GLOTTAL EXCITATION MODELING USING HMM WITH APPLICATION TO ROBUST ANALYSIS OF SPEECH SIGNAL

Akira SASOU and Kazuyo TANAKA

Electrotechnical Laboratory, AIST, MITI
1-1-4 Umezono, Tsukuba, Ibaraki 305-8568, Japan
{sasou, ktanaka}@etl.go.jp

ABSTRACT

This paper describes a robust analysis method for high fundamental frequency speech signal. In the proposed method, a Hidden Markov Model (HMM) is applied in order to represent the non-stationary property of the glottal source. Experiments are carried out using both synthetic and natural speeches to confirm the effectiveness of the method. Experimental results indicate (1) in the case of using synthetic speech in the pitch range of up to 750Hz, the proposed method can precisely estimate the original spectrum, and (2) the spectrum estimated from natural speech of pitch frequency 666Hz is less affected by the harmonics of glottal excitation, compared with result of the conventional method.

1. INTRODUCTION

The linear prediction method is widely used as the analysis of speech signal[1]. However, several problems lie in the method[2]. For example, there is a fact that the estimated vocal tract characteristics are distorted by the fundamental frequency. This deteriorates the perceived quality of re-synthesized speech and also can be the cause of speech recognition errors. In order to improve this difficulty, the DAP method has been proposed[3]. However, it is not enough to be able to analyze high fundamental frequency speech like singing voice or emotional speech.

Conventional linear prediction methods have adopted a assumption of Gaussian to the prediction error. In the case of analyzing high fundamental frequency speech, the assumption does not indicate characteristics of glottal source appropriately. The proposed method adopts an HMM model of glottal source, which can be robust to the change of fundamental frequency. We describe in the following sections a procedure for the analysis method based on the HMM glottal source model and several experimental results for confirming the feasibility of the method.

2. GLOTTAL SOURCE MODELING BY USING HMM

A conventional linear prediction method adopts the assumption of Identically Independent Distribution (IID) with normal distribution to the prediction error. On the other hand, a rapidly change of glottal volume flow waveform in the glottal closed phase causes a large prediction error. Because the prediction errors occur synchronizing with the vocal fold vibration, the occurrence probability becomes higher as the fundamental frequency increases. Then, the probability distribution of short width like normal distribution cannot represent that of the glottal source. Due to this, in the case of analyzing high fundamental frequency speech, the separation of the vocal tract characteristics from the glottal source becomes difficult.

In order to model the glottal source having non-Gaussian property, conventional method uses the assumption that the glottal source is stationary and conforms to the probability distribution of wider width than normal distribution[4]. On the other hand, the proposed method uses the model that adopts the non-stationary property of the glottal source. That is, the proposed method regards the glottal source as the non-stationary signal that stochastically moves for several stationary states. A probability distribution of each stationary state is assumed to be a single of normal distribution. In the case of analyzing periodic waveform like voiced speech, each state inside one period is assumed to be the same as the each state inside other periods. From these assumptions, each node of the HMM glottal source model is connected in a ring state as shown in the figure 1. By making the transition one direction, the model can represent the periodicity.

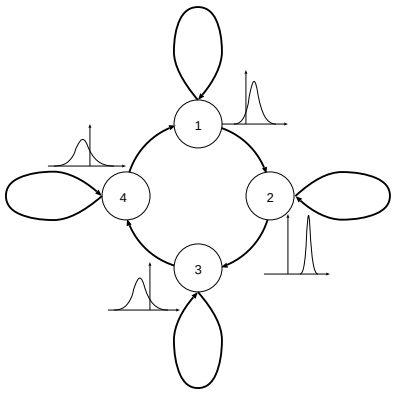
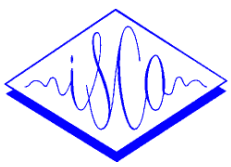


Figure 1: 4 states HMM

ISCA Archive
http://www.isca-speech.org/archive



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

3. PROPOSED METHOD

3.1. Estimating Population Parameters of Glottal Source via Maximization Likelihood Method

Separation of an observed speech signal into the vocal tract characteristics and the glottal source is a kind of blind system identification, which must be achieved in a condition of that the glottal source can not be referred to. The proposed HMM model regards glottal source as a random variable that the population parameters vary periodicity. Assuming that the observed speech signal is generated by the HMM model, we have studied about what kind of condition is necessary to separate the vocal tract characteristics and glottal source from the observation. In order to analyze the problem, the generation process of speech is modeled by using a glottal source signal and a vocal tract filter. And the analysis process is modeled by linear Prediction Error Filter (PEF). Then, we found that, the PEF coincides with the inverse filter of the vocal tract, if the population parameters of the prediction error coincide with those of the glottal source. The Markov estimation method can be used for estimating the inverse filter of the vocal tract on condition that the population parameters of the glottal source are available in advance. However, those are unknown previously.

