7N-03

教師ありモノラル音声分離のための残響音声データ内の単一話者区間を活用した転移学習

生嶋　竜実 †　　武田 龍 ‡　　駒谷 和範 ‡

† 大阪大学 工学部電子情報工学科　　‡ 大阪大学 産業科学研究所

1. はじめに

音声分離とは，複数の話者の同時発話音声から，それぞれの話者の音声を抽出する技術である．例えば，会議の話者別の議事録の書き起こしといった，複数人が話している状況の音声認識などで必要な処理である．実際の録音データはモノラル音声である場合や，室内残響が含まれることが多い．そのような環境でも音声分離が動作することが必要である．

近年，残響を含まない音声（クリーン音声）の分離には深層学習に基づく手法が盛んに研究されている [1]．一方，残響を含む音声（残響音声）の分離はあまり取り組まれていない．Maciejewski らは，シミュレーションで作成したデータで音声分離を行うモデル（分離モデル）を学習させ，残響音声に対する分離と残響除去を同時に行っている [2]．しかし，モデル構築に必要な量の残響音声を収録するには手間がかかる．そのため少量のデータでも残響音声用の分離モデルを構築できる必要がある．

本研究では残響音声の単一話者区間を活用して音声分離モデルを転移させ，同時発話区間における残響音声の分離を行う．図 1 に本研究での転移学習の概要を示す．まず分離対象の音声の単一話者区間を用いて擬似的に混合音を作成し，転移学習用データの入出力ペアとする．次に，作成した少量のデータを用いてクリーン音声の分離モデルを残響音声用の分離モデルに転移させる．この分離モデルを用いて同時発話区間の音声を分離する．転移学習の有効性を残響音声の分離実験で確認した．

2. 音声分離モデルと学習用データ

2.1 モノラル音声分離モデルと教師あり学習

深層学習に基づく教師あり手法がクリーン音声の分離で高い精度を示している．大量の音声データを学習に用いることで，高精度な音声分離モデルを構築できる．特に，残響を含まない混合音声の時間波形を入力，各話者のクリーン音声の時間波形を出力とした end-to-end モデルが主流である [1]．このモデルは，時間周波数領域モデルで対応が必要だった位相情報を考慮する必要がなく，歪みの少ない自然な音声を出力できる利点がある．

残響を含む音声の分離については，分離と残響除去を統合的に行う手法が提案されている [2]．大量のクリーンな各単一話者音声を教師データに，残響を含む混合音声を入力データとすることで，end-to-end の分離・残響除去モデルが構築できる．シミュレーションで残響音声を大量に生成することで，様々な残響音声を学習に用いている．この手法でもある程度高い分離精度を示している．

2.2 音声分離タスクにおける転移学習

転移学習とは，あるタスクの学習済みモデルを他のタスクに適用させる技術である．十分な量のデータが確保できないタスクにおいて，事前の学習で得られた知識を

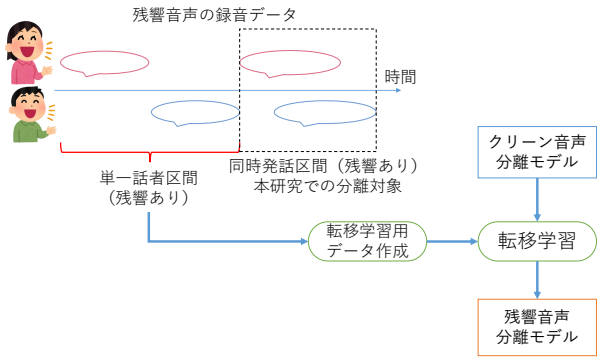

図 1: 単一話者区間を活用した転移学習の概要

目的のタスクでの学習に適用することでモデル精度が向上する．

音声分離タスクにおいて，転移学習は知る限りは使われていない．そのため，その適用方法には工夫の余地がある．例えば，従来手法 [2] に則って転移学習を行う場合，入力データに残響混合音声，教師データに各単一話者のクリーン音声が必要である．しかし，実際には適応対象の録音データはクリーンな音声ではない．モデルの入出力や転移学習に用いるデータの検討が必要である．

