7Q-01

複数スマートフォンで収録された会話音声のための グループ数推定機能を有する対話グループ検出

荒毛　祐紀†　　岩野　公司†
東京都市大学†

1. はじめに

複数のスマートフォン端末で録音された大量の多人数会話音声をクラウド上に集積し，処理を行うことができれば，高精度な会議音声認識や新しい会議可視化システムなどの実現が可能になる[1]．このような環境では，蓄積された音声データのうち，「どの音声とどの音声が同じグループでの会話か」を自動的に推定する「対話グループ検出」の技術が必要となる．

我々はこれまでに，クラスタリングに基づく対話グループ検出手法の提案を行っている[2]．この手法では，各端末で録音された音データについて，一定間隔で特徴量を抽出して連結することで高次元ベクトルに変換し，それらを k-means 法でクラスタリングすることで対話グループの同定を行う．しかし，対話グループ数を既知としていたため，実用を考えると，対話グループ数が未知の場合に対応する必要がある．

そこで本研究では，ベイズ情報量規準(BIC)に基づく対話グループ数の推定機能を有した対話グループ検出手法を提案する．また，性能改善のためのクラスタリングの高精度化についても検討を行う．

2. 提案する対話グループ検出手法

2.1 対話グループ検出の流れ

図 1 に，本研究で提案する対話グループ検出の流れを示す．

まず，蓄積された各音データの冒頭 60 秒について 1 秒間隔で音響特徴量を抽出し，それらを時間順に連結して高次元ベクトルを作成する．音響特徴量には，「12 次元 MFCC とその 1 次微分成分，対数パワーの 1 次微分成分」の計 25 次元を使用するため，高次元ベクトルの次元数は 1,500 となる．次に，検出対象となる全データのベクトルを使用して主成分分析を行い，得られた固有空間を用いて高次元ベクトルを m 次元に圧縮する．次元圧縮したベクトルを特徴量として対話グループ数の推定を行い，その結果を利用したクラスタリングを行うことで最終的な対話グループを決定する．

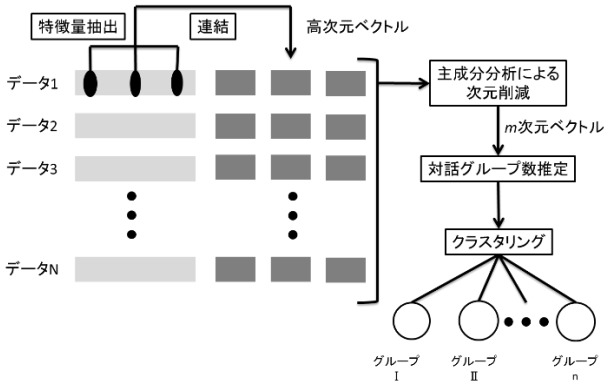

図 1. 提案する対話グループ検出の流れ

2.2 対話グループ数の推定

対話グループ数は，ベイズ情報量規準(BIC)に基づく x-means 法[3]を用いて推定する．この手法では，まず，初期値として十分小さなクラスタ数を与え，蓄積データを k-means 法でクラスタリングする．次に，各サブクラスタについて 2 分割すべきかどうかを式(1)の BIC によって判断する．

$$\mathrm{BIC} = -2\log L\left(\hat{\theta}_i; x_i \in C_i\right) + q\log n_i \qquad (1)$$

ここで，C_i は i 番目のクラスタ，x_i は C_i に含まれるデータである．L は尤度関数であり，データ x_i に対して仮定した多次元正規分布のモデルパラメータ $\hat{\theta}_i$ に対する x_i の尤度を示す．また，q はパラメータ空間の次元数，n_i は C_i に含まれるデータ数である．式(1)の第一項は現状のクラスタがデータをどれだけ精密に分類できているかを表す指標であり，第二項はデータに対し過度に細かいクラスタにならないように抑制するためのペナルティである．各クラスタについて k-means 法によって 2 分割した前後の BIC を計算し，分割後の BIC が大きくなる場合にはクラスタ数を増やし，そうでない場合には分割を止める．なお，k-means 法では，初期クラスタに対する最初のデータの割り当てがランダムで行われるため，実行の度に推定グルー

Dialogue group detection with automatic estimation of the number of groups for conversational speech recorded by multiple smartphones
Yuki Arake†, Koji Iwano†, †Tokyo City University

プ数が変化する可能性がある. そこで, 20 回推定を実行した上で, 多数決で最終的な対話グループ数を決定する.

