Acoust. Sci. & Tech. **22**, 4 (2001)

TECHNICAL REPORT

Fundamental frequency estimation of speech signals using MUSIC algorithm

Takahiro Murakami and Yoshihisa Ishida

School of Science and Technology, Meiji University,
1–1–1, Higashi-Mita, Tama-ku, Kawasaki, 214–8571 Japan
e-mail: ishida@isc.meiji.ac.jp

(*Received 31 July 2000, Accepted for publication 8 January 2001*)

Abstract: In this article a new method for fundamental frequency estimation from the noisy spectrum of a speech signal is introduced. The fundamental frequency is one of the most essential characteristics for speech recognition, speech coding and so on. The proposed method uses the MUSIC algorithm, which is an eigen-based subspace decomposition method.

Keywords: Fundamental frequency, MUSIC algorithm, Noisy speech

PACS number: 43.72.Ar

1. INTRODUCTION

The fundamental frequency of speech signals is an essential feature of human voice [1]. Its estimation is very important in various speech processing systems, especially in speaker recognizers, speech instruction systems for hearing impaired children, and analysis by synthesis speech coders. We know a lot of algorithms for estimating the fundamental frequency. However, the accurate estimation method of the fundamental frequency has not been established yet. Many engineers have been studying new methods.

In this paper, we describe a new and analytic method to accurately estimate the fundamental frequency of noisy speech signals. The proposed method uses the MUSIC (**MU**ltiple **SI**gnal **C**lassification) algorithm [2–7], which was proposed by Schmidt [8]. The MUSIC algorithm exploits the noise subspace to estimate the unknown parameters of the random process. This algorithm can estimate the frequencies of complex sinusoids corrupted with additive white noise. Andrews *et al*. [9] have already proposed the fundamental frequency determination method using the MUSIC algorithm. They increase the fundamental frequency determination capability at low signal to noise ratios by applying the singular value decomposition (SVD) to speech enhancement. On the other hand, our method can reduce greatly the number of eigenvalues to be calculated in order to use the band-limited MUSIC spectrum and shorten calculation time for estimating fundamental frequencies.

This paper is organized as follows. The principle of the MUSIC algorithm is reviewed in Section 2. In Section 3 we present an analytic method for the fundamental frequency estimation and illustrate estimation results. In Section 4 we end with the conclusion.

2. MUSIC ALGORITHM [2–7]

The MUSIC algorithm is an eigen-based subspace decomposition method for estimation of the frequencies of complex sinusoids observed in additive white noise. Consider a noisy signal vector $\boldsymbol{y}$ composed of P real sinusoids modeled as

$$\boldsymbol{y} = \boldsymbol{S}\boldsymbol{a} + \boldsymbol{n} \tag{1}$$

where

$$\boldsymbol{a} = [X_1 \quad X_2 \quad \cdots \quad X_P]^T \tag{2}$$

$$\boldsymbol{S} = [\boldsymbol{s}_1 \quad \boldsymbol{s}_2 \quad \cdots \quad \boldsymbol{s}_P] \tag{3}$$

$$\boldsymbol{s}_k = [1 \quad e^{j2\pi f_k} \quad \cdots \quad e^{j2\pi(N-1)f_k}]^T. \tag{4}$$

N is the number of samples, f_k is the frequency of the k-th complex sinusoid, X_k is the complex amplitude of k-th sinusoid and $\boldsymbol{n}$ is a zero mean Gaussian white noise vector with variance ${\sigma_n}^2$.

The autocorrelation matrix of the noisy signal $\boldsymbol{y}$ can be written as

$$\begin{aligned} \boldsymbol{R}_{yy} &= E[\boldsymbol{y}\boldsymbol{y}^H] \\ &= \boldsymbol{R}_{xx} + \boldsymbol{R}_{nn} \\ &= \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^H + {\sigma_n}^2\boldsymbol{I} \end{aligned} \tag{5}$$

where E denotes the expectation, H denotes the Hermitian

transpose and $\boldsymbol{A} = E[\boldsymbol{a}\boldsymbol{a}^H]$ is the diagonal matrix. In addition, $\boldsymbol{R}_{\boldsymbol{xx}} = \boldsymbol{S}\boldsymbol{A}\boldsymbol{S}^H$ and $\boldsymbol{R}_{\boldsymbol{nn}} = {\sigma_n}^2\boldsymbol{I}$ are the autocorrelation matrices of the signal and noise processes as

