5S-4

複数スマートフォンで収録された多人数会話音声の話者決定

米山　修平†　　齋藤　かの子†　　岩野　公司†
東京都市大学†

1．　はじめに

近年，会話や会議の音声から話者情報や発声内容などを自動的に抽出し，議事録の自動作成などに役立てることが期待されている．そのためには，「いつ誰が話したか」という情報を抽出する技術「話者決定」が必要となる．

一方，現在は 1 人が 1 台以上の携帯情報端末を所有することが通常となり，会議などの多人数会話の音声を参加者それぞれの端末で同時に録音することが容易になった．また，インターネットを介してこれらの複数データを即座に集約し，利用することができるクラウド環境も技術的に実現可能である[1].

そこで本研究では，まず，上述のような環境を想定し，多人数会話を各参加者が所有するスマートフォンで同時収録した音声データベースを構築する．次に，このような環境で収録された多人数会話音声の話者決定手法を複数提案し，構築データベースを用いて評価する．

2．　多人数会話音声データベースの構築

データベースの構築のため，スマートフォン（iPhone, Android）を利用して，合計約 7 時間の多人数会話音声データを収録した．参加者の総数は 12 名（男性 6 名，女性 6 名）で，大学内の研究室や教室において，「雑談」「会議」「発表」の 3 つのタスクで会話を行った．それぞれのタスクについて 5 セッションの会話を実施し，セッションあたりの参加者は平均で 4.0 名である．スマートフォンは各参加者の目前のテーブルの上に置くよう統一した．同一会話中で時間同期をとることを可能にするため，会話の冒頭にはベルの音が収録されている．また，収録データに対し，「参加者各自の発声区間」「雑音」「無音」といった音響イベントの正解時間ラベルを聴取によって人手で付与している．

Speaker diarization for multi-party conversational speech recorded by multiple smartphones
†Shuhei Yoneyama, Kanoko Saito, Koji Iwano, Tokyo City University

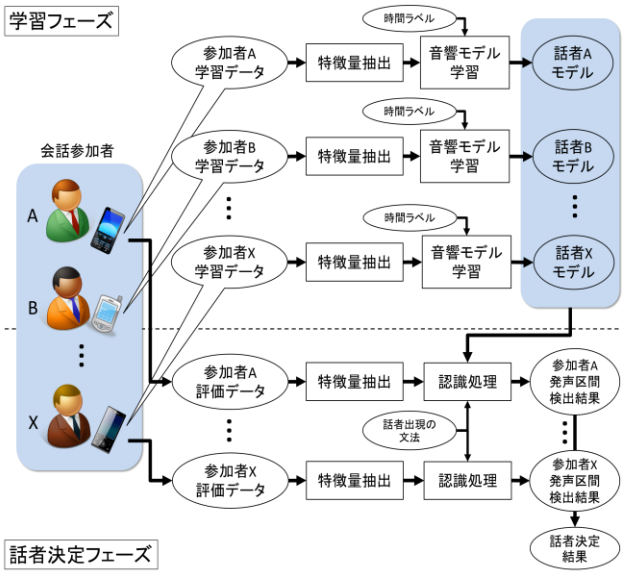

図 1　多人数会話音声の話者決定の流れ

3．　多人数会話音声の話者決定手法

3.1　話者決定の流れ

図 1 に，本研究で提案する多人数会話音声の話者決定の流れを示す．

学習フェーズにおいて，会話参加者の話者モデルの学習を行う．その際，会話参加者の端末ごとに収録された音響データを学習データとして，その端末所有者の話者モデルと無音モデルを作成する．学習データには，「所有者の発声区間」「無音区間」を示した時間ラベルを各種方法で付与し，学習に利用する．各話者モデルは 3 状態の隠れマルコフモデル（HMM）で作成し，音響特徴量には，MFCC12 次元＋ΔMFCC12 次元＋Δ対数パワーの合計 25 次元ベクトルを利用する．

話者決定フェーズでは，端末ごとの入力評価データに対し，その所有者の発声している区間の検出を行う．具体的には，学習フェーズで得られた全参加者の話者モデル，当該端末の学習データで構築された無音モデル，話者が自由な順序で出現することを許した話者出現文法を利用し，音声認識と同様の Viterbi 探索に基づく計算を行って最尤となるモデル系列を求め，所有

