# REAL-TIME SPEECH VISUALIZATION SYSTEM :KANNON – APPLYING AUDITORY CHARACTERISTICS

*Ken Nakamuro, Katsuhiro Haruki, and Sueo Sugimoto*

Department of Electrical and Electronic Engineering, Ritsumeikan University
Noji-Higashi, Kusatsu City, Shiga 525-8577 Japan
phone: + (81) 77-561-2673, FAX: + (81) 77-561-2663, email: sugimoto@se.ritsumei.ac.jp
web: www.sugimotolab.se.ritsumei.ac.jp

## ABSTRACT

We have been developing a real time speech-displaying system called “KanNon” which helps deaf person to understand speaker’s speech contents. We designed the KanNon system to display a sound spectrogram, pitch frequency and loudness of speech as well as characters by speech-recognition system as real-time scrolling image. For the purpose of displaying formant patterns clearly with high accuracy, we applied Burg method combining with the minimum cross-entropy (Burg-MCE) method, and human auditory characteristics such as an equal loudness preemphasis and mel-scale frequency to the sound spectrogram. Finally, we show more effective display for the spectrogram reading in the KanNon system.

## 1. INTRODUCTION

When the deaf person communicates with someone, they use sign language, transcript or lip reading. However there is difficulty to communicate with healthy person using these methods in terms of convenience of communication. Against this background, we have developed the KanNon system[1, 2] as a new real time visualization system, which helps deaf person to communicate with someone using the spectrogram reading. The KanNon system displays not only sound spectrogram, but also pitch frequency, loudness of the speech and characters by a speech recognition system.

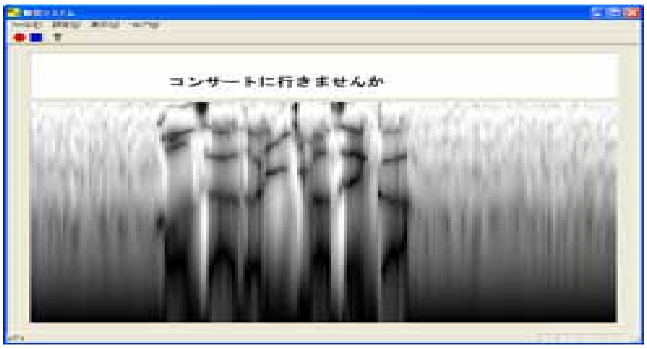
In this view, we apply the auto regressive(AR) model as the vocal tract model for the spectral estimation, and estimate AR model parameters by Burg method combining with minimum cross-entropy (Burg-MCE) method[3] with change detection using Kullback information distance[4]. According to this proposed method, we could result sound spectrogram with clear formant.

Additionally, we estimate the pitch frequency and loudness of the speech from the prediction errors and the variance of the prediction errors of the estimated AR model.

Furthermore, we apply human auditory characteristics processing such as equal loudness preemphasis and mel-scale frequency to display the sound spectrogram to emphasize the important formant pattern, and develop phoneme recognition using time delay neural network (TDNN)[5]. In the first step, we develop TDNN for Japanese vowels /a/, /i/, /u/, /e/, /o/ using 16 mel-scale filter bank coefficients from the power spectrum of AR model.

## 2. THE KANNON SYSTEM

We show the interface of the KanNon system in Fig. 1. In the past research, we developed the KanNon system displaying sound spectrogram and colored square-shaped figure image being related to the pitch frequency and loudness of a speech[1, 2]. In the present version, the KanNon system displays sound spectrogram and characters by speech-recognition system as real time scrolling image. And we also consider the displaying estimated pitch frequency and loudness of speech as font color and font size of the characters resulted by a speech recognition system.



コンサートに行きませんか

Figure 1: Interface of the KanNon system

## 3. SPECTRAL ESTIMATION

We adopted the auto regressive(AR) model known as all-pole model to estimate spectral peeks of the vocal tract.

$$x_{k,t} \quad = \quad \sum_{i=1}^{L} \phi_{k,i}^{(L)} x_{k,t-i} + e_{k,t}^{(L)}, \quad k = a, b \qquad (1)$$

For the AR model parameter estimation method, we use the Burg-MCE method developed in [3].

