話者方位推定に基づくリアルタイム発話区間検出システムの開発

西浦 敬信[1]，傳田 遊亀[2]，猿渡 洋[3]，鹿野 清宏[3]
立命館大（情報理工）[1]，立命館大院（理工）[2]，奈良先端大院（情報科学）[3]

1 はじめに

受音信号から音声が存在する区間を検出する発話区間検出（Voice Activity Detection : VAD）は，雑音下音声認識を実現するために必要不可欠な技術である．しかし，ゼロ交差検出（Zero Crossing Detection : ZCD）法 [1] や話者方位推定に基づく VAD 法 [2] などの従来手法には，高雑音環境で性能が低下してしまうという問題があった．本稿では雑音に頑健な VAD を実現するために，筆者らが提案している話者方位推定法 [3] に基づく適応 ZCD 法を提案する．さらに，提案する適応 ZCD 法を用いて，PC 上で実時間動作可能な発話区間検出システムを開発する．以下では，適応 ZCD 法と発話区間検出システムの詳細について述べる．

2 適応 ZCD 法

適応 ZCD 法の概要を図 1 に示す．適応 ZCD 法ではまず，WCSP（Weighed Cross-power Spectrum Phase）法 [3] と時系列最尤推定 [3] によって話者方位を推定し，2 つの空間特徴量：空間信頼度と空間分散度を抽出する．そして，空間特徴量に基づいて VAD 閾値を動的に制御し，音声のゼロ交差を頑健に検出する．

2.1 WCSP 法

ペアマイクロホン M_1, M_2 で受音した信号 $x_m(t)$, $m = 1, 2$ に対する WCSP 法は以下の式で表せる．

$$WCSP(k) = \mathrm{IDFT}\left[W(\omega)\frac{X_1(\omega)X_2(\omega)^*}{|X_1(\omega)||X_2(\omega)|}\right], \quad (1)$$

ここで，$WCSP(k)$ は WCSP 係数を，IDFT[·] は逆フーリエ変換を，$W(\omega)$ は音声の平均スペクトルに基づく重み係数を，$X_m(\omega)$ は $x_m(t)$ の複素スペクトルを，* は複素共役を表す．WCSP 法は音声のスペクトル特性に基づいて，各周波数の白色化相互相関に信頼度を付与することで音声に特化した方位推定を行う．

2.2 CSP 係数サブトラクション

CSP 係数サブトラクションは，音響空間に存在する定常雑音を空間的に抑圧することで雑音に頑健な方位推定を行う．CSP 係数サブトラクションではまず，式 (2) によって定常雑音の空間分布を表す雑音分布 CSP 係数 $WCSP_{\bar{n}}(k)$ を N フレーム分の非発話区間で学習する．そして，式 (3)(4) によって WCSP 係数から雑音分布 WCSP 係数を減算し，音声分布 WCSP 係数 $WCSP_{\hat{s}}(k)$ を求める．最後に，式 (5) によって到来方位 θ と到来時間差 τ_θ を推定し，さらに音声分布 WCSP 係数の最大値を空間信頼度 r_{sp} として抽出する．

$$WCSP_{\bar{n}}(k) = \frac{\sum_{n=1}^{N}\max\left(WCSP(n,k),\ 0\right)}{N}, \quad (2)$$

$$WCSP_{\hat{s}}(k) = WCSP(k) - \alpha WCSP_{\bar{n}}(k), \quad (3)$$

$$\alpha = \frac{\max_k\left(WCSP(k)\right)}{\max_k\left(WCSP_{\bar{n}}(k)\right)}, \quad (4)$$

$$[\theta, \tau_\theta, r_{sp}] = f_{max}\left(WCSP_{\hat{s}}(k)\right). \quad (5)$$

