7N-04

モノラル音源分離のための音源間類似度に基づく学習用混合信号の選択

宗像 北斗 † 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

深層学習 (DNN) に基づく音源分離に取り組む．音源分離とは，複数の音源が混合された観測信号から，それぞれの音源信号へと分離する技術である．本研究では，非音声も対象に含め，分離したい音源がモノラル音源である場合を想定する（図 1）．

非音声信号を対象とする場合，DNN の学習用混合信号の合成時に問題がある．同じ音の種類 (クラス) からなる混合信号の中には，本質的に分離が不可能な場合がある．分離不可能な学習用混合信号を含む学習データを学習に用いた場合，モデルの分離精度は低下する．従来，非音声を対象とした DNN によるモノラル音源分離の研究 [1] では，学習用混合信号の合成方法について議論されていない．

本稿では，音源間類似度を定義し，分離可能な学習用混合信号を選択する手法を提案する．これにより分離精度の向上を目指す．提案手法による学習データを学習に用いたモデルの分離精度を scale-invariant SDR [2] の改善量 (SI-SDRi) で評価する．

2. 学習用混合信号の合成時の問題

DNN による 2 音源でのモノラル音源分離の定式化は以下の通りである．i 番目の混合元信号を $\mathbf{s}_i = [s_{i,1}, s_{i,2}, \ldots, s_{i,T}]$ とすると，混合信号 $\mathbf{s}_{mix}$ は $\mathbf{s}_{mix} = \mathbf{s}_1 + \mathbf{s}_2$ と表される．T は信号の長さである．DNN に $\mathbf{s}_{mix}$ を入力し，$\mathbf{s}_i$ に対応する推定分離信号 $\hat{\mathbf{s}}_i$ を得る．$\hat{\mathbf{s}}_i$ を $\mathbf{s}_i$ に対してなるべく少ない歪みで得ることを目標とする．DNN は事前に大量の学習データ（混合信号およびその混合元信号）で学習させる．

非音声信号を対象とした場合，学習用混合信号の合成時に分離不可能な混合信号が合成されることがある．このような混合信号は学習には不適切である．不適切な学習用混合信号を含む学習データで学習させると分離精度が低下する．不適切な学習用混合信号の例として，2 つの混合元信号の音量アクティベーションの変化が類似する場合，あるいは周波数スペクトルが類似する場合が挙げられる．このような状況は同じ非音声クラスの信号の混合時に生じやすい．

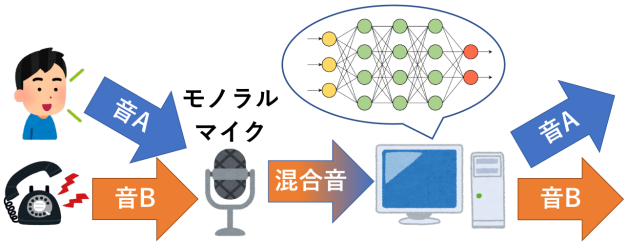

図 1: モノラルチャネルの音源分離

Arranging Mixed Audio Signals as Training Data Based on Sound Source Similarity for Monaural Sound Source Separation: *Hokuto Munakata, Ryu Takeda, and Kazunori Komatani (Osaka Univ.)*

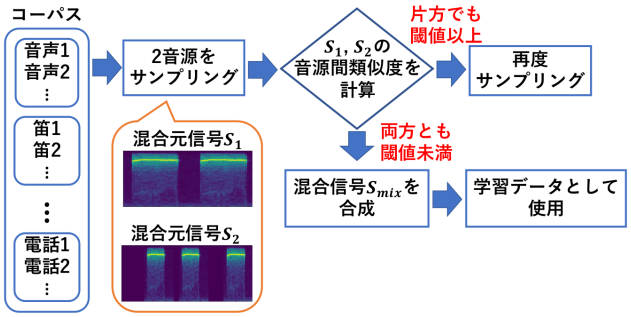

図 2: 提案手法の概要

3. 分離可能な学習用混合信号の選択

2 種類の音源間類似度 S_{act}, S_{spec} を導入することで，混合信号が分離可能か否かを判定する．2 つの混合元信号のサンプリング後，2 つの信号の長さを短い方に揃える．その後，音源間類似度を計算する．どちらの類似度も閾値未満の場合のみ混合信号を合成し，学習データに加える．超えた場合，先にサンプリングした混合元信号と同じクラスから再度サンプリングする．図 2 に一連の流れの概要を示す．