### 3.1 Burg-MCE method

The Burg-MCE method is Burg method[7] combining with the minimum cross-entropy spectral estimation method[8, 9], which estimates AR model parameters using known prior AR model parameters.

Let’s suppose that the $L$th order AR parameters in the $(s-1)$th frame are already estimated, then the $\ell$th order AR parameters in the $s$th frame are estimated by applying the Burg method recursively. And we describe $(s-1)$th and $(s)$th frame's AR models as prior and posterior AR model, respectively. According to the Burg-MCE method, for $m$ satisfying $\ell+1 \leq m \leq L$, the $m$th order posterior AR model parameters $\{\phi_{b,i}^{(m)}\}_{i=1}^{m}$ and variance of the prediction errors $\sigma_{b,m}^2$ in $s$th frame are estimated by using the $m$th order prior AR model parameter $\phi_{a,m}^{(m)}$ and variance $\sigma_{a,m}^2$ in the $(s-1)$th frame. In particular, the reflection coefficient $c_m (= \phi_{b,m}^{(m)})$ is obtained as:

if $(1 \leq m \leq \ell-1)$

$$c_m = \frac{-2\sum\limits_{t=m+1}^{L} e_{b,t}^{(m-1)} r_{b,t-1}^{(m-1)}}{\sum\limits_{t=m+1}^{L}\left\{\left[e_{b,t}^{(m-1)}\right]^2 + \left[r_{b,t-1}^{(m-1)}\right]^2\right\}} \tag{2}$$

if $(\ell \leq m \leq L)$

$$c_m = \frac{-\frac{\sigma_{a,m}^2}{\phi_{a,m}^{(m)}} + \mathrm{sgn}\left(\phi_{a,m}^{(m)}\right)\sqrt{\left(\frac{\sigma_{a,m}^2}{\phi_{a,m}^{(m)}}\right)^2 + 4\sigma_{b,m-1}^4}}{2\sigma_{b,m-1}^2} \tag{3}$$

$$\phi_{b,i}^{(m)} = \phi_{b,i}^{(m-1)} + c_m \phi_{b,m+1-i}^{(m-1)}, \ (i = 1, \ldots, m) \tag{4}$$

$$e_{b,t}^{(m)} = e_{b,t}^{(m-1)} + c_m r_{b,t-1}^{(m-1)}, \ (t = m+1, \ldots, L) \tag{5}$$

$$r_{b,t}^{(m)} = r_{b,t-1}^{(m-1)} + c_m e_{b,t}^{(m-1)}, \ (t = m+1, \ldots, L) \tag{6}$$

$$\sigma_{b,m}^2 = (1 - c_m^2)\sigma_{b,m-1}^2 \tag{7}$$

$$e_{b,t}^{(0)} = r_{b,t}^{(0)} = x_{b,t}, \quad ( \, t = 1, \ \ldots, \ L \, ) \tag{8}$$

where $|c_m| < 1$. Then $m$th AR parameters $\{\phi_{b,i}^{(m)}\}_{i=1}^{m}$, $\sigma_{b,m}^2$ are obtained from (2), (3), (4) and (7). According to the Burg-MCE method, estimation results are greatly affected by the model parameters in the prior frame. Therefore, if the model parameters in the prior frame differ considerably from the model parameters in the posterior frame, we can not get good estimated spectra. To solve this problem, we apply the Kullback information distance to AR models and detect the changed frame of data [4] in the following subsection.

