SPEECH SIGNAL ANALYSIS WITH EXPONENTIAL AUTOREGRESSIVE MODEL

Kentaro Ishizuka, Hiroko Kato and Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
{ishizuka, katohi, nak}@cslab.kecl.ntt.co.jp

ABSTRACT

This paper proposes a speech signal analysis approach that uses the exponential autoregressive (ExpAR) model. In real speech signal, the amplitude and frequency are fluctuating randomly. These fluctuations are non-Gaussian and have nonlinear dynamics. This means that they cannot be modeled adequately with linear AR models or compositions of sine/cosine waves as these analysis methods are known to be affected by such fluctuations. Our proposed approach using ExpAR model can deal with such fluctuations, and it is autoregressive in form with amplitude dependent exponential coefficients. Studies to fit the ExpAR model to real speech data have shown that AIC (Akaike's Information Criteria) values achieved by the ExpAR model are better (lower) than those obtained with a linear AR model, and that the ExpAR model provides a good model of speech fluctuations as movements of the position of its poles. The coefficients change with time depending on the amplitude of speech signals, and so this model is also capable of realizing a fine instantaneous spectral estimation. The modeling of such speech fluctuations has the potential to be used for improving the automatic speech recognition performance in clean or noisy environments, and the naturalness of synthesized speech.

1. INTRODUCTION

Linear autoregressive (AR) models have long been used in speech analysis. This approach assumes that the speech signal is an AR process with a short time frame of, for example, 25 ms, and it estimates the AR coefficients from the signal [1][2].

However, even when the frame length is this short, in practice the speech signal is not a strict AR process. For example, Fig.1 shows the short-term waveform of the vowel /o/ spoken by a Japanese male. Even in such a stable part of speech, the length of the period between glottal pulses changes with a one by one cycle (jitter), and the amount of expiration also changes randomly (shimmer) [3]. These phenomena correspond to amplitude and fundamental frequency fluctuations with non-Gaussian characteristics and nonlinear dynamics [4]. Essentially, such signals with nonlinear fluctuations cannot be modeled adequately with linear AR models or compositions of sine/cosine waves such as harmonic analysis. Indeed, these analysis methods are known to be affected by such fluctuations, and sometimes this results in an insufficient spectral shape estimation. Such an effect often degrades, for example, the performance of automatic speech recognition (ASR). With respect to speech synthesis, it is well known that such fluctuations improve the naturalness of synthesized speech [5]. In terms of psychoacoustics, these fluctuations may be one of the cues for perceiving speech in noisy environments [6]. Therefore, if modeling the fluctuations can provide a better representation of speech signals, then the model has the potential to improve the performance of ASR in clean or noisy environments, and the naturalness of synthesized speech. In addition, since such modeling can represent the characteristics of speech signals well, it can possibly be used as filtering parameters when using state-space models to enhance signals, such as with Kalman filtering.

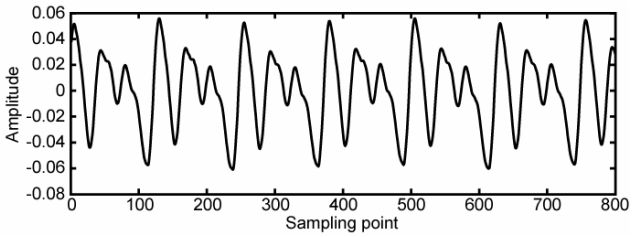

***Figure 1**: Waveform of Japanese vowel /o/.*

Haggan and Ozaki proposed the exponential autoregressive (ExpAR) model for modeling nonlinear fluctuations of time series data [7]. This model is autoregressive in form with amplitude dependent exponential coefficients. The ExpAR model can effectively predict such nonlinear fluctuation behavior as the year record of a trapped Canadian lynx by employing these coefficients. Although this model has been used for time series analysis, it has not been applied to speech signals.

