Robust LP Analysis Using Glottal Source HMM with Application to High-Pitched and Noise Corrupted Speech

Akira Sasou and Kazuyo Tanaka

National Institute of Advanced Industrial Science and Technology
1-1-4 Umezono, Tsukuba, Ibaraki 305-0045,JAPAN
a-sasou@aist.go.jp, kaz.tanaka@aist.go.jp

Abstract

This paper presents a robust feature extraction method effective to speech signal with high fundamental frequency and/or corrupted by additive white noise. The method represents the glottal source wave using HMM in order to model the non-stationary properties. The nodes of HMM are concatenated in a ring state to represent the periodicity of voiced sounds. The method can accurately extract glottal source wave and vocal tract characteristics from speech signals even in high fundamental frequency as ranging up to 750Hz. From identification theory, estimation of vocal tract characteristics from speech corrupted by additive noise requires glottal source wave that can not be observed directly, so that it needs to be estimated. Therefore, estimation accuracy of vocal tract characteristics highly depends on the estimation accuracy of glottal source wave. We apply the glottal source HMM to extracting the glottal source wave from corrupted speech, and confirmed the feasibility of the method.

1. Introduction

The linear prediction (LP) method is widely used as the analysis of speech signal[1]. However, several problems still remain to be solved. For instance, (1) local peaks of LP spectral estimate are strongly biased toward the harmonics, especially for high-pitched speech[2], (2) addition of white noise to the Auto-Regressive (AR) process drastically changes the spectral estimate[3]. These phenomena deteriorate the perceived quality of re-synthesized speech and also can cause speech recognition errors.

We already showed that the LP analysis incorporating glottal source HMM (Hidden Markov Model) has the ability to estimate the characteristics of both vocal tract and glottal source precisely from clean high-pitched speech signal [4]. In this method, the glottal source wave is modeled by HMM in order to represent its non-stationary property and the nodes of the HMM are concatenated in a ring state in order to represent the periodicity of voiced sounds. The method is able to accurately estimate original vocal tract characteristics as well as glottal source wave in the fundamental frequency range of up to 750Hz.

In this paper, we presents an extended algorithm which keeps robustness in applying to not only high fundamental frequency speech but also speech with additive white noise. In the following sections, we first describe basic formulations and specific procedures, and next show application results to confirm the feasibility of the method.

2. Glottal Source Modeling by HMM

Conventional linear prediction methods assume that the glottal source conforms to an Identically Independent Distributed (IID) normal distribution. However, especially in the case of high fundamental frequency, the actual glottal source indicate non-stationary properties. The proposed method applies an HMM to modeling the glottal source. The nodes of the HMM are concatenated in a ring state in order to represent the periodicity of voiced sounds. Fig.1 shows an example of a glottal source HMM with 4 states. In the general case that the HMM has M states, each state is identified by the unique number from 1 to M. By making the transition in one direction, the HMM can represent the periodicity of voiced speech. A probability distribution of each state is assumed to be a single, normal distribution. That is, the mth state of the HMM has the population parameters $\mu(m), \sigma^2(m)$. The analysis algorithm presented in [4] is based on maximum likelihood method.

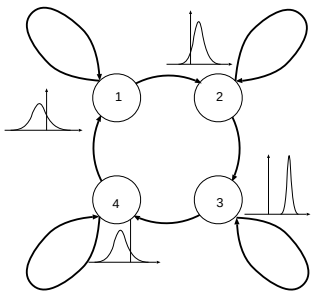

Figure 1: *Glottal source HMM with 4 states.*

3. The Extended Algorithm

3.1. LP analysis based on blind system identification

The estimation of vocal tract characteristics from an observed speech can be regarded as a kind of blind system identification problem which involves a model shown in Fig.2, where only the observation speech $y(n)$ is available for processing in the identification of vocal tract and glottal source. The additive noise $v(n)$ is assumed to be White Gaussian $N(0, \sigma_v^2)$ and the variance σ_v^2 is also unknown.

From the point of view of system identification theory, the estimation of the vocal tract characteristics from the speech corrupted by additive white noise can be achieved by using Least Square (LS) method or Instrumental Variable (IV) method. However, the estimation process of these methods require the

glottal source, which can not be observed directly and needs to be estimated. Thus, the estimation accuracy of the vocal tract characteristics highly depends on the estimation accuracy of the glottal source. The proposed method employs the glottal source HMM in order to extract glottal source from corrupted speech.

