5Q-03

話者・音素特徴に基づくマルチチャネル音声分離

Du Yicheng[1]　**關口 航平**[2,3]　**坂東 宜昭**[4]　**Aditya Arie Nugraha**[3]　**吉井 和佳**[2,3]

[1]京都大学 工学部情報学科　[2]京都大学 大学院情報学研究科　[3]理化学研究所　[4]産業技術総合研究所

1. はじめに

複数人が同時に発話する環境で音声認識を行う際には，フロントエンド処理としてマイクアレイを用いた音声分離が行われる．近年，マイク配置や音源，環境の事前情報を用いずに，観測された混合音のみから音源信号および音源信号の混合過程を推定する汎用的なブラインド音源分離技術 (BSS) が目覚ましい発展を遂げている [1,2]．

多チャネル音源分離においては，音源信号の音響的な特徴を表現する音源モデルと，音源信号の伝達特性を表現する空間モデルとが重要な役割を果たしている．例えば，代表的な BSS 手法である多チャネル非負値行列因子分解 (MNMF) [2] や独立低ランク行列分析 (ILRMA) [1] では，音源のパワースペクトル密度に対して非負値行列因子分解 (NMF) に基づく低ランクモデルを，音源のイメージ（マイク位置での多チャネルスペクトル）に対してフルランクあるいはランク 1 の共分散行列を持つ多変量複素ガウスモデルを仮定している．

最近では，低ランク性の仮定が成り立たない音声を分離する目的で，大量の音声データから変分自己符号化器 (VAE) [3] を予め教師なし学習しておき，音源モデルを学習済みの深層生成モデル（VAE のデコーダ部分）に置き換えた半教師あり音源分離手法が提案されている [4–6]．特に，[4,5] においては，通常の VAE の代わりに，話者ラベルで条件付けた Conditional VAE (CVAE) [7] を学習しておき，音声分離時には話者ラベルを推定し，話者特徴を捉えることで分離精度の向上を図っている．

本稿では，話者ラベルに加えて音素ラベルを同時に考慮することで，各時刻の音声スペクトルをより精緻にモデリングできるという仮説のもとで，両者で条件付けた CVAE に基づく音声分離手法を提案する．音素に着目した音源分離は [8] でも提案されており，有効性が確認されている．具体的には，CVAE を学習することで得られた音声の深層生成モデルとフルランク空間モデルを統合した確率モデルを定式化し，最尤推定の枠組みで話者ラベルと音素ラベルを同時に推定しながら音声分離を行う．

2. 提案法

本章では，提案手法である話者特徴と音素特徴を同時に考慮した音源モデルと，フルランク空間モデルを統合した確率モデルによる音声分離法について述べる．

2.1 音源モデルの定式化

N 個の音声があるとする．本研究では各音声 n のパワースペクトル密度 $\boldsymbol{\sigma}_n^2 = \{\sigma_{nft}^2\}_{f=1,t=1}^{F,T}$ が話者特徴と音素特徴を同時に考慮した DNN で決定されると仮定する．

$$\mathbf{z}_{nt} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{1}$$

$$\boldsymbol{\sigma}_n^2 = g_n \cdot \mathrm{DNN}_\theta\left(\{\mathbf{z}_{nt}\}_{t=1}^{T}, \mathbf{z}_n^{\mathrm{s}}, \{\mathbf{z}_{nt}^{\mathrm{p}}\}_{t=1}^{T}\right) \tag{2}$$

Multichannel Speech Separation Based on Speaker and Phoneme Features: *Du Yicheng, Kouhei Sekiguchi, Yoshiaki Bando, Aditya Arie Nugraha, and Kazuyoshi Yoshii*