$$\boldsymbol{R}_{\boldsymbol{xx}} = \sum_{k=1}^{N} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^H \tag{6}$$

$$\boldsymbol{R}_{\boldsymbol{nn}} = {\sigma_n}^2 \sum_{k=1}^{N} \boldsymbol{v}_k \boldsymbol{v}_k^H. \tag{7}$$

where λ_k and $\boldsymbol{v}_k$ are the eigenvalues and eigenvectors of the matrix $\boldsymbol{R}_{\boldsymbol{xx}}$ respectively. The autocorrelation matrix of the noisy signal may be expressed as

$$\begin{aligned} \boldsymbol{R}_{\boldsymbol{yy}} &= \sum_{k=1}^{N} \lambda_k \boldsymbol{v}_k \boldsymbol{v}_k^H + {\sigma_n}^2 \sum_{k=1}^{N} \boldsymbol{v}_k \boldsymbol{v}_k^H \\ &= \sum_{k=1}^{N} \mu_k \boldsymbol{v}_k \boldsymbol{v}_k^H \end{aligned} \tag{8}$$

where $\mu_k = \lambda_k + {\sigma_n}^2$ are the eigenvalues of the matrix $\boldsymbol{R}_{\boldsymbol{yy}}$. All the eigenvalues are the real numbers and satisfy

$$\mu_1 \geq \mu_2 \geq \cdots \geq \mu_P > \mu_{P+1} = \cdots = \mu_N = {\sigma_n}^2. \tag{9}$$

Then, the MUSIC spectrum is defined as

$$P_{XX}^{\mathrm{MUSIC}}(f) = \frac{1}{\sum_{k=P+1}^{N} |\boldsymbol{s}^H(f)\boldsymbol{v}_k|^2} = \frac{1}{\boldsymbol{s}^H(f)\boldsymbol{V}\boldsymbol{V}^H\boldsymbol{s}(f)}. \tag{10}$$

where $s(f) = [1 \quad e^{j2\pi f} \quad \cdots \quad e^{j2\pi(N-1)f}]^T$ is the complex sinusoidal vector and $\boldsymbol{V} = [\boldsymbol{v}_{P+1} \quad \cdots \quad \boldsymbol{v}_N]$ is the matrix of eigenvectors of the noise subspace.

3. BAND-LIMITED SPECTRUM AND FUNDAMENTAL FREQUENCY ESTIMATES USING THE MUSIC ALGORITHM

3.1. Band-Limited MUSIC Spectrum

In case of speech signals, the harmonic structure appears more clearly in a low-frequency domain [1]. Then, before describing the estimation method of fundamental frequencies, we consider applying the MUSIC algorithm only to the low-frequency components of a frequency spectrum. Assume that the number of samples is 256 points and the sampling frequency is 11.025 [kHz]. In consideration of the existence range of fundamental frequencies, only the frequency components below 1 [kHz] are used for the MUSIC algorithm. Therefore, the frequency components of a MUSIC spectrum are the set of those at frequencies 43 [Hz], 86 [Hz], $\cdots$, $f_k = 11025/256k$ [Hz], $\cdots$, 991 [Hz] and $1 \leq k \leq 23 (= K)$. The size of the autocorrelation matrix $\boldsymbol{R}_{\boldsymbol{yy}}$ is 256×256 and its rank will be less than or equal to K. Then, we have

$$\begin{aligned} &\mu_1 \geq \mu_2 \geq \cdots \geq \mu_P \geq \cdots \geq \mu_{K=23} > {\sigma_n}^2, \\ &\mu_{K+1} = \mu_{K+2} = \cdots = \mu_{N=256} = 0 \\ &\sum_{k=K+1}^{N} \boldsymbol{s}^H(f)\boldsymbol{v}_k = 0 \end{aligned} \tag{11}$$

