment. A weighting coefficient for the language model and a word insertion penalty were optimized for each feature of the AR-HMM-based MFCC and the conventional MFCC. Correct word and error rates were evaluated from the recognition results from the first pass. The error rate was evaluated by summing substitutions, deletions and insertions. Tables 4 and 5 present the correct word and error rates, respectively. These results show that when compared with those from the conventional MFCC, the AR-HMM-based feature brought improvements of 3.63% and 1.85% in correct word and error rates, respectively.

5.4. Model Adaptation

The recognition experiment described above was conducted under the condition that the AR-HMM-based MFCC was mismatched to the acoustic model. In this section, we prepared two adapted acoustic models. One model was adapted to half the amount of the AR-HMM-based MFCCs, which were obtained from the vocal phrases of even numbers. The other model was adapted to half the amount of the conventional MFCCs. Each Gaussian component in each acoustic model is transformed using a linear transformation estimated by means of Maximum Likelihood Linear Regression (MLLR). We iterated this adaptation process eight times.

Table 4. Correct word rate [%]

Song No.	3	4	7	11	21	27	34	37	41	44	55	74	Avg.
ARHMM	79.52	70.00	55.39	76.14	45.62	75.29	77.39	86.88	54.55	84.94	42.22	92.53	70.04
MFCC	74.10	65.45	52.45	73.30	36.41	74.12	65.22	80.00	70.91	91.63	51.85	61.49	66.41

Table 5. Error rate [%]

Song No.	3	4	7	11	21	27	34	37	41	44	55	74	Avg.
ARHMM	27.71	31.82	62.75	37.50	63.13	35.29	30.43	17.50	49.09	17.57	76.30	14.37	38.62
MFCC	33.73	36.36	62.75	32.39	71.43	33.53	46.09	21.25	30.91	12.97	60.00	44.25	40.47

Table 6. Correct word rate [%] (using adapted models)

Song No.	3	4	7	11	21	27	34	37	41	44	55	74	Avg.
ARHMM	93.37	85.45	90.20	78.41	76.50	79.41	86.96	86.88	85.45	96.23	74.07	85.63	84.88
MFCC	91.57	75.45	86.27	80.11	48.85	90.59	79.13	82.50	81.82	92.89	76.30	64.37	79.15

Table 7. Error rate [%] (using adapted models)

Song No.	3	4	7	11	21	27	34	37	41	44	55	74	Avg.
ARHMM	8.43	15.45	14.71	31.25	29.03	35.29	26.09	15.00	17.27	7.11	40.74	18.97	21.61
MFCC	11.45	24.55	19.61	27.84	56.68	10.00	31.30	18.12	18.18	11.30	31.85	40.23	25.09

Tables 6 and 7 represent the results that the eighth adapted acoustic models recognized the corresponding features, respectively. Comparing these results with those from the the conventional MFCC, the AR-HMM-based feature brought improvements of 5.73% and 3.48% in the correct word and the error rates, respectively. In comparison with the conventional MFCC recognized by acoustic models without adaptation, the AR-HMM-based MFCC with the model adaptation brought improvements of 18.47% and 18.86% in the correct word and the error rates, respectively.

6. CONCLUSIONS

In this paper, we applied AR-HMM to extraction of a feature from singing voices. The effectiveness of the AR-HMM-based feature was confirmed from acoustic assessments and singing-voice recognition experiments.

In future studies, moreover, to achieve higher recognition accuracy for a singing voice, the influence of prolonged sounds should be taken into account. We also need to consider formant variations, particular with singing voices. For instance, the singing-formant appears in the frequency band from 2.5kHz to 3kHz in singing voices. Another example is that the first formant frequency tends to shift to the fundamental frequency when that frequency is higher than the first formant frequency.

7. REFERENCES

[1] F.Itakura and S.Saito, “A statistical method for estimation of speech spectral density and formant frequencies,” Electronics and Communications in Japan, Vol.53-A, No.1, pp.36-43, January 1970.

[2] B.S.Atal and S.L.Hanauer, “Speech analysis and synthesis by linear prediction of the speech wave,” J. Acoust. Soc. Am., Vol.50, pp.637-644, 1971.

[3] J.Makhoul, “Linear Prediction: A Tutorial Review,” in Proc.of IEEE, Vol.63, No.4, pp.561-580, April 1975.

[4] S.M.Kay, “The Effects of Noise on the Autoregressive Spectral Estimator,” IEEE ASSP-27, No.5, pp.478-485, Oct. 1979.

[5] A.Sasou, M.Goto, S.Hayamizu, K.Tanaka, “Comparison of Auto-Regressive, Non-Stationary Excited Signal Parameter Estimation Methods,” Proc. of IEEE MLSP, Sep. 2004.

[6] H.Ozeki, T.Kamata, M.Goto, S.Hayamizu, “The influence of vocal pitch on lyrics recognition of sung melodies,” Proc. of Acoust. soc. Japan, Vol.1, pp.637-638, Sep. 2003(in Japanese).

[7] M.Goto, “Development of the RWC Music Database”, Proc. of ICA 2004, pp.I-553-556, April 2004.

[8] K.Itoh et al., “JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research,” J. Acoust. soc. Japan (E), Vol.20, No.3, pp.199-206, March, 1999.

[9] http://julius.sourceforge.jp/en/julius.html

[10] http://chasen.naist.jp/hiki/ChaSen/