多チャンネルマイクロホンアレイを用いた音声区間検出および音源定位の精度の向上の検討

On Improving the Accuracy of Voice Activity Detection and Sound Source Localization by Microphone Array

黄 楊暘 †, 大塚 琢馬 †, 中臺 一博 ‡, 奥乃 博 †

Yangyang Huang†, Takuma Otsuka†, Kazuhiro Nakadai‡, Hiroshi G. Okuno†

† 京都大学大学院情報学研究科, ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

†**Graduate school of Informatics, Kyoto University,** ‡**HONDA Research Institute Japan, Co., Ltd.**

†{yangyang, ohtsuka, okuno}@kyoto-u.ac.jp, ‡nakadai@jp.honda-ri.com

Abstract

In Real-World Auditory Scene Analysis concerning human-robot interaction, three types of information are essential and need to be extracted from the observation data – **WHO** speaks **WHEN** and **WHERE**. This paper presents such a system that is used to accomplish the resolution of these objects. To evaluate such a system, we formulate the use of evaluation indicators which are precision rate, recall rate, localization error and speaker ID error rate. Multiple Signal Classification (MUSIC) is a powerful method used for analysing WHEN and WHERE, more specifically, voice activity detection (VAD) and direction of arrival estimation (DOA). In this paper, we describe our system and compare its performance in VAD and DOA with MUSIC method.

1 はじめに

人とロボットが共生するためには, ロボットの聴覚機能の開発は不可欠である. 特に重要な聴覚機能としては，ロボットが人間と会話する場面を考えると，様々な人が発話する観測音の中から，いつ，どこで，誰が，何を話したかを認識する機能が挙げられる (図 2). これらの機能は, 音声区間検出, 音源定位, 音源同定や, 音源分離問題として, 様々な手法が開発されている. [Nakadai *et al.*, 2010; Tranter and Reynolds, 2006; Nakamura *et al.*, 2011].

本稿では, 上記のいつ，どこで，誰が話しているかを推定する話者ダイアライゼーション問題を取り扱う. 本問題は，マイクロホンアレイで収録された複数話者同士の自由発話に対して，各話者の音声区間検出や音声到来方向, および話者の推定を行う複合的な問題である. この問題には次の 2 点が重要である.

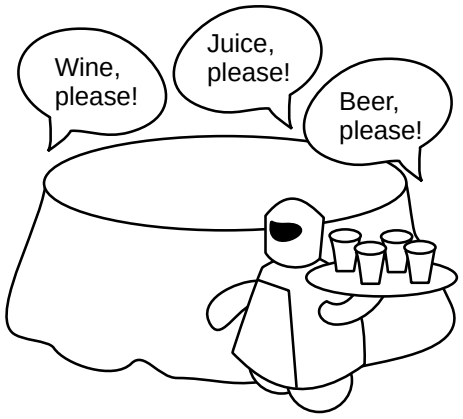


Figure 2: 例えば, 図示のカクテルパーティで接客するロボットの場合を考えて, いつ, どの方向から誰が注文を理解するのが重要である.

- 各部分問題に対してどのような要素技術を選択すれば全体の性能向上に寄与するかを明らかにすること,
- 複数の要素技術を直列につないで処理を行う場合，前段の処理の結果が後段の処理に影響するため，前段の処理は様々な観測音に対して頑健な手法が望ましい.

