# Two-Microphone Voice Activity Detection in the Presence of Coherent Interference

*Gibak Kim and Nam Ik Cho*

School of Electrical Engineering
Seoul National University, Korea
kgb@ispl.snu.ac.kr, nicho@snu.ac.kr

## Abstract

In this paper, we propose a two-microphone Voice Activity Detection (VAD) method in the presence of coherent interference. The proposed method is based on the Cross Power Spectrum Phase (CPSP) which is an implementation of the Phase Transform (PHAT) weighted cross correlation between two microphones. The PHAT weighting whitens the spectrum of input signals and makes the cross correlation dependent entirely on the phase of the cross spectrum. If we assume that the direction of desired speech signal is known and the time delay between microphones is compensated, the Averaged CPSP (A-CPSP) can be utilized as a VAD measure. In order to enhance the VAD performance in the presence of strong coherent interference from other direction, we propose a Maximum Partially Averaged Real CPSP (MPA-RCPSP) method which detects the cophased frequency region with high Signal-to-Interference Ratio (SIR). Simulation results demonstrate that the proposed MPA-RCPSP is a more reliable measure to the conventional A-CPSP in the presence of strong coherent interference.

**Index Terms**: voice activity detection, two-microphone, Cross Power Spectrum Phase.

## 1. Introduction

The VAD is used to detect desired speech signal in the presence of noise and thus an essential part in many applications such as speech recognition, speech enhancement, and speech coding. In speech recognition systems, the accuracy of start and end point detection affects the recognition rate. In speech coding, we can reduce the coding rate by allocating no bits for the speech absent periods. VAD is also critical in adaptive speech enhancement since many algorithms use the noise statistics and noisy speech statistics which are estimated from speech present/absent periods.

There have been many single-microphone VAD algorithms based on time domain or spectral domain energy, zero-crossing rate, cepstral coefficients, spectral entropy and the like [1]. However, most of these methods fail in the presence of interference which has broadband speech-like spectral characteristic. For better VAD performance, multi-channel algorithms have been introduced, which take advantage of the spatial selectivity for discriminating the desired speech signal in the presence of speech-like noise. Specifically, Le Bouquin and Faucon introduced a technique based on the coherence function [2]. They assumed that the spatial correlation between the disturbing noises is weak for all frequencies of interest while the speech signals are highly correlated. And they compare the averaged Magnitude Squared Coherence (MSC) with a threshold to decide whether the speech is present in the current segment or not. However, this method fails in the presence of correlated noise signal. More recently, Armani *et. al.* proposed an algorithm based on CPSP Coherence Measure (CPSP-CM) that is usually used for speaker location and tracking purposes [3]. They calculate the maximum CPSP-CM from the current frame data and compare with the current threshold. With the assumption that the direction of desired speech signal is known, the CPSP-CM based VAD successfully detects the speech present segments in the presence of coherent interference as well as uncorrelated noise. However, in the strong coherent interference (low SIR), the CPSP of noisy signal is much affected by the coherent interference, resulting in unreliable VAD.

We propose a more reliable two-microphone VAD measure in the presence of strong coherent interference. This method searches for the cophased frequency region of the CPSP for more reliable detection of desired speech signals. Intensive simulation results show that the proposed measure is more reliable than the CPSP-CM based VAD measure.

This paper is organized as follows. In the next section, we describe some backgrounds for the CPSP and the classification of noise field. Section 3 presents the proposed method that detects the cophased frequency region of the CPSP. Section 4 shows the simulation results and section 5 concludes the paper.

### Abbreviations

| | |
|---|---|
| VAD | Voice Activity Detection |
| CPSP | Cross Power Spectrum Phase |
| PHAT | Phase Transform |
| A-CPSP | Averaged Cross Power Spectrum Phase |
| MPA-RCPSP | Maximum Partially Averaged Real CPSP |
| SIR | Signal-to-Interference Ratio |
| MSC | Magnitude Squared Coherence |
| CPSP-CM | CPSP-based Coherence Measure |
| PSD | Power Spectral Density |

## 2. Backgrounds

### 2.1. Cross Power Spectrum Phase (CPSP)

The discrete time signals received at the two microphones are modeled as

$$x_1(n) = s(n) + v_1(n) \quad (1)$$

$$x_2(n) = \alpha s(n - D) + v_2(n) \quad (2)$$

where $s(n)$ is a desired speech signal at discrete time index $n$ and $\alpha$ reflects the attenuation which is assumed to be time invariant. The noise $v_1(n)$ and $v_2(n)$ model the ambient noise, sensor noise, or coherent interference produced by localized sources and also include the reverberation. $D$ is the time delay in samples, which

