E-034

対話グループ推定における音源放射方向の有効性
Effectiveness of Radiation Characteristics of Sound Sources for Dialogue Group Estimation

小川　広夢†　　高橋　徹†
Hiromu Ogawa　Toru Takahashi

1. はじめに

音環境理解研究の多くは、音環境内の音響イベントの属性を推定する課題として定式化されている。音響イベントの属性とは、発音している音源の情報であり、音源の位置、発音時刻、発音内容などが注目され、音源定位、音源検出、音声認識などの技術課題として扱われてきた。

我々は、音響イベントの関係を推定することが、音環境の深い理解に繋がるという信念に基づき、音環境理解を定式化している。本報告では、音源間の関係を推定する具体例として、音源が人である場合を想定し、音源間の関係として対話グループを推定する課題について述べる。この問題は、音環境内の各音源の放射方向の推定問題と、各放射方向から対話グループを推定する問題の２つから成り立っている。前者の問題は、津田ら[1]による研究があり、そちらにゆだね、本報告では、後者について述べる。

2. 対話グループ推定問題の定義

音環境が N 個の音源から構成されているとする。n 番目の音源 $s(n)$が、座標$(s_x(n), s_y(n))$にあり、放射方向 $s_\theta(n)$ ($\in$ [0,360])であるとする。n 番目の音源の属性値を $p(n)=(s_x(n), s_y(n), s_\theta(n))$と表す。また、$p(n)$はクラス $q(n)$ ($\in\{1,2,\ldots,M\}$) に属しているものとする。ただし、クラス数を M とし、m 番目のクラスの重心座標を $c(m)=(c_x(m), c_y(m))$と表す。

$\boldsymbol{P}=\{p(1),p(2),..,p(N)\}$, $\boldsymbol{Q}=\{q(1),q(2),..,q(N)\}$, $\boldsymbol{C}=\{c(1),..,c(M)\}$ とすると、本研究における対話グループ推定問題は、$\boldsymbol{P}$ が与えられたときに、$\boldsymbol{Q}$ と $\boldsymbol{C}$ を求める問題として定式化できる。音源が話者の場合を想定しており、$\boldsymbol{Q}$ は各話者がどの対話グループに所属するかを表す。本研究の目的は、$\boldsymbol{P}$ から真の対話グループ $\boldsymbol{Q}^*$を推定する問題である。

2.1 k-means 法[2]による従来の解決アプローチ

従来複数音源が存在する空間で、音源をグループ分類する問題は、音源座標の近さを基準に用いられてきた。つまり、誤差関数を座標間のユークリッド距離基準により

$$\varepsilon_{\mathrm{Baseline}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C}) = \sum_{n=1}^{N}(s_x(n)-c_x(q(n)))^2+(s_y(n)-c_y(q(n)))^2$$

の $\boldsymbol{Q}$ と $\boldsymbol{C}$ に関する最小化問題

$$\boldsymbol{Q}^*,\boldsymbol{C}^* = \arg\min_{\boldsymbol{Q},\boldsymbol{C}} \varepsilon_{\mathrm{Baseline}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C})$$

と定式化されていた。音源方向が考慮されていないため、対話グループ推定の精度が不十分であった。

Kendon[3]の対話研究によると、対話グループに参加する話者は、向き合う傾向があることが指摘されている。３名以上ではＯ型やＬ型、２名の場合は対峙するＩ型を形成するとされている。これらの研究は、対話グループの形成には、話者の向きがかかわっていることを示唆しており、位置と方向に基づく分類手法の必要性がうかがえる。

2.2 位置と方向に基づく分類手法

我々は、距離誤差$\varepsilon_{\mathrm{dst}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C})$と方向誤差$\varepsilon_{\mathrm{dir}}(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C})$を統合した基準で分類する手法を提案する。以後、$(\boldsymbol{P},\boldsymbol{Q},\boldsymbol{C})$は冗長であるため省略して記述する。我々は誤差関数を

$$\varepsilon_{\mathrm{Prop}} = \varepsilon_{\mathrm{dst}}^{\alpha}\varepsilon_{\mathrm{dir}}^{1-\alpha}$$