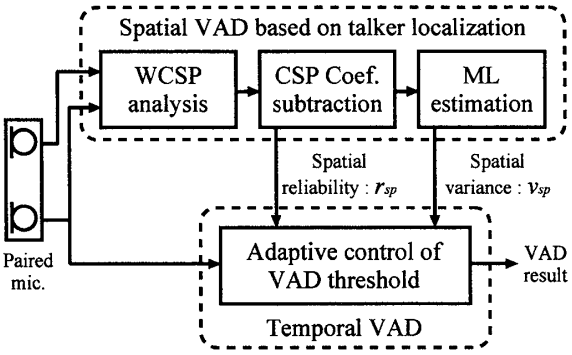


図 1: 適応 ZCD 法の概要

2.3 時系列最尤推定

過去 I フレーム（$i = 1, \cdots, I$）で推定した話者方位の系列 $\boldsymbol{\tau}_\theta = [\tau(1) \cdots \tau(I)]$ から最尤話者方位を推定し，突発的な方位推定誤りの影響を低減する．ここで，式 (6) に示すように過去 I フレームの到来時間差系列は，真の話者方位 θ_s に依存する真の到来時間差系列 $\boldsymbol{\tau}_{\theta_s}$ と平均 $\mathbf{0}$，分散 $\boldsymbol{R}$ の正規分布に従う観測誤差系列 $\boldsymbol{n}$ の和で表せると仮定する．この場合，真の話者方位から推定到来時間差系列が生起される確率，すなわち式 (7) を最大にする方位 $\hat{\theta}_s$ が最尤話者方位となる．

$$\boldsymbol{\tau}_\theta = [\tau_\theta(1) \cdots \tau_\theta(I)] = \boldsymbol{\tau}_{\theta_s} + \boldsymbol{n}, \quad (6)$$

$$P[\boldsymbol{\tau}_\theta|\theta_s] = \frac{e^{-\frac{1}{2}[\boldsymbol{\tau}-\boldsymbol{\tau}_{\theta_s}]^T\boldsymbol{R}^{-1}[\boldsymbol{\tau}-\boldsymbol{\tau}_{\theta_s}]}}{2\pi^{\frac{I}{2}}|\boldsymbol{R}|^{\frac{1}{2}}}, \quad (7)$$

$$\hat{\theta}_s = \underset{\theta}{\mathrm{argmax}}(P[\boldsymbol{\tau}_\theta|\theta_s]), \quad (8)$$

ここで，^{T} は転置ベクトルを，$^{-1}$ は逆行列を表す．そして，最尤話者方位 $\hat{\theta}$ と各フレームの推定到来方位 $\theta(i)$ の絶対誤差を空間分散度 v_{sp} として抽出する．

$$\boldsymbol{v}_{sp} = [v_{sp}(1) \cdots v_{sp}(I)] = |\hat{\theta}_s - \theta(i)|. \quad (9)$$

A development of real-time voice activity detection system based on talker localization.

1 Takanobu Nishiura, 2 Yuki Denda, 3 Hiroshi Saruwatari, and 3 Kiyohiro Shikano
1 College of Information Science and Engineering, Ritsumeikan University
2 Graduate School of Science and Engineering, Ritsumeikan University
3 Graduate School of Information Science, Nara Institute of Science and Technology

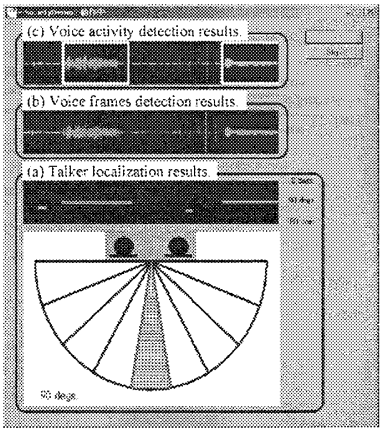
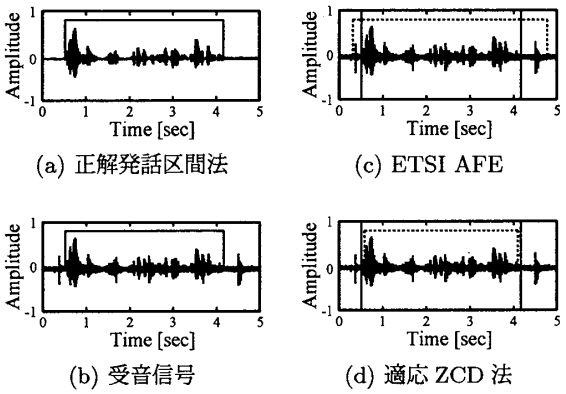


