4T-2

話者ダイアライゼーションシステムのための音声区間検出および到来方向推定の精度向上の検討

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 ‡(株) ホンダ・リサーチ・インスティチュート・ジャパン

1. はじめに

計算機により環境中に存在する様々な音の観測混合音から，それぞれの音源に関する情報を抽出することは，ロボット等が人間と音声コミュニケーションを行うことや，人間の音環境理解の支援のためには重要な技術である．本研究では，観測混合音の中からいつ，どこで，誰が，何を話したかを認識する話者ダイアライゼーション問題を扱う．話者ダイアライゼーションを構成する要素技術は，音声区間検出などが挙げられる [1, 2, 3].

実環境において頑健に話者ダイアライゼーションを行うには，次の 2 点が重要である．

1. 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，
2. 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[1] 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みが考えられる．しかし，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 (図 3) [4]. 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．

本稿では, 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．

2. 問題設定および収録データ

本稿で取り扱う問題の設定は以下のようになる：

入力: マイクロホンアレイで収録した多人数自由発話
出力: 各発話 (active speech segment 図 2) の方向時刻
仮定: 各話者が大きく動かないこと

2.1 データの収録

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 [5] を利用して音源方向の正解データを作成した．[4] 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 値を利用した．

On Improving the Accuracy of VAD and DOA for Speaker Diarization System *Yangyang Huang (Kyoto Univ.), Takuma Otsuka (Kyoto Univ.), Kazuhiro Nakadai (HRI-JP), and Hiroshi G. Okuno (Kyoto Univ.)*

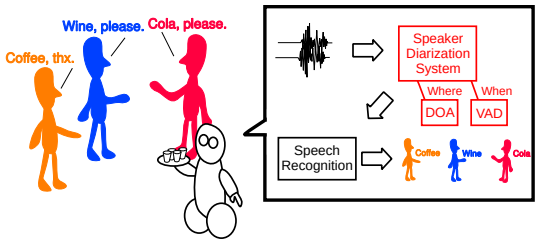

図 1: ロボット侍者にとって, どこの誰が何を注文したかのような要求を把握したい場合に, 話者ダイアライゼーションタスクが必要となる．

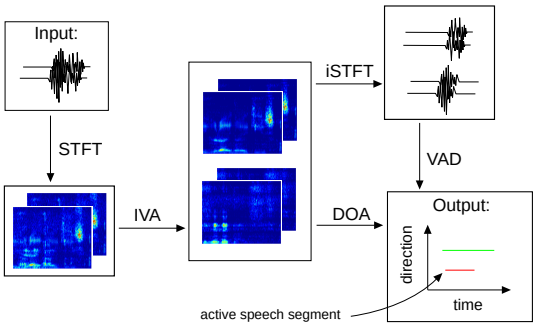

図 2: IVA を前にもって全体の処理の流れ図となる．IVA で分離した分離音声に対して, 音源定位および音声区間検出を行い, 結果を出力する．

3. 手法の概略

3.1 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．ICA にはパーミュテーション問題がある．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，パーミュテーション問題を回避している [6, 7].

3.2 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．

3.3 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

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

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 $\mathbf{e}_{b,f,m}$ と事前に測定した周波数ビン f 方向 θ に対応する伝達関数ベクトル $\mathbf{a}_{f,\theta}$ を利用する．算出式は次のようになる．^{H} は行列のエルミート転置を表す．

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

計算式では，雑音信号部分空間に対応する $N+1$ 番目から M 番目までの固有値に対応する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 [8] を参照する．

4. 実験結果

4.1 MUSIC 法に基づく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．

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 3 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 3 で示したように，精度の定量的な向上が確認できた．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．四つの 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．提案手法と MUSIC 法による結果の音源定位誤差は同じく 7.5 度程度となる．提案手法はより高い再現率を持ちつつ，より高い F 値を示している．

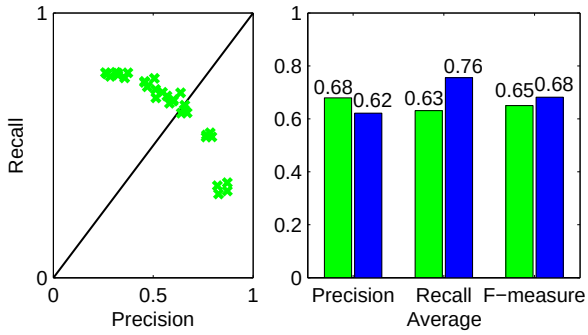

図 3: 左側はベースライン手法の適合再現率の変動．右側は手法の適合率，再現率や F 値を描いた．緑色: ベースライン手法は，青色: 提案手法

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，移動音源を含む環境や，マイクロホンを搭載したロボットが動くなどの一般的な状況への対応が今後の課題である．(2) 音声区間検出の判定は音量の閾値処理を用いているので，非音声などの雑音に対しての頑健性の強化も今後の課題として挙げられる．

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) 24220006 の支援を受けた．

参考文献

[1] K. Nakadai et al. 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.

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

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

[4] 黄ら．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, 2012.

[5] 角ら．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.

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

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

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