This paper proposes a speech signal analysis approach that uses the ExpAR model. Section 2 describes the ExpAR model in detail. In section 3, studies fitting the ExpAR model to real speech signals are described. In the last section, we provide short conclusion of this study.

2. EXPONENTIAL AUTOREGRESSIVE MODEL

The conventional linear AR model is as described below, where $x(t)$ is the observed signal at time t, ϕ_i are constant AR coefficients, $\varepsilon(t)$ is the prediction error at time t, and p is the number of the model order ($i = 1, \ldots, p$).

$$x(t) = \phi_1 x(t-1) + \phi_2 x(t-2) + \cdots + \phi_p x(t-p) + \varepsilon(t)$$

The ExpAR model employs exponential terms that depend on the amplitude of the observed signal as the AR coefficients as seen below, where γ and π_i are also constant AR coefficient parameters. γ is a scaling factor used to control the effect of the exponential terms [7].

$$\begin{aligned} x(t) = {} & (\phi_1 + \pi_1 \exp(-\gamma x(t-1)^2))x(t-1) \\ & + (\phi_2 + \pi_2 \exp(-\gamma x(t-1)^2))x(t-2) \\ & + \cdots \\ & + (\phi_p + \pi_p \exp(-\gamma x(t-1)^2))x(t-p) + \varepsilon(t) \end{aligned}$$

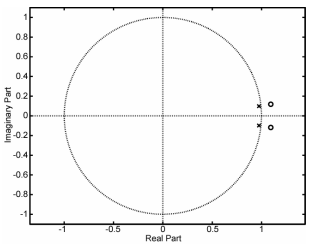

***Figure 2**: Characteristic roots of exponential AR model $x(t) = (1.9482+0.2408\exp(-1.0507x(t-1)^2))x(t-1) + (-0.9585-0.2534\exp(-1.0507x(t-1)^2))x(t-2)$ on the unit circle. The crosses and circles are roots for conditions (i) and (ii), respectively.*

Due to the presence of exponential terms with an $x(t–1)$ multiplier, this model can change its behavior depending on the amplitude of $x(t–1)$. This model is capable of realizing the amplitude-dependent frequency and limit cycle behavior. If we ignore $\varepsilon(t)$ so that the model exhibits limit cycle behavior, the necessary conditions for the solutions of this model are as shown below.

(i) the roots of $\lambda^p - \phi_1\lambda^{p-1} - \cdots - \phi_p = 0$ lie inside the unit circle

(ii) the roots of $\lambda^p - (\phi_1 + \pi_1)\lambda^{p-1} - \cdots - (\phi_p + \pi_p) = 0$ do not all lie inside the unit circle

The characteristic equation (i) corresponds to the equation of the ExpAR model when an $x(t–1)$ is at infinity. On the other hand, the characteristic equation (ii) corresponds to the equation when an $x(t–1)$ is equal to zero. The above means that for a small $x(t–1)$, the system tends to diverge due to condition (ii), while for a large $x(t–1)$, the system tends to converge towards zero due to condition (i). In addition, to avoid unstable singular points, the following condition also should be satisfied.

(iii) $(1-\sum\phi_i)/\sum\pi_i > 1$ or $(1-\sum\phi_i)/\sum\pi_i < 0$

For example, we consider the following ExpAR model [7]:

$$x(t) = (1.9482 + 0.2408e^{-1.0507x(t-1)^2})x(t-1) + (-0.9585 - 0.2534e^{-1.0507x(t-1)^2})x(t-2)$$

This model satisfies all the above three conditions. The crosses and circles shown in Fig. 2 show the characteristic roots for conditions (i) and (ii), respectively. The poles of the ExpAR model move between these two points shown in Fig. 2 depending on the amplitude (value) of $x(t–1)$.

