マイクロホンアレイを用いた複数人対話からの音声区間検出および話者方向推定の評価手法

○黄 楊暘 † 大塚 琢馬 † 中臺 一博 ‡ 奥乃 博 †
† 京都大学大学院 情報学研究科 知能情報学専攻 ‡HRI-JP

1. はじめに

人とロボットが共生するために, ロボットの聴覚機能の開発は不可欠である. 日常生活の会話場面を考えると, ロボットの聴覚機能にはいつ, どこ, 誰が何を話したかを認識する機能を有するのが望ましい. 本稿では, いつ, どこ, 誰に重点を置いて, 複数話者の音声区間検出問題について報告する.

音声区間検出, 音源定位および音源同定を用いていつ, どこ, 誰といったを推定するのに, さまざまな手法が開発されておる [1]. 音声区間検出と音源同定は, それぞれ音声特徴量に基づく音声・非音声の 2 クラス決定問題, 話者ごとの多クラス決定問題と位置づけられる. 音源定位の手法は一般に, 別々のマイクロホンまで到着するのに要する時間差に着目している.

本稿では, 複数話者の自由発話を対象とした音声区間, 音源定位および音源同定の推定結果を評価するための正解データの作成方法と性能評価指標を提案する. 正解データを作る際の課題は 2 つある. (1) 複数話者の同時発話の音声区間特定. (2) 音源の移動を考慮した時刻ごとの音源方向特定. これらの課題を解決するために, 自由発話を収録するマイクロホンアレイ以外に, 接話型のマイクロホンと話者位置の三次元座標を測定する MAC3D システムを利用して [3], リファレンスデータを作成した. 音声区間検出・音源定位・音源同定を総合的に評価するための指標として適合率や再現率, F 値, 挿入エラー比率, 削除エラー比率, 定位誤差を導入した. 正解データと評価指標を定義した後, MUltiple SIgnal Classification(MUSIC 法) に基づいた手法を説明して, ベースライン手法として評価した. その後, 音声特徴を用いて, 音声区間検出, 音源定位および音源同定を同時に推定する手法を提案して評価した.

以下, 2 章で正解データの作成法と推定結果の評価指標を導入する. 3 章でこれらの問題を同時に解決する手法を説明する. 4 章で作成した正解データと評価指標を用いた評価実験の結果を報告し, 5 章でもとめと今後の課題を述べる.

以下に本稿で扱う問題設定を示す:

- 入力：多チャンネルの音声信号
- 出力：音声区間, 音源の到来方向および話者 ID
- 条件 1：各話者の事前学習データが入手可能
- 条件 2：マイクロホンアレイの伝達関数が既知

条件 1 に関して, 音声区間と話者についての正解ラベルが与えられた音声データを用いて, 各話者クラス構築のための事前学習を行う. 条件 2 に関して, 遅延和ビームフォーマの音源分離手法を利用するために, マイクロホンアレイの伝達関数が必要である. 伝達関数は各方向からの音の伝達特性を表す.

2. 正解データおよび評価指標

本節では, 音声区間検出, 音源定位, および話者同定に関する評価指標の設計と, 実録音対話データからの正解データ作成手順を説明する. 本稿で用いる表記を表 1 にまとめた.

マイクロホンアレイの入力音声信号を, 長さが 0.5 秒のブロックに分割して, 方向ごとに, 音声区間であるかどうかおよび音源の ID を推定する. 結果の形式は, ブロック数 × 方向数の二次元アレイのデータ構造として扱う. $x_{b,\theta}$ は b 番のブロックにおいて, θ 方向の推定結果の値を表す. $x_{b,\theta}$ は 0 以上の整数である, 0 の場合は無音区間, 0 より大きい場合はその音源 ID の音声区間であることを示す.

2·1 正解データの作成

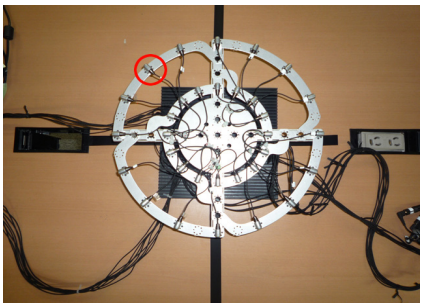
実環境の音源は, 今回収録したデータを含めて, 一般に移動する. 複数音源が時々刻々位置を変化させながら音を発したり黙ったりするデータに対して, 音源位置や音声区間の評価用フィファレンスデータが必要であるため, 今回は次の手順で正解データを作成した.

