# Voice Activity Detection with Array Signal Processing in the Wavelet Domain

Yusuke HIOKA Nozomu HAMADA

Signal Processing Lab., Department of System Design Engineering, Keio Univ.

**ABSTRACT**

In many conventional voice activity detection (VAD) methods, speech signal is assumed to be acquired in high quality. However, human-machine interface based on speech is usually employed in indoor environment where various interferences exist, therefore, the VAD performance is seriously deteriorated. In this paper, we propose a novel VAD method with array signal processing on wavelet domain, in which we utilize the time, frequency and space information in the speech signal to separate interferences. In the proposed method, speech signal acquired by microphone array is at first decomposed into appropriate subbands with wavelet packet, and then array signal processing is executed on each subbands to realize VAD system for speech signal arriving from particular direction.

## 1 Introduction

Recently, human-machine interface system based on speech attracts much interests, supporting with the rapid improvement of the CPU performance. The speech-based interface is greatly based on speech recognition, in which the information of voice activity segments (VAS) is effective to improve the recognition rate. For the voice activity detection, various methods have been proposed. They use the features of speech signal, such as transition of the power[1], harmonic structure in spectrum[2][3][4] and the existence of signal source directionality[5]. In these methods, acquired speech is usually assumed to be sufficiently clean, due to the preprocessing used in speech recognition and compression for transmission. However at indoor environments where the interface is ordinarily used, there are various localized interferences arriving from particular direction such as the sound of closing door, etc. For these nonstationary interferences, the conventional methods do not realize sufficient performance, because of stationarity and whiteness assumption to noise.

Kaneda[6][7] proposed an effective VAD method available for these nonstationary interferences, using their high performance speech emphasizing system "AMNOR(Adaptive Microphone array for NOise Reduction)". He uses microphone array to discriminate signals utilizing direction difference between speech and interference. However, target speech and interference are required to arrive from sufficiently separated direction due to the spatial resolution in AMNOR. This limitation critically restricts the applicable condition of the method. In this research, we propose a new method to be robust to the direction of interference, with microphone array signal processing in the wavelet domain to integrate the time, frequency and spatial information of speech signal.

## 2 Speech Signal Features

In the proposed method, the features of a speech signal in the following three domains are considered.

1. Temporal Information : existence of stationarity, time localization
2. Frequency Information : existence of harmonical structure, power spectrum form
3. Spatial Information : direction of arrival (DOA), existence of directionality

The conventional VAD methods[1]-[7] make use of them individually.

This study aims to integrate the information in all three domains to realize robust system for nonstationary interference arriving from the direction close to that of desired speech.

## 3 Proposed Method

Fig.1 shows the overall flow chart of the proposed method whose main parts are array processing in the wavelet domain.

### 3.1 Signal modeling and delay-sum beamforming

At first, we assume speaker of this interface system is in its front(direction angle: $0°$) and the interference arriving angle is denoted by $\theta°$. The received signal of the $i$-th sensor on the $M$ sensors microphone array is modeled as

$$x_i(n) = s(n) + d(n - \tau_i) + n_i \qquad (1)$$

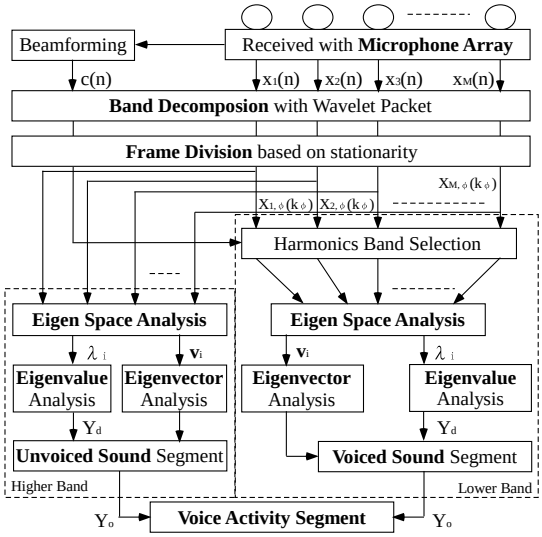


Figure 1: Proposed Method

,where $s(n)$,$d(n)$,$n_i$ are the desired speech, interference and the sensor noise respectively, and $\tau_i$ indicates the delay at $i$-th sensor. We use delay-sum beamforming for emphasizing the desired speech. In our case, it is realized by summing up all sensor signals as eq.2.