The estimation of order p, and the coefficients $\{\gamma, (\phi_i, \pi_i, i = 1, \ldots, p)\}$ in the ExpAR model essentially needs a nonlinear optimization procedure. However, this problem can be overcome by fixing parameter γ at one of a grid of values and estimating the order p and corresponding ϕ_i, π_i parameters. Then the problem becomes one of fitting a linear regression of $x(t)$ to the series $\{ x(s) ; s < t \}$ and $\{ \exp(-\gamma x(t–1)^2)x(s); s < t \}$, where $\{ x(t) \}$ is a mean deleted series. The order p of the fitted model is selected by using AIC (Akaike's Information Criteria) for a nonlinear time series [8] as below, where m is the maximum order of the model to be considered, n is the total number of observations, and the least squares estimate of the residual variance of the model $\hat{\sigma}_p^2$.

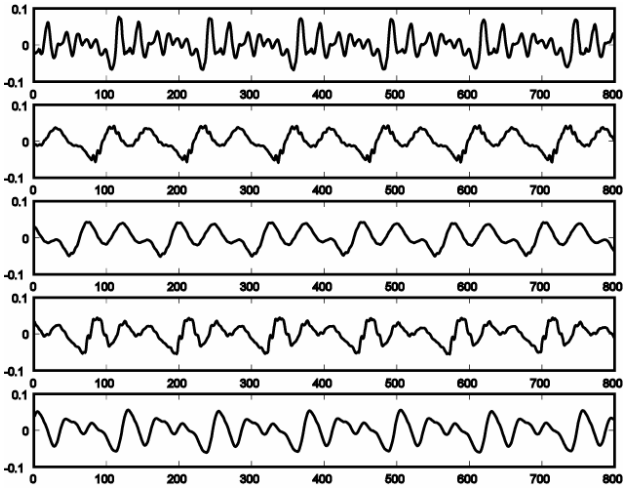

***Figure 3**: Waveforms of Japanese vowels (from the top) /a/, /i/, /u/, /e/, and /o/ spoken by a Japanese male speaker.*

$$AIC(p) = (n-m)\log\hat{\sigma}_p^2 + 2(2p+1)$$

$$\hat{\sigma}_p^2 = (\hat{\varepsilon}_{m+1}^2 + \hat{\varepsilon}_{m+2}^2 + \cdots + \hat{\varepsilon}_n^2)/(n-m)$$

The term $(2p+1)$ is the number of estimated parameters in the model, including the fitted mean. A model with a lower AIC value is better than a model with a higher one.

The models fitted for each γ can also be compared using the AIC, to find the best model over all γ. After fixing $\gamma = \gamma_0$, the estimation procedure is used to fit the model

$$x(t) = (\phi_1 + \pi_1 e^{-\gamma_0 x(t-1)^2})x(t-1) + \cdots + (\phi_p + \pi_p e^{-\gamma_0 x(t-1)^2})x(t-p) + \varepsilon(t)$$

for $t = p+1, \ldots, n$; $p = 1, \ldots, m$. The AIC is also used to choose the best order p. The least squared values of the parameters (ϕ_i, π_i) can be estimated as the values minimizing the sum of prediction errors S described below.

$$S = \sum_{s=p+1}^{n}\left[x(s) - \sum_{k=1}^{p}(\phi_k + \pi_k e^{-\gamma_0 x(s-1)^2})x(s-k)\right]^2$$

In matrix form, we can introduce α, X, and Y as below.

$$\alpha = (\phi_1, \pi_1, \phi_2, \pi_2, \cdots, \phi_p, \pi_p)^T$$

$$X = \begin{pmatrix} x(p) & x(p)e^{-\gamma_0 x(p)^2} & \cdots & x(1) & x(1)e^{-\gamma_0 x(p)^2} \\ x(p+1) & x(p+1)e^{-\gamma_0 x(p+1)^2} & \cdots & x(2) & x(2)e^{-\gamma_0 x(p+1)^2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x(n-1) & x(n-1)e^{-\gamma_0 x(n-1)^2} & \cdots & x(n-p) & x(n-p)e^{-\gamma_0 x(n-1)^2} \end{pmatrix}$$

$$Y = (x(p+1), x(p+2), \cdots, x(n))^T$$

By using these matrices, the necessary conditions whereby parameter α minimizes S can be written as follows.

