社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ＨＩ－122 (20)
2007－SLP－ 65 (20)
2007／2／10

話者方位情報とゼロ交差情報に基づく ハンズフリー発話区間検出の評価

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（Voice Activity Detection : VAD）は必要不可欠である．ゼロ交差情報などの時間特徴量に基づいた時間領域 VAD 法は，雑音によって歪みを受けた遠隔発話に対して十分な性能を得られないという問題がある．また，話者方位情報などの空間特徴量に基づいた空間領域 VAD 法は，指向性雑音環境下で大きく性能が劣化するという問題がある．本稿ではこれらの問題を解決するために，時間領域 VAD 法と空間領域 VAD 法を統合することを検討し，話者方位情報とゼロ交差情報に基づいた雑音に頑健な時間-空間領域ハンズフリー VAD 法を提案する．提案手法は，音声の到来方向推定に特化した WCSP (Weighted Cross-power Spectrum Phase) 法によって空間安定度と空間信頼度を抽出する．そして，抽出した空間特徴量に基づく適応型ゼロ交差検出法によって発話区間を頑健に検出する．実オフィス環境における評価実験の結果，提案手法は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, 適応型ゼロ交差検出, 話者方位推定, Weighted CSP 法, ハンズフリー音声認識

An Evaluation of Hands-free Voice Activity Detection Based on Talker Direction and Zero Crossing Detection

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

† Graduate School of Science and Engineering, Ritsumeikan University

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

‡ College of Information Science and Engineering, 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 indispensable for hands-free speech recognition. Time domain VAD algorithms based on time domain features such as zero crossing information cannot perform satisfactory VAD performance against distant-talking noisy speech. In addition, spatial domain VAD algorithms based on spatial domain features such as talker direction information provides degraded VAD performance due to directional interferences. To overcome these problems, in this paper, we study to integrate the time domain VAD algorithm and the spatial domain VAD algorithm; therefore, we propose the noise robust time-spatial domain VAD algorithm based on talker direction information and zero crossing information. The proposed algorithm firstly extracts two spatial features: spatial reliability and spatial stability, based on WCSP (Weighted Cross-power Spectrum Phase) analysis. Then, adaptive zero crossing detection based on extracted spatial features robustly detects voice activity frame. As a result of evaluation experiments in an actual office room, we confirmed that the performance of the proposed VAD algorithm is superior to that of the conventional VAD algorithms.

Key words Hands-free voice activity detection, Adaptive zero crossing detection, Talker Direction Estimation, Weighted CSP analysis, Hands-free speech recognition.

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で発話された音声（遠隔発話）は残響や背景雑音の影響により大きく歪んでしまい，遠隔発話の音声認識率が著しく低下するという問題がある．この問題を解決するために，マイクロホンアレー [1, 2, 3] 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 [4]．これらの研究では，ビームフォーミングを用いて発話者の方向に形成した指向特性によって目的音声と雑音を空間的に分離し，音声を高音質に受音することで音声認識率の改善を図っている．

マイクロホンアレーを用いた高品質音声受音を実現するためには以下の技術が必要不可欠である．**1.** 発話区間検出（Voice Activity Detection : VAD）．発話の有無を検出し，受音信号から発話区間を切り出す．**2.** 話者方位推定（Talker Direction Estimation : TDE）．VAD によって発話が検出された場合，音声の到来方位を推定することで話者方位を求める．**3.** ビームフォーミング．話者方位情報と音声/非音声区間の情報に基づいて指向特性を形成し，音声を高品質に受音する．我々はこれまでに，音声の到来方向推定に特化した WCSP（Weighted Cross-power Spectrum Phase）法に基づいた頑健な話者方位推定法を提案している [5]．従って，本稿では雑音に頑健なハンズフリー VAD の検討を行う．発話区間を正確に検出できない場合，話者方位推定性能が低下するばかりでなく，指向特性のミスマッチによる音声歪みや音声認識誤りの増加といった問題が発生する．