We propose the following procedure that estimates the population parameters.

1. The hypothesis population parameters of the glottal source are prepared.
2. The prediction error is estimated in order that the probability of the prediction error conditioned by the population parameters is maximized.
3. The hypothesis population parameters that maximize the likelihood to the prediction error are adopted as the estimate of the glottal source population parameters.

The reasons that such a procedure is taken into account are as follows. In the case that the hypothesis population parameters differ from the real population parameters of the glottal source, it is difficult to obtain the prediction error that matches to the probability distribution. On the other hand, in the case that the hypothesis population parameters coincide with the real distribution of the glottal source, then the most matched signal to the distribution becomes the glottal source itself, and it is possible to obtain the signal on condition that the PEF coincides with the inverse filter of the vocal tract. At this time, the likelihood of the hypothesis population parameters to the obtained prediction error is maximized in the meaning of the expectation. So, the the hypothesis population parameters that maximizes the likelihood can be thought as the estimation of the glottal source population parameters. Based on this consideration, we constructed the parameter estimation algorithm that maximizes the likelihood, which is described in the next section.

The glottal source is assumed to be modeled by an HMM with the M states. Each state has a unique number from 1 to M for identifying it. And $s_n \in S = [1, \cdots, M]$ stands for the state of the HMM at time n. The population parameters of each state are given by $\mu_m, \sigma_m^2, m \in S$. Then the glottal source e_n's population parameters at time n are given by $m_n = \mu_{s_n}, v_n = \sigma_{s_n}^2$.

The generation process of speech is assumed to be modeled with AR process of order p. By using the glottal source vector $\mathbf{e}_p = [e_p \ e_{p+1} \ \cdots \ e_{N-1}]^T$ and the coefficient vector of vocal tract filter $\mathbf{a} = [a_1 \ a_2 \ \cdots \ a_p]^T$, the observed speech vector $\mathbf{x}_p = [x_p \ x_{p+1} \ \cdots \ x_{N-1}]^T$ is given by

$$\mathbf{x}_p = \Omega\mathbf{a} + \mathbf{e}_p \tag{1}$$

where $\Omega = [\mathbf{x}_{p-1} \ \mathbf{x}_{p-2} \ \cdots \ \mathbf{x}_0]$.

The analysis process is achieved by inverse filtering with PEF of order p. Let $\mathbf{b} = [b_1 \ b_2 \ \cdots \ b_p]^T$ and $\tilde{\mathbf{e}}_p = [\tilde{e}_p \ \tilde{e}_{p+1} \ \cdots \ \tilde{e}_{N-1}]^T$ represent the coefficient vector of PEF and the prediction error vector, respectively. Then, the prediction error vector is given by

$$\tilde{\mathbf{e}}_p = f_{\tilde{e}}(\mathbf{b}) = \mathbf{x}_p - \Omega\mathbf{b} \tag{2}$$

Let $\mathbf{m}_p = [m_p \ m_{p+1} \ \cdots \ m_{N-1}]^T$ represent the expectation of the glottal source vector $\mathbf{e}_p$. The covariance matrix of $\mathbf{e}_p$ is given by $\Sigma_p = \mathrm{diag}(v_p, v_{p+1}, \cdots, v_{N-1})$, assuming that the glottal sources are mutually independent. On the other hand, the expectation vector $\tilde{\mathbf{m}}_p$ and the covariance matrix $\tilde{\Sigma}_p$ of the prediction error vector $\tilde{\mathbf{e}}_p$ are given by

$$\begin{aligned}
\tilde{\mathbf{m}}_p &= f_{\tilde{m}}(\mathbf{b}) \\
&= \mathbf{m}_p - E[\Omega](\mathbf{b} - \mathbf{a}) \\
\tilde{\Sigma}_p &= f_{\tilde{\Sigma}}(\mathbf{b}) \\
&= \Sigma_p - E[(\mathbf{e}_p - \mathbf{m}_p)(\mathbf{b} - \mathbf{a})^T\{\Omega - E[\Omega]\}^T] \\
&- E[\{\Omega - E[\Omega]\}(\mathbf{b} - \mathbf{a})(\mathbf{e}_p - \mathbf{m}_p)^T] \\
&+ E[\{\Omega - E[\Omega]\}(\mathbf{b} - \mathbf{a})(\mathbf{b} - \mathbf{a})^T\{\Omega - E[\Omega]\}^T]
\end{aligned} \tag{3}$$

From these relationships, we found that the PEF's coefficients $\mathbf{b}$ becomes equal to the vocal tract filter's coefficients $\mathbf{a}$, provided that the prediction error's population parameters $\tilde{\mathbf{m}}_p$,$\tilde{\Sigma}_p$ coincide with those of the glottal source $\mathbf{m}_p$,Σ_p.