3.1 音量アクティベーションによる音源間類似度

ある区間の 2 音源の音量アクティベーションの変化が類似する場合，その区間は 2 音源が鳴っているにもかかわらず，1 音源しか鳴っていないとも判断できる。例えば 2 つ電話の音が同じタイミングで鳴っていた場合である．この場合 1 つの和音とも捉えることができる．こうした混合音を用いた学習は DNN が調波構造を捉えることを妨げると考えられる．

2 音源間の音量アクティベーションの類似度 S_{act} を j 番目の混合元信号 $\mathbf{s}_j$ のスペクトログラム $S_{f,t}^{(j)}$ から導出する．f,t はそれぞれ周波数 bin，時間フレームである．まず音量アクティベーションを $A_t^{(j)} = \sum_f |S_{f,t}^{(j)}|$ と定義する．これは時間信号の包絡を表す．その後，各時間フレームごとにアクティベーションを閾値 $\theta_A^{(j)} = \frac{1}{10} \max A_t^{(j)}$ との大小により $A_t'^{(j)}$ と 2 値化する．$A_t'^{(j)}$ は $A_t^{(j)}$ が $\theta_A^{(j)}$ を超えた場合 1，超えなかった場合 0 をとる．次に音量アクティベーションの変化を $D_t^{(j)} = A_{t+1}'^{(j)} - A_t'^{(j)}$ とする．これを用いて，S_{act} を以下のように定義する．

$$S_{act} = \frac{2\sum_t u(D_t^{(1)} D_t^{(2)})}{\sum_t (|D_t^{(1)}| + |D_t^{(2)}|)} \quad (1)$$

ただし u はステップ関数である．式 (1) の分子は 2 音源の全体の音量アクティベーションの変化の一致回数，分母は 2 音源の音量アクティベーションの変化の回数の合計を表す．ただし最大で 1 になるように分子に 2 を掛ける．

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

表 1: 3 種類のクラスの全組み合わせに対する分離結果

学習データ	音声+音声	音声+笛	音声+電話	笛+笛	笛+電話	電話+電話	平均
従来手法	7.8	15.0	16.0	4.0	6.3	10.8	10.0
提案手法	7.5	17.6	15.7	5.4	6.5	10.0	10.5

3.2 周波数スペクトルによる音源間類似度

2 つの音源の周波数スペクトルが類似する場合，ある瞬間の音がどちらの音源から発せられたのか判別できない．例えば 2 音源が同じ音程の笛の音であった場合である．DNN はある瞬間の混合音を分離した後，それがどちらの音源によるものか分類する．こうした混合音を用いた学習は，分類能力を低下させると考えられる．

周波数スペクトルを表す音響特徴量であるメル周波数ケプストラム係数 (*MFCC*) を用いる．*MFCC* はスペクトログラムから得られる．また，スペクトログラムと比較して次元数が少なく，各次元の相関が小さい．また事前にスペクトログラムから無音部分を削除する．具体的にはスペクトログラムの各時間フレームごとにパワースペクトルを計算し，その最大値から −25 [dB] 未満の時間フレームを削除する．

2 音源間の周波数スペクトルによる類似度 S_{act} を j 番目の混合元信号 $\mathbf{s}_j$ から得られる $MFCC_{d,\tau}^{(j)}$ から導出する．d,τ はそれぞれ MFCC の次元，時間フレームである．まず *MFCC* の時間平均 $r_d^{(j)} = \frac{1}{l}\sum_\tau MFCC_{d,\tau}^{(j)}$ を求める．l は *MFCC* の時間フレームの長さである．次に次元 d ごとに

$$L_d = \sqrt{\frac{(r_d^{(1)} - r_d^{(2)})^2}{|r_d^{(1)}||r_d^{(2)}|}} \tag{2}$$