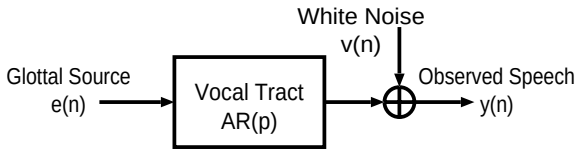

Figure 2: *Schematic of blind system identification.*

According to the model shown in Fig.2, the observation speech $y(n)$ is represented by

$$y(n) = \frac{1}{A(z^{-1})}e(n) + v(n) \tag{1}$$

where $A(z^{-1}) = 1 + \sum_{k=1}^{p} a_k z^{-k}$, $e(n)$ and $v(n)$ are the glottal source and the white noise respectively. By canceling the denominator, we can rewrite Eq.(1) as

$$y(n) + \sum_{k=1}^{p} a_k y(n-k) = e(n) + w(n) \tag{2}$$

where $w(n)$ is the filtered white noise given by

$$w(n) = v(n) + \sum_{k=1}^{p} a_k v(n-k) \tag{3}$$

Eq.(2) means that we can regard the observation speech $y(n)$ as the AR process driven by the combined excitation signal $g(n) = e(n) + w(n)$.

Let $\mathbf{g}_p$, $\mathbf{e}_p$ and $\mathbf{w}_p$ be the combined excitation, the glottal source and the filtered white noise vectors such as $\mathbf{g}_p = [g(p), \cdots, g(N-1)]^T$ and so, where N is a number of samples in an analysis frame. These are random variable vectors which conform to $N(\mathbf{m}_g, \mathbf{C}_g)$, $N(\mathbf{m}_e, \mathbf{C}_e)$ and $N(\mathbf{m}_w, \mathbf{C}_w)$, respectively.

The iterative algorithm described in [4] consists of the following 2 processes.

1. The one is maximum likelihood estimation of the AR coefficients with assuming the glottal source HMM.
2. The other is maximum likelihood estimation of the glottal source HMM with assuming the AR coefficients.

By iteration of the processes, the likelihood increases monotonously and is converging to an optimum value or local optimum value.

Due to the presence of additive noise, the extended algorithm has to consider not only the glottal source HMM but also the filtered white noise. The modification of each process is described as follows.

3.2. Estimation of AR coefficients

In this process, we assume the population parameters of the glottal source HMM $N(\mathbf{m}_e, \mathbf{C}_e)$ and the filtered white noise $N(\mathbf{m}_w, \mathbf{C}_w)$. The estimation processes of those parameters are described in the next section.

The population parameter of the combined excitation signal $N(\mathbf{m}_g, \mathbf{C}_g)$ is given by

$$\mathbf{m}_g = \mathbf{m}_e + \mathbf{m}_w, \quad \mathbf{C}_g = \mathbf{C}_e + \mathbf{C}_w \tag{4}$$

The AR coefficient estimate $\hat{\mathbf{a}} = [a_1, \cdots, a_p]^T$ can be obtained by maximizing the logarithmic likelihood such as $\hat{\mathbf{a}} = \arg\max\ l(\mathbf{g}_p(\mathbf{a}); \mathbf{m}_g, \mathbf{C}_g)$. The solution is given by

$$\hat{\mathbf{a}} = -(\mathbf{Y}^T \mathbf{C}_g^{-1} \mathbf{Y})^{-1} \mathbf{Y}^T \mathbf{C}_g^{-1} (\mathbf{y}_p - \mathbf{m}_g) \tag{5}$$

where $\mathbf{y}_p = [y(p), \cdots, y(N-1)]^T$ and $\mathbf{Y} = [\mathbf{y}_{p-1}, \mathbf{y}_{p-2}, \cdots, \mathbf{y}_0]$.

3.3. Estimation of population parameters

In this process, we assume the AR coefficient $\hat{\mathbf{a}}$ is already estimated. The previously estimated population parameters of the glottal source $N(\mathbf{m}_e, \mathbf{C}_e)$ and the filtered white noise $N(\mathbf{m}_w, \mathbf{C}_w)$ are also used for the initial estimates.