従来の VAD 法は，受音信号から抽出した音声特徴量に基づいて音声/非音声の識別を行う手法が主であり，シングル/マルチチャネル受音系における手法に大別できる．シングルチャネル系の VAD 法としては，ゼロ交差 [6]，振幅（パワー）レベル [6] や Kurtosis[7] といった時間特徴量に基づく時間領域 VAD 法や，スペクトル情報 [8]，音声 GMM（Gaussian Mixture Model）の尤度 [9] といった周波数特徴量に基づいた周波数領域 VAD 法が提案されている．しかし，従来の時間/周波数領域 VAD 法の発話区間検出性能は雑音によって歪みを受けた遠隔発話に対して低下するという問題がある．一方，マイクロホンアレーなどを用いたマルチチャネル系では，時間/周波数特徴量のみでなく，話者方位情報や音源の指向特性などの空間特徴量を得ることができる．文献 [10, 11] では，話者方位情報に基づいた空間領域 VAD 法が提案されており，高い性能を得られることが確認されている．しかし，話者方位情報は本質的な音声特徴量ではないため，空間的に相関の高い雑音が到来する指向性雑音環境下では，発話区間検出が困難になるという問題がある．また，ビームフォーミングによってあらかじめ SN 比を改善した音声から抽出した時間/周波数特徴量に基づく VAD 法が提案されている [12, 13]．これらの手法では，ビームフォーミングによる指向特性形成の際に話者方位情報が利用されているが，発話区間検出においては時間/周波数特徴量が利用されており，空間特徴量を利用することでさらに性能が改善すると考えられる．

我々はこれらの問題を解決するために，時間特徴量のみでなく空間特徴量も積極的に利用するハンズフリー発話区間検出法を提案している [14]．この手法は，WCSP 法によって推定した空間的な相関値の最大値（空間信頼度）に基づいた適応型ゼロ交差検出法によって発話区間を検出している．しかし，空間特徴量として空間信頼度のみを用いているため，突発的な指向性雑音が生じた場合に誤検出が起こるという問題がある．そこで本稿では，提案手法の性能を改善するために，空間信頼度のみでなく話者方位推定の時間方向に対する評価尺度として空間安定度を導入することを検討する．

2 従来の VAD 法

2.1 ゼロ交差情報に基づく時間領域 VAD 法

音声は短時間において，定常的に振動（ゼロ交差）を繰り返す信号であることが知られている．ゼロ交差検出（Zero Crossing Detection : ZCD）法はこの性質を利用した時間領域 VAD 法である [6]．ZCD は振幅閾値に基づいた式 (1) によって短時間フレーム内の信号のゼロ交差を検出し，式 (2) によって音声/非音声識別を行うことで発話区間を検出する．

$$ZCR = ZCD\big(x(t), TH_a\big), \tag{1}$$

$$\text{VAD} = \begin{cases} \text{Speech frame}, & ZCR \geq TH_z \\ \text{Non-speech frame} & ZCR < TH_z \end{cases}, \tag{2}$$

ここで，$x(t)$ はフレーム長 T の受音信号を，TH_a は振幅閾値を，ZCR は短時間フレーム内のゼロ交差回数を，関数 $ZCD\big(x(t), TH_a\big)$ はゼロ交差回数を返す関数を，TH_z はゼロ交差回数閾値を表す．また，関数 $ZCD\big(x(t), TH_a\big)$ 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. ・$x(t)$ と $x(t+1)$ が異符号なら 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[15] などにおいて利用されている．しかし，高雑音環境下で性能が低下するという問題がある．

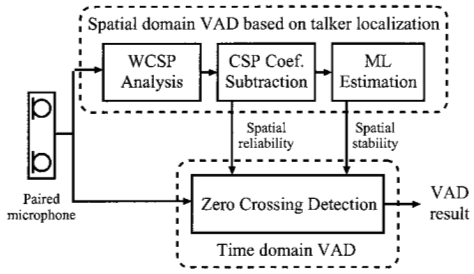


図 1: 提案手法の概要

2.2 話者方位情報に基づく空間領域 VAD 法

発話者がマイクロホンアレーに向かって発話している場合，空間的に相関の高い（指向性の高い）音源が存在することになる．文献 [10, 11] では，CSP 法を用いて空間的に相関の高い音源を検出することで VAD を行う手法が提案されている．今，マイクロホンペア M_1, M_2 で信号 $x_1(t), x_2(t)$ を受音した場合，CSP 法は以下の式で表すことができる．

$$CSP(k) = \mathrm{IDFT}\left[\frac{X_1(\omega)X_2(\omega)^*}{|X_1(\omega)||X_2(\omega)|}\right], \quad (3)$$

