7Q-02

複数スマートフォンで収録された会話音声の相互スペクトル減算を用いた話者決定

小平　優希†　　篠田　浩一‡　　岩野　公司†

東京都市大学†　　東京工業大学‡

1. はじめに

我々は，参加者各自が所有する複数のスマートフォン端末で録音された多人数会話音声に対し，各参加者の発声区間を推定（話者決定）する手法の提案を行っている[1]．従来研究[1]では，端末ごとに事前収録した所有者の単独発声により各参加者の発声モデル（話者モデル）と無音モデルを構築し，話者決定時にはそれらを利用して，対象会話音声に対する最尤モデル系列を探索することで，端末ごとに所有者の発声区間を推定する手法を提案している．しかし，その検出性能は6割以下に留まっており，更なる性能改善が望まれている．性能が不十分な要因の1つとして，対象音声には，事前収録音声と異なり，他者音声の混入が生じていることが挙げられる．

そこで本研究では，対象音声に「相互スペクトル減算（CCSS: Cross-channel spectral subtraction）[2]」を適用して他者の声を取り除き，さらに，そのデータで再学習された話者モデルを利用する話者決定手法を提案し，その性能を評価する．

2. 相互スペクトル減算

相互スペクトル減算（CCSS）[2] は，多人数会話において参加者各自が所有するマイクロフォンで収録された各音声信号から，他者の音声信号を取り除くのに有効な手法である．

会話参加者およびマイクの数を N とするとき，i 番目（i = 1, 2, …, N）の参加者の所有するマイクで観測される信号の時刻 t における短時間スペクトル分析の結果を $X_i(f,t)$ とする．このとき，マイク i で観測される参加者 i の音声のスペクトルを $Y_i(f,t)$ とすると，その推定値 $\hat{Y}_i(f,t)$ のパワースペクトルは式(1)で求められる．

$$|\hat{Y}_i(f,t)|^2 = |X_i(f,t)|^2 - \sum_{j\neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j(f,t)|^2 \quad (1)$$

Speaker diarization using cross channel spectral subtraction for conversational speech recorded by multiple smartphones
Yuki Kodaira†, Koichi Shinoda‡, and Koji Iwano†,
†Tokyo City University, ‡Tokyo Institute of Technology

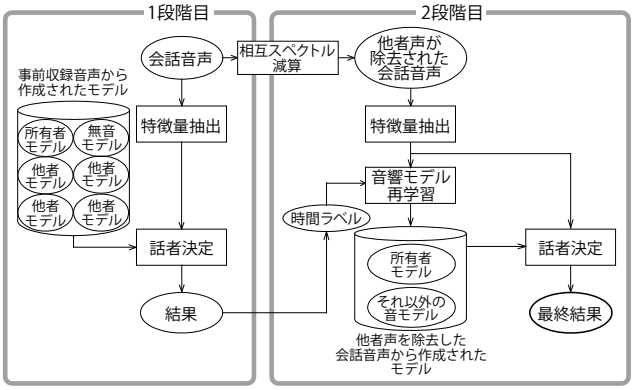


図1　提案する話者決定手法の処理の流れ

ここで，$H_{ij}(f,t)$ は話者（マイク）i, j 間の伝達関数であり，話者 j のみが話している時刻における観測スペクトルX_iとX_jの比から推定される．その際，伝達関数は時刻に対して連続的に変化すると考えられるので，式(2)のように逐次更新を行って推定する．

$$|\hat{H}_{ij}(f,t)|^2 = \rho_h |\hat{H}_{ij}(f,t-1)|^2 + (1-\rho_h)\frac{|X_i(f,t)|^2}{|X_j(f,t)|^2} \quad (2)$$

$\rho_h \in [0,1]$は忘却係数である．最終的な分離信号は，式(3)に基づく n 回の反復推定により求められ，α_nは各反復の減算係数である．

$$|\hat{Y}_i^{(n)}(f,t)|^2 = |X_i(f,t)|^2 - \alpha_n \sum_{j\neq i} |\hat{H}_{ij}(f,t)|^2 |\hat{Y}_j^{(n-1)}(f,t)|^2 \quad (3)$$

3. 相互スペクトル減算を用いた話者決定手法

提案する話者決定手法の流れを図1に示す．

1段階目では，端末ごとに事前収録された音声を学習データとして「端末所有者の話者モデル」と「無音モデル」を構築する．音響特徴量には「12次元 MFCC＋12次元ΔMFCC＋Δ対数パワー」の25次元ベクトルを利用し，モデルには3状態の隠れマルコフモデル（HMM）を利用する．次に，各端末で収録された会話音声に対し，端末