The combined excitation vector $\hat{\mathbf{g}}_p$ is evaluated by using the AR coefficient estimate $\hat{\mathbf{a}}$ and Eq.(2). In order to update the population parameters $N(\mathbf{m}_e, \mathbf{C}_e)$ and $N(\mathbf{m}_w, \mathbf{C}_w)$, we have to decompose the $\hat{\mathbf{g}}_p$ into the glottal source vector $\hat{\mathbf{e}}_p$ and the filtered white noise vector $\hat{\mathbf{w}}_p$. The decomposition can be achieved by maximizing the logarithmic likelihood such as $\hat{\mathbf{e}}_p, \hat{\mathbf{w}}_p = \arg\max\ (l(\mathbf{e}_p; \mathbf{m}_e, \mathbf{C}_e) + l(\mathbf{w}_p; \mathbf{m}_w, \mathbf{C}_w))$ on condition that $\mathbf{e}_p + \mathbf{w}_p = \hat{\mathbf{g}}_p$. The solution is given as follows. In the case of $\det(\mathbf{C}_e) \geq \det(\mathbf{C}_w)$,

$$\hat{\mathbf{e}}_p = (\mathbf{I} + \mathbf{C}_w \mathbf{C}_e^{-1})^{-1} (\hat{\mathbf{g}}_p - \mathbf{m}_w + \mathbf{C}_w \mathbf{C}_e^{-1} \mathbf{m}_e) \tag{6}$$

In the case of $\det(\mathbf{C}_e) < \det(\mathbf{C}_w)$,

$$\hat{\mathbf{e}}_p = (\mathbf{I} + \mathbf{C}_e \mathbf{C}_w^{-1})^{-1} (\mathbf{C}_e \mathbf{C}_w^{-1} (\hat{\mathbf{g}}_p - \mathbf{m}_w) + \mathbf{m}_e) \tag{7}$$

$$\hat{\mathbf{w}}_p = \mathbf{g}_p - \hat{\mathbf{e}}_p \tag{8}$$

The new population parameter $N(\hat{\mathbf{m}}_e, \hat{\mathbf{C}}_e)$ of the glottal source is obtained as follows.

(a1) All the population parameters $\hat{\mu}(m), \hat{\sigma}^2(m), m \in [1, \cdots, M]$ of the HMM are estimated from $\hat{\mathbf{e}}_p$ by the Baum-Welch algorithm.

(a2) The state transition sequence $s_p, s_{p+1}, \cdots, s_{N-1}$ is estimated by the Viterbi algorithm.

(a3) The updated population parameter $N(\hat{\mathbf{m}}_e, \hat{\mathbf{C}}_e)$ is given by

$$\hat{\mathbf{m}}_e = [\hat{\mu}(s_p), \hat{\mu}(s_{p+1}), \cdots, \hat{\mu}(s_{N-1})]^T$$
$$\hat{\mathbf{C}}_e = \mathrm{diag}\left(\hat{\sigma}^2(s_p), \hat{\sigma}^2(s_{p+1}), \cdots, \hat{\sigma}^2(s_{N-1})\right) \tag{9}$$

The new population parameter $N(\hat{\mathbf{m}}_w, \hat{\mathbf{C}}_w)$ of the filtered white noise is obtained as follows. From $v(n) \sim N(0, \sigma_v^2)$ and Eq.(3), the population parameter is represented by the updated white noise variance $\hat{\sigma}_v^2$ and the AR coefficients $\hat{\mathbf{a}}$ as

$$\hat{\mathbf{m}}_w = \mathbf{0}, \quad \hat{\mathbf{C}}_w = \hat{\sigma}_v^2 \hat{\mathbf{C}}'_w \tag{10}$$

where $\hat{\mathbf{C}}'_w = \left(\sum_{k=0}^{p} \hat{a}_k \hat{a}_{k-|i-j|}\right)$ and $\hat{a}_0 = 1$. The new estimate $\hat{\sigma}_v^2$ can be obtained by maximizing the logarithmic likelihood such as $\hat{\sigma}_v^2 = \arg\max\ l(\hat{\mathbf{w}}_p; \hat{\mathbf{m}}_w, \sigma_v^2 \hat{\mathbf{C}}'_w)$. The solution is given by

$$\hat{\sigma}_v^2 = \frac{1}{N-p} (\hat{\mathbf{w}}_p - \hat{\mathbf{m}}_w)^T (\hat{\mathbf{C}}'_w)^{-1} (\hat{\mathbf{w}}_p - \hat{\mathbf{m}}_w) \tag{11}$$