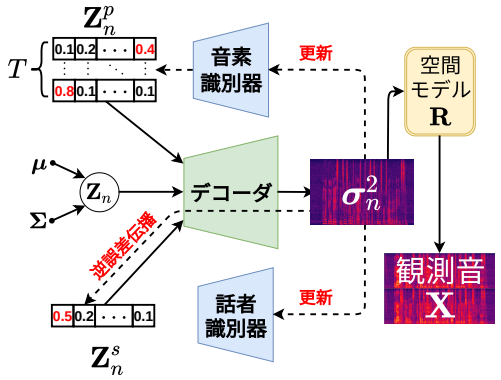

図 1: 話者・音素特徴を考慮した音源分離法

ここで，$\mathbf{z}_n^{\mathrm{s}} \in \{0,1\}^{K^{\mathrm{s}}}$（$K^{\mathrm{s}}$ は話者数）は音声 n の話者ラベルを表す one-hot ベクトル，$\mathbf{z}_{nt}^{\mathrm{p}} \in \{0,1\}^{K^{\mathrm{p}}}$（$K^{\mathrm{p}}$ は音素数）は音声 n の時刻 t での音素ラベルを表す one-hot ベクトル，$\mathbf{z}_{nt} \in \mathbb{R}^D$（$D$ は次元数）は音声の持つ話者・音素特徴以外の音響的特徴を表す潜在変数で，g_n は話者の音量を表すスケールパラメータであり，θ は DNN のパラメータである．

式 (2) の DNN は話者ラベルと音素ラベルで条件付けされた CVAE [7] のデコーダに相当しており，事前に大量のラベル付き音声データを用いて CVAE を学習し，パラメータ θ を求めておく．

2.2 空間モデルの定式化

M 個のマイクがあるとし，音声信号と混合音の複素スペクトログラム，音声 n のイメージをそれぞれ $\mathbf{s}_{ft} \in \mathbb{C}^N$, $\mathbf{x}_{ft} \in \mathbb{C}^M$, $\mathbf{c}_{ftn} \in \mathbb{C}^M$ とする（$\mathbf{x}_{ft} = \sum_{n=1}^{N} \mathbf{c}_{ftn}$）．反射や残響のない理想的な環境では，次式が成立する．

$$\mathbf{c}_{ftn} = \mathbf{a}_{fn} s_{ftn} \tag{3}$$

ここで，$\mathbf{a}_{fn} \in \mathbb{C}^M$ はステアリングベクトルである．音源の時間周波数ビン s_{ftn} は球対称複素ガウス分布 $\mathcal{N}_{\mathbb{C}}(0, \sigma_{ftn}^2)$ に従うと仮定すると，次式が成立する．

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \textstyle\sum_{n=1}^{N} \sigma_{ftn}^2 \mathbf{R}_{fn}\right) \tag{4}$$

ここで，$\mathbf{R}_{fn} = \mathbf{a}_{fn}\mathbf{a}_{fn}^H \in \mathbb{C}^{M\times M}$ は空間相関行列であり，ランク 1 行列である．式 (3) は現実に厳密には成り立たないため，提案法では $\mathbf{R}_{fn}$ をフルランク行列とした空間モデルを用いる．

2.3 最尤推定に基づく音声分離

我々の目標は前節の確率モデルを用いて，観測データ $\mathbf{X} = \{\mathbf{x}_{ft}\}_{f=1,t=1}^{F,T}$ が与えられたときに，空間相関行列 $\mathbf{R} = \{\mathbf{R}_{fn}\}_{f=1,n=1}^{F,N}$，潜在表現 $\mathbf{Z} = \{\mathbf{z}_{nt}\}_{n=1,t=1}^{N,T}$，話者ラベル $\mathbf{Z}^{\mathrm{s}} = \{\mathbf{z}_n^{\mathrm{s}}\}_{n=1}^{N}$，音素ラベル $\mathbf{Z}^{\mathrm{p}} = \{\mathbf{z}_{nt}^{\mathrm{p}}\}_{n=1,t=1}^{N,T}$，各話者の音量 $\mathbf{g} = \{g_n\}_{n=1}^{N}$ を求めることである．具体的には，式 (4) の尤度関数を最大化するよう反復最適化を行う．$\mathbf{R}$ と $\mathbf{g}$ は従来法 [5] と同様に，$\mathbf{Z}$，$\mathbf{Z}^{\mathrm{s}}$ および $\mathbf{Z}^{\mathrm{p}}$ は誤差逆伝播法で更新した．収束後，多チャネルウィー

