# A FEATURE FOR VOICE ACTIVITY DETECTION DERIVED FROM SPEECH ANALYSIS WITH THE EXPONENTIAL AUTOREGRESSIVE MODEL

*Kentaro Ishizuka and Hiroko Kato*

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

## ABSTRACT

This paper proposes a feature for voice activity detection (VAD) obtained from a speech signal analysis that uses the exponential autoregressive (ExpAR) model. This model employs exponential terms that depend on the amplitude of observed signals in the AR coefficients part. Since these terms can model the nonlinearity of speech caused by the nonlinear fluctuation of vocal cord vibration, this model can provide a better fit for speech signals. A parameter in the exponential terms of the ExpAR model called 'the scaling parameter,' is directly associated with the degree of nonlinearity of analyzed signals. Therefore, the scaling parameter changes when observed signals include speech signals. Based on this property, this parameter is usable as a feature for VAD under noisy conditions. An experiment using noisy speech data confirmed the potential performance of the proposed feature by comparing receiver operating characteristics curves obtained from the proposed feature and conventional robust features. Another experiment was conducted by comparing recalls, precisions, and F-measures for speech interval detection achieved by our proposed VAD algorithm, that utilized only the proposed feature, and two widely used standardized algorithms. The result showed that the proposed method could achieve better performance than those of the standardized algorithms.

## 1. INTRODUCTION

Voice activity detection (VAD) in the presence of environmental noise is a crucial aspect of speech signal processing techniques such as automatic speech recognition [1], speech coding [2], and speech enhancement [3]. Since these techniques depend strongly on the VAD accuracy or assume ideal VAD, insufficient VAD accuracy seriously affects their practical performance. This fact has created a need for effective VAD in real environments [4]. In general, VAD consists of two parts: an 'acoustic feature extraction' part, and a 'decision mechanism' part. Although both parts influence the VAD performance, this paper focuses particularly on the inherent performance of the features.

The short-term signal energy and zero-crossing rate have long been used as simple acoustic features for VAD [5]. Although these features are indeed effective under high signal-to-noise ratio (SNR) conditions, they are degraded easily by environmental noise. Therefore, many robust features have been proposed. Some of these features employ the periodicity of speech signals. These are autocorrelation function based features [6][7], spectrum based features that utilize harmonicity [8], and pitch based features [9]. Other features are based on power in the band-limited region [10][11][12]. Spectral shapes such as mel-frequency cepstral coefficients [7], delta line spectral frequencies (LSF) [11], and whole spectra [12] have been also used to extract speech signals from other sounds. However, these speech features sometimes become similar to those of other sound signals. Therefore, there is still a need for more robust and effective features that represent the characteristics of speech.

Recently, we proposed a stochastic time series modeling approach for speech signal analysis, which uses the exponential autoregressive (ExpAR) model [13]. This model is a class of nonlinear AR model, and can provide a better fit for speech signals that inherently include nonlinear fluctuations [14]. The ExpAR model was originally proposed in [15] for modeling data generated by nonlinear dynamics, where the nonlinear force depending on the amplitude acts and the system is perturbed by noise. The scaling parameter of the ExpAR model is directly associated with the degree of nonlinearity of analyzed signals. This nonlinearity may relate to the inherent characteristics of the speech production system. Using the property, we achieved a significant VAD performance improvement in comparison with previous methods.

This paper first provides a brief explanation of the speech signal analysis using the ExpAR model, and then describes the relationships between the ExpAR model parameters and the physical characteristics of speech production, and the potential availability of the scaling parameter for VAD in Section 2. In Section 3, evaluation experiments confirm the efficiency of the proposed parameter in the presence of real environmental noise at a low SNR. Section 4 concludes this study.