The procedure is repeated by regarding $\hat{\mathbf{m}}_e, \hat{\mathbf{C}}_e, \hat{\mathbf{m}}_w$ and $\hat{\mathbf{C}}_w$ as $\mathbf{m}_e, \mathbf{C}_e, \mathbf{m}_w$ and $\mathbf{C}_w$ respectively until the likelihood

$$L(\hat{\mathbf{e}}_p; \hat{\mathbf{m}}_e, \hat{\mathbf{C}}_e) L(\hat{\mathbf{w}}_p; \hat{\mathbf{m}}_w, \hat{\mathbf{C}}_w) \quad (12)$$

is converged.

3.4. Flow of the extended algorithm

The extended algorithm consists of the following processes.

1. **Estimation of the initial population parameters**
   The initial AR coefficients $\mathbf{a}^{(0)}$ and the prediction errors $\mathbf{g}_p^{(0)}$ are evaluated by covariance LP analysis. $N(\mathbf{m}_w^{(0)}, \mathbf{C}_w^{(0)})$ is then given from Eq.(10) by using the initial variance $\sigma_v^{2,(0)}$, which, for instance, can be evaluated by $\sigma_v^{2,(0)} = (\mathbf{g}_p^{(0)})^T \mathbf{g}_p^{(0)} / (N-p)(1+|\mathbf{a}^{(0)}|^2)$. The moving averages of the 10 samples from the squared prediction errors are also evaluated in order to estimate $\mathbf{C}_e^{(0)}$ by using Eq.(9). The $\mathbf{m}_e^{(0)}$ is set to $\mathbf{0}$. Iterate the following processes from $i = 0$.
2. **Estimation of the AR coefficients**
   The new AR coefficients $\mathbf{a}^{(i+1)}$ is evaluated as described in the section 3.1.
3. **Estimation of the population parameters**
   The new population parameters $N(\mathbf{m}_e^{(i+1)}, \mathbf{C}_e^{(i+1)})$ and $N(\mathbf{m}_w^{(i+1)}, \mathbf{C}_w^{(i+1)})$ are estimated as described in the section 3.2.
4. If the likelihood in Eq.(12) has converged, the process ends up. Otherwise, repeat the processes from step2 as $i \Rightarrow i+1$.

4. Experiment with Synthetic Speech

The clean speeches of fundamental frequency 100Hz and 750Hz were prepared, which were synthesized by using impulse trains and AR coefficients of order 16 extracted from a male's vowel /a/. The fundamental frequency 750Hz is the highest one of all the clean speeches from which the method described in [4] can accurately extract the vocal tract spectra. In this experiment, the noise corrupted speeches were generated by adding white Gaussian noise to the clean synthetic speech in the SNR range from -5dB to 50dB. The sampling frequency is 16kHz. The glottal source HMM contains 2 states and each state has 2 paths. For the purpose of comparison, the same experiment is carried out by auto-correlation LP analysis using Hanning window. In both the methods, the prediction order and the analysis frame width were set to 16 and 30ms.

Fig.3 and 4 show the vocal tract spectra estimated from the synthetic speeches of fundamental frequency 100Hz by the proposed method and auto-correlation LP analysis respectively. Fig.3 shows the proposed method can extract the 1st and 2nd formants from the corrupted speeches in the SNR range of down to -5dB. In the case of the fundamental frequency 750Hz, Fig.5 shows the proposed method can extract the vocal tract spectra in the SNR range down to 30dB. On the other hand, as shown in Fig.6, the spectral estimate of auto-correlation LP are highly biased toward the harmonics and any of the 1st and 2nd formants cannot be observed.

5. Experiment with Natural Speech

In this experiment, the proposed method was applied to feature extraction from a natural high-pitched speech of vowel sound /a/, which was uttered by a female and sampled at 16kHz. The fundamental frequency is 666Hz. The speech signal is pre-emphasized with the coefficient 0.99. After that, the corrupted speeches were generated by adding white Gaussian noises to the pre-emphasized natural speech. Thus, in this case, we assume that the spectral gradient of background noise is approximately -6dB/Oct. The number of states in the glottal source HMM was set to 15. For the purpose of comparison, the same experiment is carried out by auto-correlation LP analysis using Hanning window. In both the methods, the prediction order and the analysis frame width were set to 16 and 30ms.

