マイクロホンアレーを用いた時間 / 空間情報に基づくハンズフリー発話区間検出の検討

傳田 遊亀 †　田中 貴雅 †　中山 雅人 †　西浦 敬信 ‡　山下 洋一 ‡

† 立命館大学大学院 理工学研究科

〒 525-8577 滋賀県草津市野路東 1-1-1

‡ 立命館大学 情報理工学部

〒 525-8577 滋賀県草津市野路東 1-1-1

E-Mail:{gr021052@se, rs012019@se, gr020040@se, nishiura@is, yama@media}.ritsumei.ac.jp

あらまし　ハンズフリー音声認識において発話区間検出 (VAD : Voice Activity Detection) は必要不可欠である．受音信号の時間情報のみに基づいた従来の時間領域 VAD 法は，雑音の影響の少ない近接発話では高い性能を得ることが出来る一方，雑音によって大きく歪みを受けた遠隔発話では十分な性能を得られないという問題がある．そこで本稿では，時間情報に基づいた時間領域 VAD 法と空間情報に基づいた空間領域 VAD 法を統合することを検討し，雑音に頑健な時間 / 空間領域ハンズフリー VAD 法を提案する．提案手法では，ゼロ交差検出 (ZCD : Zero Crossing Detection) 法に基づく時間領域 VAD 法と，音声信号の到来方向推定に特化した WCSP (Weighted Cross-power Spectrum Phase) 法に基づく空間領域 VAD 法を統合することで発話区間を検出する．実騒音環境における評価実験の結果，提案手法はハンズフリー環境において，従来手法よりも高い発話区間検出性能を得られることを確認した．

キーワード　ハンズフリー発話区間検出, ゼロ交差検出法, Weighted CSP 法, ハンズフリー音声認識, マイクロホンアレー

A Study of Hands-free Voice Activity Detection Based on Time / Spatial Information Using a Microphone Array

Yuki Denda†　Takamasa Tanaka†　Masato Nakayama†　Takanobu Nishiura‡　Yoichi Yamashita‡

† Graduate School of Sience and Engineering, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577 JAPAN

‡ College of Information and Scienece, Ritsumeikan University

1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577 JAPAN

E-Mail:{gr021052@se, rs012019@se, gr020040@se, nishiura@is, yama@media}.ritsumei.ac.jp

Abstract　Voice activity detection (VAD) is necessary for hands-free speech recognition. Conventional time-domain VAD algorithms based on only time-sequence information of captured signals works well with closed-talking speech. However, it can not achieve the satisfactory VAD performance with noisy distant-talking speech. To overcome this problem, in this paper, we study to integrate the time-domain VAD algorithm based on the time-sequence information and spatial-domain VAD algorithsm based on spatial-sequence information, and propose the noise robust time / spatial-domain VAD algorithm. The zero crossing detection (ZCD) method is employed as a time-domain VAD algorithm and the weighted cross-power spectrum phase (WCSP) analysis proposed for noise robust direction of arrival estimation of the target talker's speech is employed as a spatial-domain VAD algorithms. The proposed hands-free VAD algorithm then performs the hands-free VAD by integrating both ZCD method and WCSP analysis. As a result of evaluation experiments in an actual room, we confirmed that the performance of the proposed VAD is superior to that of the conventional VAD in hands-free environments.

Key words　Hands-free voice activity detection, Zero crossing detector, Weighted CSP analysis, Hands-free speech recognition, Microphone array.

1 はじめに

近年，ヘッドセットマイクなどを装着せずに，自由に移動しながら発話された音声を認識するハンズフリー音声認識の需要が高まっている．しかし，マイクロホンから離れた位置の発話者の音声を認識する必要があるハンズフリー音声認識においては，残響や背景雑音の影響により受音した音声が大きく歪むため，音声認識率が著しく低下するという問題がある．近年，この問題を解決するために，マイクロホンアレー [1] をハンズフリー音声認識に応用するための研究が盛んに行われている [2]．これらの研究では，マイクロホンアレーを用いて発話者の方向に指向特性を形成 (ビームフォーミング) し，目的音声と雑音を空間的に分離することで高音質な音声の受音を実現している．しかし，マイクロホンアレーを用いるためには以下の技術を実現することが必要不可欠である．