図 2: 発話区間検出システムの動作画面



(a) 正解発話区間法　(c) ETSI AFE

(b) 受音信号　(d) 適応 ZCD 法

図 3: 発話区間検出結果（SNR10 dB）

2.4 空間特徴量に基づく VAD 閾値の動的制御

適応 ZCD 法は空間特徴量に基づいて VAD 閾値を動的に制御し，音声/非音声フレームを識別する．

$$TH_a(i) = \begin{cases} TH_{a_L} & r_{sp}(i) \geq \bar{r}_n \\ TH_{a_H} & r_{sp}(i) < \bar{r}_n \end{cases}, \quad (10)$$

$$z(i) = ZCD\big(x(t), TH_a(i)\big), \quad (11)$$

$$TH_z(i) = \begin{cases} 20 & v_{sp}(i) \leq \varepsilon_{sp_r} \\ 60 & v_{sp}(i) > \varepsilon_{sp_r} \end{cases}, \quad (12)$$

$$VAD(i) = \begin{cases} Speech & z(i) \geq TH_z(i) \\ Non\text{-}speech & z(i) < TH_z(i) \end{cases}, \quad (13)$$

ここで，$\bar{r}_n$ は雑音区間の平均空間信頼度を，$ZCD(\cdot,\cdot)$ は $x(t)$ のゼロ交差回数を返す関数を，ε_{sp_r} は空間分散度の許容誤差を表す．式 (10) より，空間信頼度が高い場合に音声を受音しているとみなし，小さい振幅閾値 TH_{a_L} を用いて音声のゼロ交差を頑健に検出する．さらに式 (12) より，空間分散度が大きい場合は音声を受音していないとみなし，大きいゼロ交差回数閾値 TH_{a_H} を用いて雑音フレームの誤検出を防ぐ．

2.5 発話区間形成

フレームレベルの音声/非音声識別によって生じる湧き出し誤りやショートポーズなどを棄却するために，検出した音声フレームに基づいて発話区間を形成する．まずはじめに，100 ms 以内に検出された複数の音声フレームを接続し，1つの音声区間候補を形成する．次に，言語的な意味を含む音声の最低継続時間長を 350 ms と仮定し，その継続長より短い音声区間候補を棄却することで最終的な発話区間を決定する．

3 リアルタイム発話区間検出システム

提案した適応 ZCD 法を用いて，PC 上で実時間動作可能な発話区間検出システムを開発した．システムのオーディオインタフェースに低レイテンシで動作可能な Steinberg 社の ASIO（Audio Stream Input Output）を用い，サンプリング周波数 48 kHz，量子化 16 ビットで信号を受音する．ペアマイクロホンの素子間隔は 148.75 mm とした．システムの動作画面を図 2 に示す．図 2(a) は話者方位推定結果表示部を，(b) は発話フレーム表示部を，(c) は発話区間表示部を表す．

4 評価実験

4.1 実験条件

暗騒音レベル 39.6 dBA，室内残響（$T_{[60]}$）0.41 s の実オフィス環境において発話区間検出実験を行った．80 度方向から到来する ATR 音素バランス 216 単語（女性 3 名，男性 3 名）を評価用音声データとして使用した．また，150 度方向から到来するサーバ音を定常指向性雑音として，140 度方向から到来する拍手音を突発指向性雑音として使用した．そして，標準化手法 ETSI AFE（Advanced Front End）の VAD[4] と提案する適応 ZCD 法を SNR10 dB の条件で比較した．

4.2 実験結果

CPU:Intel Pentium M 1.4 GHz，メモリ:512 MByte のノート PC を使用して ETSI AFE と適応 ZCD 法の処理時間（Real Time Factor : RTF）を計測した．その結果，ETSI AFE の RTF が 0.18，適応 ZCD 法の RTF が 0.17 となり，適応 ZCD 法は ETSI AFE と同程度の処理時間で VAD を行えることが確認できた．

発話区間検出結果を図 3 に示す．図中の実線は目視によって付与した正解発話区間を，点線は検出された発話区間を表す．図より，適応 ZCD 法は ETSI AFE よりも正確に発話区間を検出できた．

5 まとめ

本稿では，話者方位推定に基づく適応ゼロ交差検出法を提案し，ノート PC 上で実時間動作可能な発話区間検出システムを開発した．今後は，音声/非音声の遷移確率を考慮した発話区間形成法について検討する．

謝辞 本研究の一部は文科省 LP“e-Society”，グローバル COE プログラムおよび科研費 17700216，17200014 による支援を受けて行われた．

参考文献

[1] R.P. Venkatesha et al., Proc. ISCC02, 2002.
[2] L. Armani et al., Proc. Eurospeech03, 2003.
[3] Y. Denda et al., IEICE Trans. on Inf. & Sys., 2006.
[4] ETSI ES 202 050 v.1.1.5, 2007.