例えば，ロボット聴覚システム HARK[Nakadai *et al.*, 2010] では，全体の処理を multiple signal classification (MUSIC) 法による音源定位を行い，その音源方向推定結果に基づいて音源分離など，各音源に関する処理を行う. 本話者ダイアライゼーション問題についてもまず各話者の方向を推定し，その結果を用いて話者同定などを行う枠組

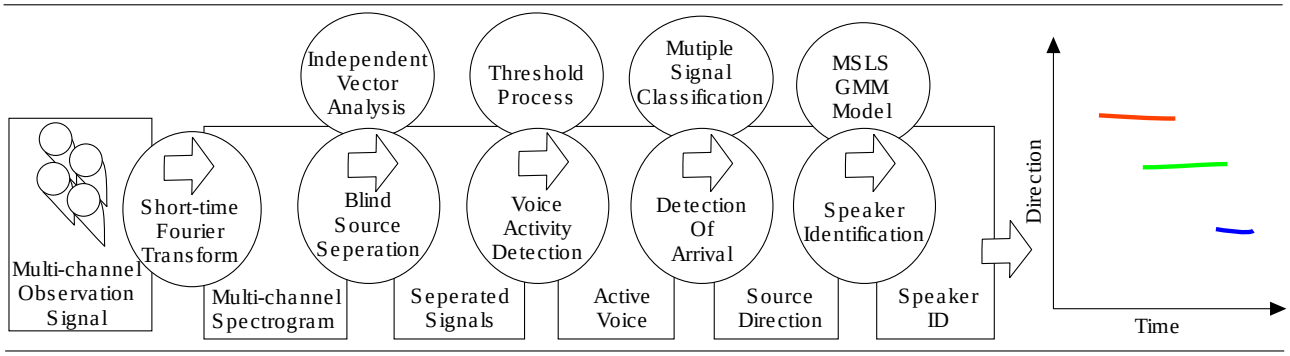


Figure 1: 処理の流れおよび出力結果の図示.
ひとつひとつの発話を線分で時間ー方向座標系で示すように, 色は音源 ID の違いを指している.

みが考えられる．しかし，表 1 でも示すように，MUSIC 法には入力音に依存した音源数や閾値などのパラメータにより，出力が大きく変わるという問題がある．従って，システム全体を最適化するには，注意深く MUSIC 法で用いるパラメータを選択する必要があるという問題があった．

上記に示して 2 点に関する本稿の貢献は次の通りである．

- 効率的な性能評価のため，収録した発話に対して正解データを付与し，話者ダイアライゼーション問題に対して性能評価指標を定義し，
- 前処理に頑健性の高い音源分離手法である independent vector analysis (IVA) を用いることで全体の性能を改善した．

正解データは，各話者に接話型マイクを使用して音声区間を決定した他，話者の位置を計測する MAC 3 D システム [角康之 *et al.*, 2008] を利用して音源方向の正解データを作成した．また，話者ダイアライゼーションシステムの評価指標としては，各話者の音声区間に対しては適合率，再現率を用いて F 値を定義し，音源定位誤差も導入した．

本稿は次のように構成されている．2 節では問題設定と提案手法の処理の流れ，および各要素技術を示す．3 節では評価用データの収録環境と正解データ作成法を説明し，および評価指標を定義し，4 節では評価実験結果を報告する．5 節で本稿をまとめる．

2　問題設定とシステム構成

本節では，話者ダイアライゼーションシステムの問題設定を示した後，提案手法の枠組みを示し，利用するそれぞれの要素技術を概説する．本稿で扱う問題設定は次の通りである．

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, MUSIC による音源定位では, 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.

提案手法は図 1 に従って処理する. 入力である多チャンネル音声信号を短時間フーリエ変換の後に，音源分離手法 IVA を適用する．得られた各話者の分離音声に対してパワーの閾値処理による音声区間検出を行う．また，各分離音声に対して MUSIC 法を用いて各話者の方向推定と，mel-scale log spectrum (MSLS) 音声特徴量を用いた話者同定を行う．話者同定には，混合ガウスモデル (GMM) による判別を行う．

2.1　IVA を用いたブラインド音源分離

IVA は多チャンネルの時間周波数領域における音源分離法であり，独立成分分析 (independent component analysis; ICA) の拡張手法である．本節ではまず ICA について概観し，IVA への拡張を簡潔に説明する．