## 2. EXPONENTIAL AUTOREGRESSIVE MODEL

### 2.1. Model and parameter estimation

Unlike conventional linear AR models, the ExpAR model employs exponential terms that depend on the amplitudes of observed signals in the AR coefficients part as seen below:

$$x_t = \sum_{k=1}^{p} \left( \phi_k + \pi_k e^{-\gamma x_{t-1}^2} \right) x_{t-k} + \varepsilon_t$$

where $x_t$ is the observed discrete signal at time $t$ ($t = 1, .. , N$), $\{\phi_i, \pi_i, \gamma\}$ are constant parameters, $\varepsilon_t$ is assumed to be a white Gaussian sequence at time $t$, $p$ is the model order ($i = 1, .. , p$), and $\gamma$ is a scaling parameter used to control the effect of the exponential terms. The model order $p$, the coefficients $\{\phi_i, \pi_i\}$, and the scaling parameter $\gamma$ are estimated as follows [13]:

1. Fix the AR order $p = p_0$ as minimizing Akaike information criterion (AIC) by applying linear AR models in advance, and initial value of the scaling parameter $\gamma = \gamma_0$ as below:

   $$\gamma_0 = -\log \varepsilon \Big/ \underset{1 \le t \le N}{Max}(x_t^2)$$

   where $\varepsilon$ is a very small number.

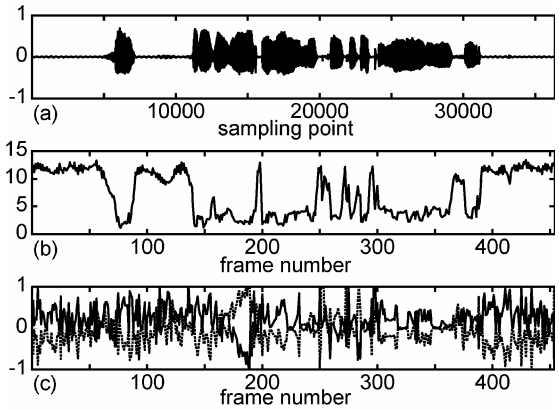


***Figure 1****: (a) Speech waveform under silent conditions. (b) Log $\gamma$ parameter. (c) $\pi_{1,2}$ parameters.*

2. With $p$ and $\gamma$ fixed, the least squared values of the parameters $\{\phi_i, \pi_i\}$ are estimated as the values that minimize the sum of the prediction errors $S$ described below:

$$S = \sum_{s=p+1}^{N} \left[ x_s - \sum_{k=1}^{p} \left( \phi_k + \pi_k e^{-\gamma x_{s-1}^2} \right) x_{s-k} \right]^2$$

3. Numerical optimization is used to find the optimal $\gamma$ parameter with the constraint of fixing the estimated parameters $\{\phi_i, \pi_i\}$. This optimization employs a sequential quadratic programming to maximize the log likelihood of the prediction using the ExpAR model as below:

$$\log L = \log f\left(x_1, x_2, \cdots, x_p \middle| \phi_i, \pi_i, \gamma, \hat{\sigma}^2\right)$$
$$= -\frac{N-p}{2}\log 2\pi - \frac{N-p}{2}\log \hat{\sigma}^2 - \frac{N-p}{2}$$

where $f(\cdot|\cdot)$ is the probability density function and $\hat{\sigma}^2 = S/(N-p)$.

4. Iterate steps 2 and 3 until the $\gamma$ parameter converges to an optimal value.

### 2.2. Relationship with nonlinearity of speech signal

In this section we mention the reason for the ExpAR model's effectiveness with speech signals.

In previous models, the speech production system was described as consisting of two subsystems: the vocal cord vibration system and the vocal tract resonance system. In many cases, a two-path model has been used, that is, a physical oscillation model such as a secondary differential equation and linear predictive coding (LPC) have been applied to the vocal cord oscillation system and the vocal tract system, respectively. Theoretically, the vibration of the vocal cords is generated by phenomena whereby the volume velocity of the airflow and the impedance at the glottis change in time, and this has a nonlinear effect on the vocal tract. Therefore, it is necessary to model speech signals by combining interaction between the nonlinear dynamic mechanisms for the vocal cords and vocal tract. Unlike the previous modeling approach, the ExpAR model can be used to describe both vocal cord and vocal tract characteristics.

