SINGING VOICE RECOGNITION CONSIDERING HIGH-PITCHED AND PROLONGED SOUNDS

Akira Sasou

National Institute of Advanced Industrial Science and Technology (AIST)
a-sasou@aist.go.jp

ABSTRACT

A conventional Large Vocabulary Continuous Speech Recognition (LVCSR) system has difficulty recognizing singing voices accurately because both the high-pitched and prolonged sounds of singing voices tend to degrade its recognition accuracy. We previously described an Auto-Regressive Hidden Markov Model (AR-HMM) and an accompanying parameter estimation method. We demonstrated that the AR-HMM accurately estimated the characteristics of both articulatory systems and excitation signals from high-pitched speech. In this paper, we describe an AR-HMM applied to feature extraction from singing voices and propose a prolonged-sound detection and elimination method.

1. INTRODUCTION

Conventional Large Vocabulary Continuous Speech Recognition (LVCSR) systems can recognize declarative sentences with high accuracy. However, they are still subject to (1) the presence of background noise and/or reverberant environments and (2) variation of utterance style, for instance, spontaneous speech, emotional speech, singing, etc. In particular, there have been few studies on singing-voice recognition.

Ozaki et al.[1] demonstrated that recognition accuracy for a singing voice drastically deteriorates in comparison with that for normal speech and that the deterioration is caused by high-pitched as well as prolonged sounds. In [1], 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 because harmonic components of high-pitched speech become sparse.

The linear prediction (LP) method is widely used for analyzing speech signals [2, 3]. LP methods assume that the excitation signal conforms to an Identically Independent Distributed (IID) Gaussian. However, actual excitation signals exhibit non-stationary properties, especially for a high fundamental frequency. As a result, local peaks in the LP spectral envelope estimated from high-pitched speech are strongly biased toward harmonics, as is the case with the MFCC. To correct this, we proposed an Auto-Regressive Hidden Markov Model (AR-HMM) and an accompanying parameter estimation method [5] in which the HMM was introduced as a non-stationary excitation model. We also demonstrated that the proposed method could accurately estimate the characteristics of both articulatory systems and excitation signals from high-pitched speech.

To overcome the difficulties of singing-voice recognition, we apply a speech analysis method based on the AR-HMM and propose a prolonged-sound detection and elimination method. We then evaluate the recognition accuracy of the AR-HMM-based feature extraction method and the prolonged sound detection and elimination method with respect to singing voices.

2. FEATURE EXTRACTION METHOD

2.1 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.

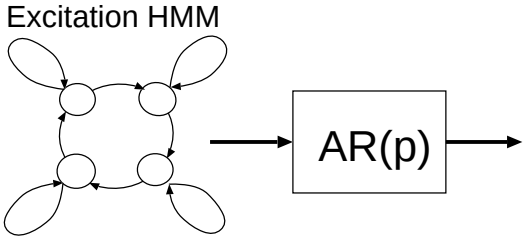

Figure 1: Example of AR-HMM.

2.2 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:

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

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:

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

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

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

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).

2.3 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.

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

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,

$$u(m) = -\log \left| 1 - \sum_{i=1}^{P} a(i) \exp(-i 2\pi m / N) \right|, \qquad (6)$$

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.

3. PROLONGED SOUND DETECTION AND ELIMINATION

Spectral envelope deformation during a prolonged sound is expected to be small. We adopted delta features in order to detect such small deformations of the spectral envelopes. We then eliminated the frame features identified in a prolonged sound and used the remaining frame features for recognition. This method is referred to as the derivative method[6]. The details of this process are as follows.

Let $\Delta c(n, i)$ represent ith delta Cepstral coefficient at frame n. We first evaluate the following quantity.

$$s(n) = \sum_i \{\Delta c(n, i)\}^2 \qquad (7)$$

The $s(n)$ is then smoothed with a moving-average filter.

$$l(n) = \frac{1}{2M+1} \sum_{m=-M}^{M} s(n+m) \qquad (8)$$

When the $l(n)$ of successive N_r frames drops below the threshold value l_{thr}, the group of frames is identified as a prolonged sound. The following frames are eliminated as long as the quantities remain below the threshold value. The number of frames identified in a prolonged sound is thereby restricted to N_r.

Figure 2 illustrates an example of prolonged-sound detection. The top of the figure represents a vocal waveform. The second figure represents its spectral envelopes. The third figure represents $l(n)$ evaluated from the AR-HMM-based MFCC. The bottom figure represents the durations within which frames are eliminated. In this experiment, the frame shift size was 10ms, and l_{thr}, N_r and M were set to 10.