$$(X^T X)\alpha - X^T Y = 0$$

We can then obtain the least squared estimated values $\hat{\alpha}$ of parameter α as follows.

$$\hat{\alpha} = (X^T X)^{-1} X^T Y$$

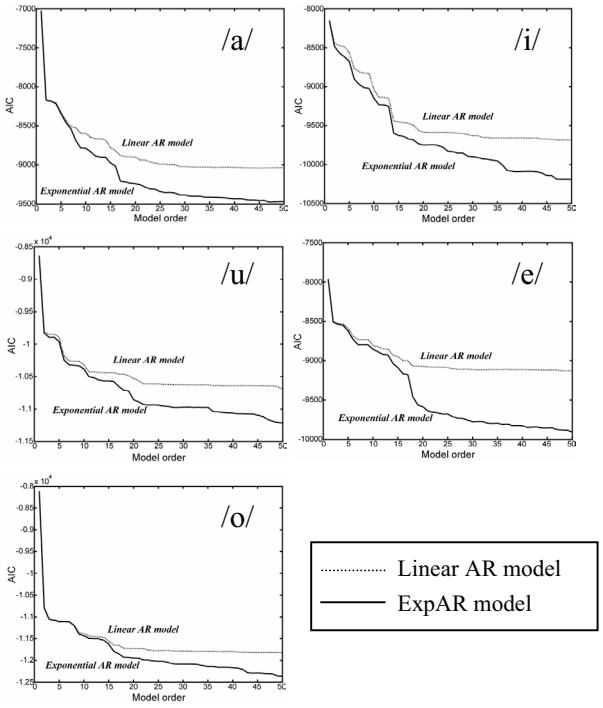

***Figure 4**: AIC values of the linear AR model and the ExpAR model as a function of the model order.*

Note that $\hat{\alpha}$ is only the least squared estimated value, therefore, it does not always satisfy conditions (i)-(iii). If these conditions are not satisfied, certain additional numerical optimizations are needed for satisfying these conditions.

3. FITTING MODEL TO SPEECH

We performed some studies to fit the ExpAR models to real speech data. As speech data, we used five Japanese vowels /a/, /i/, /u/, /e/, and /o/ spoken by a Japanese male. These vowels were recorded at a sampling rate of 48 kHz, and down-sampled to 16 kHz. Figure 3 shows the waveforms of these five vowels. These signals are largely stable, although jitter and shimmer can be observed.

3.1. Fitting ExpAR model

First, we compared the ExpAR model and the linear AR model in terms of AIC. AIC is a criteria that evaluates the accuracy of a model, and it is widely used to select an adequate structure of a statistical model [9]. AIC values are calculated for five vowels, where each frame length was 800 sampling points. In this AIC calculation for the ExpAR models, the γ value was fixed as 440. Figure 4 shows the AIC values for five vowels as a function of the order p (p = 1, …, 50). As shown in Fig. 4, AIC values achieved with the ExpAR model were lower than those obtained with the linear AR model. This confirms that the ExpAR model is more suitable for speech data than the linear AR model. Henceforth, since the linear AR model achieves enough small AIC value around when p = 25, only the 25-order ExpAR model and the 25-order linear AR model were used. In this paper, we used only the AIC criterion to select model. MDL, BIC and other statistical criteria are not considered.