3. 分離対象音声を用いた転移学習

本研究では録音データにおける同時発話区間中の音声データを分離するモデルを構築する（図 1）．ここで，従来のように分離と残響除去を同時には行わず，残響を含む音声を話者毎に出力するモデルを扱う．実際の録音ではクリーンな音声を得ることは難しく，分離と残響除去は同時に行うモデルを直接的に構築できないからである．自然な音声の分離を狙い，end-to-end 分離モデルの一つである Dual Path RNN-TasNet (DPRNN) [1] を用いた．

転移学習を適用するにあたり，分離対象となる音声データは次の性質を持つと想定する．まず分離対象の音声は特定の数名による対話に限定し，データ内で話者が変化しないとする．また対象の音源は大きく移動しない，つまり，データ内での残響特性の変動は小さいとする．例えば会議の録音データの場合，参加者の数は途中で変わらず，あまり移動もしないため残響も変化しない．これらの想定により，学習すべき残響や話者のパターンを限定でき，対象音声への分離モデルの転移が容易になる．

単一話者区間における残響音声を用いて転移学習用データを作成する．一定の長さの分離対象の音声から人手で単一話者区間を切り出すことで，音声を準備する．普通は対話音声の中に単一話者区間が含まれ，数分程度であれば人手で切り出せる．単一話者音声を混合した残響音声と単一話者それぞれの残響音声を転移学習データの入出力ペアにする．混合する音声の組み合わせはランダムに行い，組み合わせパターンを変えることでデータ

Transfer Learning Using Single Speaker Sections in Reverberant Speech Data for Supervised Monaural Speech Separation: *Tatsumi Ikushima, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

	ベースライン		提案手法
	転移学習なし	事前学習なし	
SDR	0.03	4.03	**5.22**
SDRi	-0.17	3.83	**5.02**

表 1: 音声分離精度の比較 (dB)

量を増やす．分離に必要な音声の時間は実験で確認する．

4. 評価実験

4.1 実験設定

分離対象の音声を活用した転移学習の有効性を確認する．クリーンな音声の分離モデルを作成して訓練済みモデルとした．英文の新聞記事読み上げコーパス (WSJ0) の 2 話者を混合させ，音声分離タスク向けに整備されたデータセット (WSJ0-2mix) を学習に用いた．サンプリング周波数は 8,000 [Hz] で，学習用データの総量は約 30 時間，訓練用データでは約 10 時間である．

単一話者区間に切り分けた後を想定して，仮定に沿った残響音声データを作成した．まず，WSJ0 から 2 話者の音声データを抜粋した．サンプリング周波数は同じく 8,000 [Hz] である．次に，部屋とマイクの位置が同じで音源位置が異なる 2 つの残響インパルス応答を各話者の音声に畳み込んで混合し，混合音声データとした．80% を学習用データに，残り 10%ずつを検証用とテスト用データにした．このとき学習用データは混合の組み合わせを変えてデータ数を増やしたため，学習用データが約 10 分，検証用データとテスト用データはそれぞれ約 1 分になった．

また，転移学習に用いるデータ量と分離性能との関係を調査した．約 10 分，5 分，2.5 分，1 分の学習用データで転移学習を行い，それぞれの分離精度を比較した．検証用やテストデータ，使用した訓練済みモデルは転移学習の有効性を確認した実験と同様とした．

学習の損失関数は SI-SDR (Scale Invariant Signal-to-Distortion Ratio) [3] を，最適化には Adam を用いた．実装には機械学習ライブラリの PyTorch や教師あり音源分離ライブラリの Asteroid [4] を用いた．

4.2 実験結果

転移学習に関する実験結果を表 1 に示す．それぞれの値は各分離結果における，単一話者音声と混合音声の平均 SDR (Signal-to-Distortion Ratio) と，SDR の混合音声からの改善量 (SDRi) を示している．転移学習を行わない場合や分離対象となるデータのみで学習した場合と比べ，分離対象の音声で転移学習を行った場合が最も数値の高い結果となった．