1. 発話の有無 (発話区間) を検出する．
2. (発話があった場合は) 発話者の位置を推定する．
3. ビームフォーミングによって音声を強調する．

我々はこれまでに，雑音環境下においても頑健に発話者の位置を推定することを目的として，音声の到来方向推定に特化した WCSP (Weighted Cross-power Spectrum Phase) 法に基づいた発話者位置推定法を提案している [3]．本稿では，次のステップとして発話区間検出 (VAD : Voice Activity Detection) について検討を行う．発話区間を正確に検出できない場合，発話者位置推定性能が低下するばかりでなく，音声認識性能も著しく低下する．また，適応型マイクロホンアレー [4, 5] のフィルタを更新するためには，発話区間検出は非常に重要である．

従来の VAD 法としては，受音した信号から音声の各種特徴量を抽出し，音声 / 非音声の識別を行うことで発話区間を検出する手法が数多く提案されている．例えば，ゼロ交差 [6]，振幅 (パワー) レベル [6] といった時間情報に基づく時間領域 VAD 法や，スペクトル情報 [7]，音声 GMM (Gaussian Mixture Model) の尤度 [8] といった周波数情報に基づいた周波数領域 VAD 法が提案されている．さらに，複数の特徴量を統合して発話区間を検出する手法も提案されている [9]．

従来の時間情報や周波数情報のみに基づいた VAD 法は，主に接話マイクロホンで受音した近接発話を想定している．しかし，マイクロホンアレーで受音した遠隔発話は雑音によって歪みを受けているため，従来手法の発話区間検出性能が大きく低下するという問題がある．本稿ではこの問題を解決するために，従来の時間情報のみでなく空間情報も積極的に利用することを検討し，雑音に頑健な時間 / 空間領域ハンズフリー VAD 法を提案する．提案手法では，ゼロ交差検出 (ZCD : Zero Crossing Detection) 法 [6] に基づく時間領域 VAD 法によって時間情報を抽出する．一方，マイクロホンアレーで音響信号を受音した場合，空間内の音環境を把握することが可能である．そこで本稿では，発話者の位置情報を推定する WCSP 法に基づいた空間領域 VAD 法によって空間情報を抽出する．最後に時間 / 空間情報を統合することで，時間 / 空間領域において正確に発話区間を検出することを検討する．

2 従来の時間領域 VAD 法

2.1 ゼロ交差検出法

ゼロ交差検出 (ZCD : Zero Crossing Detection) 法は，振幅閾値とゼロ交差閾値の二つの閾値に基づいて，主に母音などの比較的大きいエネルギーを持つ音声区間を検出する時間領域 VAD 法である [6]．ZCD は以下の式で表すことができる．

$$z = ZCD\big(x(t), Th_a\big), \tag{1}$$

$$\mathrm{VAD} = \begin{cases} \text{Speech frame}, & z \geq Th_z \\ \text{Non–speech frame} & z < Th_z \end{cases}, \tag{2}$$

ここで $x(t)$ はフレーム長 T の短時間フレームにおける受音信号を，Th_a は振幅閾値を，z はゼロ交差回数を，Th_z はゼロ交差回数閾値を，関数 $ZCD(x(t), Th_a)$ はゼロ交差回数を返す関数を表す．また，関数 $ZCD(x(t), Th_a)$ は以下の手順で実現できる．

Step 1. z を 0 で初期化．
Step 2. $x(t)$ が Th_a 以上なら **Step 3.** へ．
$x(t)$ が Th_a より小さければ **Step 4.** へ．
Step 3. $x(t)$ が正，かつ $x(t+1)$ が負なら z をインクリメントして **Step 4.** へ．
$x(t)$ が負，かつ $x(t+1)$ が正なら z をインクリメントして **Step 4.** へ．
Step 4. t が T より小さければ t をインクリメントして **Step 2.** へ．
t が T 以上なら終了．

ZCD は計算量が少なく簡便な手法であるため，音声認識エンジン Julius[10] などにおいて広く利用されている．しかし，式 (2) の振幅閾値が固定されているため雑音環境の変化に脆弱であるという問題がある．この問題を解決するために，非発話区間で式 (3) によって振幅閾値を更新する方法が提案されている [6]．

$$Th_{a_{i+1}} = (1-p)Th_{a_i} + p\overline{|x(t)|}, \tag{3}$$

ここで p は更新係数を，$\overline{|x(t)|}$ は受音信号の平均振幅を表す．しかし，振幅閾値の更新が受音信号の振幅 (時間情報) のみに基づいているため，高雑音環境下においては性能がそれほど改善されないという問題がある．

