LE-006

音響信頼度に基づく動的特徴量統合を用いた全方位マルチモーダル話者方位推定の検討

A Study of Omnidirectional Audio-Visual Talker Localization with Dynamic Feature Fusion Based on Audio Reliability

傳田 遊亀† 西浦 敬信‡ 山下 洋一‡
Yuki Denda Takanobu Nishiura Yoichi Yamashita

1．まえがき

近年，複数のユーザが存在するビデオ会議の自動構造化に関する研究が広く行われている[1]．会議の構造化を行うためには，発話中のユーザ（発話者）を検出し，その方位（話者方位）を推定することが必要不可欠である．しかし，音響情報に基づく話者方位推定法は室内残響や背景雑音の影響で性能が低下し，画像情報に基づく話者方位推定法は照明条件の急激な変化などに脆弱であるという問題がある．これらの問題の解決策として，音響情報と画像情報の統合に基づくマルチモーダル話者方位推定法が提案されている[2,3,4]．我々は，マイクロホンアレーとアクティブビデオカメラを用いる手法を提案している[2]．しかし，アクティブビデオカメラの視野外に話者が存在する場合には画像情報を用いることができず性能が低下するという問題がある．文献[3]では，分散ビデオカメラによって得られる広範囲画像情報を用いた手法が提案されているが，カメラキャリブレーションが必要になるためシステムが複雑化するという問題がある．特徴量統合の観点からマルチモーダル方位推定法を分類した場合，事前統計量に基づくベイジアンネットワーク[4]などを用いた手法によって高い性能が得られている．しかし，統計量を事前に学習する必要があり，大量の学習データ収集が必要になるという問題がある．

本論文ではこれらの問題を解決するために，事前統計量を用いない動的特徴量統合に基づく全方位マルチモーダル話者方位推定法を提案する．提案手法は，正三角形マイクロホンアレーを用いた音声の到来方位推定によって全方位音響特徴量を，全方位ビデオカメラを用いた人物位置検出によって全方位画像特徴量を抽出する．次に，音響/画像特徴量に基づく観測頻度分布によって各特徴量の妥当性を統計的に評価する．最後に，指向性音源観測基準に基づいて評価した音響特徴量の信頼度（音響信頼度）を統合重みとして用いる動的特徴量統合を行い，話者方位を推定する．

2．提案手法

提案する全方位マルチモーダル話者方位推定法の概要を図 **1** に示す．提案手法は，正三角形マイクロホンアレーを用いた WCSP（Weighted Cross-power Spectrum Phase）法[5]と CSP 係数サブトラクション[5]によって全方位音響特徴量を，全方位ビデオカメラを用いた背景差分[6]と肌色検出[7]によって全方位画像特徴量を抽出する．ここで，

† 立命館大学大学院 理工学研究科

‡ 立命館大学 情報理工学部

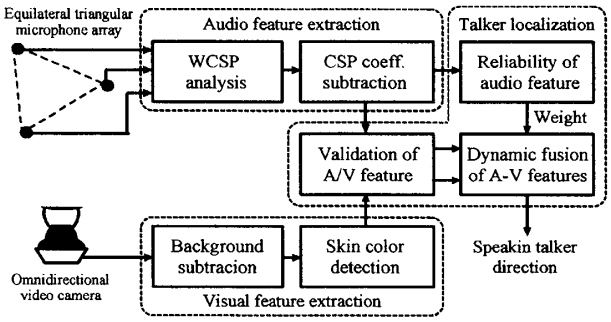


図 1：全方位マルチモーダル話者方位推定法の概要

短時間の観測時間内では話者が同じ方位に連続して観測されるため，音響/画像特徴量に基づく観測頻度分布によって各特徴量の妥当性を統計的に評価することができる．最後に，指向性音源観測基準に基づいて音響特徴量の信頼度を評価する．音響特徴量の信頼度が低い区間においては，画像特徴量に対する動的統合重み係数を大きくすることで耐雑音性を改善できる．