ごとに所有者の発声区間の推定を行う．所有者以外の話者モデルを各端末から集め，音声認識と同様の探索手法で最尤モデル系列を推定することで，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．この第1段階は従来手法[1]と同様である．

第2段階では，会話音声にCCSSを施し，各端末の収録音から他者の音声を取り除く．1段目の検出結果を正解時間ラベルとして，端末ごとにCCSS処理後のデータを用いてモデルの再学習を行う．CCSS処理後の音声には他者の音声はほとんど含まれておらず，所有者発声区間以外は無音に近い状態になると考えられるため，複数の他者モデルと無音モデルは1つの「所有者発声以外のモデル」として融合し，「所有者の話者モデル」との2モデルで話者決定を再実行して，最終結果を得る．

4. 評価実験

4.1 実験条件

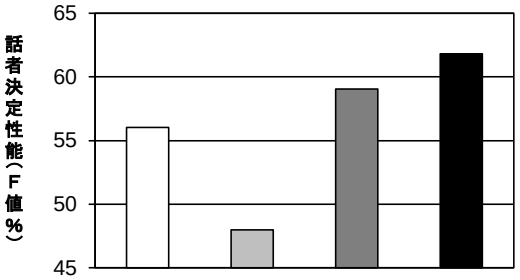
評価には，従来研究[1]においてスマートフォンで収録した雑談5セッションの音声データを使用する．1セッションあたりの平均参加人数は4.0，時間長は5〜10分である．1段階目で使用するモデルの学習データ（事前収録音声）には，音素バランス文の読み上げ音声（1人につき50文）を使用した．なお，CCSSでは $\rho_h = 0.98$ として伝達関数の更新を行い，式(3)の反復回数 n は2 $(\alpha_1 = 1, \alpha_2 = 4)$ として，分離信号を求めた．

比較のため，「CCSSの使用」「話者モデルの再学習」「他者モデルの融合」の有無のパターンが異なる実験を複数行った．全て行わないものが「従来手法」，全て行うものが「提案手法」となる．比較対象として用意したものは，順に「有，無，無」としたものと，「有，有，無」とした2種類で，それぞれの処理の効果を確認することができる．

4.2 提案手法の評価結果

各実験における，各話者の発声区間検出性能をフレームあたりのF値で算出した．図2に結果を示す．実験の結果，従来手法に比べ提案手法により約6%の検出性能の向上が確認され，提案手法の性能は61.8%となった．

CCSSのみを行い，モデル再学習と他者モデルの融合を行わない場合には，従来手法からの性能劣化が確認された．これはCCSSにより，所有者の発声区間においてスペクトル減算による歪みが発生し，CCSS後の音響データと第一段階で構築されたモデルとの間に隔たりが生じている



CCSS使用	無	有	有	有
再学習	無	無	有	有
他者モデル融合	無	無	無	有

（従来手法）　（提案手法）

図2　話者決定性能の比較

ことが原因であると考えられる．その隔たりを解消するために，モデルの再学習を行うことで，約11%の性能向上が見られ，従来手法の性能を3.0%上回ることができる．さらに，他者モデルを融合することで，2.8%の性能向上が得られることも分かる．これは，CCSS後の他者単独の発声区間が無音に近づき，元々の無音区間を含めて同じ音響現象としてモデル化することが可能となることを示唆している．

なお，モデルの再学習の代わりに，MAP推定[3]による教師無し適応を用いた実験も行ったが，再学習に比べて十分な性能は得られなかった．

5. まとめ

本研究では，複数スマートフォンで収録された会話音声を対象とした，相互スペクトル減算を用いた話者決定手法を提案し，性能評価を行った．その結果，従来手法に比べ，提案手法による検出性能の改善が得られ，その有効性を確認することができた．今後は，深層学習の利用による話者モデルの高精度化や，発言数が極端に少ない話者に対する対処手法の検討などにより，さらなる話者決定性能の改善が望まれる．

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．

参考文献

[1] 岩野他，“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.
[2] Y. Nasu et al., “Cross-channel spectral subtraction for meeting speech recognition,” Proc. ICASSP, pp. 4812-4815, 2011.
[3] C.H. Lee and J.L. Gauvain, “Speaker adaptation based on MAP estimation of HMM parameters,” Proc. ICASSP93, pp.558-561, 1993.