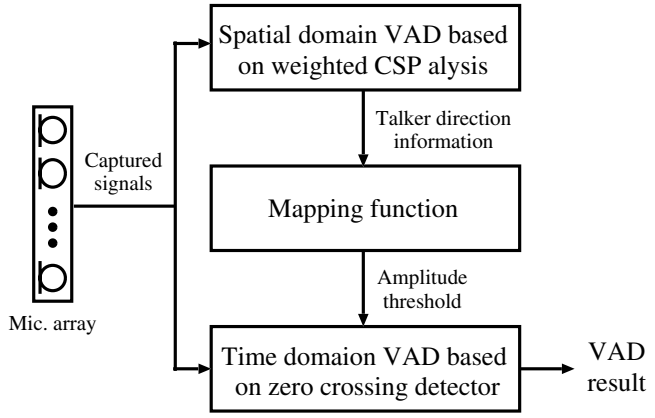


図 1: 提案手法の概要

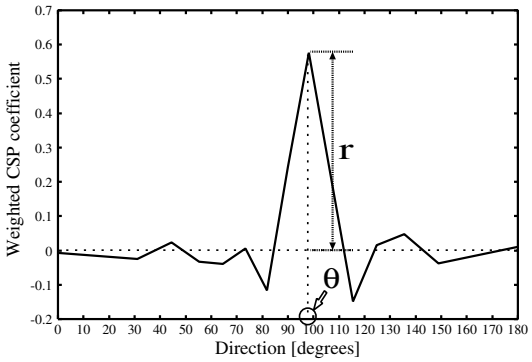


図 2: WCSP 係数の一例

3 時間 / 空間領域におけるハンズフリー VAD

提案する時間 / 空間領域 VAD 法の概要を図 1 に示す．提案手法ではまずはじめに，マイクロホンアレーで受音した音声信号の到来方向 (DOA : Direction Of Arrival) を WCSP (Weighed Cross-power Spectrum Phase) 法によって推定し，発話者の位置情報 (空間情報) を抽出する．次に，マッピング関数に基づいて空間情報を ZCD 法の振幅閾値にマッピングする．最後に，求めた振幅閾値とゼロ交差閾値に基づいて音声 / 非音声区間の識別を行い，発話区間を推定する．

3.1 WCSP 法に基づく空間領域 VAD 法

ハンズフリー音声認識においては，目的信号を音声に限定することができる．さらに，目的音声と雑音のスペクトルに相関がないと仮定する．我々は以上の仮定に基づいて，音声の DOA 推定に特化した WCSP 法をこれまでに提案している [3].

$x_1(t), x_2(t)$ がマイクロホン M_1, M_2 で受音した時間領域信号を表すとすると，WCSP 法は以下の式で表すことができる．

$$WCSP(k) = \mathrm{IDFT}\left[W(\omega)\frac{X_1(\omega)X_2(\omega)^*}{|X_1(\omega)||X_2(\omega)|}\right], \quad (4)$$

$$[r, \theta] = f_{max}\big(WCSP(k)\big), \quad (5)$$

ここで $WCSP(k)$ は WCSP 係数を，IDFT[·] は逆離散フーリエ変換 (IDFT : Inverse Discrete Fourier Transform) を，$W(\omega)$ は音声の平均スペクトルに基づいた重み係数を，$X_{[\cdot]}(\omega)$ は時間領域信号 $x_{[\cdot]}(t)$ の周波数表現を，* は複素共役を，r は WCSP 係数の最大値を，θ は r に対応する角度を，f_{max} は r と θ を返す関数を表す．式 (4) より，WCSP 法は音声信号の平均スペクトル特性 $W(\omega)$ を音声の DOA 推定の信頼度基準として，各周波数に重み付けを行う．従って，WCSP 法は音声の推定に特化した到来方向推定法として解釈することができる．なお本稿では，ATR 日本語音素バランス 503 文 [11] から女性話者 14 名，男性話者 6 名の発話を使用してあらかじめ計算した重み係数を使用した [3].