2.1 音声の到来方位推定に基づく音響特徴量

2.1.1 WCSP 法

我々は，音声信号と雑音信号のスペクトルに相関がないと仮定し，平均音声スペクトルに基づいて各周波数の位相情報に信頼度を付与することで雑音に頑健な音声の到来方位推定を行う WCSP 法を提案している[5]．WCSP 法は，ペアマイクロホン M_1, M_2 で受音した信号 $x_1(t), x_2(t)$ に基づいて以下の式によって表せる．

$$WCSP(k) = \mathrm{IDFT}\left[W_S(\omega)\frac{X_1(\omega)X_2(\omega)^*}{|X_1(\omega)||X_2(\omega)|}\right] = \mathrm{IDFT}\left[W_S(\omega)e^{-j\omega(\varphi_2-\varphi_1)}\right], \quad (1)$$

$$WCSP(\theta) = \mathrm{F}(WCSP(k)), \quad (2)$$

$$\mathrm{F}:\ \theta = \cos^{-1}\left(\frac{ck}{dF_s}\right), \quad (3)$$

ここで，$WCSP(k)$ は時間領域の WCSP 係数を，$\mathrm{IDFT}[\cdot]$ は逆フーリエ変換を，$W_S(\omega)$ は平均音声スペクトルに基づく重み係数を，$X_{[\cdot]}(\omega)$ は $x_{[\cdot]}(t)$ の周波数表現を，* は複素共役を，$\varphi_{[\cdot]}$ は $x_{[\cdot]}(t)$ の到来時間を表す．また式(3)は，時間領域の WCSP 係数から角度領域の WCS

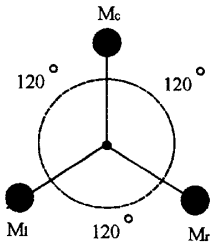


図 2: 正三角形マイクロホンアレーの配置

P 係数（$WCSP(\theta)$）への変換式を，cは音速を，dはマイク間隔を，F_Sはサンプリング周波数を表す．次に，ペアマイクロホンを用いた WCSP 法を，正三角形マイクロホンアレーを用いた全方位 WCSP 法に拡張する．図 2 に示す正三角形マイクロホンアレー（M_c, M_l, M_r）の 3 組のマイクペア $(M_c, M_l), (M_l, M_r), (M_r, M_c)$ を用いて 3 個の WCSP 係数 $WCSP_{c,l}(\theta), WCSP_{l,r}(\theta), WCSP_{r,c}(\theta)$ を求める．次に，3 組のペアマイクの位置関係に基づいて，3 個の WCSP 係数から全方位 WCSP 係数 $WCSP_{omni}(\theta)$ を合成する．

$$WCSP_{omni}(\theta) = WCSP_{l,r}(\theta) + WCSP_{r,c}(\theta+120) + WCSP_{c,l}(\theta-120). \quad (4)$$

なお，以降は全て全方位 WCSP 係数を用いるため，$WCSP_{omni}(\theta)$ を $WCSP(\theta)$ と記述する．

2.1.2 CSP 係数サブトラクション

我々は，定常雑音の空間的な分布を除去することによって雑音に頑健な音声の到来方位推定を行う CSP 係数サブトラクションを提案している[5]．CSP 係数サブトラクションは，式(5)によって雑音の空間分布に相当する雑音分布 WCSP 係数 $WCSP_{\bar{N}}(\theta)$ を雑音区間において学習する．次に，式(6),(7)によって全方位 WCSP 係数から雑音分布 WCSP 係数を減算することで，音声の空間分布に相当する音声分布 WCSP 係数 $WCSP_{\hat{S}}(\theta)$ を求める．

$$WCSP_{\bar{N}}(\theta) = \sum_{l=1}^{L} \max(WCSP(l,\theta), 0), \quad (5)$$

$$WCSP_{\hat{S}}(\theta) = WCSP(\theta) - \alpha WCSP_{\bar{N}}(\theta), \quad (6)$$