転移学習を行ったときの分離音声のスペクトログラムの比較を図 2 に示す．横軸は時間，縦軸は周波数ビンである．分離音声を聴いて確認した際，主観的には他方の話者の声はあまり強くは聞こえなかった．完全ではないが転移学習を用いた音声分離はできている．一方で音声が原信号よりも歪んでいるように感じられた．スペクトログラム上でも全体的には音声は分離できているように見えるが，周波数ビンの 1000〜2000 [Hz] 周辺など一部の残響が消えており出力音声が歪んでいる．

転移学習用のデータ量を変更した際の分離性能を図 3 に示す．データ量 0 分は転移学習を行わない場合の結果である．出力の歪みの平均の改善量 SDRi を示している．

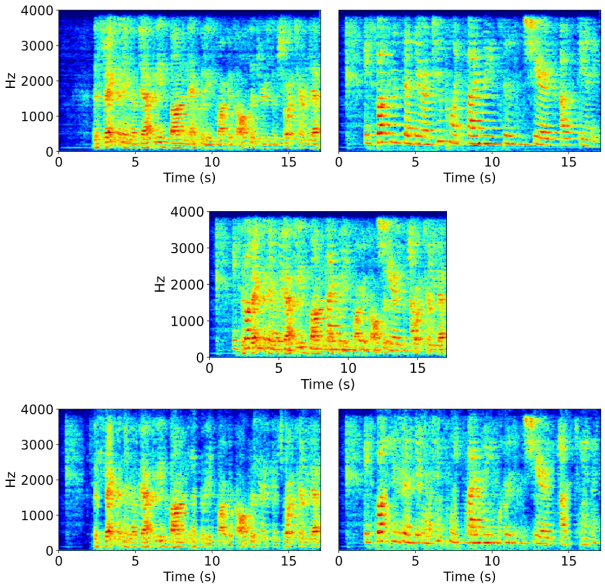

図 2: 単一話者音声（上段），混合音声（中段）および分離音声（下段）のスペクトログラム

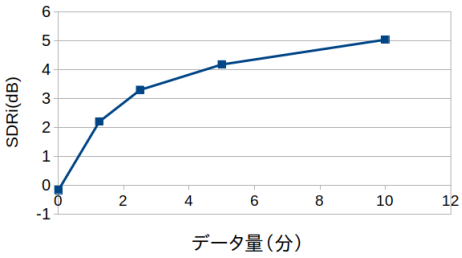

図 3: 学習用データ量と分離精度

数分程度の少ないデータでも転移学習が動作することがわかった．また，データ量が 5 分までの結果を比較したとき分離性能の向上は顕著である一方，10 分と 5 分を比較した際は増やしたデータの量に対して性能向上幅が小さい．転移学習用データには限りがあるので性能向上のためにさらにデータを増やすのは難しい．限られたデータでも分離性能を向上する必要がある．

5. まとめ

本研究では，残響のある音声の分離を行うため，分離対象となる音声に仮定を導入し，対象の音声を用いて転移学習を行った．音声分離実験では，転移学習の有効性が確認できた．一方，分離音声にはまだ歪みが含まれている．音声認識などへの応用には，歪みの低減や分離した音声に含まれる残響への対応も必要である．

参考文献

[1] Y. Luo *et al.* Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation. In *Proc. of ICASSP 2020*, pp. 46–50, 2020.

[2] M. Maciejewski *et al.* WHAMR!: Noisy and reverberant single-channel speech separation. In *Proc. of ICASSP 2020*, pp. 696–700, 2020.

[3] J. L. Roux *et al.* SDR – half-baked or well done? In *Proc. of ICASSP 2019*, pp. 626–630, 2019.

[4] M. Pariente *et al.* Asteroid: the PyTorch-based audio source separation toolkit for researchers. In *Proc. of Interspeech 2020*, pp. 2637–2641, 2020.