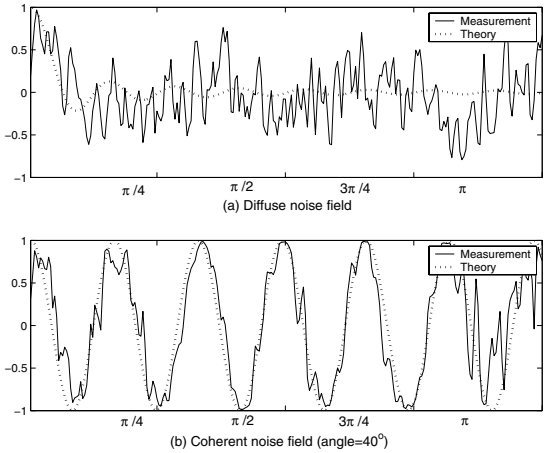


Figure 1: *Normalized cross PSD. (a) Diffuse noise field. (b) Coherent noise field.*

depends on the direction of desired speech signal. The speech signal is assumed to be uncorrelated with noise signals except the severe reverberant signals.

The information of phase difference between the two microphone signals can be revealed by the normalized cross Power Spectral Density (PSD). The cross PSD between $x_1(n)$ and $x_2(n)$ is written as

$$G_{x_1x_2}(e^{j\omega}) = \alpha G_{ss}(e^{j\omega})e^{j\omega D} + G_{v_1v_2}(e^{j\omega}) \tag{3}$$

where $\omega$ is the normalized frequency and $G_{ss}(e^{j\omega})$ is the auto PSD of desired speech signal, and $G_{v_1v_2}(e^{j\omega})$ is the cross PSD of noise signals. If the noise signals are uncorrelated ($G_{v_1v_2}(e^{j\omega}) = 0$), the cross PSD normalized by the magnitude is a complex exponential revealing the time delay $D$, i.e.,

$$\frac{G_{x_1x_2}(e^{j\omega})}{|G_{x_1x_2}(e^{j\omega})|} = e^{j\omega D}. \tag{4}$$

Practically, the normalized cross PSD is estimated by CPSP that is calculated through the discrete time Fourier transform applied to the windowed segments of $x_1(n)$ and $x_2(n)$ as

$$\phi(e^{j\omega}) = \frac{X_1(e^{j\omega})X_2^*(e^{j\omega})}{|X_1(e^{j\omega})X_2^*(e^{j\omega})|}. \tag{5}$$

### 2.2. Noise field

The noise signal can be classified into incoherent/coherent/diffse noise according to the coherence of a sound field [4]. The properties of noise fields are summarized in terms of normalized cross PSD as follows.

#### *2.2.1. Incoherent noise*

The normalized cross PSD is zero at all frequencies in the incoherent noise-field:

$$\frac{G_{v_1v_2}(e^{j\omega})}{|G_{v_1v_2}(e^{j\omega})|} = 0, \forall\omega. \tag{6}$$

Thermal noise in each microphone can be modeled as the incoherent noise.

#### *2.2.2. Diffuse noise*

The diffuse noise field is the noise model that is generated by infinite noise sources in every direction of a hypothetical sphere with an infinite radius. So it has uniform energy flow in all directions. The diffuse noise field model well describes the practical noise field in cars. In practice, noise sources arrive from all directions within a given time window in diffuse noise field. Consequently, the reverberant room can also be modeled as a diffuse field. The normalized PSD in a theoretical diffuse noise field is dependent on the distance between two microphones $d$ and the signal frequency given by

$$\frac{G_{v_1v_2}(e^{j\omega})}{|G_{v_1v_2}(e^{j\omega})|} = \frac{\sin(\omega f_s d/c)}{\omega f_s d/c} \tag{7}$$

where $c$ denotes the speed of sound and $f_s$ is the sampling frequency. The theoretical and measured normalized PSD for diffuse noise field are depicted in Fig. 1(a). Note that the normalized cross PSD of signals at the frequency higher than $\omega = \pi c/(f_s d)$ can be ignored in practice. Therefore, the diffuse noise field can be considered as the incoherent noise field as long as the distance between two microphones is sufficiently large.