者の話者モデルに割り当てられた区間を所有者の発声区間として検出する．

3.2　モデル構築手法

話者モデルの作成方法として，異なる使用状況を想定した以下の手法を検討し，それぞれの性能比較を行う．

3.2.1　話者情報を事前利用

話者決定対象となる会話を行う時点で，参加者の話者モデルが準備されていない状況を想定する．したがって，話者モデルの学習データには対象会話音声そのものを利用する．学習データに対する時間ラベルは，「各端末の音響信号には，所有者の発声が比較的大きな音量で収録されている」という仮定のもと，波形パワー値のしきい値処理で作成する．具体的には，各端末の音響データごとに波形パワー値を時系列で抽出し，平均 0，分散 1 となるように標準化した上で，しきい値がθ_v 以上となる区間を所有者の発声区間，θ_s 以下になる区間を無音区間としてラベリングする．θ_v, θ_s は予備実験によりそれぞれ 0.5，－0.5 とした．

3.2.2　話者情報を事前利用

端末ごとに所有者の発声が事前に録音・蓄積されており，話者モデルの事前作成が可能な状況を想定する．本研究では，音素バランス文 50 文の読み上げ音声と，「会議」「発表」タスクに含まれる所有者の単独発声部分（1 話者あたり約 2 分）を，事前に録音してあるデータとみなし，それを用いて話者モデルの構築を行う．時間ラベルは聴取によって付与した．

3.2.3　環境適応化

評価データの所有者以外の話者モデルについては，他人の端末で収録された音声で学習されているものであるため，収録環境の音響特性が異なる．そこで，評価用音声を用いた教師なしモデル適応を行う．適応化は，最尤線形回帰（MLLR）法[2]を利用する．

4．　話者決定性能の評価実験

4.1　実験条件

評価データには，「雑談」タスクの収録データ（約 2 時間）を利用する．話者決定性能を以下の 4 通りの手法に対して実施し，性能の比較を行った．なお，評価は所有者発声区間のフレーム単位の検出率（F 値）で行う．

- 手法 A: 話者モデルを使用せず，3.2.1 節で説明した波形パワー値のしきい値処理のみで所有者発声の区間を決定するもの．
- 手法 B: 3.2.1 節で述べた，事前情報を利用しない話者モデル構築法を用いたもの．

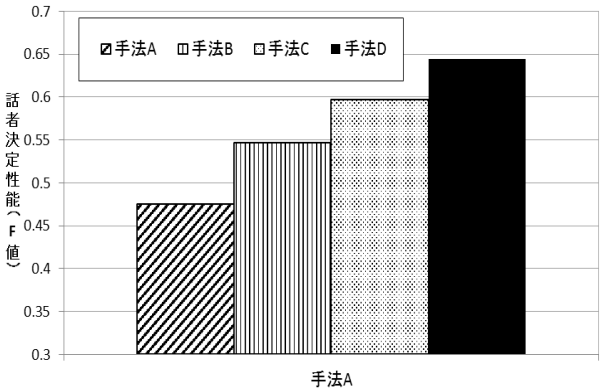

図 2　話者決定手法による性能比較

- 手法 C: 3.2.2 節で述べた，事前情報を利用する話者モデル構築法を用いたもの．
- 手法 D: 手法 C で構築されたモデルに対し，さらに 3.2.3 節の環境適応化を行ったモデルを利用するもの．

4.2　実験結果

図 2 に 4 手法の話者決定性能を示す．それぞれの検出性能は 47.5，54.7，59.7，64.4%となった．手法 A と B の比較から，波形パワー値のみによる話者決定より，その情報に基づいて学習した話者モデルの利用による性能向上が確認された．しかし，手法 C が B より良好なことから，事前のデータ収集の効果がそれ以上に大きいことがわかる．また，手法 C と D の比較から，モデル適応による環境適応効果が大きいことが確認された．

5．　まとめ

本研究では，複数スマートフォンで収録された多人数会話音声に対し，様々な状況を想定した話者決定手法の提案を行い，その評価を行った．その結果，会話に先立って収録した各所有者の音声を利用することの重要性や，モデルの環境適応化の有効性が確認され，最終的に 64.4%の話者決定性能が得られた．今後は，学習データの増強や，笑い声・重複会話区間への対策を講じることで，更なる性能の改善が望まれる．

参考文献

[1] 秋葉他，“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, no.4, pp. 1-7, 2012.

[2] C.J. Leggetter and P.C. Woodland, “Maximum likelihood liner regression for speaker adaptation of continuous-density hidden Markov models,” Computer Speech and Language, vol.9, pp. 171-185, 1995.