ICA は時間周波数領域における多チャンネル観測信号 $\mathbf{Z}_{t,f} = [z^1_{t,f}, ..., z^M_{t,f}]^T$ が次式の観測モデルで表す.

$$\mathbf{Z}_{t,f} = \mathbf{A}_f \mathbf{Y}_{t,f}$$

ただし，$\mathbf{Y}_{t,f} = [Y^1_{t,f}, ..., Y^M_{t,f}]^T$ は時間フレーム t，周波数ビン f における各音源の信号で，$\mathbf{A}_f$ は混合行列である．このとき，ICA は観測信号 $\{\mathbf{Z}_{t,f}\}_{t=1}^T$ から，

$$\hat{\mathbf{Y}}_{t,f} = \mathbf{W}_f \mathbf{Z}_{t,f}$$

に従って計算される $\hat{\mathbf{Y}}_{t,f}$ の各成分が統計的に独立になるよう分離行列 $\mathbf{W}_f$ を求める．これは，元音源である $\mathbf{Y}_{t,f}$

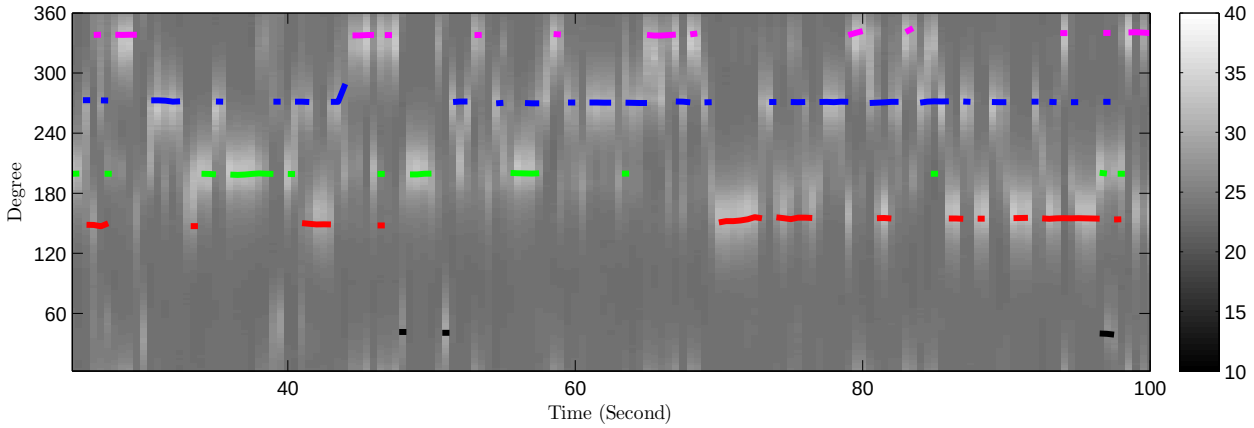


Figure 3: 作成した正解データと MUSIC スペクトルを重ね合わせて描いた図. MUSIC スペクトルのピークが音声区間の対応関係を確認できる.

の各成分が統計的に独立であるという仮定に従って音源分離に適用されている.

ICA における問題点は，式に従って各周波数ビン f ごとに計算された $\hat{\mathbf{Y}}_{t,f}$ の各成分は，必ずしも元の $\mathbf{Y}_{t,f}$ と同じ順になっていないというパーミュテーション問題である．そのため，元の音声信号を復元する際には，各周波数ビンごとに同一音源に属する成分を正しく選ぶ必要があった．それに対して IVA は，$\{\mathbf{Y}_{t,f}\}_{f=1}^{F}$ の各成分を F 次元のベクトルとみなして，全周波数ビンに関して $\{\mathbf{W}_f\}_{f=1}^{F}$ を最適化することで，パーミュテーション問題を回避している [Lee *et al.*, 2007; Ono, 2011].