$$\alpha = \frac{\max(WCSP(\theta))}{\max(WCSP_{\bar{N}}(\theta))}, \quad (7)$$

ここで，αはサブトラクション係数を表す．そして，求めた音声分布 WCSP 係数を音響特徴量 $F_A(\theta)$ とする．

2.2 人物位置検出に基づく画像特徴量

2.2.1 正規化距離に基づく背景差分

正規化距離に基づく背景差分（以下，背景差分）は，画像内の人物候補領域を検出するための一手法であり，照明条件の変化に頑健な手法である[6]．背景差分はまず，入力パノラマ画像を $N \times N$ サイズのブロックに分割し，各ブロック中の画素の輝度値に対応した N^2 次元の輝度ベクトルを求める．ここで，uを画像水平方向の画素位置，vを画像鉛直方向の画素位置，$\boldsymbol{I}(u,v), \boldsymbol{I}_B(u,v)$をそれぞれ入力画像と背景モデルの輝度ベクトル，$|\cdot|$をベクトルのノルムとすると，正規化距離は次式で表せる．

$$ND(u,v) = \left| \frac{\boldsymbol{I}(u,v)}{|\boldsymbol{I}(u,v)|} - \frac{\boldsymbol{I}_B(u,v)}{|\boldsymbol{I}_B(u,v)|} \right|. \quad (8)$$

位置 u,v の領域における正規化距離が閾値以上の場合，対象領域を人物候補領域とみなし，次節の肌色検出処理を行う．本論文ではブロックサイズ N を 8 とした．

2.2.2 肌色検出

画像に含まれる肌色情報は人物位置検出のための重要な情報である．文献[7]では，肌色分布をさまざまな色空間において正規分布でモデル化した結果，TSL（Tint, Saturation, Luminance）色空間の TS 平面で構築した肌色モデルがモデル化誤差を最も減少させると報告されており，本論文でも同様に TS 平面で肌色検出を行う．ここで，RGB 値から T，S 値への変換は以下の式で表せる．

$$r = R/(R+G+B), \quad (9)$$

$$g = G/(R+G+B), \quad (10)$$

$$S = \sqrt{(9.0/5.0)(r^2+g^2)}, \quad (11)$$

$$T = \begin{cases} \tan^{-1}(r/g)/2\pi + 1/4, & g < 0 \\ \tan^{-1}(r/g)/2\pi + 3/4, & g > 0 \\ 0 & g = 0 \end{cases}. \quad (12)$$

次に，人物候補領域内の画素の肌色対数尤度を求め，鉛直方向に積分することで画像特徴量 $F_V(\theta)$ を抽出する．

$$\boldsymbol{I}(\theta,\phi) = [S(\theta,\phi), T(\theta,\phi)], \quad (13)$$

$$\boldsymbol{I}_S = [\bar{S}, \bar{T}], \quad (14)$$

$$\lambda = [\boldsymbol{I}(\theta,\phi) - \boldsymbol{I}_S], \quad (15)$$

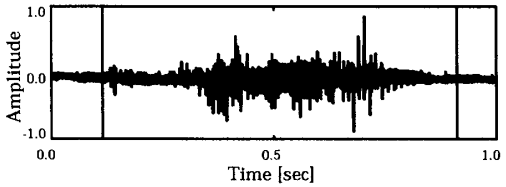
$$P[\boldsymbol{I}(\theta,\phi) \mid S] = -\ln\left(e^{-\frac{1}{2}(\lambda^T \boldsymbol{R}_S \lambda)} \Big/ 2\pi |\boldsymbol{R}_S|^{\frac{1}{2}} \right), \quad (16)$$

$$F_V(\theta) = \sum_{\phi=0}^{\phi_I} P[\boldsymbol{I}(\theta,\phi) \mid S], \quad (17)$$