$$[r, \tau] = f_{max}\big(CSP(k)\big), \quad (4)$$

$$\theta = cos^{-1}\left(\frac{c\tau}{dF_s}\right), \quad (5)$$

ここで，$CSP(k)$ は CSP 係数を，IDFT$[\cdot]$ は逆離散フーリエ変換（Inverse Discrete Fourier Transform : IDFT）を，$X_{[\cdot]}(\omega)$ は信号 $x_{[\cdot]}(t)$ の周波数表現を，* は複素共役を，r は CSP 係数の最大値を，τ は到来時間差を，θ は τ に対応する到来方位を，f_{max} は r と τ を返す関数を，c は音速を，d はマイクロホン間隔を，F_s はサンプリング周波数を表す．式 (3) より，CSP 係数は受音信号の正規化相互相関に基づいた空間的な相関（空間相関情報）を表す．このため，マイクロホンアレーに音声が到来している場合は CSP 係数の最大値 r の値が大きくなるため，発話区間を検出することができる．しかし，指向性雑音環境下では発話の有無に関わらず r が大きくなるため，発話区間検出性能が低下してしまうという問題がある．

3 提案手法

提案する話者方位情報とゼロ交差情報に基づく時間-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，WCSP（Weighed Cross-power Spectrum Phase）法に基づいて話者方位を推定し，空間信頼度と空間安定度の 2 つの空間特徴量を抽出する．そして，空間特徴量に基づいて振幅閾値とゼロ交差回数閾値を制御する適応型ゼロ交差検出法によって頑健に発話区間を検出する．

3.1 WCSP 法

式 (3) より，CSP 法では正規化相互相関を全周波数において求めている．しかし，音声の周波数特性を考慮することで CSP 法の性能をさらに性能を改善することができると考えられる．我々はこれまでに，音声の周波数特性を考慮した WCSP 法を提案している [5]．WCSP 法は以下の式で表すことができる．

$$WCSP(k) = \mathrm{IDFT}\left[W(\omega)\frac{X_1(\omega)X_2(\omega)^*}{|X_1(\omega)||X_2(\omega)|}\right], \quad (6)$$

ここで，$WCSP(k)$ は WCSP 係数を，$W(\omega)$ は音声信号の平均スペクトル特性に基づいた重み係数を表す．式 (6) より，正規化相互相関に音声の平均スペクトル特性 $W(\omega)$ が重み付けされるため，WCSP 法は音声の到来方位推定に特化した手法として解釈することができる．

3.2 CSP 係数サブトラクション

CSP 係数サブトラクションは，空間的に定常な指向性雑音環境においても頑健に話者方位推定を行える手法である．[5]．CSP 係数サブトラクションでは，あらかじめ非発話区間において定常指向性音雑音の空間的な分布を表す雑音 CSP 係数を式 (7) によって学習する．次に，式 (6) で計算した WCSP 係数から雑音 WCSP 係数を式 (8)(9) によって減算することで，雑音に頑健な音声 WCSP 係数を求める．そして，空間信頼度，到来時間差と到来方位を式 (10)(11) によって推定する．

$$WCSP_n(k) = \frac{\sum_{n=1}^{N}\max\big(WCSP(n,k),\ 0\big)}{N}, \quad (7)$$

$$WCSP_{\hat{s}}(i,k) = WCSP(i,k) - \alpha \cdot WCSP_n(k), \quad (8)$$

$$\alpha = \frac{\max\big(WCSP(i,k)\big)}{\max\big(WCSP_n(k)\big)}, \quad (9)$$

$$[r(i), \tau(i)] = f_{max}\big(WCSP_{\hat{s}}(i,k)\big), \quad (10)$$

$$\theta(i) = cos^{-1}\left(\frac{c\tau(i)}{dF_s}\right), \quad (11)$$

ここで，n は非発話区間におけるフレームインデックスを，$WCSP_n(k)$ は雑音 WCSP 係数を，i は発話区間検出時におけるフレームインデックスを，$WCSP_{\hat{s}}(i,k)$ は音声 WCSP 係数を，α はサブトラクション係数を，$r(i)$ は空間信頼度を，$\tau(i)$ は到来時間差を，$\theta(i)$ は到来方位を表す．

ここで，発話者が 100 度方向に存在してる状況において，音声/非音声区間でそれぞれ得られた WCSP 係数の一例を図 2 に示す．図 2 より，発話区間においては空間信頼度が大きな値を持つこと，非発話区間においては小さな値を持つことが確認できる．