One example of typical nonlinear vibration is the following van der Pol equation used in electrical circuit theory:

$$\ddot{x} + \left(x^2 - 1\right)\dot{x} + bx = 0 \ , \qquad (1)$$

where for $x^2 < 1$ the system has a negative damping force and starts to oscillate and diverge. On the other hand, for $x^2 > 1$ the system has a positive damping force and it starts to damp out. This is known as limit cycle behavior, and such process is that of amplitude-dependent frequency. Furthermore, an equation of motion for free oscillation has been proposed for modeling vocal cords vibration [16]:

$$\ddot{z} + K_t \dot{z} + \omega_0^2 z = 0 \ , \qquad (2)$$

where $K_t$ is the time-varying energy loss within a pitch cycle, and $\omega_0$ is the resonance angle frequency at $K_t$ = 0. The oscillation state is determined according to the size of $K_t$. Since $K_t$ is the function associated with the square of the amplitude within one pitch, the vibration process is considered to be similar the process generated by equation (1).

When considering practical nonlinear phenomena, we employ the following stochastic differential equation model $\ddot{x} + f(\dot{x}) + g(x) = \xi$ , where $\xi$ is a continuous Gaussian white noise sequence. As regards the nonlinear restoring force it can be seen that the frequency increases (decreases) as the amplitude increases (decreases). Equation (1) is also extended to the stochastic van der Pol process $\ddot{x} + \left(x^2 - 1\right)\dot{x} + bx = \xi$ , which describes perturbed limit cycle behavior.

The second order ExpAR model was easily applied to such nonlinear force systems [15]:

$$x_t = \left\{ \phi_1 + \pi_1 e^{-\gamma x_{t-1}^2} \right\} x_{t-1} + \left\{ \phi_2 + \pi_2 e^{-\gamma x_{t-1}^2} \right\} x_{t-2} + \varepsilon_t \qquad (3)$$

For the relationship between the oscillation and roots, if the coefficients satisfy the condition, which is such that the roots $\lambda_0$ and $\overline{\lambda_0}$ of $\Lambda^2 - (\phi_1 + \pi_1)\Lambda - (\phi_2 + \pi_2) = 0$ lie outside the unit circle, then $x_t$ starts to oscillate and diverge for small $x_{t-1}$. On the other hand, if the coefficients satisfy another condition such that the roots $\lambda_\infty$ and $\overline{\lambda_\infty}$ of $\Lambda^2 - \phi_1\Lambda - \phi_2 = 0$ lie inside the unit circle, then $x_t$ starts to damp out when $x_{t-1}$ becomes too large. Therefore, the ExpAR model can describe the characteristics of such amplitude-dependent nonlinear oscillations as (1) and (2). Furthermore, as for stationary condition, with according to Tweddie's theory [17], the ExpAR process is stationary even if the roots $\lambda_0$ and $\overline{\lambda_0}$ lie outside the unit circle.

### 2.3. Efficiency of scaling parameter

In the ExpAR model, scaling parameter $\gamma$ can adapt the degree of the data amplitude in terms of how it affects the nonlinearity of the model. To confirm this property of scaling parameter $\gamma$, we show an example when we apply the secondary order ExpAR model of equation (3) to speech signals. The speech data was spoken by a male speaker under silent conditions (Fig.1 (a)), and its sampling rate was 8 kHz. First, the speech signals were analyzed by 25 ms-length frames with a 15 ms overlap. Then, the ExpAR model was applied to each frame, and the $\gamma$ parameter was obtained for each frame. Figure 1 (b) shows the estimated $\gamma$ parameter for the speech signals.

Note that the nonlinear term coefficients $\pi_{1,2}$ and $\gamma$ are available to counterbalance each other. Figure 1 (c) shows the coefficients $\pi_{1,2}$. Because the estimated $\pi_{1,2}$ values do not become 0, we can confirm that a nonlinear term certainly exists in the model. The estimated $\gamma$ becomes small only during the period in which human speech is produced. We consider the nonlinearity to be large in the region where the estimated $\gamma$ becomes small. This result suggests that parameter $\gamma$ can be an effective feature for VAD. In this section, although we deal with

only secondary order models, higher orders may be necessary to improve the model's goodness-of-fit.