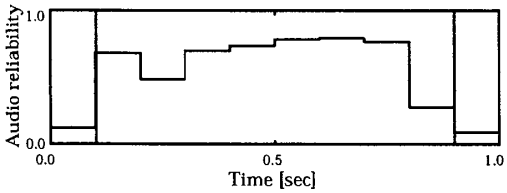
ここで，$\boldsymbol{I}(\theta,\phi)$は座標$(\theta,\phi)$の画素における$T,S$値を要素に持つベクトルを，$\boldsymbol{I}_S$は肌色ガウスモデルの平均$T,S$値を要素に持つベクトルを，$\boldsymbol{R}_S$は肌色ガウスモデルの分散共分散行列を，^{T}は転置行列を，$^{-1}$は逆行列を，ϕ_Iはパノラマ画像の鉛直方向の解像度を表す．本論文では，日本人 11 名（女性 2 名，男性 9 名）の顔画像を使用して肌色モデルを学習した．

2.3 音響/画像特徴量の妥当性評価

本論文では，短時間の観測時間内では発話者は同じ方位に連続して観測されると仮定する．従って，音響/画像特徴量に基づく観測頻度分布を用いて各特徴量の妥当性を統計的に評価できる．ここで，式(18)は妥当性評価基準 $V_{A\,or\,V}(i,\theta)$ を，式(19)は観測頻度分布 $lc_{A\,or\,V}(i,\theta)$ を表す．式(18),(19)より，特徴量の妥当性は音響/画像特徴量に基づいて話者方位を推定した場合の観測ヒストグラムを



(a) 観測音響信号



(b) 統合重み係数

図 3：音響信号と統合重み係数

を用いて評価される．

$$V_{A\,or\,V}(i,\theta)=\frac{\sum_{t=1}^{T} lc_{A\,or\,V}(i-t,\theta)}{T}, \quad (18)$$

$$lc_{A\,or\,V}(i,\theta)=\begin{cases}1, & F_{A\,or\,V}(i,\theta)\geq TH_{A\,or\,V}\\ 0, & F_{A\,or\,V}(i,\theta)<TH_{A\,or\,V}\end{cases}. \quad (19)$$

2.4 音響特徴量の信頼度評価に基づく動的統合

発話区間では音声が指向性音源として観測され，逆に非発話区間では指向性音源が観測される頻度は非常に少ない．従って本論文では，式(20)による指向性音源観測基準に基づいて音響特徴量の信頼度を評価し，動的統合の重み係数を決定する．次に，式(21)によって特徴量の動的統合を行い，式(22)によって話者方位を推定する．

$$R_A(i)=\frac{\max(F_A(i,\theta))}{\sum_{\theta=0}^{\theta_I} F_A(i,\theta)}, \quad (20)$$

$$F_{AV}(i,\theta)=R_A(i)V_A(i,\theta)F_A(i,\theta)+(1-R_A(i))V_V(i,\theta)F_V(i,\theta), \quad (21)$$

$$Talker(i,\theta)=\begin{cases}Presence, & F_{AV}(i,\theta)\geq TH\\ Absence, & F_{AV}(i,\theta)<TH\end{cases}, \quad (22)$$

ここで，$R_A(i,\theta)$ は音響特徴量の最大値と総和の比によって決定される統合重み係数を，θ_I は画像水平方向の解像度を表す．図 3 は，観測音響信号(a)と統合重み係数(b)の一例である．図 3 より，統合重み係数は発話区間において自動的に高い値になり，非発話区間において低い値になる．換言すれば，重み係数は発話/非発話区間を示す一指標であるといえる．従って，発話区間においては重み係数が自動的に大きくなるため音響特徴量を重要視して話者方位推定を行い，非発話区間では積極的に画像特徴量を用いることで話者方位を推定できる．