1. 今回の複数話者による発話データは図 1-a のように収録した. 机の上に 16 チャンネルのマイクロホンアレイ (図 1-b) を設置し, 机の周りに, 五人の話者が座った. 各話者が着席した状態でマイクロホンアレイに向いて発話を行った. 話者の首の動きなどによる音源移動はあるが, 席替えなどの音源方向の大きな変動は今回のデータには含まれない.
2. 音声区間のリファレンスデータは, 各話者の襟元につけた接話型マイクロホンによる録音データと収録時に同時に録画されたビデオを元に手動で作成した.
3. 各話者の位置の正解データはリアルタイム光学式モーションキャプチャシステム (MAC3D システム) を利用して取得した. このシステムは, 図 1-c のように各話者の肩と頭部に付けられたマーカーとカメラアレイによって各話者の位置を追跡する. 本システムにより得られた, 各話者を天井から見下ろした場合の, マイクロホンアレイを減点とする x-y 座標をプロットすると, 図 3 のようになる. 話者の x-y 座標から，マイクロホンアレイからみた

表 1 本稿の記号表記一覧

$b = 1, \cdots, B$	時間領域の分割となるブロックの番号
$\theta = 1, \cdots, 144$	円周を 2.5 度で分割した各方向の番号
$x_{b,\theta}$	ブロック b の θ 方向の推定結果
θ_p	音源定位結果の許容誤差範囲
R_p, R_r	推定結果の適合率及び再現率
E_I, E_D	推定結果の挿入エラー及び削除エラー比率
E_{dir}	推定結果の音源定位誤差
E_{ID}	音源 ID の誤推定率
$P_{b,\theta}$	MUSIC スペクトルの値
N	MUSIC スペクトルを計算する時の音源数

(a) 録音風景



(b) マイクロホン配置, 今回は外側の 16 個のマイクロホンが収録したデータを利用した, 赤い丸で囲んだのはその一つである.

(c) MAC3D システムマーカー, 帽子と肩にある白い円状物がマーカーである.

図 1 実験風景

その話者の方向も容易に計算が可能である. 今回のデータでは各話者は大きく移動しないため, マイクロホンアレイからの話者方向の範囲で話者 ID を定め, 線を色分けした.

4. 2. で作成した音声区間は,3. で付与した音声 ID と対応付けることで, $x_{b,\theta}$ を作成した.

正解データを図に描くと, 図 3 の線にあたる.

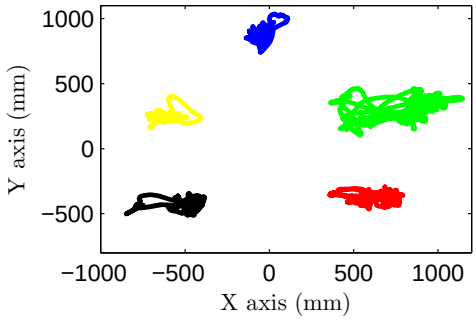


図 2 MAC3D で話者の頭の中心の座標の軌跡図である. 中心は収音マイクロホンアレイの中心となる.

2·2 評価指標

音声区間検出, 音源定位, 音源同定の結果について, 以下の評価指標を設計する.

2·2.1 音源 ID を考慮しない場合

音声区間検出の評価指標には, 挿入エラーと削除エラーを用いる. 挿入エラーは, 正解データでは無音区間となっている区間に対して, 音声を検出する誤認識のことである. それに対し削除エラーは, 正解データでは発話区間であることを示しているのに, アルゴリズムが発話を検出しないという誤認識である. 挿入・削除エラーの計算には音源方向にある程度の誤差を許容する. たとえば, 正解データでは 30° 方向に音声が存在するのに, 35° 方向に音声を検出した場合を考えた時, 30° 方向の音に対する削除エラーに加えて 35° 方向への挿入エラーが生じたとみなすのではなく, 定位誤差はあるものの挿入・削除は生じなかったとみなす. 具体的には, 許容誤差が θ_p で正解データではブロック b にて θ 方向に音源があるとき, $[x_{b,\theta-\theta_p}, x_{b,\theta+\theta_p}]$ の範囲内に存在する $x_{b,\theta}$ の値が 0 より大きい場合は, 発話区間検出については正解とみなす. ただ, 一つの音源方向の許容範囲に複数の推定結果が含まれる場合は, 挿入エラーとなる.