## 3. EXPERIMENT

In this section, we evaluate the validity of the proposed feature for VAD under noisy conditions by comparing it with other conventional features. This section first explains the property of noisy speech data for evaluation experiments, then shows the behavior of the proposed feature under noisy conditions, and shows the results of two evaluation experiments: the discriminative power and the VAD performance evaluations.

### 3.1. Noisy speech data for evaluation

Speech data mixed with environmental noise were used in this evaluation. We used travel arrangement dialogue data spoken in Japanese (SDB-L in [18]). The data consists of 2,292 utterances spoken by 178 speakers. The utterance duration is between 1.4 to 12.1 seconds. We down-sampled the sampling rate of the data from 48 to 8 kHz. Correct VAD data were generated based on the hand labeled transcription for SDB-L, which includes onset, offset, and pause information. Examples of speech data and their correct VAD data are shown in Fig. 2 (a) and (b), respectively.

As noise data, we recorded real environmental sounds at an airport arrival gate in Tokyo and in the street in the Shinjuku area of Tokyo. The recording equipment consisted of omni-directional microphones (Sony ECM-77B) and portable IC recorders (Marantz PMD670), and the data were sampled at 48 kHz. The data were down-sampled to 8 kHz, and added to the above speech data at an SNR of 0 dB. Because environmental sounds are not stationary, we adjusted the SNR so that the power peaks of the speech and noise data within the period of an utterance were the same. Different noise intervals were added to different utterances. Figure 2 (c) shows an example speech data shown in Fig. 2 (a) mixed with the street sounds.

### 3.2. Behavior of scaling parameter

When we apply the ExpAR model to observed signals as described in Section 2.3, the $\gamma$ parameters have different values for different frames, and these values change based on the degree of nonlinearity of the analyzed signal characteristics. The nonlinearity of the speech signals may differ from that of silence or other environmental sounds, therefore this $\gamma$ parameter behavior suggests its availability as a feature for VAD. To confirm the potential availability of scaling parameter $\gamma$ as a robust speech feature for VAD, we investigated the behavior of the parameter for speech signals under noisy conditions.

The estimated $\gamma$ parameter values for the speech data shown in Fig. 2 (c). Figure 2 (d) show the estimated $\gamma$ parameter values. In this case, we set the model order $p = 12$. The result shows that the $\gamma$ parameters of a frame that includes speech signals are certainly smaller than those of the other frames. Therefore, this result suggests that the $\gamma$ parameter is available as a speech feature for VAD under noisy conditions.

### 3.3. Discriminative power evaluation and results

To assess the discriminative power of the proposed feature, we compared receiver operating characteristics (ROC) curves [10] obtained from the proposed feature and conventional robust features. The ROC curves were generated from 'false accept' (ratios of the frames mis-detected as speech to the non-speech frames) and 'false reject' (ratios of the frames mis-rejected as non-speech to the speech frames) calculations with various thresholds. Superior features achieve both a lower false accept and a lower false reject. The conventional features compared with the proposed feature were spectral entropy [8], the maximum autocorrelation peak [6], and the maximum LPC residual autocorrelation peak [7][19]. In this evaluation, we only compared raw features that could be obtained from within one frame, and did not compare features obtained after post-processing (e. g. smoothing) the raw features because the performance of such post-processed features depends strongly on the performance of the raw features.

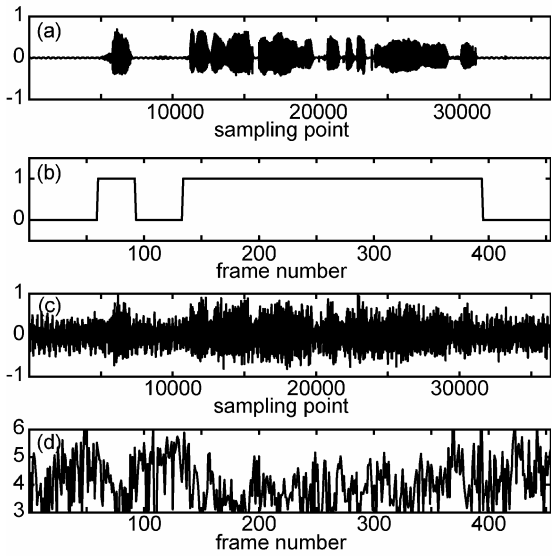


***Figure 2**: (a) Speech waveform under silent conditions. (b) Correct VAD data for (a). (c) Speech waveform (a) mixed with street noise at an SNR of 0 dB. (d) Log $\gamma$ parameter for noisy speech (c).*