In order to estimate the population parameters $\mathbf{m}_p$, Σ_p, we introduce a diagonal covariance matrix $\tilde{\tilde{\Sigma}}_p$ defined as

$$\tilde{\tilde{\Sigma}}_p = \mathrm{diag}(\tilde{v}_p^2, \tilde{v}_{p+1}^2, \cdots, \tilde{v}_{N-1}^2) \tag{4}$$

where $\tilde{v}_p^2, \tilde{v}_{p+1}^2, \cdots, \tilde{v}_{N-1}^2$ are the diagonal elements of $\tilde{\Sigma}_p$. Then the expectation of the likelihood $L(\tilde{\mathbf{e}}_p|\tilde{\mathbf{m}}_p, \tilde{\tilde{\Sigma}}_p)$ is given by

$$E[L(\tilde{\mathbf{e}}_p|\tilde{\mathbf{m}}_p, \tilde{\tilde{\Sigma}}_p)] = \int L(\tilde{\mathbf{e}}_p|\tilde{\mathbf{m}}_p, \tilde{\tilde{\Sigma}}_p) L(\tilde{\mathbf{e}}_p|\tilde{\mathbf{m}}_p, \tilde{\Sigma}_p) d\tilde{\mathbf{e}}_p \tag{5}$$

which is maximized when $\tilde{\tilde{\Sigma}}_p = \tilde{\Sigma}_p$. So, we can derive the glottal source population parameters when maximizing the equation (5).

3.2. Parameter Estimation Algorithm

The parameter estimation algorithm consists of the following processes.

1. The initial hypothesis population parameters to the glottal source are prepared as $\tilde{\mathbf{m}}_p^{(0)} = \mathbf{0}$, $\tilde{\tilde{\Sigma}}_p^{(0)} = \sigma\mathbf{I}$.
2. The PEF $\mathbf{b}$ and the prediction error $\tilde{\mathbf{e}}_p^{(i+1)}$ are evaluated in order to maximize $L(\tilde{\mathbf{e}}_p^{(i+1)}|\tilde{\mathbf{m}}_p^{(i)}, \tilde{\tilde{\Sigma}}_p^{(i)})$
3. The population parameters $\tilde{\mathbf{m}}_p^{(i+1)}$, $\tilde{\tilde{\Sigma}}_p^{(i+1)}$ are estimated in order to maximize $L(\tilde{\mathbf{e}}_p^{(i+1)}|\tilde{\mathbf{m}}_p^{(i+1)}, \tilde{\tilde{\Sigma}}_p^{(i+1)})$.
4. If the likelihood has converged, the process ends up. Otherwise, the process is repeated from step 2.

The estimated likelihood in each step has the relationship as follows.

$$\begin{aligned} L(\tilde{\mathbf{e}}_p^{(i)}|\tilde{\mathbf{m}}_p^{(i)}, \tilde{\tilde{\Sigma}}_p^{(i)}) &\leq L(\tilde{\mathbf{e}}_p^{(i+1)}|\tilde{\mathbf{m}}_p^{(i)}, \tilde{\tilde{\Sigma}}_p^{(i)}) \\ &\leq L(\tilde{\mathbf{e}}_p^{(i+1)}|\tilde{\mathbf{m}}_p^{(i+1)}, \tilde{\tilde{\Sigma}}_p^{(i+1)}) \end{aligned}$$

By repeating the procedure, the likelihood increases monotonously and is converging to a certain value (an optimum value or local optimum value).

In the step 2, the PEF that maximizes the logarithm probability is given by

$$\hat{\mathbf{b}} = [\Omega^T \tilde{\tilde{\Sigma}}_p^{-1} \Omega]^{-1} \Omega^T \tilde{\tilde{\Sigma}}_p^{-1} (\mathbf{x}_p - \tilde{\mathbf{m}}_p) \tag{6}$$

In the step 3, estimation of population parameters is achieved as follows.