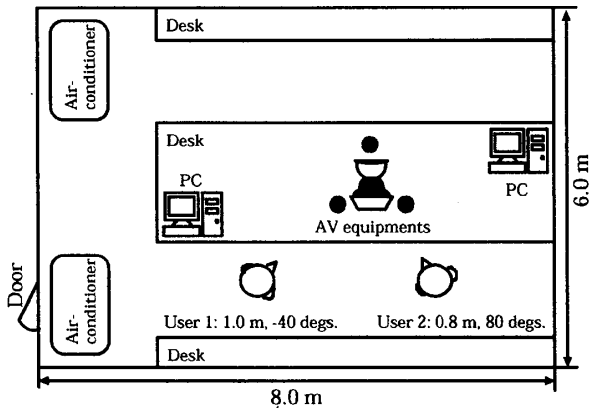


図 4：実験環境

表 1：実験条件

音響信号収録条件	
マイクロホンアレー	150.0 mm 間隔
サンプリング周波数	16 kHz
室内残響	0.41 秒
室内騒音	44.1 dBA
画像信号収録条件	
パノラマ画像	660×140 ピクセル
フレームレート	15fps
画像雑音	茶色クラフト封筒
方位推定実験条件	
User1	−40 度, 1.0 m
User2	+80 度, 0.8 m
フレーム長	2/3 秒
フレームシフト	1/3 秒

3．実環境評価実験

3.1 実験条件

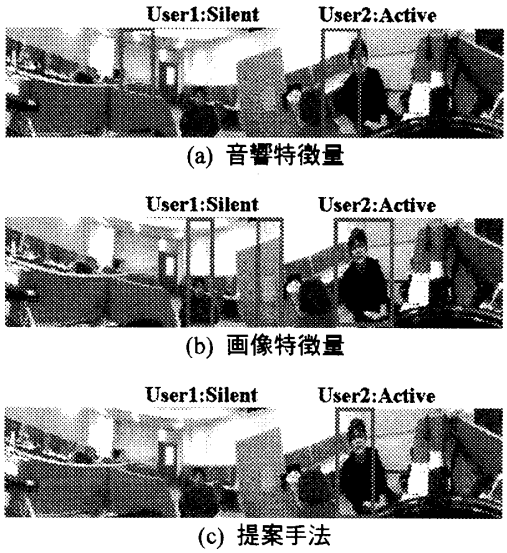
提案手法の有効性を検証するために空調機や PC などの雑音源が存在する実オフィス環境において評価実験を行った．図 4 に実験環境を，表 1 に実験条件を示す．

音響信号は，素子間隔 150 mm の正三角形マイクロホンアレーを用いて音響サンプリング 16kHz で収録した．実験環境の暗騒音レベルは 44.1dBA，室内残響（$T_{[60]}$）は 0.41 秒であった．画像信号は，660×140 ピクセルのパノラマ画像を撮影できる全方位ビデオカメラを用いて画像フレームレート 15fps で収録し，茶色クラフト封筒を画像雑音として使用した．背景モデル学習データには，封筒とユーザが存在しない状況で撮影した画像を使用した．

2 人のユーザを機器の周囲に配置し，1 人目のユーザ（User1）は機器から 1.0 m，−40 度の位置から，2 人目のユーザ（User2）は 0.8 m，+80 度の位置から，それぞれ日本語音素バランス単語を交互に 50 単語発話した． 収録した映像全体の長さは約 300 秒であり，そのうち発話区間は 100 秒，非発話区間は 200 秒である．方位推定のフレーム長は 2/3 秒，フレームシフトは 1/3 秒とした．

3.2 実験結果

図 5 に User1 が非発話中，User 2 が発話中の状況における話者方位推定結果を示す．図 5(a)に音響特徴量のみを用



(a) 音響特徴量

(b) 画像特徴量

(c) 提案手法

図 5: 話者方位推定の実験結果

表 2: 話者方位推定性能

	誤棄却率 [%]	誤検出率 [%]
固定統合重み：		
0.0 (画像特徴量のみ)	8.8	69.6
0.25	9.1	67.3
0.5	11.9	19.8
0.75	12.8	16.2
1.0 (音響特徴量のみ)	13.2	15.9
動的統合重み： 提案手法	4.2	7.6