と定式化する。ただし、

$$\varepsilon_{\mathrm{dst}} = \varepsilon_{\mathrm{Baseline}}$$

$$\varepsilon_{\mathrm{dir}} = \sum_{n=1}^{N}(s_\theta(n)-\tan^{-1}\frac{s_y(n)-c_y(q(n))}{s_x(n)-c_x(q(n))})^2$$

である。また、$\alpha \in [0,1]$は距離誤差と方向誤差の重みを調整するパラメータである。$\alpha = 1$のとき提案手法は、従来手法と等価となる。

提案手法は、誤差関数$\varepsilon_{\mathrm{Prop}}$の最小化問題になっている。ただし、この最小化問題は、解析的には解くことができない。我々は、２つの仮定を置き最小化問題を近似的に解く方法を提案する。１つ目は、近似$\tan\theta \cong \theta$が成り立つこと、２つ目は、方向誤差最小化フェーズにおいて、クラスタ内の音源は、クラスタの重心から等距離にあると仮定することである。これらの仮定を置くことで、距離誤差と方向誤差を交互に最小化することで、距離誤差と方向誤差を同時に最小化できる。最小化時のパラメータ更新式は、

$$c_x(m) = \frac{\sum_{n\in\{n|q(n)=m\}} s_\theta(n)s_x(n)}{\sum_{n\in\{n|q(n)=m\}} s_\theta(n)} - \frac{\sum_{n\in\{n|q(n)=m\}} s_y(n)-c_y(m)}{\sum_{n\in\{n|q(n)=m\}} s_\theta(n)}$$

$$c_y(m) = \frac{\sum_{n\in\{n|q(n)=m\}} s_y(n)}{N} - \frac{\sum_{n\in\{n|q(n)=m\}} s_\theta(n)(s_x(n)-c_x(m))^2}{N}$$

† 大阪産業大学　デザイン工学部　情報システム学科
Faculty of Design Engineering, Information Systems,
Osaka Sangyo University

である。これらの更新式を用いることで、

$$\boldsymbol{Q}^*, \boldsymbol{C}^* = \arg\min_{\boldsymbol{Q},\boldsymbol{C}} \varepsilon_{\mathrm{Prop}}(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{C})$$

を近似的に解くことができる。

3. 対話グループ推定（提案手法）

3.1 評価用データセット

対話グループ推定精度の評価には、確率的な手法で生成した仮想のデータベースを用いる。実録音に基づくデータベース作成には多大な費用と労力を要するため、本研究では仮想的な音環境、すなわち話者と対話グループを仮想的に配置したものを仮想音環境として評価に用いる。具体的には、話者位置と話者方向を、対話研究の知見[3,4]に基づき、確率的に生成した。

仮想データベースには、各データに 4 から 16 グループ含み，各グループに含まれる話者数は 2 から 8 名とした。30m×20m の仮想平面上に、グループの中心座標が一様分布に従うように生成した。直近のグループが近すぎて重なり合うことを許容した。ただし、直近の壁まで 2 m以上離れるように条件付けした。その後、当該グループ内の話者数を 2 から 8 の一様分布に従うよう生成した。話者の座標は、グループ中心座標から θ 度方向で、かつ、グループ中心座標から γ m の座標に割り当てた。θ は 0 から 360 度の一様分布に従って生成し、γ はガンマ分布（形状母数 4、尺度母数 0.25）に従って生成した。この分布は、0.75m 付近にピークを持つ、正の値をとる分布である。以上のように仮想データを生成し、最終的な仮想データベースは、各グループ数に対して、10 種類ずつの音環境を含み、全 130 の仮想音環境で構成されている。

仮想データは確率的に生成されているため、生成された話者の座標と話者の向きが与えられている。また、本研究では、生成過程から各話者の所属グループは明らかである。我々は、これを真の所属グループとして扱う。