3.1 The Baum-Welch algorithm estimates the population parameters $\tilde{\mu}_m, \tilde{\tilde{\sigma}}_m^2, m \in S$ of each state.

3.2 The Viterbi algorithm estimates the state transition sequence $s_p, s_{p+1}, \cdots, s_{N-1}$ that maximizes the likelihood of the HMM to the prediction error.

3.3 The expectation vector $\tilde{\mathbf{m}}_p$ and the diagonal covariance matrix of the prediction error $\tilde{\tilde{\Sigma}}_p$, are estimated by using the state transition sequence as follows.

$$\tilde{\mathbf{m}}_p = [\tilde{\mu}_{s_p}, \tilde{\mu}_{s_{p+1}}, \cdots, \tilde{\mu}_{s_{N-1}}]$$

$$\tilde{\tilde{\Sigma}}_p = \mathrm{diag}[\tilde{\tilde{\sigma}}_{s_p}^2, \tilde{\tilde{\sigma}}_{s_{p+1}}^2, \cdots, \tilde{\tilde{\sigma}}_{s_{N-1}}^2]$$

4. EXPERIMENTS

4.1. Experiment with Synthetic Speech

The experiment with synthetic speech is carried out. The glottal source used in the synthesis consists of impulse train and noises. The amplitude value of impulse is set at 50 and is located at every fundamental period. White noises of normal distribution are also added between the impulses. Accordingly, the 2 states HMM can be used to represent the glottal source. The synthetic speeches are synthesized with AR process of order 16, driven by the glottal source of different pitch frequencies ranging from 100Hz to 900Hz. The AR parameters used in the synthesis are extracted from natural speech of vowel sound /a/ uttered by a Japanese speaker. The HMM glottal source model used for the analysis has 2 states. Each node has two paths, that is, one is connected to itself and another path is for transition to another node. Also, the order of the PEF is 16. For the purpose of comparison, the same experiment is achieved by using the auto-correlation linear prediction method with Hanning window. The prediction order is 16. The analysis frame width is 30ms.

The vocal tract spectra extracted by the proposed and conventional methods are shown as in the figure 2 and 3 respectively. The vocal tract spectra extracted by the proposed method have the similar formant structure to the original one in the fundamental frequency range of up to 750Hz. In the case of the conventional method, the extracted spectra differ from the original one in the fundamental frequency range of higher than 400Hz. And the lowest frequency peak of the extracted spectra in this range indicates not the 1st formant of the original spectrum but the fundamental component of the pitch.

4.2. Experiment with Natural Speech

The natural speech of vowel sound /a/ uttered by female was recorded digitally with sampling rate of 48kHz and quantized with 16bits. After that, the digital speech signal was re-sampled with the rate of 16kHz. The pitch frequency was about 666Hz. The speech signal is pre-emphasized with the coefficients 0.99 before analyzed.The analysis frame width is 30ms. First, the algorithm must select the state number of the HMM. This is achieved by using the Akaike information criteria[5]. From the evaluated results, the state number of the HMM glottal source model is set at 15.

The vocal tract spectra extracted by the proposed and the conventional methods are shown in the figure 4 and 5 respectively. The lowest frequency peak appears at about 700Hz in the vocal tract spectrum estimated by the conventional method, which corresponds with the fundamental frequency component of the pitch. On the other hand, the vocal tract spectrum estimated by the proposed method does not contain such a peak of the frequency. The figure 6(b) shows the glottal source waveform estimated by the proposed method, in which the evident glottal source pulses are observed in the interval of the fundamental period. On the other hand, the glottal source waveform estimated by the conventional method, as shown in the figure 6(c), does not contain any evident glottal source pulse. This means that the PEF predicts the changes caused by the glottal source. In other words, the characteristics of the PEF contain the components corresponding with the glottal source.

5. CONCLUSION

We have applied an HMM to modeling glottal source in order to capture the non-stationary property, and described the analysis method of speech signal based on the HMM glottal source model. The experimental results for the synthetic speech show that the proposed method can extract the characteristics of the vocal tract precisely from the synthetic speeches of fundamental frequencies up to 750Hz. The spectrum estimated from a natural speech of pitch frequency 666Hz is less affected by the harmonics of glottal excitation, compared with result of the conventional method. From these results, we confirm that the proposed analysis method is robust to high fundamental frequency speech.