を求める．これは次元ごとのスケールを考慮した 2 音源間の違いの大きさを表す．これを用いて S_{freq} を以下のように定義する．

$$S_{spec} = \frac{1}{\sum_d L_d} \tag{3}$$

分母は L_d の総和であり，その逆数である S_{spec} は違いが大きいほど小さくなる．

4. 評価実験

音声を含む 3 種類のクラスでの分離精度の比較を行う．これは実環境に近く，かつ分離不可能な学習用混合信号が合成されやすい状況を想定している．

4.1 実験条件

音声，笛の音および電話の音のデータから学習データおよび評価データを生成した．音声は WSJ0 コーパス，笛，電話の音は RWCP-SSD コーパスのデータを用いた．電話，笛の音は 3.0 秒以上になるように繰り返す前処理を施した．この際，冒頭および繰り返し時に 0.0〜1.0 秒のランダムな無音を挟んだ．また，学習時間短縮のため全て 8000 [Hz] までダウンサンプリングした．学習データ，評価データはいずれもランダムに −5〜+5 [dB] になるように混合した．

学習データの条件を示す．学習データは約 40 時間分合成した．更に 4:1 の割合で学習用と検証用に分割した．混合元信号をコーパスからランダムにサンプリングした場合，クラスごとに学習データに含まれるデータ量が異なる．そのためクラスごとにサンプリングされる確率が等しくなるようにした．前処理を行った後の段階でデータ量は音声は約 70 時間，笛，電話の音は約 6 分であった．さらに少量データクラスにおいて何度も同じ信号がサンプリングされることを防ぐため，データ拡張を施した．データ拡張にはオープンソースソフトウェア SoX を用いた．笛，電話のデータに対して，前処理を加えた後に pitch を −500〜500 の 10 段階，tempo を 0.5〜2.0 の 10 段階，合計 20 パターンのデータ拡張を施した．さらにデータ拡張後に冒頭に 0.0〜1.0 秒のランダムな無音を追加した．

評価データには学習データに含まれていないデータのみを使用した．笛，電話の音は学習データに含まれる音とは音色が大きく異なる．前処理を行った後の段階でデータ量は音声は約 70 時間，笛の音は約 12 分，電話の音は 14 分であった．

モデルには Deep Clustering [3] を用いた．学習は検証ロスが 5 エポック連続で下がらなくなるまで行った．

4.2 実験結果および考察

各クラスの組み合わせに対する分離結果を表 1 に表す．横軸は評価データに用いた混合元信号のクラスである．縦軸は 100 件ずつの SI-SDRi の平均値であり，最右列は評価データ全体の SI-SDRi の平均値である．提案手法は，音量アクティベーションによる類似度の閾値 $\theta_{act} = 0.25, 0.35, 0.45$，周波数スペクトルによる類似度の閾値を $\theta_{freq} = 0.7, 1.3, 1.9$ とした合計 9 パターンで試した．表 1 の数値は最も平均が高かった閾値を $\theta_{act} = 0.25$, $\theta_{freq} = 0.7$ によるものである．この閾値の設定では，サンプリングされた 2 つの混合元信号は 11% で $S_{act} > \theta_{act}$，35% で $S_{freq} > \theta_{freq}$ を満たし，41% で再サンプリングされた．学習データ内には分離不可能な混合信号は含まれていなかった．また，再サンプリングされた混合信号の中には主観的に見て分離可能なものが含まれていた．

提案手法は従来手法と比べて評価データ全体での分離精度がわずかに改善されたが，クラス単位で見ると分離精度が低下したクラスがある．このことから，提案手法は分離精度を上昇させるような学習用混合信号までも学習データから除外してしまっている可能性がある．

5. おわりに

今後はクラス数を増やした場合を確かめる．さらに音源間類似度を活用し，学習に有利な混合信号を選択する手法を考える．

参考文献

[1] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, et al. Universal Sound Separation. In *Proc. WASPAA, 2019*.

[2] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey. SDR - half-baked or well done? In *Proc. ICASSP, 2018*.

[3] J. R. Hershey, Z. Chen, et al. Deep clustering: Discriminative embeddings for segmentation and separation. In *Proc. ICASSP, 2016*.

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.