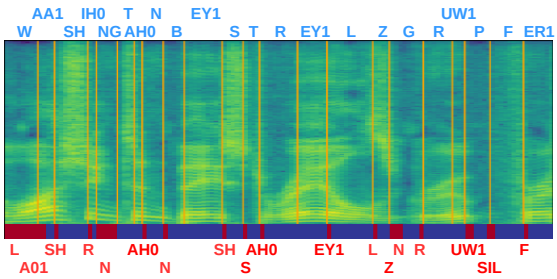

図 2: 提案手法で得た分離音の一例．青い部分は音素が正しく推定されたフレーム，赤い部分は音素の推定が誤ったフレームである．青色の文字は正解の音素ラベル，赤色の文字は誤った推定結果を表す．

ナーフィルタ [2] で分離音 $\mathbf{C} = \{\mathbf{c}_{ftn}\}_{f=1,t=1,n=1}^{F,T,N}$ を推定する．別の方法として，最適化の途中で分離音 $\mathbf{C}$ を得て，事前学習した CVAE のエンコーダに入力することで $\mathbf{Z}$ を更新することもできる．同様に，クリーンな音声データから話者ラベルあるいは音素ラベルの識別器をそれぞれ事前に学習しておけば，分離音 $\mathbf{C}$ を入力として直接 $\mathbf{Z}^{\mathrm{s}}$ および $\mathbf{Z}^{\mathrm{p}}$ を推定できる．

3. 評価実験

音声分離における話者特徴と音素特徴の有用性を検証するために行った評価実験について報告する．

3.1 実験条件

2 話者の音声分離タスクにおいて，提案法を補助情報を用いない VAE，話者ラベル条件付きの CVAE，音素ラベル条件付きの CVAE と比較した．評価尺度は信号対歪比 (SDR) 及び話者識別率，音素識別率とした．すべての手法に対し，ILRMA で得られた分離音をエンコーダと識別器に入力することで，$\mathbf{Z}$，$\mathbf{Z}^{\mathrm{s}}$ および $\mathbf{Z}^{\mathrm{p}}$ の初期化を行った．また，ILRMA で推定した分離行列から得たランク 1 の空間相関行列で $\mathbf{R}$ を初期化した．潜在変数 $\mathbf{Z}$ は誤差逆伝播法，$\mathbf{Z}^{\mathrm{s}}$ および $\mathbf{Z}^{\mathrm{p}}$ は誤差逆伝播法あるいは識別器を用いて最適化した．誤差逆伝播法には，学習率 0.002 の AdamW [10] を用いた．また，潜在変数 $\mathbf{Z}$ を事前分布のに近づけるために 0.2 の荷重減衰を適用した．識別器を用いる場合には，10 反復ごとに分離音を入力とした識別器の出力で潜在変数を更新した．反復回数は 150 回とした．話者・音素識別器，CVAE のエンコーダ・デコーダにはゲート付き CNN [9] を用いた．

音声信号には，WSJ 英語読み上げコーパス（7138 発話，サンプリング周波数 16kHz）を使用した（話者数 $K^{\mathrm{s}} = 83$）．テストデータとして 1 話者につき 2 発話を選び，83 個の 2 話者混合音を作成した（RT60 = 129 ms）．その他の発話は訓練データとして CVAE と話者・音素識別器の学習に用いた．短時間フーリエ変換には，フレーム長 64 ms，シフト幅 16 ms のハン窓を用いた．音素ラベルは，学習済みの GMM-HMM トライフォン音響モデルを用いて強制アライメントを行い，モノフォンに変換して得た（音素数 $K^{\mathrm{p}} = 72$）．CVAE は学習率 0.001 の AdamW [10] を用いて，2500 エポック学習した．