音源 ID を考慮しない場合には, マイクロホンアレイ処理によって検出された発話区間, すなわち $x_{b,\theta} > 0$ の数. その内の推定結果が正しい (挿入エラーでない) 数を S_c とする. また, 正解データ中の発話区間 $x_{b,\theta} > 0$ の数を S_d とする. 音源方向について, 正解データと推定結果の誤差の絶対値の和を Δ_{dir} とする. これらを用いて, 発話区間検出における評価指標は次のように定義される.

$$\text{適合率}: R_p = \frac{S_c}{S_a} \quad \text{再現率}: R_r = \frac{S_c}{S_d}$$

$$\text{挿入エラー比率}: E_I = \frac{S_a - S_c}{S_c}$$

$$\text{削除エラー比率}: E_D = \frac{S_d - S_c}{S_c}$$

$$\text{音源定位誤差}: E_{dir} = \frac{\Delta_{dir}}{S_c} \quad F \text{ 値}: F = \frac{2R_pR_r}{R_p + R_r}$$

2·2.2 音源 ID を考慮する場合

音源 ID を考慮する場合では, 音声区間と音源定位の推定結果が正しいにも関わらず, 音源 ID の付与が間違った場合がある. [6] ここで, 推定結果と正解データの同じ音源 ID である部分を取り出して, 各音源に対して, 前節の指標で評価することができる. この評価方法は音源 ID が正しい推定されたことを仮定して, 評価を行う. 音源 ID を考慮した発話区間検出・音源定位精度の評価指標としては, 推定された音源 ID の正解データについて前節の評価指標を適用することが考えられる. この手法は容易に評価計算を行えるが, 音源 ID の誤推定が評価スコアを著しく低下させる要因となる. したがって, 音源 ID の誤推定を定量的に評価するのが望ましい. ここで, 音源ごとに評価する時, 推定結

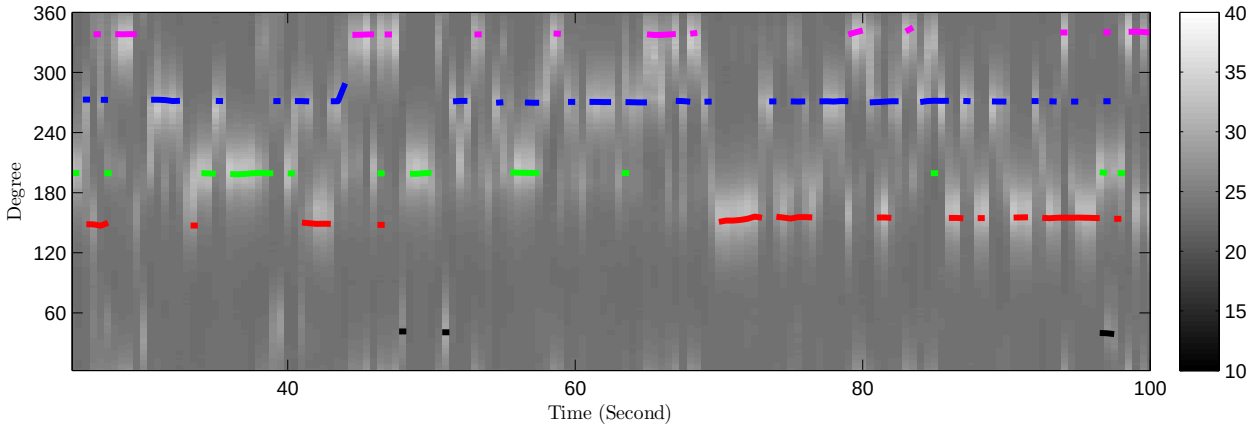


図 3 作成した正解データと MUSIC スペクトルを重ね合わせて描いた図. MUSIC スペクトルのピークが音声区間の対応関係を確認できる.

果が正しいと考えられる数をすべて足しあわせて, その総数を S_e とする. 音源 ID の誤推定率 E_{ID} を次のように定義する.

$$E_{ID} = \frac{S_c - S_e}{S_c}$$

3. ベースライン手法

ベースライン手法は MUSIC 法を利用する, MUSIC 法は音声信号の部分区間と雑音信号の部分区間が直交することを利用して, 高い精度の音源定位ができている. MUSIC スペクトルが得られたら, 事前に閾値を設定する. 閾値より以上の値が出た場合に, 音源定位と音声区間検出の同時推定ができる. MUSIC 法は, 観測信号に対して MUSIC スペクトル $P_{b,\theta}$ と呼ばれる, 各ブロック b, 方向 θ に対応するエネルギを計算し, 一定以上の $P_{b,\theta}$ を持つ方向に音源が存在するという閾値処理を行うことで音源定位を行う. MUSIC スペクトル $P_{b,\theta}$ の計算には多チャンネル観測信号のほか, 収録に用いるマイクロホンアレイの各周波数ビン f, 方向 θ に対応する伝達関数 $a_{f,\theta}$ と, 音源数 N のパラメータを元に行う. 評価実験では, MUSIC 法に与える音源数や, MUSIC スペクトルに対する閾値を変化させた場合のふるまいの違いを述べる. MUSIC 法の詳細は [4].