発話者が 100 度方向から発話してる状況で得られた WCSP 係数の一例を図 2 に示す．図 2 より，音声の到来する 100 度方向の WCSP 係数が非常に大きな値を持っている．一方，100 度以外の他の角度方向の WCSP 係数は小さな値になっている．従って，各方向の WCSP 係数の値はその方向から音声が到来している，発話者が存在している信頼度基準として捉えることができる．さらに，WCSP 係数が全方位にわたって小さい値を持つ場合，どの方向からもマイクロホンアレーに音声が到来していない，つまり非発話区間であると識別することができる．

3.2 時間領域 VAD 法と空間領域 VAD 法の統合

提案するハンズフリー VAD 法では，ZCD 法に基づく時間領域 VAD 法と WCSP 法に基づく空間領域 VAD 法を以下の式によって統合する．

$$Th_{\hat{a}} = f_{map}(r), \quad (6)$$

$$z = ZCD\big(x(t), Th_{\hat{a}}\big), \quad (7)$$

$$\mathrm{VAD} = \begin{cases} \text{Speech frame}, & z \geq Th_z \\ \text{Non–speech frame} & z < Th_z \end{cases}, \quad (8)$$

ここで $Th_{\hat{a}}$ は，マッピング関数 $f_{map}(r)$ によって決定される振幅閾値を表す．

図 3 に本稿で使用したマッピング関数を示す．音声信号到来の信頼度基準である r が大きな値を持つ場合は発話区間であると考えられる．従って，図 3 に示すように振幅閾値 $Th_{\hat{a}}$ を小さな値にマッピングすることで，発話区間におけるゼロ交差を頑健に検出することができる．一方, r が

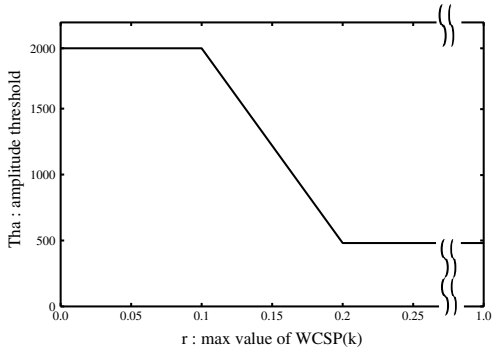


図 3: マッピング関数

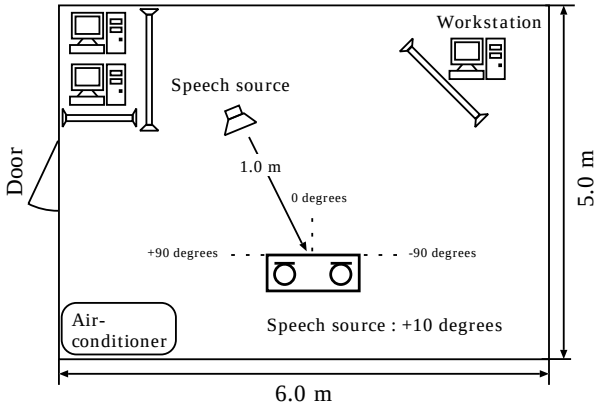


図 4: 実験環境

小さな値を持つ場合は非発話区間であると考えられるため，振幅閾値 $Th_{\hat{a}}$ を大きな値にマッピングすることで非発話区間におけるゼロ交差の誤検出を防ぐことができる．なお本稿では，**4. 評価実験**と同じ環境における予備実験に基づいてマッピング関数を設計した．

4 評価実験

4.1 実験条件

本稿では，図 4 に示す防音室において評価実験を行い，提案手法の有効性を検証した．表 1 に示すデータ収録条件を示す．背景雑音は 46.1 dBA，室内残響は 0.44 sec ($T_{[60]}$) である．マイクロホンアレーに対して距離 1.0 m，10° 方向に設置したスピーカからテストデータを再生し，サンプリング周波数 16 kHz にて収録を行った．テストデータには ATR 音素バランス 216 単語 [11] (女性話者 3 名，男性話者 3 名) を使用した．

上記の環境において，信号対雑音比 (SNR : Signal to Noise Ratio) を $0, \sim, 20$ dB まで変化させた場合の提案手法と従来手法の発話区間検出性能を比較した．ゼロ交差閾値は提案手法，従来手法ともに 60 回とした．なお，従来法における振幅閾値は初期値を 2000 とし，更新係数 0.0 (2000 固定) と 0.2 の 2 条件で更新した．