3.2 実験結果

表 1 に示す通り，誤差逆伝播法では音素ラベルを推定できず，識別器による推定が有効であることが分かった．

モデル	更新法	SDR	識別率 (%)
VAE	−/−	14.65 ± 3.44	−/−
CVAE	C/−	14.68 ± 3.49	78.92/−
(話者のみ)	B/−	14.73 ± 3.48	70.48/−
CVAE	−/C	14.70 ± 3.45	−/54.29
(音素のみ)	−/B	14.46 ± 3.49	−/5.57
提案法	B/C	14.77 ± 3.47	62.65/54.73
(話者 + 音素)	C/C	14.70 ± 3.47	72.29/54.78

表 1: 評価結果．更新法は，$\mathbf{Z}_{\mathrm{s}}$更新法/$\mathbf{Z}^{\mathrm{p}}$更新法 を示し，B は誤差逆伝播法，C は識別器による更新を表す．識別率 (%) は，話者識別率/音素識別率 を表す．

$\mathbf{Z}^{\mathrm{s}}$ を誤差逆伝播法，$\mathbf{Z}^{\mathrm{p}}$ を識別器で推定する提案モデルは，ほかのどの手法よりも分離性能が高い傾向がみられるが，改善量はあまり多くない．話者ラベルは識別器で推定したほうが話者識別率を改善できるが，SDR は若干低下する．これは話者ラベル $\mathbf{Z}^{\mathrm{s}}$ と他の潜在変数との間の分離度が低いことが原因であると考えられる．話者特徴を $\mathbf{Z}^{\mathrm{p}}$ と $\mathbf{Z}$ から完全に分離することができなければたとえ話者を正しく推定できたとしても，DNN はその話者らしいパワースペクトル密度を出力することができず，話者特徴を有効活用できないため，提案手法の分離性能のボトルネックとなっていると考えられる．

4. おわりに

本稿では話者特徴と音色特徴を同時に考慮した深層音源モデルに基づく音源分離手法を提案した．評価実験の結果，提案法が導入した話者特徴と音素特徴は音源分離に有効であることが確認された．今後は 3 話者の混合音などの複雑なデータで実験するとともに，話者特徴と音素特徴をより活かすことで分離性能の向上を図るべく，潜在変数間の分離度を上げる仕組みを作る予定である．

謝辞 本研究の一部は，科研費 No. 19H04137， NII CRIS-Line 共同研究の支援を受けた．

参考文献

[1] D. Kitamura *et al.*: “Determined Blind Source Separation Unifying Independent Vector Analysis and Nonnegative Matrix Factorization,” *TASLP*, vol.24, no.9, pp.1626–1641, 2016.

[2] H. Sawada *et al.*: “Multichannel Extensions of Non-Negative Matrix Factorization with Complex-Valued Data,” *TASLP*, vol.21, no.5, pp.971–982, 2013.

[3] D. Kingma *et al.*: “Auto-Encoding Variational Bayes,” *ICML*, 2014.

[4] H. Kameoka *et al.*: “Supervised Determined Source Separation with Multichannel Variational Autoencoder,” *Neural Computation*, vol.31, no.9, pp.1891–1914, 2019.

[5] S. Seki *et al.*: “Underdetermined Source Separation Based on Generalized Multichannel Variational Autoencoder,” *IEEE Access*, vol.7, pp.168104-168115, 2019.

[6] K. Sekiguchi *et al.*: “Semi-supervised Multichannel Speech Enhancement with a Deep Speech Prior,” *TASLP*, vol.27, no.12, pp.2197–2212, 2019.

[7] D. Kingma *et al.*: “Semi-supervised Learning with Deep Generative Models,” *NIPS*, pp.3581–3589, 2014.

[8] Z. Wang *et al.*: “Phoneme-specific Speech Separation,” *ICASSP*, pp.146–150, 2016.

[9] Y. Dauphin *et al.*: “Language Modeling with Gated Convolutional Networks,” *ICML*, pp.933–941, 2017.

[10] I. Loshchilov *et al.*: “Decoupled Weight Decay Regularization,” *ICML*, 2019.