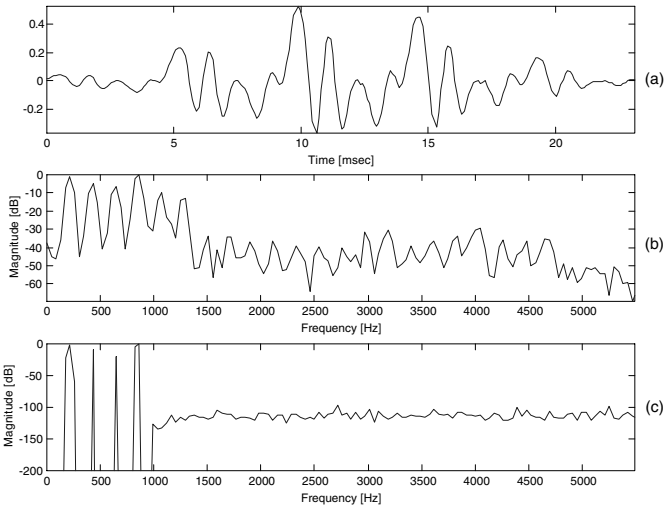
and Eq. (10) can be written as

$$P_{XX}^{\mathrm{MUSIC}}(f) = \frac{1}{\sum_{k=P+1}^{N} |\boldsymbol{s}^H(f)\boldsymbol{v}_k|^2} = \frac{1}{\sum_{k=P+1}^{K} |\boldsymbol{s}^H(f)\boldsymbol{v}_k|^2} \tag{12}$$

where $K < N$ and calculation time can be shortened greatly.

Figure 1 shows the FFT and MUSIC spectra for a Japanese female vowel /a/. Figure 2 shows the eigenvalues μ_k. It is seen that the MUSIC spectrum has sharp peaks and the influence of band-limitation appears in a high-frequency domain more than 1 [kHz]. On the other hand, the calculation time has been shortened to about 1/7 of those in case of no band-limitation. Hence we can expect the realization of a fundamental frequency estimation method, which is not affected easily by additive noise and reduces the calculation time, by using the band-limited MUSIC spectrum.

In Fig. 2, K is set to 23 and the value of P is set up so that the set of eigenvalues $\{\mu_k; k = P+1, \cdots, K\}$ corresponding to the eigenvectors $\{\boldsymbol{v}_{P+1}, \cdots, \boldsymbol{v}_K\}$ used to estimate the spectrum satisfy $\mu_1/10 > \mu_k \geq \mu_K$. If the number of sinusoids contained in speech signals is known, we can set up the value of P. However, P is unknown in general. If P is too large, the number of harmonics contained in the spectrum will increase and come to be affected easily by the noise. Oppositely, if it is too small, the cepstrum will become smooth and the estimation error



(a) Speech signal. (b) FFT spectrum. (c) MUSIC spectrum.

Fig. 1 Analysis results for a Japanese female vowel /a/.

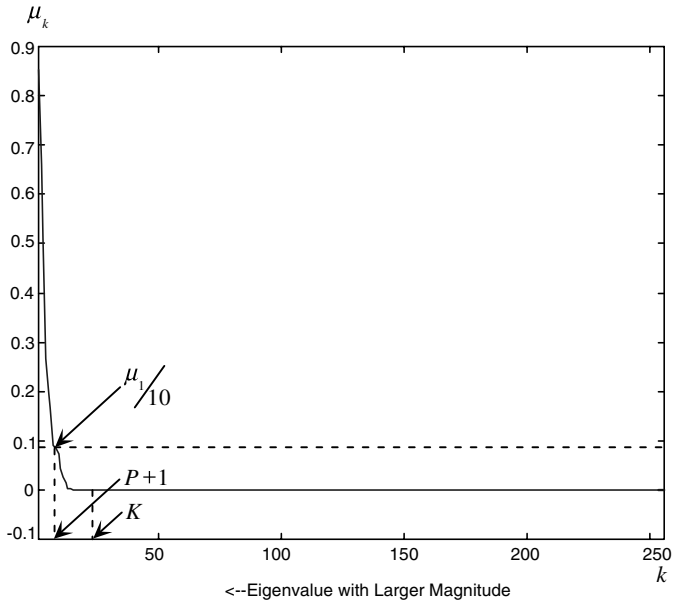


Fig. 2 Eigenvalues for a Japanese female vowel /a/.

of fundamental frequencies will increase. From experimental results, we use the set of eigenvalues $\{\mu_k; \mu_1/10 > \mu_k \geq \mu_K\}$ as mentioned above. In Fig. 2 the horizontal dotted line indicates the magnitude of the eigenvalue $\mu_1/10$ and P is set to 8.