#### *2.2.3. Coherent noise*

A directional wavefront coming from a localized sound source is modeled as the coherent noise field. In principle, given two microphone signals, one of them is a scaled and delayed signal of the other. In this case, the normalized cross PSD is given as

$$\frac{G_{v_1v_2}(e^{j\omega})}{|G_{v_1v_2}(e^{j\omega})|} = e^{j\omega f_s d\cos\theta/c}. \tag{8}$$

The real part of the normalized PSD for coherent noise is described in Fig. 1(b), with the example of noise coming from the angle of $40^\circ$.

### 2.3. VAD based on Cross Power Spectrum Phase

In [5], the inverse discrete time Fourier transform of $\phi(e^{j\omega})$ is used as the coherence measure for finding the time delay between two microphone signals as

$$C(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi}\phi(e^{j\omega})e^{j\omega\tau}d\omega. \tag{9}$$

The coherence measure $C(\tau)$ has the cross correlation peak at the time delay $D$. In [3], the CPSP-CM was used to detect start-end point for distant-talking speech recognition. They calculate the CPSP-CM from the current frame, and compare the maximum value with the threshold for the detection of coherent directional speech. However, in the presence of competing coherent noise, the CPSP-CM at the time delay $D$ may not be the maximum any longer. In this case, we need the information about the direction of desired speech signal or the corresponding time delay between two microphones. If the time delay for the desired speech signal is equal to $D$, $C(D)$ is compared with the threshold. For simplicity, we assume zero time delay ($D = 0$) without loss of generality. Note that $C(0)$ is the inverse discrete time Fourier transform calculated at $\tau = 0$ in (9), which can be considered as the Averaged CPSP (A-CPSP) over all frequencies. For the high SIR, CPSP $\phi(\omega)$ is close to one at all frequencies and A-CPSP approaches one. As the power of coherent interference becomes higher than that of desired speech signal, the CPSP oscillates according to the

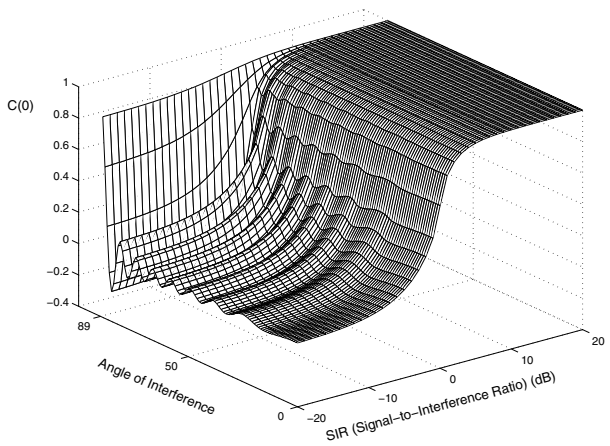


Figure 2: *Averaged Cross Power Spectrum Phase (A-CPSP) according to the angle of interference and SIR.*

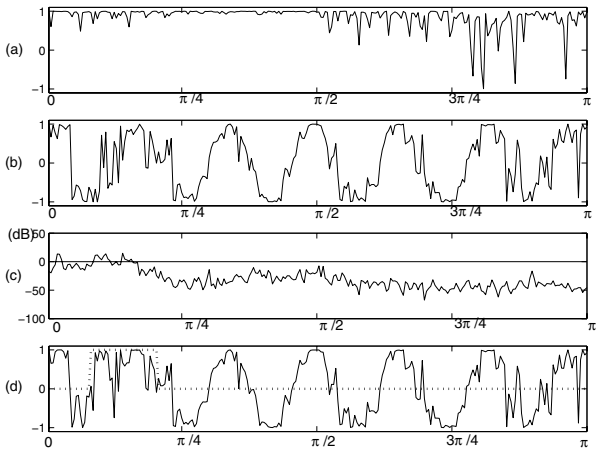


Figure 3: *Detection of cophased frequency region. (a) Real CPSP of desired speech signal. (b) Real CPSP of coherent interference. (c) SIRs at each frequency bin of the noisy signal. (d) Real CPSP of the noisy signal and $P$-point frequency window (dotted line).*