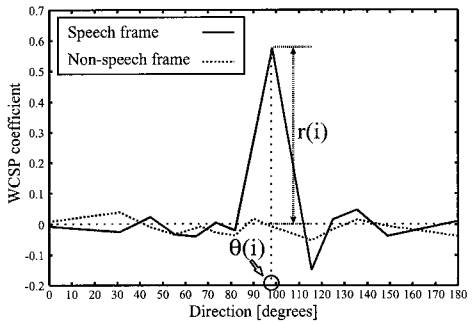


図 2: WCSP 係数の一例

3.3 話者方位最尤推定

突発的な指向性雑音に対しては，その空間的分布をあらかじめ学習することは非常に困難であり，CSP 係数サブトラクションの有効性が低下するという問題がある．この問題を解決するために，式 (10) で推定した到来時間差の観測系列 $\boldsymbol{\tau} = [\tau(1), \cdots, \tau(I)]$ に基づいて話者方位の最尤推定を行い，突発的な推定誤差の影響を減少させる．今，到来時間差の観測系列が真の到来時間差系列と観測誤差系列の和で表せると仮定する．

$$\boldsymbol{\tau} = [\tau(1), \cdots, \tau(I)] = \boldsymbol{\tau}_s(\theta_s) + \boldsymbol{n}, \tag{12}$$

ここで，θ_s は真の話者方位を，$\boldsymbol{\tau}_s(\theta_s)$ は真の到来時間差系列を，$\boldsymbol{n}$ は観測誤差系列を表す．さらに，観測誤差系列が平均 $\mathbf{0}$，分散 $\boldsymbol{R}$ の正規分布と仮定すると，観測到来時間差系列が観測される確率は次式で表せる．

$$P[\boldsymbol{\tau}|\theta_s] = \frac{e^{-\frac{1}{2}[\boldsymbol{\tau}-\boldsymbol{\tau}_s(\theta_s)]^T \boldsymbol{R}^{-1}[\boldsymbol{\tau}-\boldsymbol{\tau}_s(\theta_s)]}}{2\pi^{\frac{I}{2}}|\boldsymbol{R}|^{\frac{1}{2}}}, \tag{13}$$

$$H = [\boldsymbol{\tau} - \boldsymbol{\tau}_s(\theta_s)]^T \boldsymbol{R}^{-1}[\boldsymbol{\tau} - \boldsymbol{\tau}_s(\theta_s)]. \tag{14}$$

ここで，$[\cdot]^T$ は転置ベクトルを，$[\cdot]^{-1}$ は逆行列を表す．従って，式 (13) の観測確率を最大にする，つまり，式 (14) を最小にする話者方位 $\hat{\theta}_s$ が最尤話者方位となる．本稿では，勾配法によって最尤話者方位 $\hat{\theta}_s$ を推定した [5]．最後に，最尤話者方位と各フレームの推定話者方位との誤差を空間安定度 $s(i)$ として定義する．

$$\boldsymbol{s} = [s(i), \cdots, s(I)] = |\hat{\theta}_s - \theta(i)|, \tag{15}$$

図 3 に空間安定度の一例を示す．図 3 より，発話区間においては安定した方位推定が行えているため，空間安定度が小さな値を持つことが分かる．

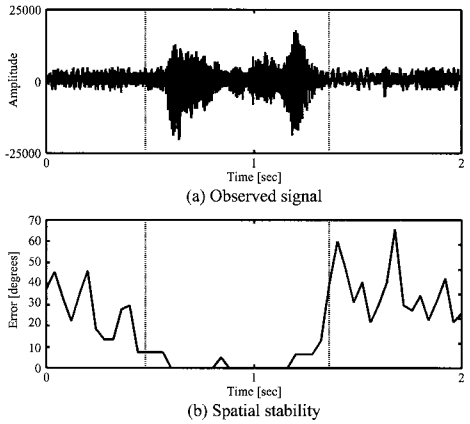


図 3: 話者方位推定の空間安定度

3.4 空間特徴量に基づく適応型ゼロ交差検出

提案手法では，空間信頼度と空間安定度に基づいた適応型ゼロ交差検出法によって発話区間を検出する．

$$TH_a = \begin{cases} TH_a(low) & r(i) \geq \bar{r}_n. \\ \bar{x}_n(t) & r(i) < \bar{r}_n. \end{cases}, \tag{16}$$

$$TH_z = \begin{cases} 20 & s(i) \leq \varepsilon. \\ 60 & s(i) > \varepsilon. \end{cases}, \tag{17}$$

$$ZCR = ZCD\big(x(t), TH_a\big), \tag{18}$$

$$\text{VAD} = \begin{cases} \text{Speech frame,} & ZCR \geq TH_z \\ \text{Non-speech frame} & ZCR < TH_z \end{cases}, \tag{19}$$