3.2 推定精度の定義

本研究では、クラスタリングの性能を話者のクラスタリング精度で評価する。音環境中の各話者が、各手法でクラスタリングされた後、正しいクラスに分類されている状態を正解とする。音環境中の全話者数を N，正しいクラスに分類された話者数を S とする．本研究におけるクラスタリング精度(Acc)は、

$$A_{\mathrm{cc}} = \frac{S}{N} \times 100$$

と定義される．

3.3 推定精度の比較実験

$\varepsilon_{\mathrm{Baseline}}$の最小化手法を Baseline と呼び、$\varepsilon_{\mathrm{Prop}}$の最小化手法（提案法）を Prop と表し、推定精度を比較する。Prop を実行するにあたって$\alpha = 0.25$とした。

図 1 に各手法で分類した時の推定精度を示す。横軸は、仮想データベース内の音環境に含まれるグループ数である。各手法ともグループ数は既知であるとして分類している。黄色バー（左）が、従来手法(Baseline)の精度を表し、青色バー（右）が提案手法(Prop)の精度を表している。グループ数に依らず、いずれの条件においても提案手法の平均精度は従来手法よりも高い。この結果から、話者の座標と方向情報を用いた分類（提案手法）が、話者の座標のみを用いた分類（従来手法）より高い精度を達成可能なことを確認できる。この結果は、話者の方向情報が対話グループ推定において有効であることを示唆している。

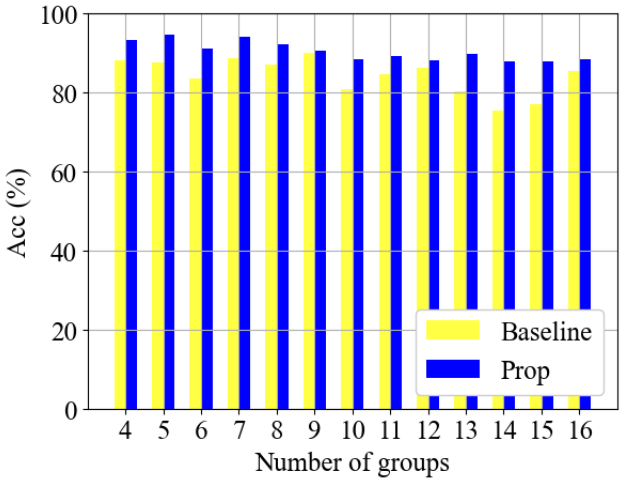

図 1　平均推定精度

4. おわりに

本稿では、対話グループ推定における音源放射方向の有効性を確認するシミュレーション実験について述べた。音源位置のみを用いた話者分類手法に比べて、音源の位置と方向情報を用いた話者分類手法が、高い対話グループ推定を実現することを実験的に示した。実験結果から、音源方向情報が対話グループ推定に有効であることを示唆した。また、我々は、音源位置と音源の方向を扱う分類のための目的関数を提案した。目的関数は、解析的には解けないため、変分近似と目的関数の近似による更新式を導出した。

今後、実対話音声から求めた話者位置と話者方向を用いた評価を実施する必要がある。一方で、実データを大量に収集することは困難であるため、継続してシミュレーション実験による解析も必要である。特にパラメータについての検討が必要である。話者の配置を確率的に与えているため、多数の分布のパラメータが存在しおり、検討の余地がある。結合重みαの最適化は最も重要であると考えられる。

また、全 130 条件実施したシミュレーションの規模も、パラメータの自由度と比較して十分ではないという課題も残っている。

謝辞

本研究は JSPS 科研費 JP21H03488 の助成を受けた。

参考文献

[1] Tsuda Ryusei, et al., “Quasi-real-time estimation of a loudspeaker direction from sound pressure level ratio among four channels”, Proc. Internoise 2023, (2023).
[2] Allen Gersho, Robert M. Gray, “Vector Quantization and Signal Compression”, Kluwer Academic Publishers, (1991).
[3] A. Kendon, “Conducting Interaction-Patterns of Behavior in Focused Encounters”, Cambridge University Press, (1990).
[4] 橋本, 他, 実験による対人距離からみた心理的領域の平面方向の拡がりに関する考察日本建築学会計画系論文集, 61(485), pp,|.135-142, (1996).