$$c(n) = \frac{1}{M}\sum_{i=1}^{M} x_i(n) \tag{2}$$

### 3.2 Wavelet Packet Decomposition

Generally, speech signal is classified into voiced and unvoiced sound based on the structure of utterance. Voiced sound contains harmonical structure in its spectrum and its power concentrates on the harmonical frequencies in lower band. Even though speech is a nonstationary signal, we can assume short-term stationarity for the voiced sound. On the other hand, power of unvoiced sound spreads over wide band, and it is difficult to assume any stationarity to it. From these features, high frequency resolution in lower band and high temporal resolution in higher band are desirable.
Applying the wavelet packet, the input signal $x_i(n)$ and the delay-sum beamforming output $c(n)$ are decomposed into subbands $X_{i,\phi}(k_\phi)$ and $C_\phi(k_\phi)$ with appropriate time and frequency resolution. The indices $\phi$ and $k_\phi$ mean the subband number and the sample point in the subband $\phi$. Lower band(under $2kHz$) is decomposed into equi-width subbands whose bandwidth is set to be smaller than the fundamental frequency of a voiced sound, to make one subband to contain only one harmonical component. On the other hand, higher band(above $2kHz$) is decomposed into octave bands, which is normally adopted in dyadic wavelet transform.

### 3.3 Frame Division

Because the speech signal is nonstationary, the wavelet packet coefficients $X_{i,\phi}(k_\phi)$ and $C_\phi(k_\phi)$ are divided into frames whose length are short enough to be able to assume stationarity. The time intervals, which can be supposed to be stationary for voiced and unvoiced sound, are different. Thus, this is taken into consideration on the frame division. In the proposed method, the frame lengths set to be about $20-30ms$ in lower band and shorter than $10ms$ in higher band. For simplification, we omit the index denoting the frame number in the following statement.

### 3.4 Voiced Sound Segment Detection

#### *3.4.1 Harmonics band selection*

For the result of subband decomposition in lower band, the harmonical components of voiced sound are supposed to be concentrated in particular subbands. These subbands are extracted by selecting $N$ largest subbands shown in eq.3,

$$\phi_p = arg\max_{\phi} \overline{|C_\phi|^2} \tag{3}$$

where $\phi$ indicates subband in lower band except for the lowest band that contains the DC component. $\overline{|C_\phi|^2}$ means the geometric mean of the coefficients $|C_\phi(k_\phi)|^2$ in each frame. In the selection of eq.3, subbands containing stationary voiced sound are prior to that of containing nonstationary interference.
The number of selected subbands $N$ is determined as 5, because normally the pitch(fundamental) frequency is lower than $400Hz$, therefore at least 5 harmonical frequencies must be in lower band.

#### *3.4.2 Eigenspace analysis*

Extracting the two spatial information, DOA and directionality, the eigenspace analysis of the array input is applied. In the case that $M$ sensors array receives signals from $m$ independent sources, the covariance matrix $\mathbf{R}$ of received signal $\mathbf{x}$ is decomposed into as followings by using its eigenvalue $\lambda_i$ and eigenvector $\mathbf{v}_i$.

$$\mathbf{x} = \sum_{l=1}^{m} f_l(t)\mathbf{s}_l + \mathbf{n} \tag{4}$$

$$\begin{aligned}\mathbf{R} &= E\left[\mathbf{x}\mathbf{x}^H\right] \\ &= [\mathbf{v}_1 \ \cdots \ \mathbf{v}_M]\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_M\end{bmatrix}\begin{bmatrix}\mathbf{v}_1^H\\ \mathbf{v}_2^H\\ \vdots\\ \mathbf{v}_M^H\end{bmatrix} \\ &\quad (\lambda_1 > \lambda_2 > \cdots > \lambda_M)\end{aligned} \tag{5}$$

$\mathbf{s}_l$ is the direction vector of a signal $f_l(t)$,

$$\mathbf{s}_l = \left[\ 1 \quad e^{-j\psi_1^l} \quad e^{-j\psi_2^l} \quad \cdots \quad e^{-j\psi_{M-1}^l}\ \right]^T \tag{6}$$

where $\psi_i^l$ is the phase delay of $f_l(t)$ at $i$-th sensor. As shown in Fig.2, each eigenvalue $\lambda_i$ is classified into two groups, signal subspace $\mathcal{S}$ and noise subspace $\mathcal{N}$ satisfying $\mathcal{S} \perp \mathcal{N}$, depending on its magnitude, and the corresponding eigenvector $\mathbf{v}_i$ becomes a basis of each subspace.

$$\mathcal{S} = span\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_m\} \quad (7)$$

$$\mathcal{N} = span\{\mathbf{v}_{m+1}, \mathbf{v}_{m+2}, \cdots, \mathbf{v}_M\} \quad (8)$$

At this point, the direction vector $\mathbf{s}_l$ is contained in the signal subspace $\mathcal{S}$.