Figure 3 shows the ROC curves achieved under airport and street noise conditions, respectively. The proposed feature achieved better ROC curves than the other features. At such a low SNR, spectral entropy could not achieve good performance because the speech spectra had fatally deteriorated. The features based on the autocorrelation function were more robust than the spectral entropy because of their inherent robustness as regards noise. However, the results indicate that the proposed feature is more robust than these autocorrelation based features. This fact suggests that the proposed feature captures the nonlinearity as characteristics of speech signals well, and that the nonlinearity of speech is not easily deteriorated by environmental noise. Furthermore, these results confirmed the potential performance of the proposed feature.

### 3.4. VAD performance evaluation and results

In this section, we propose a preliminary VAD algorithm that utilizes only the proposed feature, and compare the performance obtained from the proposed algorithm with two widely used standardized VAD algorithms. A "VAD algorithm" means the combination of feature extraction, feature post-processing, and decision mechanisms. Our proposed VAD algorithm is as following:

1. Estimate $\gamma$ parameters for each frame by applying the ExpAR to speech signals as above.

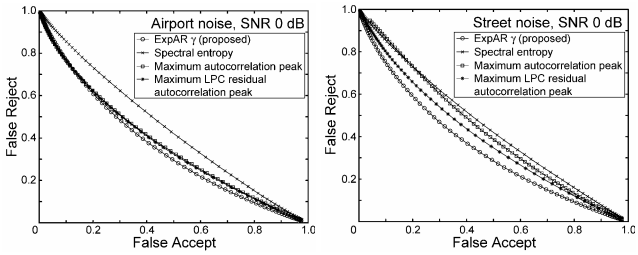


***Figure 3**: ROC curves for speech features in the presence of airport (left) and street (right) noise.*

2. Take moving averages of the $\gamma$ parameter across 6 frames for smoothing.
3. Take the long-term mean of the smoothed parameters and their standard deviation across 400 frames regardless of the existence of target speech signals, and sets the threshold as the sum of the mean and half of the standard deviation. This threshold was selected experimentally.
4. Determine a frame whose smoothed parameter obtained is below the threshold to include speech signals.

We evaluated the performance of this algorithm by comparing it with two standardized VAD algorithms. One was ITU-T G.729 Annex B VAD [11], which simultaneously utilizes differentials in LSF, full-band energy, low-band energy, and zero-crossing rate. The other is ETSI WI008 Advanced Front-end (AFE) VAD for frame dropping [12], which simultaneously utilizes the whole spectra to design Wiener filters, the spectral sub-region, and spectral variance. The test data sets were the same as those described in Section 3.1.

To evaluate the performance of the VAD algorithms, we introduced three criterions: recall (ratios of the frames correctly detected as speech to the speech frames; i.e. 1.0 – false reject = 'speech hit rate' in [10]), precision (ratios of the frames correctly detected as speech to the frames detected as speech; i.e. 1.0 – false accept), and F-measure (harmonic mean of precision and recall). Table 1 shows the results obtained from the proposed VAD algorithms, ITU-T G.729 Annex B VAD, and ETSI AFE VAD. Although only one feature is utilized, the proposed method achieved better performance in terms of F-measure than the other two standardized methods, which utilize over three features. This is a promising result suggesting that the feature integrated the other features to the proposed feature and more sophisticated decision mechanisms must offer more robust and effective VAD in the presence of noise in future.