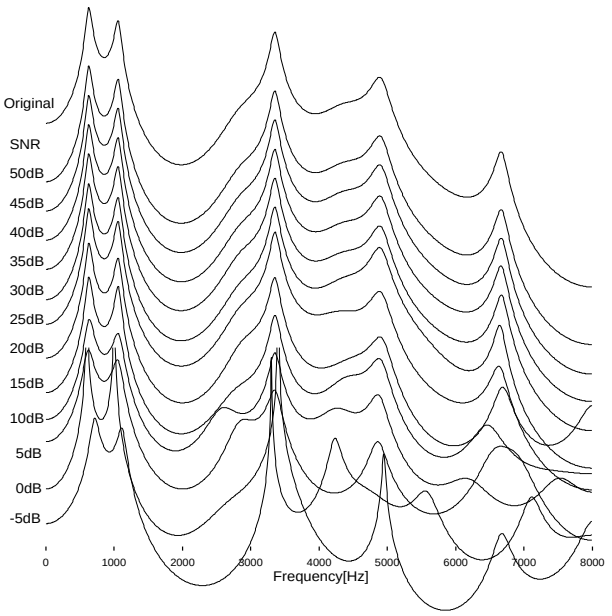

Figure 3: *Vocal tract spectra extracted by using the proposed method.* ($f_0 = 100Hz$)

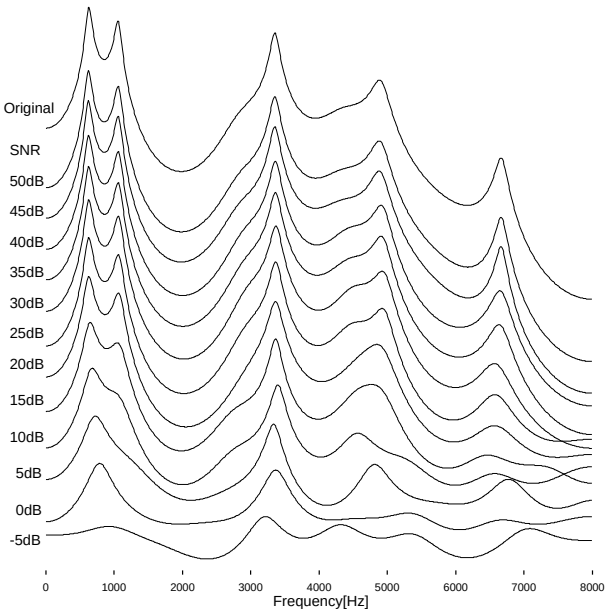

Figure 4: *Vocal tract spectra extracted by using auto-correlation LP.* ($f_0 = 100Hz$)

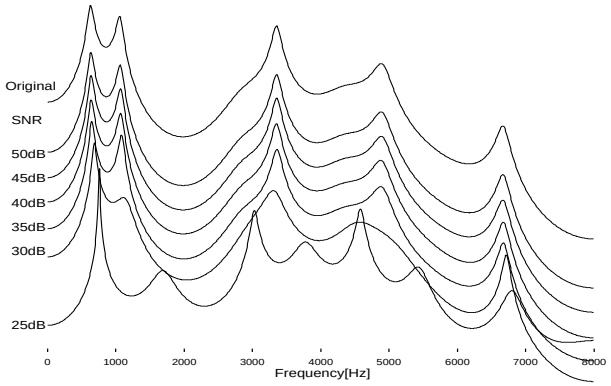

Figure 5: *Vocal tract spectra extracted by using the proposed method.* $(f_0 = 750Hz)$

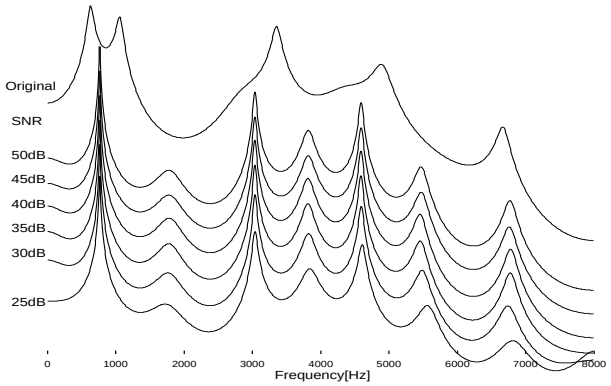

Figure 6: *Vocal tract spectra extracted by using auto-correlation LP.* $(f_0 = 750Hz)$