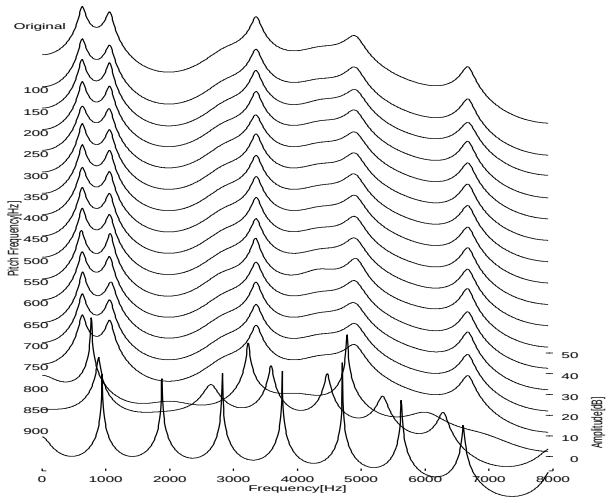


Figure 2: Vocal tract spectra estimated from synthetic speeches by using proposed method.

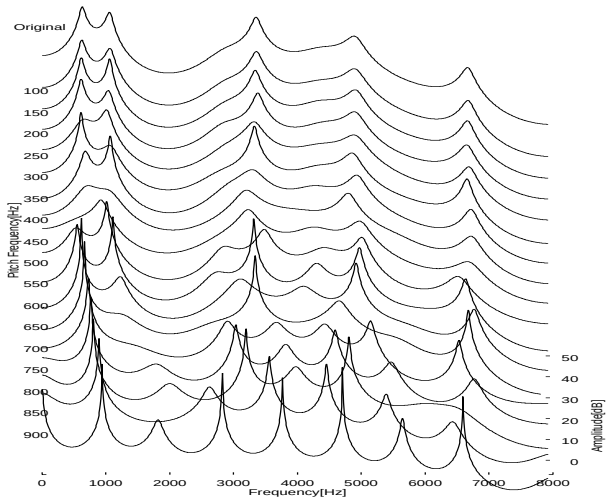


Figure 3: Vocal tract spectra estimated from synthetic speeches by using conventional method.

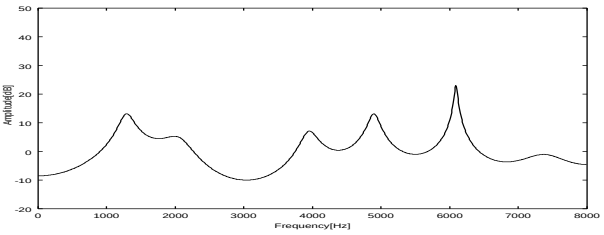


Figure 4: Vocal tract spectrum estimated from natural speech of pitch frequency 666Hz by using proposed method

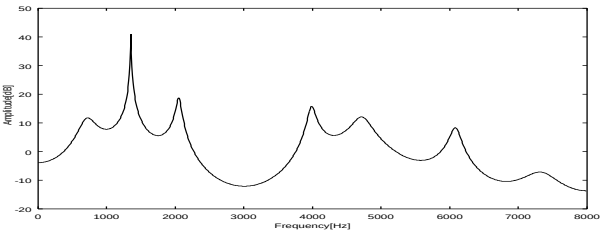


Figure 5: Vocal tract spectrum estimated from natural speech of pitch frequency 666Hz by using conventional method

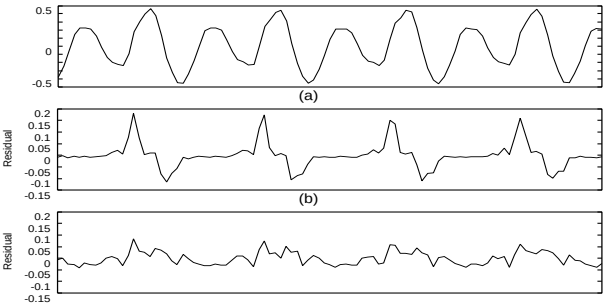


Figure 6: Estimated glottal sources, (a)natural speech waveform, (b)result of proposed method, (c)result of conventional method.

REFERENCES

[1] 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

[2] J.Makhoul,"Linear Prediction: A Tutorial Review,"Proc.of IEEE, Vol.64, No.4, April 1975

[3] A.El-Jaroudi and J.Makhoul, "Discrete All-Pole Modeling," IEEE Trans. SP39, pp.411-423, 1991

[4] Chin-Hui Lee, "On robust linear prediction of speech,"IEEE Trans. ASSP-36, 1988

[5] H.Akaike,"A New-Look at the Statistical Model Identification,"IEEE Trans. on Automatics Control. AC-19-6, pp.716-723, 1974