発話区間検出性能は，以下の式による再現率，適合率と F 値によって評価した．

$$\text{再現率} = \frac{\text{正検出したフレーム数}}{\text{正解発話フレーム総数}}, \tag{9}$$

$$\text{適合率} = \frac{\text{正検出したフレーム数}}{\text{検出したフレーム総数}}, \tag{10}$$

$$\text{F 値} = \frac{2 \times \text{再現率} \times \text{適合率}}{\text{再現率} + \text{適合率}}. \tag{11}$$

再現率，適合率と F 値による評価は，フレーム単位と区間単位で行った．発話区間は，発話フレームの前後にマージンを確保することで検出した．これは，ZCD 法は主に母音区間を検出するため，検出された発話フレームの前後にマージンを確保することで，文頭や文末などの子音区間を発話区間に含めることができるためであることや，音声認識を行うためには発話区間の前後に非音声区間が含まれている必要があるためである．本稿では前方に 0.3 sec，後方に 0.4 sec のマージンを確保した．

表 1: 収録条件

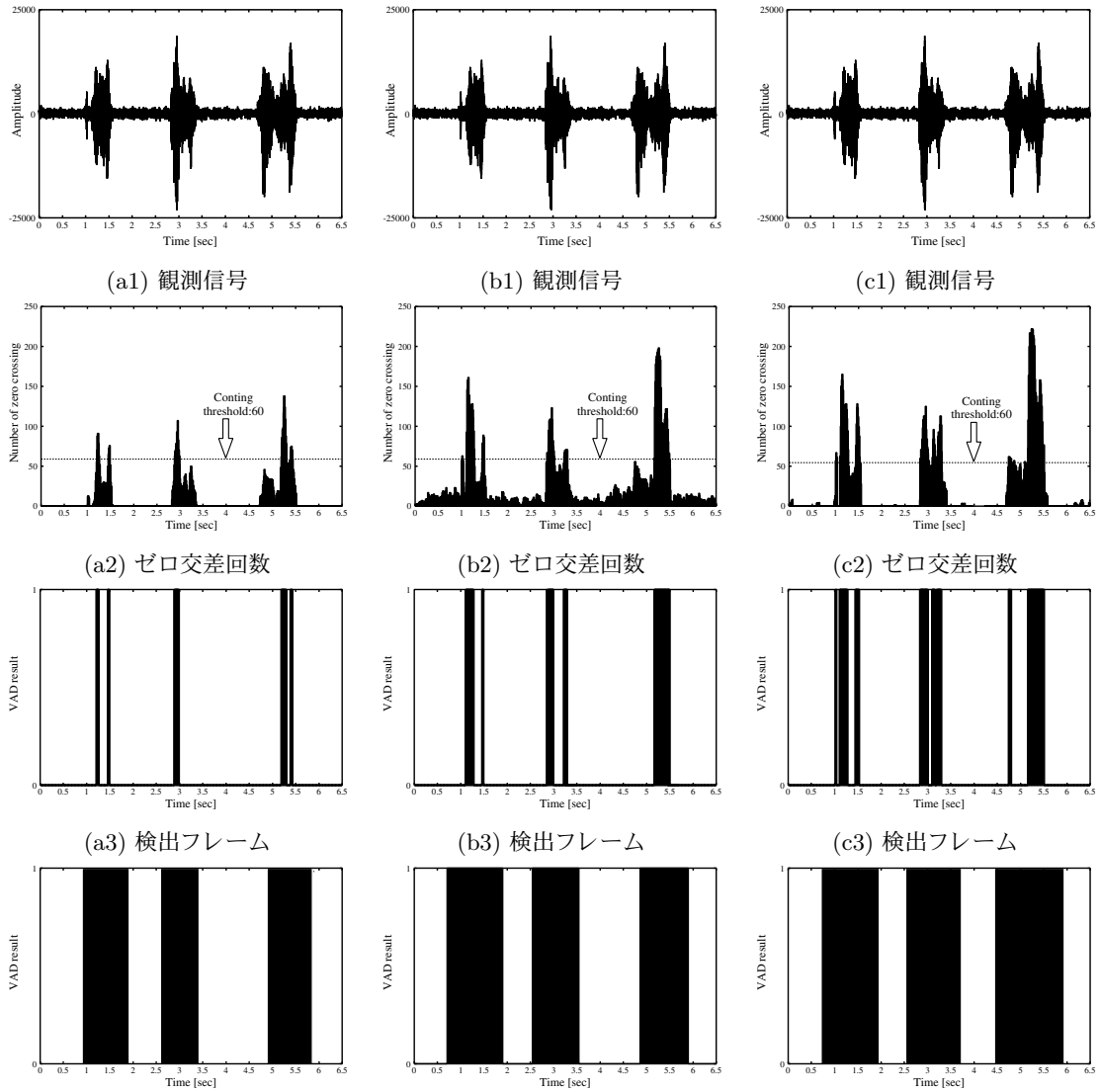
マイクロホンアレー	2 素子, 148.75 mm 間隔
サンプリング周波数	16 kHz
室内残響 $T_{[60]}$	0.44 sec
室内騒音	46.1 dBA
SNR(信号対雑音比)	0 dB，～，20 dB，clean