### 3.2 The change detection method using the Kullback information distance

We consider two adjoining frames in the observed signals. Considering two neighboring frames $(j = a, b)$ which mean prior and posterior frames, respectively, [4]. We calculate Kullback information distance of the prior and the posterior $\ell$th AR models in the recursive process of the Burg-MCE method.

$$K[a,b] = \frac{1}{2\sigma_{b,\ell}^2}(\sigma_{a,\ell}^2 + \boldsymbol{\theta}^T \boldsymbol{R}_a \boldsymbol{\theta}) + \frac{1}{2\sigma_{a,\ell}^2}(\sigma_{b,\ell}^2 + \boldsymbol{\theta}^T \boldsymbol{R}_b \boldsymbol{\theta}) - 1$$

$$\boldsymbol{R}_j = \begin{bmatrix} R_{j,0} & R_{j,1} & \cdots & R_{j,\ell-1} \\ R_{j,1} & R_{j,0} & \cdots & R_{j,\ell-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{j,\ell-1} & R_{j,\ell-2} & \cdots & R_{j,0} \end{bmatrix}, \quad (j = a, \ b) \tag{9}$$

Where $\boldsymbol{\theta} = \boldsymbol{\phi}_a - \boldsymbol{\phi}_b$, $\boldsymbol{\phi}_a = [\phi_{a,1}^{(\ell)}, \ \cdots, \ \phi_{a,\ell}^{(\ell)}]^T$, $\boldsymbol{\phi}_b = [\phi_{b,1}^{(\ell)}, \ \cdots, \ \phi_{b,\ell}^{(\ell)}]^T$. And auto correlation function $\{R_{j,k}\}_{k=0}^{\ell-1}$ is derived from inverse Levinson-Durbin's algorithm [10] as follows.

$$\begin{cases} R_{j,k+1} = -\sigma_{j,k}^2 c_{k+1} - \sum\limits_{i=1}^{k} \phi_{j,i}^{(k)} R_{j,k+1-i}, \\ \qquad\qquad\qquad\qquad (k = 0, \ \cdots, \ \ell-1) \\ R_{j,0} = \sigma_{j,1}^2/(1 - c_1^2) \end{cases} \tag{10}$$

From (9), the change detection is executed by

$$\begin{cases} K[a,b] \leq \eta & \text{then Burg-MCE method} \\ K[a,b] > \eta & \text{then Burg method} \end{cases} \tag{11}$$

where $\eta$ is a threshold value.

We set the threshold $\eta$ in (9), We choice methods between Burg-MCE method and Burg method in the recursive process step $\ell \leq m \leq L-1$.

## 4. PITCH FREQUENCY ESTIMATION

Speech signals are typically divided into two broad classes: voiced signal (; sound source is vibration of the vocal cord), and unvoiced signal (; sound source is turbulence eddy flow by the vocal tract stricture). We assume quasi-periodic impulse series and white noises as the sources of voiced and unvoiced signals, respectively. These generation's structures of speech signals are modeled by all-pole filters. So the relation between the input signal $u_t$ (: sound sources), and output signal $x_t$(: speech signal) is expressed by the following difference equation

$$x_t + \sum_{i=1}^{m} \phi_i^{(m)} x_{t-i} = G u_t, \tag{12}$$

where $G$ is constant. Here we assume the input $u_t$ for the voiced signal as

$$u_t = \sum_{k=0}^{\infty} \delta_{n-kT_p}, \qquad t = 0, 1, 2, \ldots \tag{13}$$

where $T_p$ is the period of impulse series. If AR coefficients $\{\hat{\phi}_{b,i}^{(m)}\}_{i=1}^{m}$ in (1) are completely correspond to the coefficient $\phi_i^{(m)}$ in all-pole model parameter with voiced source $Gu_t$ in (12), the following relation holds from (1) and (12).

$$\tilde{e}_t = G u_t, \tag{14}$$

where, $\tilde{e}_t$ is the prediction error of the AR model which is estimated by the Burg-MCE method. (14) shows that the prediction error $\tilde{e}_t$ obtained from AR model is proportional to the constant $G$ of input signal $u_t$. Therefore, if the signal is a voiced signal, the prediction error $\tilde{e}_t$ has the periodicity with the period $T_p$. According to this consideration, the pitch frequency can be estimated from the prediction errors $\tilde{e}_t$ which are estimated by applying the Burg-MCE method. We explain here a method of estimating the pitch frequency as follows.