ここで，式 (16) は振幅閾値の決定式を，$\bar{x}_n(t)$ は非発話区間における雑音平均振幅を，$TH_a(low)$ は $\bar{x}_n(t)$ より小さい振幅閾値を設定することを，$\bar{r}_n$ は非発話区間における雑音平均空間信頼度を，式 (17) はゼロ交差回数閾値の決定式を，ε は空間安定度の許容誤差を表す．式 (16) より，空間信頼度が高い場合は，振幅閾値を小さくすることでゼロ交差を頑健に検出できる．また式 (17) より，空間安定度が大きい場合はゼロ交差回数閾値を大きくすることで発話区間の誤検出を防ぐことができる．

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．評価用テスト音声データには ATR 音素バランス 216 単語（女性話者 3 名，男性話者 3 名）を使用した．テスト

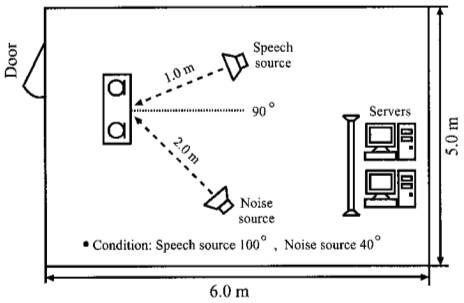


図 4: 実験環境

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，10 dB

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 3 名，男性 3 名)
	100°，1.0 m
定常無指向性雑音	サーバ音
定常指向性雑音	サーバ音 40°, 2.0 m
突発指向性雑音	拍手 40°，2.0 m
発話区間検出	
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：20 msec
最尤推定	224 msec
	（WCSP 法 10 フレーム）
空間安定度許容誤差	20°

データはマイクロホンアレーに対して距離 1.0 m，100° 方向に設置したスピーカから再生した．次に，無指向性雑音環境をシミュレートするために，部屋の四隅の壁に向けて設置したスピーカからサーバ音を再生した．また，定常指向性雑音および突発指向性雑音環境をシミュレートするために，マイクロホンアレーに対して距離 2.0 m，40° 方向にスピーカを設置し，定常指向性雑音としてサーバ音を，突発性指向性雑音として拍手の音を再生した．

上記の環境において，信号対雑音比（Signal to Noise Ratio : SNR）を 0, 10 dB と変化させ，ゼロ交差検出法 [6], CSP 法 [11] と提案手法の発話区間検出性能を比較した．評価は非音声フレームを音声フレームとして誤検出した割合を表す FAR（False Acceptance Rate）と音声フレームを非音声フレームとして誤棄却した割合を表す FRR（False Rejection Rate）に基づく ROC 曲線（Reciever Operating Curve）によって行った．ゼロ交差検出法では，ゼロ交差回数閾値を 20 に固定し，振幅閾値を変化させることで ROC 曲線をプロットした．CSP 法では，空間信頼度に対する閾値を変化させることで ROC 曲線をプロットした．提案手法では，$TH_a(low)$ を最大 $\overline{x}_n(t)$ まで変化させることで ROC 曲線をプロットした．また，式 (7) の雑音 WCSP 係数，式 (16) の雑音平均振幅と雑音平均空間信頼度は，入力信号の先頭 300msec には音声が存在しないと仮定し，その区間において学習した．なお，突発指向性雑音は先頭 300msec には含まれないものとする．