$$\mathbf{s}_l \in \mathcal{S} \quad (9)$$

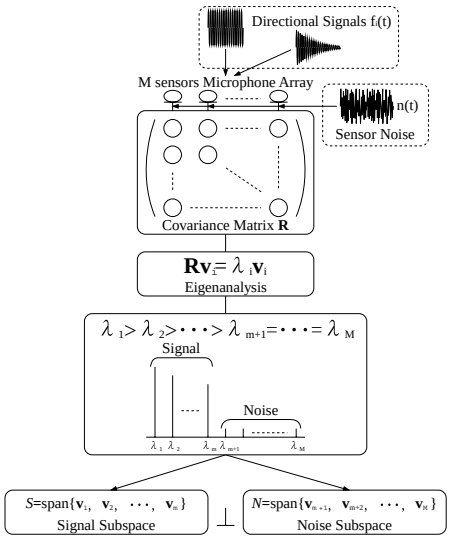


Figure 2: Eigenspace Analysis

### *3.4.3 Detection of directional signal segment*

For detecting the segment of directional signal(we call it "directional signal segment"), we use the distribution of the eigenvalue $\lambda_i$ at each frame. While any directional signal is received, larger eigenvalue components related to it exist in particular subspace $\mathcal{S}$, otherwise, they distribute uniformly. The distribution of the eigenvalues is classified by the normalized entropy $E$ about $|\lambda_i|$.

$$E = -\sum_{i=1}^{M} \frac{|\lambda_i|}{\sum_{i=1}^{M} |\lambda_i|} \log_M \frac{|\lambda_i|}{\sum_{i=1}^{M} |\lambda_i|} \quad (10)$$

Applying a threshold $E_{Th}$, which is decided from the input SNR and number of sensors, we can detect the directional signal segment $Y_d$ for $E \leq E_{Th}$.

### *3.4.4 Signal segment detection from desired direction*

Because the directional signal segments may contain not only desired speech segments but also that of interference with directionality, we use the DOA information. With an assumption that more than two interferences do no exist simultaneously, we use the orthogonality of $0^\circ$ direction vector $\mathbf{s}_0$ and secondary eigenvector $\mathbf{v}_2$ as follows.

- One directional signal case
  In this case, $\mathbf{v}_2$ belongs to $\mathcal{N}$. If $\mathbf{s}_0$ belongs to $\mathcal{S}$, namely the directional signal is desired speech, $\mathbf{s}_0$ and $\mathbf{v}_2$ are orthogonal.
  - speech segment
    $$\mathbf{s}_0 \in \mathcal{S}, \quad \mathbf{v}_2 \in \mathcal{N} \quad \Longrightarrow \quad \mathbf{s}_0 \perp \mathbf{v}_2$$
  - interference (not in front) segment
    $$\mathbf{s}_0 \notin \mathcal{S}, \quad \mathbf{v}_2 \in \mathcal{N} \quad \Longrightarrow \quad \mathbf{s}_0 \not\perp \mathbf{v}_2$$
- Two directional signal case (speech and interference overlapped)
  In this case, both $\mathbf{s}_0$ and $\mathbf{v}_2$ belong to $\mathcal{S}$, and they are not orthogonal. In addition, it is proved from simulation that desired signal component concentrates on a particular eigenvector. Besides, because the eigenvectors $\mathbf{v}_i$ are orthogonal each other, we can use the followings.
  - Speech superior segment
    $$\mathbf{s}_0 \in span\{\mathbf{v}_1\} \quad \Longrightarrow \quad \mathbf{s}_0 \perp \mathbf{v}_2$$
  - Interference superior segment
    $$\mathbf{s}_0 \in span\{\mathbf{v}_2\} \quad \Longrightarrow \quad \mathbf{s}_0 \not\perp \mathbf{v}_2$$

Applying a threshold $V_{Th} \approx 0$, the desired signal segment $Y_o$ is detected as $\mathbf{s}_0 \cdot \mathbf{v}_2 \leq V_{Th}$.

## 3.5 Unvoiced sound segment detection

For unvoiced sound segment detection, the eigenspace analysis is performed on the $X_{i,\phi}(k_\phi)$ in higher bands.

### *3.5.1 Threshold $E_{Th}$ determination*

The threshold $E_{Th}$ in higher band is set to be larger than that of in lower band, because the SNR is lower in higher band.

### *3.5.2 Utilization of all eigenvectors*

In higher band, the orthogonality of $\mathbf{s}_0$ and $\mathbf{v}_2$ decreases because the signal spectrum spreads in wideband. To suppress the effect, all of the eigenvectors $\mathbf{v}_i$ except for $\mathbf{v}_1$ are adopted for the orthogonality detection in higher band. This modification classifies an overlapped segment of unvoiced sound and interference into interference segment. It is not a fatal result because the unvoiced sound is small and highly localized, and is mostly covered with the interference.