2.2 音量閾値処理による音声区間検出

入力の多チャンネルスペクトログラムを波形信号 $y_{t,d}$ に変換して, 時間領域の信号 $Y_{t,d}$ に対して, 一定長 Δt の区間中において, 絶対値が閾値 T_v 以上の波形のサンプル数が T_s を超える場合に, 音声区間と見なす. 各分離音声の音声区間に含まれた部分をこれ以後の処理を続きます. 多チャンネルのスペクトログラムを算出した音声区間で切り出して出力する.

2.3 MUSIC 法による音源定位

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. 本手法では, MUSIC 法を音源定位に使う. 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ーを計算し，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．その算出の手順が次のようになる. 入力スペクトログラム $z_{t,f}$ の自己相関形式

$$R_{b,f} = \sum_{t=(b-1)*\Delta T}^{b\Delta T} \mathbf{z}_{t,f}\mathbf{z}_{t,f}^{H}$$

を取って, 安定の定位結果を得るために, フレーム ΔT 分の自己相関行列を足し合わせる, 一つのブロックと見なす. 各時間ブロック b と周波数ビン f の $R_{b,f}$ に対して固有値分解を行なって, チャンネル数と同じ M 個の固有値と固有ベクトルが得られる $\{\lambda_{b,f,m}, \mathbf{e}_{b,f,m}\}$. 固有値の大きい順から, 固有値と固有ベクトルを並べる. その時間ビンと周波数ビンの MUSIC エネルギーは算出された固有ベクトル $\mathbf{v}_{b,f,m}$ と事前に測定した伝達関数 $a_{f,\theta}$ を利用する. 算出式は次のようになる.

$$P_{b,f,\theta} = \frac{\|a_{f,\theta}a_{f,\theta}^{H}\|}{\sum_{m=N+1}^{M}|a_{f,\theta}\mathbf{e}_{b,f,m}^{H}\|}$$

計算式では, $N+1$ 番目の固有ベクトルから, $N-m$ 個の固有ベクトルを利用する. 周波数ビンの統合は周波数ビン $1,\cdots,F$ に対して, 最大の固有値 $\lambda_{t,f,1}$ の平方根による重み付け和によって行う.

$$P_{b,\theta} = \sum_{f=1}^{F}\sqrt{\lambda_{t,f,1}}P_{b,f,\theta}$$

MUSIC 法の詳細は[Schmidt, 1986]を参照する.

2.4 MSLS 特徴量の計算

本稿では, 話者同定の音声特徴量として 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[Yamamoto *et al.*, 2007].

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 $|V_{f,i}|(f = 0...256)$ を 13 次元の特徴ベクトル **r** 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. 周波数領域で等間隔各成分の窓をかけて, 得られた各成分に対して, 対数値を取って, **h** が得られる.
3. 13 次元のベクトル $h(i)$ を以下のように $r(i)$ 正規化する. $i = 1,\cdots,13$.

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

2.5 GMM のパラメータ学習

GMM による識別は，IVA で分離した事前にラベルを付けた各話者からの 20s 程度の分離音声を学習データとして. 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, μ^l, Σ^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, **r** は音声特徴量ベクトルを指す.

$$Class = \underset{c}{\operatorname{argmax}} \sum_{l=1}^{3} g_c^l N(\mathbf{v}|\mu_c^l, \Sigma_c^l)$$

3 実験データ収録環境

本節では, 実録音対話データからの正解データ作成手順と, 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を説明する.

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 を目標として, 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.

3.1 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 4-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 4-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 4-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その話者の方向も容易に計算が可能である．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.

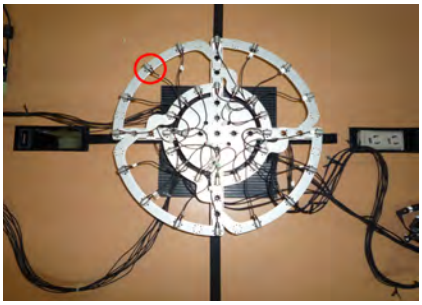
3.2 評価指標

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.