Second, we fitted a 25-order ExpAR model to these speech data. Figure 5 shows the estimated waveforms with the ExpAR

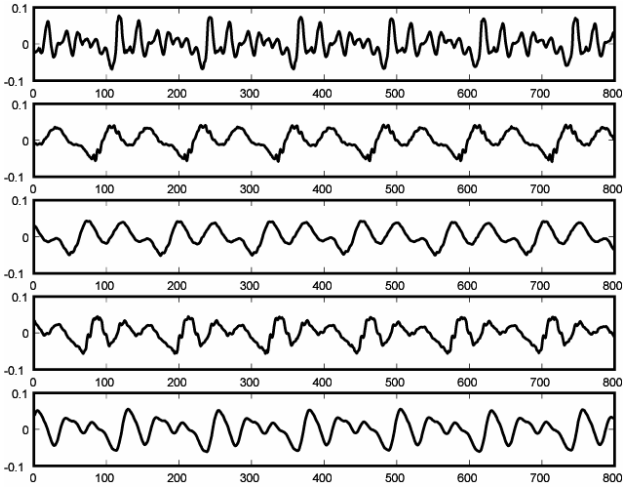

***Figure 5**: Vowel waveforms estimated by fitting the ExpAR model to vowels. (from the top) /a/, /i/, /u/, /e/, and /o/.*

models. As shown in the figure, the fitting confirmed that this model can predict the speech signal well (see also Fig. 3). Next, we compared the ExpAR model and the linear AR model in terms of the prediction errors. For this comparison, we used 25-order ExpAR and determined a fixed γ for each vowel as the value that minimizes the AIC, for example, 559.42 and 24.6 for vowel /a/ and /o/, respectively. Table 1 shows the result. The prediction errors were always lower than those estimated by using a conventional 25-order linear AR model.

3.2. Pole behavior

Thirdly, we observed the behavior of the poles (eigenvalues) of the ExpAR model. In the ExpAR model, the coefficients of the AR model change with time depending on the signal amplitude, therefore, we can obtain the characteristic roots and instantaneous spectrum for each sampling point. The poles, which are calculated at each sampling point, are plotted in Fig. 6. Figure 6 shows that the positions of the poles move with respect to time within certain ranges. As described above, these ranges are fixed by eigenvalues calculated from characteristic equations described in necessary conditions (i) and (ii). This behavior corresponds to the speech sound fluctuations. These characteristics of the ExpAR model may provide a good representation of the speech signal fluctuations.

3.3. Instantaneous spectrum

Lastly, we estimated instantaneous spectrum for each sampling point. Figure 7 shows 125 estimated instantaneous spectra

***Table 1**: Prediction errors achieved by the ExpAR model and the linear AR model for each vowel. The errors achieved by the ExpAR is always lower than those achieved by the linear AR.*

	ExpAR model	Liner AR model
/a/	1.8259	2.3682
/i/	1.2506	1.5872
/u/	0.6009	0.8029
/e/	1.429	2.2171
/o/	0.2899	0.3698

(10^{-3})