## 3.6 Voice activity segment detection

Finally, the desired voice activity segment is derived from the voiced sound segment $Y_o^{Low}$ and unvoiced sound segment $Y_o^{High}$ as follows.

$$Y = Y_o^{Low} \cup Y_o^{High} \quad (11)$$

Table 1: Parameters in the simulation

| Band($Hz$) | Low($0-2000$) | High($2000-8000$) |
|---|---|---|
| $Fs(Hz)$ | 16000 | |
| SNR(dB) | 30 | |
| SIR(dB) | 0 | |
| $\theta(deg)$ | 10 | |
| $E_{Th}$ | 0.005 | 0.9 |
| $V_{Th}$ | 0.002 | 0.18 |

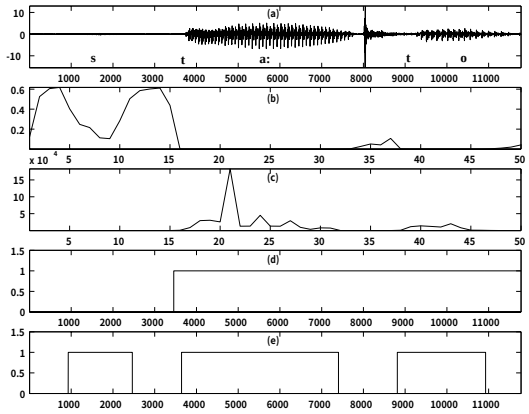


Figure 3: Voice and interference separately exist(a)Input signal,(b)$E^l$ in lower band,(c)$V^l$ in lower band,(d)Result(Conventional),(e)Result(Proposed)

## 4 Simulation Results

For the linear microphone array, 5 sensors are equally ($2cm$) spaced. A male speech "start" in Japanese and crap sound as interference received by a microphone are delayed to make the array input data virtually. The parameters in this simulation are shown in Tab.1. For the conventional method, we adopt [6], using delay-sum beamformer instead of AMNOR, to make it impartial comparison.

The result for the case that the voice and interference do not exit simultaneously is shown in Fig.3. In the result of conventional method, we can find the miss detection of interference due to the close direction of it, though the proposed method succeeds to discriminate it. On the other hand, the result for the case that the voice and interference simultaneously exist is shown in Fig.4, the overlapped segment of low SIR is removed correctly by the proposed method, though the conventional method failed to remove it.

Furthermore, the proposed method has the ability to detect the low power unvoiced sound segment, which is hard to detect by conventional method. This is because the proposed method looks for it in the subbands.

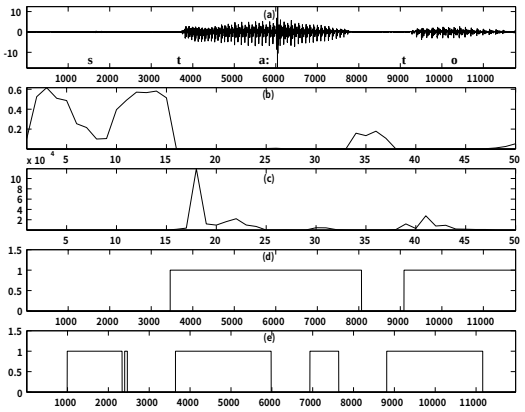


Figure 4: Voice and interference exist simultaneously

## 5 Conclusion

We proposed a VAD method with array signal processing in the wavelet domain to utilize the time, frequency and the spatial information in the desired speech signal. Applying the eigenspace analysis to the covariance matrix of the array received signal, the proposed method keeps its VAD precision as the direction of the interference closes to that of the desired speech.

From the simulation, the proposed method succeeds to detect correct segment under the close interference direction.

## References


[1] S.Furui,"Digital Speech Processing"Tokai University Pub.,1985 (in Japanese)

[2] I.Abdallah, S.Montresor, M.Baudry, "Robust Speech/Non Speech Detection in Adverse Condition Using an Entropy Based Estimator", 1997 13th Int. Conf. on DSP Proc., vol.2, pp757-760., 1997.

[3] Agbinya J I, "Discrete Wavelet Transform Techniques in Speech Proccesing.", IEEE TENCON, vol.2, pp514-519., 1996.

[4] S.Kadambe, G.F.Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals", IEEE Trans. Information Theory, vol.38, no.2, pp.917-924, 1992.

[5] J-F.Chen, W.Ser, "Speech Detection Using Mircophone Array", Electronic Letters, vol.36, no.2, pp181-182, 2000.

[6] Y.Kaneda, "Speech Period Detection Using a Microphone Array under Noisy Environments", Trans. IEICE A-73, no.8, pp1391-1398, 1990. (in Japanese)

[7] K.Kiyohara, Y.Kaneda, et al.,"A Mircophone Array System for Speech Recognition", IEEE ICASSP 97,vol.1, pp215-218,1997.