表 2: 実験条件

テストデータ	
音声	ATR 音素バランス 216 単語 (女性話者 3 名，男性話者 3 名)
雑音	空調雑音，PC 雑音 (定常雑音)
発話フレーム検出	
ZCD 法	フレーム長：25 msec
	フレーム周期：10 msec
WCSP 法	フレーム長：64 msec
	フレーム周期：40 msec
ゼロ交差閾値	60
振幅閾値	初期値：2000
	更新係数：0.0，0.2
発話区間検出	
前方マージン	0.3 sec
後方マージン	0.4 sec

4.2 実験結果

図 5 に SNR15 dB における発話区間検出結果を示す．図 5(a) は従来手法 (更新係数 0.0) の結果を，図 5(b) は従来手法 (更新係数 0.2) の結果を，図 5(c) は提案手法の結果を表す．また，図 5(a1)(b1)(c1) は観測信号の波形を，図



(a1) 観測信号　(b1) 観測信号　(c1) 観測信号

(a2) ゼロ交差回数　(b2) ゼロ交差回数　(c2) ゼロ交差回数

(a3) 検出フレーム　(b3) 検出フレーム　(c3) 検出フレーム

(a4) 検出区間
従来手法 (更新係数 0.0)

(b4) 検出区間
従来手法 (更新係数 0.2)

(c4) 検出区間
提案手法

図 5: 発話区間検出の実験結果

5(a2)(b2)(c2) はゼロ交差回数を，図 5(a3)(b3)(c3) は音声フレームとして検出されたフレームを，図 5(a4)(b4)(c4) は音声区間として検出された区間を表す．図 5(a2)(b2) より，時間情報のみに基づく従来手法では，振幅閾値を更新することによって発話区間におけるゼロ交差検出性能が改善されていることが確認できる．しかし，非発話区間におけるゼロ交差検出回数も同時に増加しているため，突発雑音などが生じた場合に性能が低下することが考えられる．次に図 5(b2)(c2) より，時間 / 空間情報に基づく提案手法では，発話区間における頑健なゼロ交差検出と，非発話区間におけるゼロ交差の非検出を両立できていることが確認できる．最後に図 5(a3)(b3)(c3)(a4)(b4)(c4) より，提案手法は従来手法よりも正確に音声フレーム / 区間を検出できていることが確認できる．

次に，表 3 に再現率に基づく評価結果を，表 4 に F 値に基づく評価結果を示す．表 3，表 4 の “従来手法 (1)” は従来手法 (更新係数 0.0) の実験結果を，“従来手法 (2)” は従来手法 (更新係数 0.2) の実験結果を，“提案手法” は提案手法の実験結果ををを表す．なお適合率については，SNR，VAD 手法に依存せず全実験条件下で 100%であったため表記していない．これは，非音声区間を音声区間と誤検出したフレームや区間がなかったことを意味する．原因としては，今回使用した雑音が空調機雑音や PC ファンノイズといった定常雑音であったためと考えられる．表 3，表 4 より，再現率 / F 値を区間単位で評価することによって，性能が改善される傾向があることが確認できる．これは，検出されたフレームの前後にマージンを確保することで，検出されなかった文頭や文末などの子音区間などを発話区間として検出することができたためめであると考えられる．次に，従来手法と提案手法を比較した場合，例えば SNR0 dB に