First, we perform center clipping processing under the following condition with the variance of the prediction error $\hat{\sigma}_e^2$ for removing prediction error factor in $\tilde{e}_t$ .

$$\hat{e}_t = \begin{cases} \tilde{e}_t - c\hat{\sigma}_e & (\tilde{e}_t \geq \hat{\sigma}_e) \\ \tilde{e}_t + c\hat{\sigma}_e & (\tilde{e}_t \leq -\hat{\sigma}_e) \\ 0 & \text{otherwise} \end{cases}, \quad t = 0, \ldots, N-m-1, \tag{15}$$

where, $c$ is a positive constant. Then we apply the method of pitch frequency estimation in consideration of time continuity to $\hat{e}_t$. A method of pitch frequency estimation in consideration of time continuity

### 4.1 A method of pitch frequency estimation in consideration of time continuity

Now we explain a method to select the most suitable pitch from some candidates utilizing its time continuity proposed in [6].

When we extract the peak value from the auto correlation functions of the prediction error sequence setting by the threshold, we can obtain several candidates of pitch frequencies. We try to correct the discontinuous points, which are caused by the estimation failures using several candidates of pitch frequencies and based on the fact that pitch frequency is changing slowly. In particular, we will show how to obtain the candidate near true pitch frequency in the several candidates on the present frame effectively by referencing the information on the pitch frequencies of past several frames, through the weighted matrix and considering the time continuity of pitch frequencies.

We denote $F_s$ and $N$ as the sampling frequency and the frame data size respectively, then the short-time auto-correlation function $R_{s,k}$ of prediction errors $\hat{e}_{s,t}$ in the $s$th frame is given by

$$R_{s,k} = \frac{1}{N-m} \sum_{t=0}^{N-m-1-k} \hat{e}_{s,t}\hat{e}_{s,t+k}, \quad k = 0, \ldots, N-m-1. \tag{16}$$

In (16), we choose $P$ peaks in descending order except $R_{s,0}$ and ones below the threshold value. Then, we define candidates of the delay time of the auto-correlation function as $k_i(s)$ $(i = 1, \ldots, P)$ from 1st to the $P$th in the $s$th frame and define two functions as follows.

$$\begin{aligned} f(i,s) &= \frac{F_s}{k_i(s)} \\ g(i,s) &= R_{s,k_i(s)}, \quad i = 1, \ldots, P, \end{aligned} \tag{17}$$

where $f(i,s)$ in (17) shows the pitch frequency which is $i$th candidate in the $s$th frame, and (18) show the value of the auto-correlation function which is the $i$th candidate's delay time $k_i(s)$ in the $s$th frame. We calculate these functions of the frames from the $(s-M+1)$th to the $s$th and store the values on the history matrix shown in Fig. 2.

Then we denote $p_i$ $(i = 1, \ldots, P)$ as the score of each candidate in the $s$th frame, and add points for $p_i$ using the history matrix and the criterion as follows.

For $i = 1, \cdots, P,$ find $l$ and $j$ $(1 \leq l \leq P, 1 \leq j \leq M)$ such that

$$f(i,s)(1-\alpha) \leq f(l,s+1-j) \leq f(i,s)(1+\alpha) \tag{18}$$

then

$$p_i = p_i + w_{lj}\, g(l, s+1-j) \tag{19}$$

Where $w_{lj}$ is the weight for each element in the history matrix given by

$$w_{lj} = (P+1-l)(M+1-j), \tag{20}$$

and $\alpha$ $(0 \leq \alpha \leq 1)$ decides frequency range to vote score in reference to the history matrix. So the $\alpha$ means continuity of pitch frequency. After we apply above addition of points to all elements of the history matrix, we calculate

$$\gamma = \underset{r=\{1,\ldots,P\}}{\arg\max}\ (p_r) \tag{21}$$

and finally we obtain the estimated pitch frequency $f_p(s) \equiv f(\gamma, s)$ in the $s$th frame. However, if there is no peak beyond the threshold or the obtained pitch frequency is outside the range defined beforehand, we regard the $s$th frame as a unvoiced frame or a silent frame and $f_p(s)$ is made into 0[Hz].