CPSP of interference and A-CPSP approaches zero. Theoretical analysis for A-CPSP shows that A-CPSP rapidly varies around 0dB SIR. It approaches one for the SIR above 0dB and approaches zero for SIR below 0dB. When the sampling frequency is $16kHz$ and the distance between two microphones is $34cm$, and the interference is coming from the source at the angle of $40^\circ$, the theoretical A-CPSP values are 0.935, 0.894, 0.820, 0.645, 0.453, 0.350, 0.276 for the SIRs of 3dB, 2dB, 1dB, 0dB, -1dB, -2dB, -3dB (Fig. 2) respectively. It can be observed that A-CPSP rapidly goes to zero as SIR goes below 0dB, and thus speech present segment is apt to be missed in low SIR. Moreover, even when the SIR is higher than 0dB, A-CPSP tends to go to zero when the frequency region with SIR below 0dB is wider than that with SIR above 0dB. The speech signal usually has resonances called "formants" and most acoustic energies are concentrated on the narrow frequency bands. Therefore, if the interference has broadband spectrum, even moderate interference can yield wide frequency region with SIR below 0dB.

In order to alleviate the above mentioned problems of A-CPSP, we propose an alternative method that searches for the maximum of partially averaged real CPSP (MPA-RCPSP) and use it as a measure for detecting speech present segments. The MPA-RCPSP for the current frame is obtained by shifting a $P$-sized frequency window and averaging real CPSPs within this window, and then finding the maximum averaged value. This process aims to detect a partial cophased frequency region where the SIR is higher than 0dB, thereby detecting speech present segments more reliably in the presence of strong coherent interference. Fig. 3 illustrates the proposed method. The CPSPs of the cophased speech signal and the coherent interference with angle of $40^\circ$ are described in Fig. 3(a),(b). Fig. 3(c) shows SIRs at each frequency bin of the noisy signal with -12.7dB SIR for the whole frequency bins which is defined as

$$\mathrm{SIR(dB)} = 10\log_{10}\frac{\int_{-\pi}^{\pi}\left|X_1^s(e^{j\omega})\right|^2 d\omega}{\int_{-\pi}^{\pi}\left|X_1^v(e^{j\omega})\right|^2 d\omega} \tag{10}$$

where $X_1^s(e^{j\omega}), X_1^v(e^{j\omega})$ are the speech/noise components of the discrete time Fourier transform of the signal at the first microphone. Even though the SIR is very low (-12.7dB) in this example, there exists a cophased frequency region with SIR above 0dB at low frequencies, and this region can be detected by finding MPA-RCPSP which is close to one (Fig. 3(c)). Therefore, the cophased speech signal can be detected using this measure as long as there exists a frequency region with higher SIR than 0dB during $P$-sized frequency window. This method is based on the fact that the A-CPSP approaches one with just slightly higher than 0dB (0.820 for 1dB) (Fig. 2).

The window size $P$ is generally chosen as the multiples of the CPSP period corresponding to the interference in order that the measure is close to zero for the interference-only periods.

$$\tau = \left[\frac{d\cos\theta}{c} f_s\right] \tag{11}$$

$$P = m\frac{2\pi}{|\tau|}, \;\; m = 1, 2, \cdots, |\tau| \tag{12}$$

where $[\cdot]$ denotes the nearest integer to the argument. The window size $P$ less than the period of CPSP results in increasing the MPA-RCPSP during the interference-only periods. To suppress the MPA-RCPSP during the interference-only periods, we subtract the CPSP due to interference from the CPSP of noisy signal such as

$$\tilde{\phi}(e^{j\omega}) = \phi(e^{j\omega}) - C(\tau)e^{-j\omega\tau} \tag{13}$$

where $\tau$ is the time delay between two microphones for the interfering signal.

## 3. Simulation Results

In the simulations, we generate multichannel noisy signals by adding noise to the speech. The reverberant multichannel signals are generated by the convolution of dry source (sound data measured in an anechoic room) with acoustic impulse responses from the RWCP Sound Scene Database [6]. The RWCP Sound Scene Database is a common database collected in real acoustic environment for research in varied fields of application such as speech

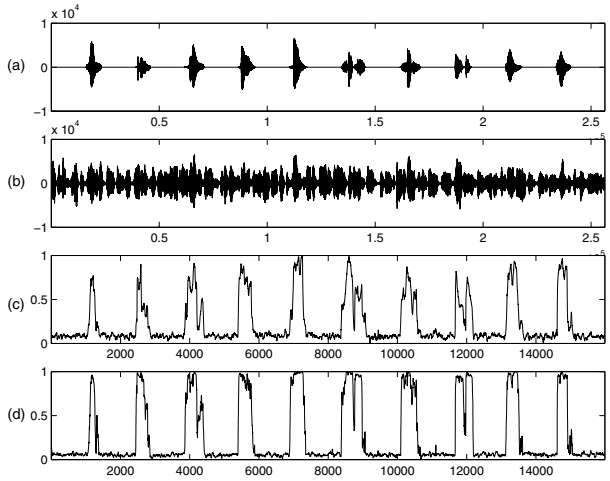


Figure 4: *Interference of a competing speech. (a) Desired speech signal. (b) Noisy signal contaminated by directional competing speech. (c) Normalized A-CPSP. (d) Normalized MPA-RCPSP.*