3.2. Estimation Algorithm of Fundamental Frequency and Experimental Results

Figure 3 shows a MATLAB program for fundamental frequency estimation using the MUSIC algorithm. In this figure, a MATLAB function ''eigs'' computes only a few selected eigenvalues and eigenvectors. The proposed method estimates the fundamental frequency of speech signals by taking the FFT of the logarithm of the band-limited MUSIC spectrum like the cepstral method.

The analysis procedure is summarized as follows:

(1) The analyzed speech signal is sampled by 11.025 [kHz] and a 256-point Hamming window is applied.

(2) The autocorrelation matrix $\boldsymbol{R}_{yy}$ of the speech signal is computed from its power spectrum obtained by the FFT. We use only the frequency components below 1 [kHz] in consideration of the existence range of fundamental frequencies.

(3) The eigenvalues and eigenvectors of $\boldsymbol{R}_{yy}$ are computed using a MATLAB function ''eigs''. Each number of eigenvalues and eigenvectors is set to $K = 23$.

(4) The MUSIC algorithm computes a band-limited spectrum for the speech signal. The set of eigenvalues $\{\mu_k\}$, which span the noise subspace and are used for spectral estimation, are chosen so as to satisfy $\mu_1/10 > \mu_k \geq \mu_K$.

(5) The FFT is applied to the logarithmic power spectrum

```
% Fundamental Frequency Estimation Using MUSIC Algorithm
function main
clear
% File Name
FNAME='hirai_aiueo';
% Length of Data
N=256;
NN=fix(N/2);
% Sampling Frequency
FS=11025;
% Cut-off Frequency
FC=1000;
CN=fix(N*FC/FS);
% Start Point
NS=6500;
% Time Vector
t=(0:N-1)*1000/FS;
% Input of Speech Signal
voice=wavread(FNAME);
signal=voice(NS+1:NS+N);
% Hamming Window
signal=signal.*hamming(N);
% MUSIC Algorithm
musicsignal=func_music(signal,N,CN,FS);
tmp=max(musicsignal);
musicsignal=20*log10(musicsignal/tmp);
% DFT of MUSIC Spectrum
fftmusicsignal=musicsignal-min(musicsignal);
fftmusicsignal(1)=0;
fftmusicsignal=real(fft(fftmusicsignal,N));
fftmusicsignal=fftmusicsignal(1:NN);
% Fundamental Frequency Estimation
for k=1:NN-1
    if fftmusicsignal(k)<0
        break
    end
end
maxnum=k;
for k=maxnum+1:NN
    if fftmusicsignal(k)>fftmusicsignal(maxnum)
        maxnum=k;
    end
end
tmp=fftmusicsignal(maxnum-1:maxnum+1);
maxnum=maxnum-1+(tmp(1)-tmp(3))/(2*(tmp(1)-2*tmp(2)+tmp(3)));
pitchfftmusicsignal=FS/maxnum
```

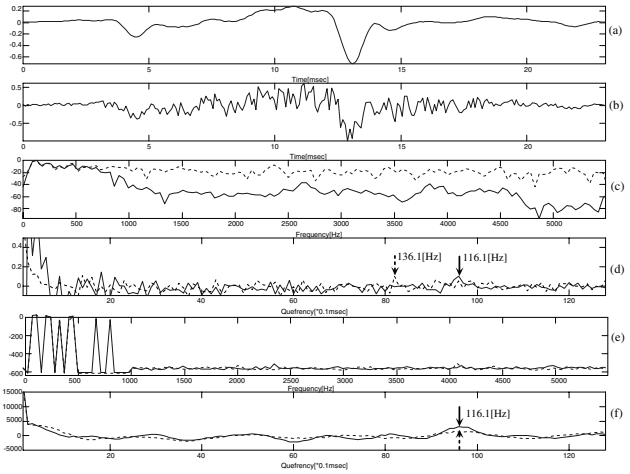
Continued

```
function[musicsignal]=func_music(signal,N,CN,FS)
% FFT
fftsignal=abs(fft(signal));
% Autocorrelation Matrix
A=zeros(CN-1);
for k=2:CN
    A(k-1,k-1)=(fftsignal(k)/N)*(fftsignal(k)/N);
    S(1:N,k-1)=exp(j*2*pi*(0:N-1)*(k-1)/N).';
end
Ryy=S*A*S';
% Eigenvalues and Eigenvectors
[V,D]=eigs(Ryy,CN);
D=abs(D);
PARAM=max(max(D))*1e-1;
num=0;
for k=1:CN
    if D(k,k)<PARAM
        num=num+1;
        Vf(1:N,num)=V(:,k);
    end
end
% MUSIC Algorithm
for k=1:N
    sf=exp(j*2*pi*(0:N-1)*(k-1)/N).';
    musicsignal(k)=1/abs(sf'*Vf*Vf'*sf);
end
```