(frame)

| | $s$ | $s-1$ | $s-2$ | $s-3$ | $s-4$ |
|---|---|---|---|---|---|
| $1st$ | $f(1,s)$<br>$g(1,s)$ | $f(1,s-1)$<br>$g(1,s-1)$ | $f(1,s-2)$<br>$g(1,s-2)$ | $f(1,s-3)$<br>$g(1,s-3)$ | $f(1,s-4)$<br>$g(1,s-4)$ |
| $2nd$ | $f(2,s)$<br>$g(2,s)$ | $f(2,s-1)$<br>$g(2,s-1)$ | $f(2,s-2)$<br>$g(2,s-2)$ | $f(2,s-3)$<br>$g(2,s-3)$ | $f(2,s-4)$<br>$g(2,s-4)$ |
| $3rd$ | $f(3,s)$<br>$g(3,s)$ | $f(3,s-1)$<br>$g(3,s-1)$ | $f(3,s-2)$<br>$g(3,s-2)$ | $f(3,s-3)$<br>$g(3,s-3)$ | $f(3,s-4)$<br>$g(3,s-4)$ |
| $4th$ | $f(4,s)$<br>$g(4,s)$ | $f(4,s-1)$<br>$g(4,s-1)$ | $f(4,s-2)$<br>$g(4,s-2)$ | $f(4,s-3)$<br>$g(4,s-3)$ | $f(4,s-4)$<br>$g(4,s-4)$ |
| $5th$ | $f(5,s)$<br>$g(5,s)$ | $f(5,s-1)$<br>$g(5,s-1)$ | $f(5,s-2)$<br>$g(5,s-2)$ | $f(5,s-3)$<br>$g(5,s-3)$ | $f(5,s-4)$<br>$g(5,s-4)$ |

(candidate)

Figure 2: The history matrix($P = M = 5$)

## 5. POWER SPECTRUM OF A SPEECH

Power spectrum of the AR model is given by following equation.

$$P(\omega) = \frac{\sigma_{b,L}^2}{\left| \sum_{k=0}^{L} \Phi_{b,k}^{(L)} e^{-j\omega k T_s} \right|^2}, \ \omega = 2\pi f \tag{22}$$

where $\Phi_{b,k}^{(L)}$ is estimated by Burg-MCE method described in 3.1, $f$ is frequency and $T_s$ is sampling period. Next we consider the human auditory characteristics into the power spectrum directly estimated from the observation signal.

### 5.1 Equal loudness pre-emphasis

Human hearing is not equally sensitive to loudness of sounds at each frequency. The next equation represents the sensitivity response of the human ear.

$$E(\omega) = \frac{(\omega^2+\beta_1)\omega^4}{(\omega^2+\beta_2)^2(\omega^2+\beta_3)(\omega^6+\beta_4)}. \qquad (23)$$

$$\begin{cases} \beta_1 = 56.8 \times 10^6, & \beta_2 = 6.3 \times 10^6 \\ \beta_3 = 0.38 \times 10^9, & \beta_4 = 9.58 \times 10^{26} \end{cases}$$

Here we calculate the power spectrum multiplying $E(\omega)$ to the $P(\omega)$ in (22).

### 5.2 Mel-scale spectrum

The formant transition of phonemes appears in the low frequency range. According to the auditory mechanism, we have higher spectral resolution in the lower frequency than one in the higher frequency. So we adopt mel-scale sound spectrogram to apply this auditory characteristic to the KanNon system.

The mel-scale was proposed by Stevens, Volkman and Newman in 1937[11] is a scale of pitches judged by listeners to be equal in distance from one to another. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 [mel]. Above about 500 [Hz], larger and larger intervals are judged by listeners to produce equal pitch increments.