2.3 クラスタリング手法の変更による改善

2.2 節の手法により，グループ数推定と同時に k-means 法によるクラスタリングも行われるが，提案手法では得られたグループ数に基づいて異なる手法による再クラスタリングを行うことで高精度化を図る．具体的には，k-medoid 法に基づく PAM 法[4]を用いる．k-medoid 法はクラスタの代表点を「クラスタ内の点のうち，他の点との距離の総和が最小となる点」とする手法で，セントロイドを平均で与える k-means より計算量が大きくなるが，ノイズや外れ値に対する耐性が高いことが報告されている．

3. 性能評価

3.1 実験データ

実験には合計 15 対話グループの音響データを使用する．各グループあたり，平均 4.0 人が参加しており，音データの総数は 60 となる．なお，次元数 m は事前の実験により 15 と設定した．

3.2 評価実験の結果

図 2 に，初期クラスタ数を変化させながら，200 回ずつ対話グループ数推定を行った際に，正しいグループ数（15）を判定できた割合を正解率として示す．実験の結果，初期クラスタ数を 3 以上に設定すればグループ数を 100%正しく推定できることが確認された．

図 3 に，推定された正解グループ数 15 を与えて k-means 法と PAM 法でクラスタリングを行った場合の，対話グループ検出性能の比較を示す．性能はクラスタの純度（Purity）で表す．結果として，PAM 法により性能が向上し，100%正しくグループを推定できることが確認された．

4. まとめ

本研究では，対話グループ数を未知とした場合でも対応可能な，対話グループ検出手法の提案を行った．提案手法を評価した結果，グループ数が正しく推定され，対話グループが 100%正しく検出されることが確認された．今度の課題としては，随時アップデートされる音データへの対応や，実システムの開発の検討などがあげられる．

謝辞 本研究の一部は JSPS 科研費 基盤研究（B）25280058 の助成を受けたものです.

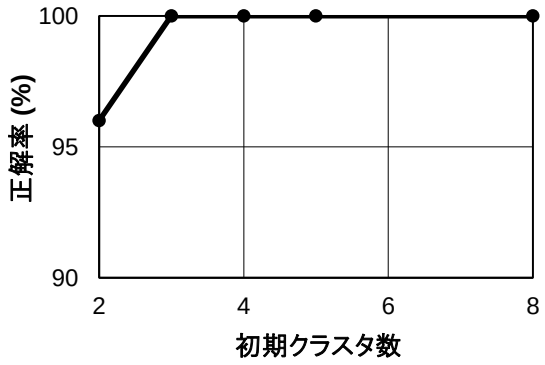

図 2. 初期クラスタ数の変化に対する対話グループ数の正解率

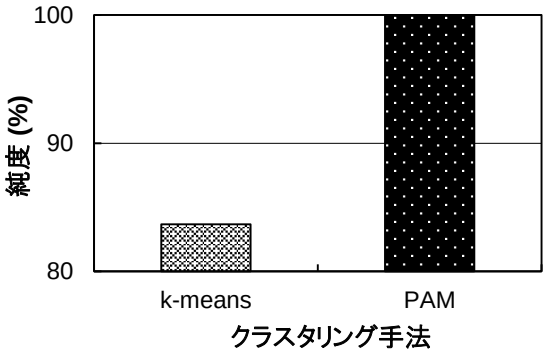

図 3. 対話グループ検出性能の比較

参考文献

[1] 秋葉ら，“クラウド時代の新しい音声研究パラダイム，”情報処理学会研究報告，vol. 2012-SLP-92, vo.4, pp.1-7, 2012.

[2] 岩野ら，“複数スマートフォンで収録された多人数会話音声における対話グループ検出と話者決定，”信学技報, vol. 114, no. 151, pp. 47-52, 2014.

[3] D. Pelleg, and A. Moore, “X-means: Extending K-means with Efficient Estimation of the Number of Clusters,” Proc. ICML2000, pp. 727-734, 2000.

[4] L. Kaufman, P. J. Rousseeu, “Partitioning Around Medoids (Program PAM),” in Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience, New Jersey, Chapter 2, pp. 68-125, 2005.