***Table 1**: Recalls, precisions, F-measures obtained from the proposed VAD, ITU-T G.729 Annex B VAD* [11]*, and ETSI Wi008 Advanced Front-end VAD* [12].

| Method | Measure | Environmental noise | |
|---|---|---|---|
| | | Airport | Street |
| Proposed VAD | Recall | **0.772** | **0.775** |
| | Precision | **0.738** | 0.744 |
| | F-measure | **0.755** | **0.759** |
| ITU-T G.729 Annex B VAD | Recall | 0.605 | 0.639 |
| | Precision | 0.671 | 0.682 |
| | F-measure | 0.636 | 0.660 |
| ETSI WI008 AFE VAD | Recall | 0.536 | 0.624 |
| | Precision | 0.736 | **0.772** |
| | F-measure | 0.620 | 0.690 |

## 4. CONCLUSION

We proposed a feature for VAD based on a scaling parameter obtained by applying the ExpAR model to speech signals. Because the purpose of this scaling parameter is to control the nonlinear oscillation associated with the physical characteristics of speech production, the value of the parameter changes when the observed signals include speech signals. An evaluation experiment confirmed the potential performance of the proposed feature under noisy conditions. A second evaluation experiment showed that our proposed VAD algorithm utilizing only the proposed feature could achieve better performance than two standardized VAD methods, and confirmed the validity of using the proposed feature for VAD.

**Acknowledgements:** The authors thank Dr. Takehiro Moriya for offering us to use ITU-T G.729 Annex B VAD, and thank Dr. Atsushi Nakamura for his useful suggestions about speech database.

## REFERENCES

[1] Junqua, J.-C., Mak, B., and Reaves, B. "A robust algorithm for word boundary detection in the presence of noise," IEEE Trans. on Speech and Audio Processing, 2, 406-412, 1994.

[2] Srinivasan, K. and Gersho, A. "Voice activity detection for cellular networks," *Proc. of IEEE Workshop on Speech Coding for Telecommunications*, 85-86, 1993.

[3] Le Bouquin-Jeannès R. and Faucon, G. "Study of voice activity detector and its influence on a noise reduction system," Speech Communication, 16, 245-254, 1995.

[4] Karray, L. and Martin, A. "Towards improving speech detection robustness for speech recognition in adverse conditions," Speech Communication, 40, 261-276, 2003.

[5] Rabiner, L. R. and Sambur, M. R. "An algorithm for determining the endpoints of isolated utterances," The Bell System Technical Journal, 54, 297-315, 1975.

[6] Basu, S. "A linked-HMM model for robust voicing and speech detection," *Proc. of ICASSP*, 1, 816-819, 2003.

[7] Kristjansson, T., Deligne, S., and Olsen, P. "Voicing features for robust speech detection," *Proc. of Interspeech*, 369-372, 2005.

[8] Shen, J.-L., Hung, J.-W., and Lee, L.-S. "Robust entropy-based endpoint detection for speech recognition in noisy environments," *Proc. of ICSLP*, 1998.

[9] Tucker, R. "Voice activity detection using a periodicity measure," IEE Proceedings-I, 139, 377-380, 1992.

[10] Marzinzik, M. and Kollmeier, B. "Speech pause detection for noise spectrum estimation by tracking power envelope dynamics," IEEE Trans. on Speech and Audio Processing, 10, 109-118, 2002.

[11] ITU-T Recommendation G.729, Annex B., 1996.

[12] ETSI standard document. ETSI ES 202 050 V1.1.3, 2003.

[13] Ishizuka, K., Kato, H., and Nakatani, T. "Speech signal analysis with exponential autoregressive model," *Proc. of ICASSP*, 1, 225-228, 2005.

[14] 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.

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

[16] Omura, H. and Tanaka, K. "Speech analysis and synthesis system based on the nonlinear energy damping model," Bulletin of the Electrotechnical Laboratory, 62, 11-21, 1998.

[17] Tweedie, R. L. "Sufficient conditions for ergodicity and stationarity of Markov chains on a general state space," Stochastic Process. Appl., 3, 385-403, 1975.

[18] Nakamura, A., Matsunaga, S., Shimizu, T., Tonomura, M., and Sagisaka, Y. "Japanese speech databases for robust speech recognition," *Proc. of ICSLP*, 1998.

[19] Merkel, J. D. "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. on Audio and Electroacoustics, AU-20, 367-377, 1972.