To convert $m$ [mel] into $f$ [Hz] use:

$$f = 700 \exp\left(\frac{\mathrm{m}}{1127.01048} - 1\right) \qquad (24)$$

And the inverse:

$$\mathrm{m} = 1127.01048 \ln\left(\frac{1}{700}f - 1\right) \qquad (25)$$

## 6. SPEECH RECOGNITION

In previous KanNon system, we have build the speech recognition system using Microsoft speech API which is based on Hidden Markov Model (HMM) in the KanNon system. For further work, we are developing phoneme recognition system to built quicker displaying of the characters by speech recognition. For this purpose, we adapt a phonemic level speech recognition system using time delay neural network (TDNN) architecture. The TDNN achieved a higher recognition rate in the phoneme recognition than HMM in [5]. In the first step, we develop TDNN for Japanese vowels /a/, /i/, /u/, /e/, /o/ using 16 mel-scale filter bank coefficients from the power spectrum of AR model.

## 7. CONCLUSIONS

In our research, Burg-MCE method with change detection resulted sound spectrogram with clear formant comparing with result by Burg method. Applying the mel scale to the sound spectrogram, we could emphasize the important formant transition. For the future work, we visualize pitch and loudness as font color and font size of characters by speech recognition.

## REFERENCES

[1] T. Nakamoto, Y. Saruta, K. Horii, and S. Sugimoto: The KanNon System - The Visualization System of Speech Signals. *Proc. of IASTED Int. Conf. Signal and Image Processing*, (2001).

[2] K. Nakamuro, Y. Togawa, K. Tanaka, and S. Sugimoto: The Speech - Displaying System: KanNon. *Proc. of 35th ISCIE Int. Symp. on Stochastic Systems Theory and Its Application*, pp. 253–258, Ube, Japan (2003).

[3] K. Nakamuro, K. Tanaka, Y. Togawa and S. Sugimoto: The Speech-displaying System: Kannon - Applying Minimum Cross Entropy Method. *Proc. 34th ISCIE Int. Symp. on Stochastic Systems Theory and Its Application*, pp. 36–41, Fukuoka, Japan (2002).

[4] S. Oe, T. Soeda, and T. Nakamizo: A Method of Predicting Failure or Life for Stochastic System by Using Autoregressive Models. *Int. Journal of Systems Science*, Vol. 11, No. 10, pp. 1177–1188, (1980).

[5] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang: Phoneme Recognition: Neural Networks vs. Hidden Markov models. *IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pp.107-110, (April 1988).

[6] A. Ogihara and S. Yoneda: A Method for Selecting the Most Suitable Pitch from Some Candidates Utilizing Its Time Continuation. *Trans. IEICE(A)*, Vol. J74-A, No. 07, pp. 948–956, (1991-7) (in Japanese).

[7] J. P. Burg: Maximum Entropy Spectral Aanlysis. *Ph. D Dissertation, Stanford Univ., Stanford, CA*, (1975).

[8] John E. Shore: Minimum Cross - Entorpy Spectral Analysis. *IEEE Trans. Acoust., Speech, Signal Processing*, Vol. ASSP-29, pp. 230–237, (1981).

[9] S. Sugimoto, T. Wada, and E. Nakatomi: Minimum Cross Entropy and Informational Approaches for a Spectrum Estimation. *Proc. of 7th IFAC/IFORS Symp. on Identification and System Parameter Estimation*, pp. 1727–1732, (1985).

[10] T. Nakamizo: *Signal Analysis and System Identification (in Japanese).* Corona Pub. Co. Tokyo, Japan, (1988).

[11] S. S. Stevens, J. Volkmann, E. B. Newman: The Mel Scale Equates The Magnitude of Perceived Differences in Pitch at Different Frequencies *J. Acoust. Soc. Am. 8*, 185 (1937).

[12] T. Wada, K. Nakamuro, and S. Sugimoto: A Spectral Estimation Algorithm Based on Minimum Cross Entropy Method. *Proc. of SICE Annual Conf. 2002 in Osaka*, pp. 262–267, (June 2002).