3.2.1 音源 ID を考慮しない場合

音声区間検には, 挿入エラーと削除エラーを考えて. それらは適合率と再現率で定量的に評価する．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，正解データでは音声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

(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.

Figure 4: 実験風景

Table 1: MUSIC スペクトルに基づくベースライン手法の適合値率・再現率評価. 行: MUSIC スペクトルで音が存在すると判定する閾値. 列：MUSIC スペクトル計算時に仮定する音源数. パラメータの変化に伴って, 精度が大きく変わることがみられる.

	1		2		3	
	P	R	P	R	P	R
25	0.541	0.679	0.268	0.770	0.155	0.719
27	0.641	0.621	0.323	0.766	0.155	0.719
29	0.766	0.539	0.457	0.742	0.156	0.719
31	0.827	0.317	0.600	0.667	0.179	0.711

存在する $x_{b,\theta}$ の値が 0 より大きい場合は，音声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音声区間，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，正解データ中の音声区間 $x_{b,\theta} > 0$ 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, 音声区間検出における評価指標は次のように定義される.

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

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

3.2.2　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. [高橋徹 *et al.*, 2009] 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 を考慮した音声区間検出・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. $E_{ID} = \frac{S_c - S_e}{S_c}$

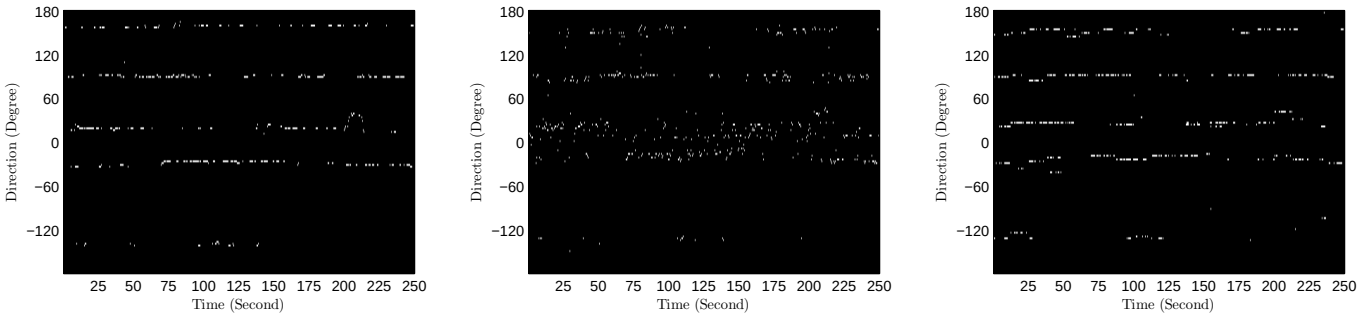
4　実験結果

4.1　ベースライン手法の評価

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 $\Delta_\theta (= 15°)$ 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 1 にまとめる.

4.2　提案手法の評価

IVA 音源分離処理では, 音源数をその場にいた話者数 5 に設定している. 音声区間検出の閾値処理の部分では, T_v を 0.01 に設定して, T_s を 8000 サンプル中の 100 に設定している. 音声区間検出と音源定位の推定結果について, 図 5 で示したように, 評価実験の結果と MUSIC 法による結果の比較を行った. リファレンスデータに対して, 提案手法がより精度の高い結果が得られることがわかった. 数値的な評価に関しては, 図 6 で示したように, 精度の定量的な向上が確認できた. 図 6 の左辺では, MUSIC 法のデータ点が多い理由は閾値と音源数を変えて結果 MUSIC 法を評価



(a) リファレンスデータの音声区間と音源方向　(b) MUSIC 法による推定結果　(c) 提案手法による推定結果

Figure 5: 提案手法と MUSIC 法に基づいたベースライン手法の比較, 提案手法のほうの推定結果が MUSIC 法だけを利用した手法より精度が高いことがわかる.