Fig. 3 Fundamental frequency estimation using the MUSIC algorithm.

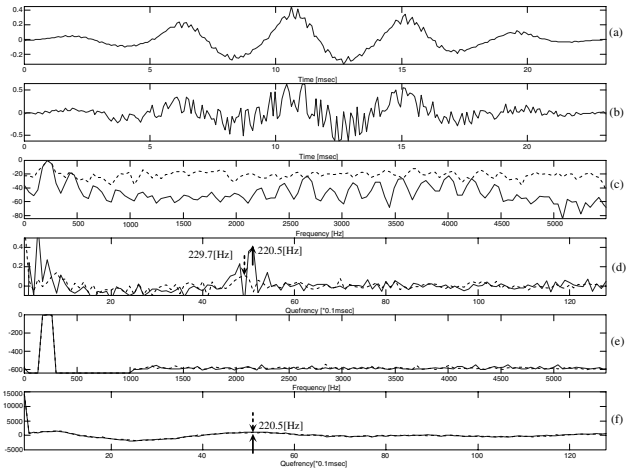
and the fundamental frequency is estimated from the peak location of the time-domain signal (i.e., cepstrum) obtained by its transformation using peak picking [10].

Japanese male and female vowels, /a/ and /i/, are tested in both noise free and noisy environments. In the experiment, the additive noise is Gaussian. We compare the proposed method with the cepstral method, which is commonly used for estimating the fundamental frequencies. In Figs. 4 and 5, the experimental results for the Japanese male vowel /a/ and the Japanese female vowel /i/ are shown respectively. In each figure, (a) shows the original speech signal, (b) the speech signal corrupted with additive noise (SNR = 0.63 [dB] and −0.19 [dB], respectively), (c) FFT spectrum of the speech signal, (d) cepstrum obtained by the FFT, (e) MUSIC spectrum and (f) cepstrum by the MUSIC algorithm. In (c)–(f) of Figs. 4 and 5, the solid lines denote the noise free environment and the dotted lines denote the noisy environment, respectively.



(a) Original speech signal. (b) Noisy speech signal. (c) FFT spectrum. (d) Cepstrum obtained by the FFT. (e) MUSIC spectrum. (f) Cepstrum by the MUSIC algorithm.

Fig. 4 Analysis results for a Japanese male vowel /a/ (SNR = 0.63 [dB]).



(a) Original speech signal. (b) Noisy speech signal. (c) FFT spectrum. (d) Cepstrum obtained by the FFT. (e) MUSIC spectrum. (f) Cepstrum by the MUSIC algorithm.

Fig. 5 Analysis results for a Japanese female vowel /i/ (SNR = −0.19 [dB]).

Table 1 Average value of absolute error rates of estimated fundamental frequencies.

	Male speakers					(%)
	/a/	/i/	/u/	/e/	/o/	Average
Cepstral method	22.5	40.7	3.6	2.5	1.8	14.2
MUSIC algorithm	0.5	3.9	2.2	3.9	0.5	2.2
	Female speakers					(%)
	/a/	/i/	/u/	/e/	/o/	Average
Cepstral method	0.9	3.9	4.2	2.0	3.3	2.9
MUSIC algorithm	0.7	3.8	1.6	0.8	0.7	1.5

These numerical values represent the average value for each vowel.

In case of the cepstral method, the estimated fundamental frequencies of the Japanese male vowel /a/ are 116.1 [Hz] for the noise free speech and 136.1 [Hz] for the noisy speech, respectively. In contrast, the fundamental frequencies estimated by the MUSIC algorithm are 116.1 [Hz] for both cases. For the Japanese female vowel /i/, the estimated fundamental frequencies by the cepstral method are 220.5 [Hz] and 229.7 [Hz], respectively. The fundamental frequencies by the MUSIC algorithm are 220.5 [Hz] for both cases.