4. 提案手法

本手法では, 前節の方法により定位された方向をもとに, 各方向の分離音声を遅延和ビームフォーマにより抽出した上で Mel-Scale Log Spectrum (MLSS) 音声特徴量を抽出し, 有声/無声区間推定と話者 ID 推定を行う. 話者 ID の推定は混合ガウス分布 (GMM) による教師あり学習に基づく分類問題として扱う. ブロックで各方向の分離音声に対して, 無音・音声の推定を行って, 音源定位を達成することを想定している.

4·1 遅延和ビームフォーマによる音声分離

特定方向の分離音声の時間周波数領域成分 $V_{f,dir,i}$ を以下のように遅延和ビームフォーマの手法で計算する.

$$V_{f,dir,i} = \frac{\mathbf{a}_{f,dir}^H \mathbf{o}_{f,i}}{\mathbf{a}_{f,dir}^H \mathbf{a}_{f,dir}}$$

$\mathbf{o}_{f,i}$ は時間周波数領域に変換した音声信号である. $\mathbf{o}_{f,i}$ と伝達関数の成分 $\mathbf{a}_{f,dir}$ の次元数はチャンネル数に等しい. $\mathbf{a}_{f,dir}^H$ は $\mathbf{a}_{f,dir}$ のエルミート転置を表す. 遅延和ビームフォーマによる分離信号は, 指定された方向から到来する方向のほか, 別の方向からの漏れノイズも混入する. したがって, 音源 ID に付与には漏れノイズの影響を受けにくい MSLS 音声特徴量を利用する.

4·2 MSLS 特徴量の計算

本稿では, 音声特徴量として, MSLS 特徴量を利用する. MSLS 特徴量は, 人間の聴覚機能を反映した対数周波数軸上のパワーに基づく特徴量である. MSLS 特徴量は音源分離時に生じた漏れノイズに対する頑健性が期待でき, たとえば分離音声の音声認識などに利用されている [5].

MSLS 特徴量の計算の手順は次のようになる. メル周波数窓を使って,257 次元の線形周波数軸の分離音声の絶対値 $|V_{f,\theta,i}|(f = 0...256)$ を 13 次元の特徴ベクトル $y_{\theta,i}$ に変換する.

1. メル周波数と周波数の関係の計算式は次のようになる.

$$m = 1127 \log(1.0 + \frac{f}{700.0})$$

2. 各成分の対数値を取って, $\mathbf{x}$ が得られる.
3. 13 次元のベクトル $x(i)$ を以下のように $y(i)$ 正規化する. $i = 1, \cdots, 13$.

$$y(i) = \frac{1}{13} \sum_{p=0}^{12} \left\{ \sum_{r=1}^{1} 3 \left\{ x(r) \cos(\frac{\pi p(r-0.5)}{13}) \right\} \cos(\frac{\pi p(i-0.5)}{13}) \right\}$$

表 2 音源数を 1 に, 閾値を 27 に設定した時のベースライン手法の評価結果 (%).

閾値	R_p	R_r	E_I	E_D	$E_{dir}(^\circ)$
27	70.6	72.1	41.8	39.0	5.85

表 3 MUSIC スペクトルに基づくベースライン手法の F 値の評価. F 値が 1 に近いほど, 結果の精度が高い. 太字は各列の最大値である. 行の閾値の変化と列の音源数設定変化のが結果への影響が見られる.

	1	2	3	4	5
25	0.672	0.462	0.319	0.269	0.208
27	**0.713**	0.529	0.319	0.269	0.208
29	0.700	0.640	0.320	0.269	0.208
31	0.529	**0.706**	0.359	0.269	0.208
33	0.127	0.666	0.591	0.281	0.208
35	0	0.215	**0.696**	0.545	0.208
37	0	0	0.255	**0.661**	**0.271**

4·3 GMM のパラメータ学習

正解データがあるため, 分離音声に対して, 音声/非音声と話者 ID のラベルをつけられる. 学習に使われた正解データは, 録音ファイルから, 五名の話者ごとに 20 秒の分離音声の音声特徴量と音源が存在しない部分の 20 秒の分離音声を選んだ.