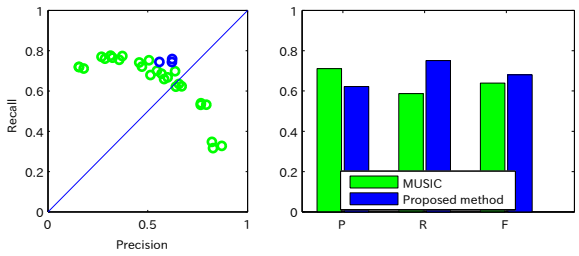


Figure 6: 比較結果の図である. 左辺は適合率ー再現率分布で. 右辺は適合率 (P), 再現率 (R), F 値 (F) 評価の比較である.

した結果をプロットしている. 右辺はそれぞれの手法の適合率, 再現率と F 値評価を示している. 三つの 4 分程度の実録音データに対して, MUSIC 法の評価結果については, 各録音データに対して, F 値の高いほうを選んで各指標の平均を取っている. また, 提案手法の音源 ID の誤推定率 E_{ID} は 0.23 である. 提案手法と MUSIC 法による結果の音源定位誤差が同じく 7.5 度ぐらいとなる.

4.3 考察

実験を通じて, 提案手法はより高い再現率と F 値を示した. しかし, 本手法には次の制約が存在する. (1)IVA 音源分離は音源が動かない前提で分離行列を推定しているので, 本手法の移動音源への対応が必要となる. (2) 音声区間検出の閾値処理だけでは, 環境雑音に対して頑健性が足りないと予想している.

5 まとめ

本稿では, いつ, どこで, 誰が話しているかを推定する話者ダイアライゼーションシステムの構成を述べた. 話者ダイアライゼーション問題は複合的な問題なので, 様々な処理を直列につないで対処するが, 本手法は様々な観測音に対して頑健な IVA を前処理とすることで, 全体のパフォーマンスの改善に寄与している. 評価実験では, MUSIC 法をベースとした手法により音声区間検出と音源定位精度の向上を確認した.

謝辞: 本研究の一部は科研費基盤 (S) の支援を受けた.

参考文献

[Lee *et al.*, 2007] I. Lee, T. Kim, and T.W. Lee. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing*, 87(8):1859–1871, 2007.

[Nakadai *et al.*, 2010] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. Design and implementation of robot audition system'hark'—open source software for listening to three simultaneous speakers. *Advanced Robotics*, 24(5-6):739–761, 2010.

[Nakamura *et al.*, 2011] K. Nakamura, K. Nakadai, F. Asano, and G. Ince. Intelligent sound source localization and its application to multimodal human tracking. In *In Proceedings of the IEEE/RSJ International Conference on IROS*, pages 143–148. IEEE, 2011.

[Ono, 2011] N. Ono. Stable and fast update rules for independent vector analysis based on auxiliary function technique. In *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 189–192. IEEE, 2011.

[Schmidt, 1986] R. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Transactions on Antennas and Propagation*, 34(3):276–280, 1986.

[Tranter and Reynolds, 2006] S.E. Tranter and D.A. Reynolds. An overview of automatic speaker diarization systems. *In Proceedings of the IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2006.

[Yamamoto *et al.*, 2007] S. Yamamoto, J. Valin, K. Nakadai, M. Nakano, H. Tsujino, K. Komatani, T. Ogata, and HG Okuno. Simultaneous speech recognition based on automatic missing feature mask generation by integrating sound source separation. *Journal of the Robotics Society of Japan*, 25(1):92, 2007.

[角康之 *et al.*, 2008] 角康之, 西田豊明, 坊農真弓, and 來嶋宏幸. Imade: 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**, 49(8):945–949, 2008.

[高橋徹 *et al.*, 2009] 高橋徹, 中臺一博, 石井 Carlos 寿憲, Jani Even, and 奥乃博. 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**, 29(1F3-3), 2009.