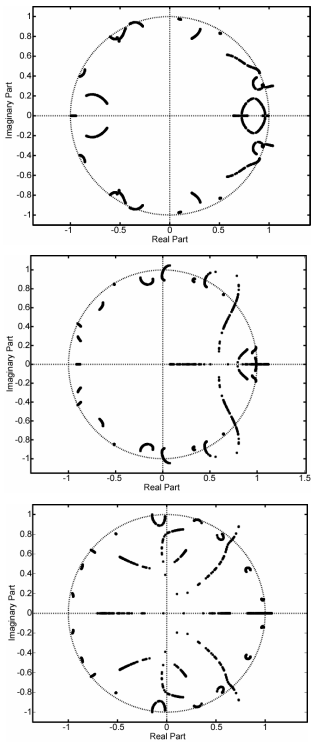

***Figure 6**: Behavior of poles with time. (from the top) fitting ExpAR models to vowel /a/, /i/, and /u/.*

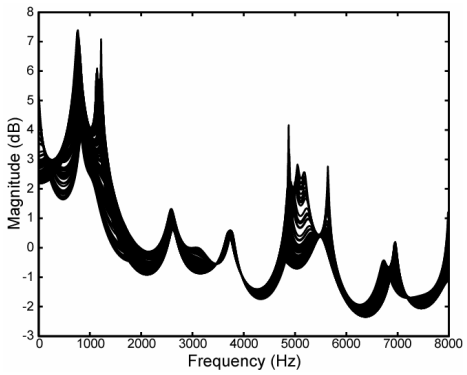

***Figure 7**: Instantaneous spectrum with time estimated by fitting the 25-order ExpAR model to vowel /a/.*

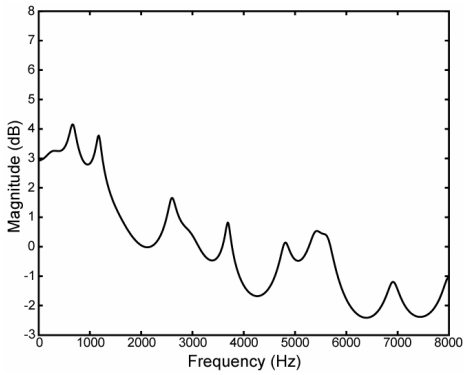

***Figure 8**: Spectrum estimated by fitting the25-order linear AR model to vowel /a/.*

(corresponding to about one speech signal cycle) simultaneously. To compare with the spectrum estimated by the linear AR model, Fig. 8 shows the spectrum. As shown in the figures, although the parameters are estimated from 800 points speech signal, the ExpAR model can estimates fine instantaneous spectra at each sampling point. On the other hand, the spectrum estimated by the linear AR model is affected by the speech fluctuation, and then it results in an averaged spectrum.

4. CONCLUSION

This paper proposed a speech analysis approach that uses the ExpAR model to deal with random fluctuations in speech signals. The ExpAR model can represent amplitude-dependent behavior using exponential terms. Some studies designed to fit ExpAR models to speech signals confirmed that this model achieves lower AIC values than linear AR models. In addition, the studies showed that the ExpAR model well represents the fluctuations of the speech signal, and that it can estimate fine instantaneous spectra at each sampling point. The ExpAR model is highly suitable for use with speech signals and is therefore potentially applicable as a feature parameter extraction method for ASR. It could also be applied to the parameters themselves to synthesize more natural speech.

Acknowledgements: The authors thank Dr. P. Zolfaghari and Dr. H. Solvang for valuable comments on a draft of this paper.

REFERENCES

[1] Itakura, F. and Saito, S. "Analysis synthesis telephony based on the maximum likelihood method," *Reports of the 6th Int. Cong. Acoust.*, C-5-5, 1968.

[2] Atal, B. S. and Schroeder, M. R. "Predictive coding of speech signals," *Reports of the 6th Int. Cong. Acoust.*, C-5-4, 1968.

[3] Horii, Y. "Fundamental frequency perturbation observed in sustained phonation," J. Speech Hear. Res. **22**, 5-19, 1979.

[4] Aoki, N. and Ifukube, T. "Analysis and perception of spectral 1/*f* characteristics of amplitude and period fluctuations in normal sustained vowels," J. Acoust. Soc. Am. **106**, 423-433, 1999.

[5] Klatt, D. H. and Klatt, L. C. "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. **87**, 820-857, 1990.

[6] Ishizuka, K. and Aikawa, K. "Effect of F0 fluctuation and amplitude modulation of natural vowels on vowel identification in noisy environments," *Proc. of ICSLP*, 1633-1636, 2002.

[7] Haggan, V. and Ozaki, T. "Modelling nonlinear random vibrations using an amplitude-dependent autoregressive time series model," Biometrica, **68**, 1, 189-196, 1981.

[8] Ozaki, T. and Oda, H. "Non-linear time series model identification by Akaike's Information Criterion," In *Information and Systems*, Oxford, Pergamon, Ed. Dubuisson, B., 83-91, 1978.

[9] Akaike, H., "Information measures and model selection," *Proc. of 44th Session of the International Statistical Institute*, **1**, 277-291, 1983.