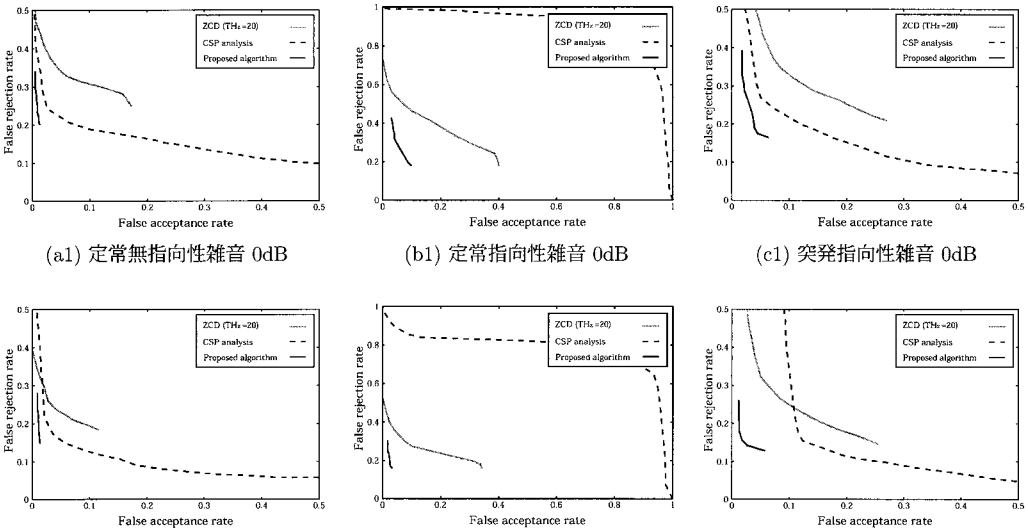
4.2 実験結果

図 5 に発話区間検出実験の結果を示す．図 5(a) は定常無指向性雑音環境における ROC 曲線を，図 5(b) は定常指向性雑音環境における ROC 曲線を，図 5(c) は突発指向性雑音環境における ROC 曲線を表す．なお，定常指向性雑音環境における ROC 曲線は，CSP 法が他の提案手法の ROC 曲線と大きく離れているため，定常無指向性と突発指向性雑音環境における ROC 曲線とは異なるスケールで描画している．

まず最初に，全体的な傾向として提案手法は CSP 法と比較して FAR が低く FRR が高い傾向にあることが確認できる．これは，音声のゼロ交差が雑音によって歪みを受けてしまい正確に検出できなくなっているためであると考えられる．次に，図 5(a) の定常無指向性雑音環境において，CSP 法の性能は SNR の変化の影響をあまり受けないことが確認できる．これは，無指向性雑音はパワーが大きくなっても空間的な相関はそれほど高くならないためであると考えられる．また，提案手法は平均 FRR が 0.3 程度と CSP 法の平均 FRR0.2 程度よりも高いが，これは子音などのゼロ交差を検出できていないためであると考えられる．しかし，FAR はほぼ 0 となっており雑音の誤検出割合を大きく減少できている．最後に，図 5(b)(c) の定常指向性雑音環境と突発指向性雑音環境において，CSP 法の性能が大きく劣化していることが確認できる．それに対して提案手法は，定常無指向性雑音環境下と同様，平均 FRR0.2 程度，平均 FAR0.05 と高い検出率を達成できている．

5 まとめ

本稿では，話者方位情報とゼロ交差情報にに基づいた時間-空間領域ハンズフリー VAD 法を提案した．実オフィス



(a1) 定常無指向性雑音 0dB　(b1) 定常指向性雑音 0dB　(c1) 突発指向性雑音 0dB

(a2) 定常無指向性雑音 10dB　(b2) 定常指向性雑音 10dB　(c2) 突発指向性雑音 10dB

図 5: FAR と FRR による ROC 曲線

環境における評価実験の結果，提案手法は従来の時間/空間領域 VAD 法よりも高い発話区間検出性能を得ることが確認できた．今後の課題として，最小発話継続時間や空間安定度の安定継続時間などの指標に基づいた Hang-over 処理について検討する．

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, 1985.

[2] L.J. Griffiths, et al., "An alternative approach to linearyly constrained adaptive beamforming," IEEE Trans. AP, vol.AP-30, pp.27–34, 1982.

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

[4] 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（秋）, vol.I, pp.515–518, 1998.

[5] 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, 2006.

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

[7] J.M. Gorriz, et al., "An improved voice activity detection using higher order statistics," IEEE Trans. SAP, vol.SAP-13, no.5, pp.965–874, 2005.

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

[9] 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.

[10] 藤本雅清 他, "マイクロフォンアレイと 2 段階 MLLR 適応による実環境下ハンズフリー音声認識 –対話型テレビのフロントエンドシステムの構築–," 音講論（秋）, vol.I, pp.69–70, 2002.

[11] L. Armani, et al., "User of a CSP-based voice actitvity detector for distant-talking ASR," Proc. Eurospeech03, pp.501–504, 2004.

[12] 金田豊, "マイクロホンアレーを用いた雑音下での音声区間検出", 信学論, vol.J73-A, no.8, pp.1391–1398, 1990.

[13] M.W. Hoffman, et al., "GSC-based spatial voice activity detection for enhanced speech coding in the presence of competing speech," IEEE Trans. SAP, vol.SAP-9, no.2, pp.175–179, 2001.

[14] 傳田遊亀 他, "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," 情報処理学会研究報告, 2006-SLP-62, pp.7-12, 2006

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