GMM による識別は，各音源 ID に加えて無声状態を考慮した多クラス識別問題として各パラメータの学習を行う. ラベル付けた音声特徴量データを EM アルゴリズムで混合ガウス分布の各混合の重み, 平均と分散 $g^l, \mu^{\mathbf{l}}, \mathbf{\Sigma^l}$ を学習する. $l(=1, \cdots, 3)$ は各混合のインデックスを表す, 本稿では混合数を 3 にした. c をクラスの番号として, クラスの決定は次の式で行う. $\mathcal{N}$ はガウス分布の確率密度関数で, $\mathbf{v}$ は音声特徴量ベクトルを指す.

$$Class = \arg\max_{c} \sum_{l=1}^{3} g_c^l \mathcal{N}(\mathbf{v}|\mu_{\mathbf{c}}^{\mathbf{l}}, \mathbf{\Sigma_c^l})$$

5. 実験結果

5·1 ベースライン手法の評価

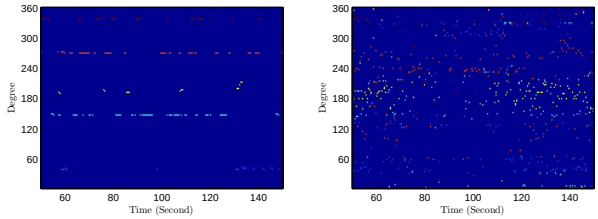
MUSIC 法による音源定位結果は，MUSIC スペクトルに対応する閾値と音源数パラメータを変化させた場合の F 値を評価する. MUSIC スペクトルに対して, 以下の処理を順に行って, 音声区間検出, 音源定位を行う. MUSIC スペクトルでは, 閾値以下の範囲である部分を無音区間と見なす. 一つのブロックにおいて, 連続の方向区間 $\Delta_\theta(=15^\circ)$ 内に連続で閾値より大きい場合, そのなかの最大値が位置する $x_{b,\theta}$ を音源の方向にして, 区間内の他の $x_{b,\theta}$ を無音区間と見なす. 以上の手順で計算された MUSIC 法による音源定位結果を表 2, 3 にまとめる.

5·2 提案手法の評価

提案手法による推定結果を表 4 と図 4 にまとめた. 挿入エラー, 削除エラーが多く起こった. また, 音源 ID の誤推定率が 42.9 である. 提案手法による無音・音声区間の識別が失敗している. さらに, 音源方向が正しく推定されないのも, 音源 ID の推定率の低下の原因ともなる.

表 4 提案手法の評価結果 (%)

R_p	R_r	E_l	E_D	$E_{dir}(^\circ)$	E_{ID}
11.2	43.3	792	130	7.49	42.9



(a) 正解データ　(b) 提案手法の評価結果

図 4 正解データと提案手法の推定結果, 挿入エラーが多く起こったが確認できる.

6. まとめと今後の課題

本稿では, 音声区間検出，音源定位および音源同定という問題に対する研究を加速させるため，複数人対話の実録音データの整備を行った. ベースライン手法と提案手法の評価実験からは, 今の段階では, 提案手法の性能がよくない. また，音声特徴量を考慮した上記 3 つの問題に対する同時解法の開発による性能向上や, 場面や収録環境に対する頑健性の向上も今後の課題である.

謝辞: 本研究の一部は科研費特別研究員奨励金/基盤 (S) の支援を受けた.

参考文献

[1] S. Tranter, "An overview of automatic speaker diarization systems," *IEEE Trans. on Audio, Speech, and Language Processing* vol. 14, No. 5, pp. 1557-1565, 2006

[2] 石塚健太郎, 藤本雅清, 中谷智広：“音声区間検出技術の最近の研究動向”, 日本音響学会誌, vol.52, no.10, pp.537-543, 2009.

[3] 角ら, " 会話の構造理解とコンテンツ化のための実世界インタラクション研究基盤", 情報処理学会誌, Vol. 49, No. 8, pp. 945-949, 2008

[4] R. O. Schmidt, "Multiple Emitter Location and Signal Parameter Estimation," *IEEE Trans. on Antennas and Propagation,* vol.34, no. 3, pp. 276-280, 1986.

[5] S. Yamamoto et al., "Simultaneous speech recognition based on automatic missing feature mask generation by integrating sound source separation", *RSJ.* vol.25, no. 1, pp. 99-102, 2007.

[6] 高橋ら, " 実環境したでの音源定位・音源検出の検討", 第 29 回日本ロボット学会学術講演会, Vol. 29, 1F3-3, 2009