いた話者方位推定法の結果を，図 5(b)に画像特徴量のみを用いた話者方位推定法の結果を，図 5(c)に提案するマルチモーダル話者方位推定法の結果を示す．図 5(a)より，音響特徴量のみを用いた場合，発話中の User 2 の方位+80 度に対して推定結果が+66 度,−113 度となり，実験環境内の雑音の影響により話者が存在しない方位を誤推定していることが確認できる．次に図 5(b)より，画像特徴量のみを用いた場合の推定結果は+85 度, +8 度, −51 度となった．User2 の方位に関しては音響特徴量を用いた場合よりも正確に推定できているが，発話していない User1 や背景モデル学習データに含まれていない封筒の方位も誤推定していることが確認できる．最後に図 5(c)より，提案手法の推定結果は+76 度となり，音響/画像特徴量を単独で使用した場合と比較して，User 2 の方位のみをより正確に推定できることが確認できる．

次に，発話区間における話者方位の誤棄却率と誤検出率に基づいて話者方位推定性能の評価を行った．誤棄却率は検出すべき話者方位を検出できなかった割合を，誤検出率は検出すべき話者方位以外を誤って検出した割合を表す．本論文では目視によって発話区間を検出した．また，推定された話者方位と正解話者方位の誤差が 5 度以内の場合に話者方位が正しく推定されたとした．実験結果を表 2 に示す．表 2 より，音響雑音の影響を受けない画像特徴量の重みを大きくすることで誤棄却率を改善できることが確認できる．しかし，検出したユーザが発話中であるのかを画像特徴量に基づいて識別することは困難であるため，誤検出率が高くなることが確認できる．最後に，音響/画像特徴量を単独で用いた場合や固定統合重み係数を用いた場合と比較して，動的特徴量統合に基づく提案手法によって話者方位推定性能を改善できた．

4．むすび

本論文では，音響特徴量の信頼度基準に基づく動的特徴量統合を用いた全方位マルチモーダル話者方位推定法を提案した．提案手法は，正三角形マイクロホンアレーを用いた WCSP 法と CSP 係数サブトラクションによって音響特徴量を，全方位ビデオカメラを用いた背景差分法と肌色検出によって画像特徴量を抽出する．次に，音響/画像特徴量に基づく観測頻度分布を用いて各特徴量の妥当性を統計的に評価する．最後に，指向性音源観測基準に基づいて評価した音響特徴量の信頼度を統合重み係数とする動的特徴量統合を行い，頑健な話者方位推定を実現する．実オフィス環境における評価実験の結果，音響/画像特徴量を単独で用いた場合や固定統合重み係数に基づく統合法と比較して，提案手法は高い話者方位推定性能を得られることが確認できた．今後は，さらに詳細な評価実験を行い提案手法の有効性を検証する．

5．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．

6．参考文献

[1] T. Hain, J. Dines, G. Garau, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals, "Transcription of conference room meetings: an investigation," Proc. Eurospeech05, pp.1611-1614, 2005.

[2] Y. Denda, T. Nishiura, H. Kawahara, and T. Irino "A design of audio-visual talker tracking system based on CSP analysis and frame difference in real noisy environments," Proc. IEEE MMSP04, pp.63-66, 2004.

[3] K. Wilson, N. Checka, D. Demirdjian, and T. Darrell, "Audio-video array source separation for perceptual user interfaces," Proc. WPUI01, pp.1-7, 2001.

[4] 陳彬, 目黒光彦, 金子正秀, "ベイジアンネットワークに基づく視聴覚情報の統合を用いた画像からの３次元音源位置推定," 電気学会論文誌 C, vol.C24, no.3, pp.720-728, March, 2004.

[5] Y. Denda, T. Nishiura, and Y. Yamashita, "Robust talker direction estimation based on weighted CSP analysis and maximum likelihood estimation," IEICE Trans. on Information and Systems, vol.E89-D, no.3, pp.1050-1057, 2006.

[6] 長屋茂喜, 宮武孝文, 藤田武洋, 伊藤渡, 上田博唯," 電子情報通信学会論文誌 D-II, vol.J79, no.4, pp.568-576, April, 1996.

[7] J.C. Terrillon and S. Akamatsu, "Comparative performance of different chrominance spaces for color segmentation and detection of human faces in complex scene images," Proc. ICVI, pp.180-187, 1999.