Table 1 shows the average value of absolute error rates for Japanese 5 vowels uttered by 5 male and 5 female speakers in the noisy environment (SNR = 0.69 [dB]). We define the absolute error rate as

$$\text{absolute error rate} \triangleq \left| \frac{f_{\mathrm{M}} - f_{\mathrm{T}}}{f_{\mathrm{T}}} \right| \times 100\%. \qquad (13)$$

where f_{T} and f_{M} are true and estimated fundamental frequencies, respectively. The true fundamental frequencies were directly estimated from original speech waveforms. In this example, the average absolute error rate of the cepstral method for male speakers is 14.2% and that of the MUSIC algorithm is 2.2%. In addition, the average absolute error rates for female speakers are 2.9% and 1.5%, respectively. Though all the average values are large because of the low SNR, Table 1 suggests that the proposed method is superior to the conventional cepstral method for estimating the approximately true fundamental frequency.

4. CONCLUSION

We have proposed a new method to estimate the

fundamental frequency of noisy speech signals. Although the MUSIC algorithm is used briskly in the field of mobile communications, it seems that it is seldom used in the field of speech analysis. This research is very fundamental as application to speech signal processing of the MUSIC algorithm. However, we confirm that the feature of the method has been used efficiently.

ACKNOWLEDGEMENT

The authors are grateful to the anonymous reviewers for their helpful suggestions in improving the quality of this paper.

REFERENCES

[1] W. Hess, *Pitch Determination of Speech Signals* (Springer-Verlag, New York, 1983).

[2] M. Kaveh and A. J. Barabell, ''The statistical performance of the MUSIC and the minimum-norm algorithms in resolving plane waves in noise'', *IEEE Trans.* **ASSP-34**, 331–341 (1986).

[3] M. Egawa, T. Kobayashi and S. Imai, ''Instantaneous frequency estimation in low SNR environments using improved DFT-MUSIC'', *1996 IEICE General Conference*, A-158 (1996) (in Japanese).

[4] Y. Ogawa and K. Itoh, ''High-resolution estimation using the MUSIC algorithm'', *Trans. IEE Jpn.* **116**, 671–677 (1996) (in Japanese).

[5] S. L. Marple, *Digital Spectral Analysis with Applications* (Prentice-Hall, New Jersey, 1987).

[6] S. V. Vaseghi, *Advanced Signal Processing and Digital Noise Reduction* (Wiley, New York, 1996).

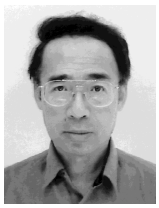
[7] N. Kikuma, *Adaptive Signal Processing with Array Antenna* (Science and Technology Publishing Company, Tokyo, 1999) (in Japanese).

[8] R. O. Schmidt, ''Multiple emitter location and signal parameter estimation'', *IEEE Trans.* **AP-34**, 276–280 (1986).

[9] M. S. Andrews, J. Picone and R. D. Degroat, ''Robust pitch determination via SVD based cepstral methods'', *ICASSP* '90, 253–256 (1990).

[10] J. D. Markel and A. H. Gray, *Linear Prediction of Speech* (Springer-Verlag, New York, 1976).

Takahiro Murakami was born in Chiba, Japan, on February 8, 1978. He received the B.E. degree in Electronics and Communication from Meiji University, Kawasaki, Japan, in 2000. He is currently working toward the M.E. degree at Graduate School of Electrical Engineering, Meiji University. He is interested in speech signal processing. He is a member of IEICE.



Yoshihisa Ishida was born in Tokyo, Japan, on February 24, 1947. He received the B.E., the M.E., and the Dr. Eng. Degrees in Electrical Engineering from Meiji University, Kawasaki, Japan, in 1970, 1972, and 1978, respectively. In 1975 he joined the Department of Electrical Engineering, Meiji University, as a Research Assistant and became a Lecturer and an Associate Professor in 1978 and 1981, respectively. He is currently a Professor at the Department of Electronics and Communication, Meiji University. His current research interests are in the area of digital signal processing, speech analysis. He is a member of ASJ, IEEE, and IEICE.