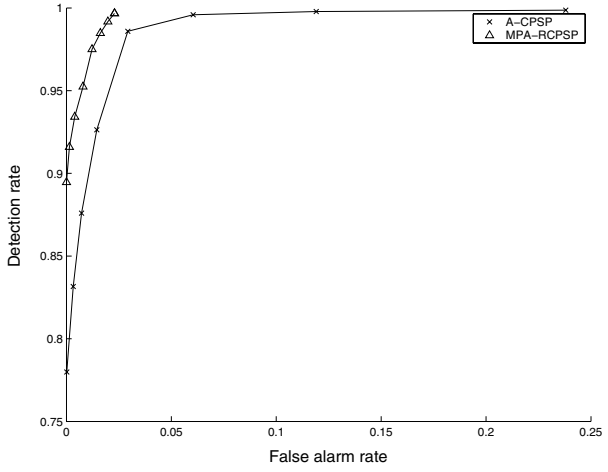


Figure 5: *Speech detection rate vs. false alarm rate (ROC curves).*

recognition, echo canceller, active noise control and so on. Numerous impulse responses with microphone array are measured in real environments to simulate the various environments by the convolution with dry sources. The impulse responses are measured at several positions which are 2m distance from the microphone array with reverberation time of 300 ms. Speech signal is convolved with the impulse response measured at the fore side of the microphone array. In order to test the algorithm with coherent interference, the competing speech is convolved with the impulse response measured at the angle of $40^{\circ}$.

The simulation results in the presence of directional competing speech are displayed in Fig. 4. The interference is added to the utterances of ten digits with 0dB SIR. The sampling rate is 16kHz and 512 points FFT is applied to the hanning-windowed data. Fig. 4(a) shows the desired speech signal. The noisy speech signal corrupted by directional interference is shown in Fig. 4(b). The A-CPSP and MPA-RCPSP are normalized so that the maximum is 1 and the minimum is 0 for the purpose of comparison (Fig. 4(c),(d)). The figure shows that the proposed method provides more reliable VAD measure than the conventional method. The receiving operating characteristic (ROC) curves for the speech detection rate vs. false alarm rate are also described to compare the performance of the proposed MPA-RCPSP with the conventional A-CPSP (Fig. 5). These curves show the trade-off between the correct speech detection rate and false alarm rate depending on the threshold and support that the proposed MPA-RCPSP outperforms the conventional A-CPSP.

## 4. Conclusions

We have proposed a new two-microphone VAD algorithm based on CPSP. The proposed method searches for the maximum of partially averaged real CPSP to find the cophased frequency region and uses it as a measure for detecting speech present segments instead of the averaged CPSP over all frequencies. The cophased speech signal can be detected using this measure as long as there exists a frequency region with higher SIR than 0dB during $P$-sized frequency window. In the simulation results, the ROC curves for speech detection rate vs. false alarm rate supports that the proposed method outperforms the conventional method.

## 5. Acknowledgements

This research was performed for the Intelligent Robotics Development Program, one of the 21st Century Frontier R&D Programs funded by the Ministry of Commerce, Industry and Energy of Korea.

## 6. References

[1] B.-F. Wu, "Robust endpoint detection algorithm based on the adaptive band-partitioning spectral entropy in adverse environments," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, September 2005.

[2] R. Le Bouquin Jeannes and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system", Speech communication vol. 16, pp. 245–254, 1995.

[3] L. Armani, M. Matassoni, M. Omologo, and P. Svaizer, "Use of a CSP-based voice activity detector for distant-talking ASR," In *Eurospeech*, pp. 501-504, 2003.

[4] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Communication 25*, pp. 75-95, 1998.

[5] M. Omologo and P. Svaizer, "Acoustic event localization using a crosspower-spectrum phase based technique," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 273-276, 1994.

[6] "RWCP Sound Scene Database in Real Acoustical Environments," Real World Computing Partnership, (c)1998-2001.