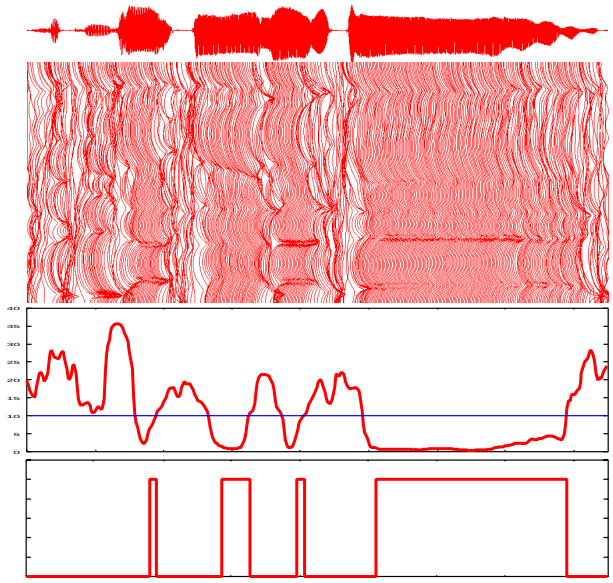

Figure 2: Example of prolonged sound detection

4. EXPERIMENTS

4.1 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.

Table 1: Fundamental frequencies [Hz]

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

4.2 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 2.3. In addition, a conventional MFCC was evaluated for comparison.

Table 2: Number of Nodes Selected

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

Table 3: Acoustic scores evaluated by forced alignment

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

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 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.

4.3 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. 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 2.06% and 1.49% in correct word and error rates, respectively.

Table 4: Correct word rate [%]

Song No.	3	4	7	11	21
ARHMM	76.51	73.64	59.80	77.84	55.76
MFCC	72.29	61.82	54.41	75.57	40.55
Song No.	27	34	37	41	44
ARHMM	70.59	76.52	86.25	61.82	85.77
MFCC	71.76	73.04	86.25	75.45	90.79
Song No.	55	74	Avg.		
ARHMM	46.67	95.98	72.26		
MFCC	49.63	90.80	70.20		

Table 5: Error rate [%]

Song No.	3	4	7	11	21
ARHMM	34.94	30.00	63.73	44.89	55.76
MFCC	41.57	46.36	69.61	38.64	76.96
Song No.	27	34	37	41	44
ARHMM	50.00	40.87	22.50	44.55	20.08
MFCC	37.65	46.09	18.75	28.18	15.06
Song No.	55	74	Avg.		
ARHMM	79.26	14.94	41.79		
MFCC	79.26	21.26	43.28		

Table 6: Correct word rate [%]

Song No.	3	4	7	11	21
ARHMM	81.93	87.27	65.20	76.14	50.23
MFCC	78.92	65.45	62.25	76.70	43.78
Song No.	27	34	37	41	44
ARHMM	88.82	79.13	89.38	64.55	91.21
MFCC	77.65	74.78	88.75	76.36	89.54
Song No.	55	74	Avg.		
ARHMM	54.07	96.55	77.04		
MFCC	57.78	92.53	73.71		

Table 7: Error rate [%]

Song No.	3	4	7	11	21
ARHMM	24.70	14.55	50.49	39.20	58.99
MFCC	30.12	39.09	52.45	31.82	72.35
Song No.	27	34	37	41	44
ARHMM	14.12	31.30	17.50	38.18	14.64
MFCC	29.41	38.26	14.37	28.18	17.57
Song No.	55	74	Avg.		
ARHMM	57.78	7.47	30.74		
MFCC	56.30	16.09	35.50		

4.4 Recognition Results for both High-Pitched and Prolonged Sounds

The recognition experiment from our earlier study described above only consider high-pitched sound. In this section, we consider both high-pitched and prolonged sounds. In order to eliminate the effects of prolonged sounds, we applied the prolonged-sound compensation method described in section 3 to the AR-HMM-based and conventional MFCCs obtained from the previous experiment. Tables 6 and 7 present the recognition results. The prolonged-sound compensated, AR-HMM-based MFCCs improved the correct word rate by 3.33% and the errror rate by 4.76% over similarly compensated conventional MFCCs, and by 6.84% and 12.54% over the baseline performance in which the conventional, uncompensated MFCCs were used for recognition.

5. CONCLUSIONS

In this paper, we applied the AR-HMM to feature extraction from singing voices. We also described a method to detect and eliminate prolonged sounds. The results of acoustic assessments and singing-voice recognition experiments confirmed the effectiveness of the AR-HMM-based feature extraction and the prolonged-sound compensation method.

Furthermore, to achieve higher recognition accuracy for singing voices, future studies will need to consider all the formant variations peculiar to a singing voice. For instance, the singing formant appears in the frequency band from 2.5kHz to 3kHz. Another example is that the first formant frequency tends to shift to the fundamental frequency when the fundamental frequency becomes higher than the first formant frequency.

Acknowledgment

The author would like to thank Dr. M.Goto for providing database and valuable discussions.

REFERENCES

[1] 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).

[2] 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.

[3] 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.

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

[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] P.L.Cerf, D.V.Compernolle, "A New Variable Frame Rate Analysis Method for Speech Recognition," IEEE Signal Porcessing Letters, Vol.1, No.12, Dec. 1994.

[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/