表 3: 再現率に基づく評価結果

フレーム単位					
SNR [dB]	0	5	10	15	20
従来手法 (1)	4.1	6.8	8.3	8.8	9.0
従来手法 (2)	2.9	7.4	11.8	15.0	16.8
提案手法	9.3	15.9	20.4	23.0	24.1
区間単位					
SNR [dB]	0	5	10	15	20
従来手法 (1)	50.9	65.6	71.6	72.0	72.2
従来手法 (2)	38.9	66.4	78.9	83.6	86.0
提案手法	79.9	89.4	93.1	93.9	94.7

表 4: F 値に基づく評価結果

フレーム単位					
SNR [dB]	0	5	10	15	20
従来手法 (1)	7.9	12.8	15.3	16.2	16.5
従来手法 (2)	5.6	13.8	21.1	26.1	28.8
提案手法	17.0	27.4	33.9	37.4	38.8
区間単位					
SNR [dB]	0	5	10	15	20
従来手法 (1)	67.0	79.2	83.5	83.7	83.8
従来手法 (2)	56.0	79.8	88.2	91.1	92.0
提案手法	87.5	94.4	96.4	96.8	97.3

おける区間単位の F 値が 20 程度，SNR20 dB における区間単位の F 値が 14 程度改善されていることが確認できる．従って提案手法によって，音声フレーム，音声区間検出性能を大幅に改善することができた．

以上の評価実験結果より，提案する時間 / 空間領域 VAD 法は従来の時間領域 VAD 法よりも正確に発話区間を検出できることが確認できた．従って，提案手法によってハンズフリー環境環境下における雑音に頑健なハンズフリー発話区間検出を実現できる．

5 まとめ

本稿では，マイクロホンアレーを用いたハンズフリー音声認識を実現するために必要不可欠な発話区間検出について検討を行い，時間情報に基づいた ZCD 法と空間情報 (発話者の位置情報) に基づいた WCSP 法を統合した時間 / 空間領域 VAD 法を提案した．実騒音環境下における評価実験の結果，提案手法はハンズフリー環境において，従来手法よりも正確に発話区間を検出できることが確認できた．今後の課題として，WCSP 係数の振幅情報のみでなく，推定した話者位置情報も利用することを検討する．また，マッピング関数の環境適応についても検討を行う予定である．

謝辞

本研究の一部は，文科省リーディングプロジェクト e-Society および科研費 17700216 と 17200014 による研究助成を受けた．

参考文献

[1] J. L. Flanagan, et al., “Computer-steered microphone arrays for sound transduction in large rooms,” J. Acoust. Soc. Am., vol.78, no.5, pp.1508–1518, Nov. 1985.

[2] 中村 哲, “音声認識系へのマイクロホンアレーの応用,” 音講論, vol.I, pp.515–518, 1998.

[3] Y. Denda, et al., “Robust talker direction estimation based on weighted CSP Analysis and maximum likelihood estimation,” IEICE Trans. on Inform. and Sys., vol.E89-D, no.3, pp.1050–1057, Mar. 2006.

[4] Y. Kaneda, et al.,, “Adaptive microphoned-array system for noise reduction,” IEEE Trans. ASSP, vol. 34, no. 6, pp. 1391–1400, Dec. 1986.

[5] S.U. Pillai, “Array signal processing,” Springer-Verlag, New York, 1989.

[6] R.P. Venkatesha, et al., Gaurav, “Comparison of voice activity detection algorithms for voip,” Proc.ISCC02, pp.530-535, 2002.

[7] P.N. Garner, et al., “A differential spectral voice activity detector,” Proc. ICASSP04. vol. 1, pp. 597–600, 2004.

[8] A. Lee, et al., “Noiser robust real world spokne dialog system using GMM based rejection of unintended inputs,” Proc.ICSLP04, vol.1, pp.173–176, 2004.

[9] Y. Kida, et al., “Voice activity detection based on optimally weighted combination of multiple features,” Proc.EUROSPEECH05, pp.2621–2624, 2006.

[10] T. Kawahara, et al., “Japanese dictation toolkit,” J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.233–239, May. 1999.

[11] K. Takeda, et al., “Acoustic-phonetic labels in a Japanese speech database,” Proc.EUROSPEECH87, vol.2, pp.13–16, 1987.