The vocal tract spectra estimated by the proposed method and auto-correlation LP are shown in Fig.7 and Fig.8. In the spectra estimated by the auto-correlation LP, it is likely that the peak appeared at the frequency 700Hz is the fundamental frequency component. The other peaks also tend to correspond with the harmonics. Compared to that, the vocal tract spectra in the SNR of down to 25dB estimated by the proposed method are less affected by the harmonics of glottal excitation.

6. Conclusions

We applied an HMM to modeling glottal source in order to achieve accurate estimation of AR coefficients from high-pitched and/or noise corrupted speech. The experimental results indicated that the extended algorithm keeps robustness to not only high fundamental frequency speech but also speech with additive white noise.

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

[4] A.Sasou,K.Tanaka,"Glottal excitation modeling using HMM with application to robust analysis of speech signal," Proc. of ICSLP2000, pp.704-707, 2000

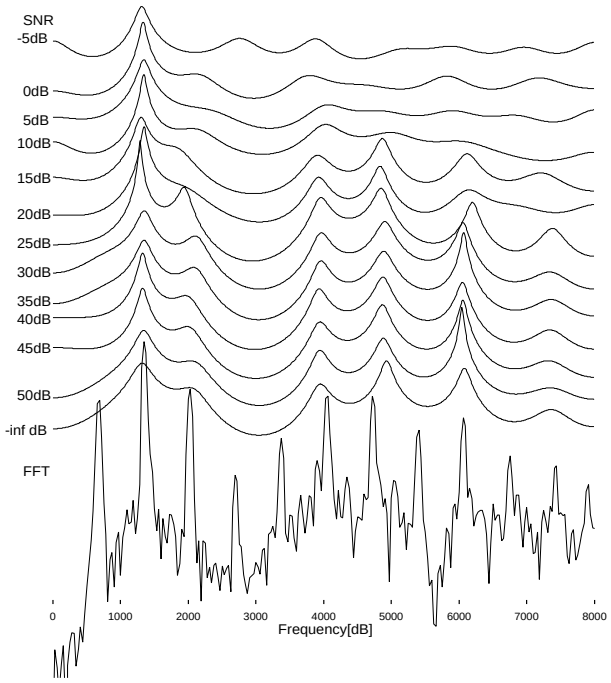

Figure 7: *Vocal tract spectra extracted from a natural speech by using the proposed method.* $(f_0 = 666Hz)$

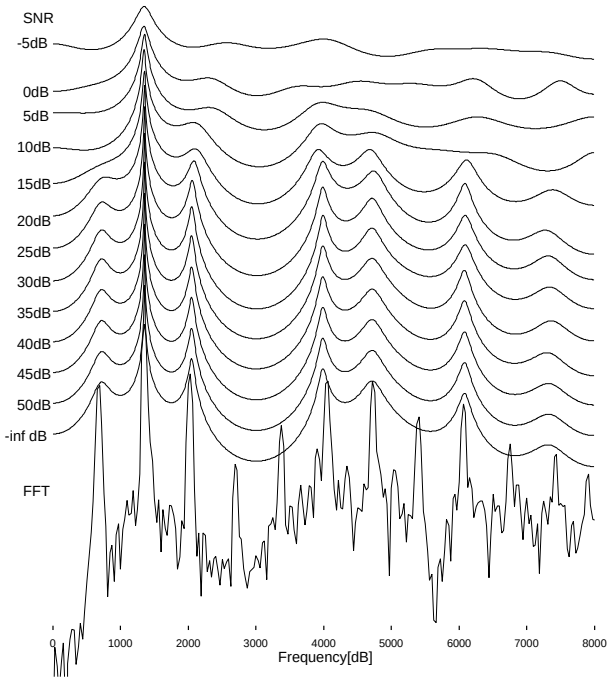

Figure 8: *Vocal tract spectra extracted from a natural speech by using auto-correlation LP.* $(f_0 = 666Hz)$