1R-02

# 深層フルランク空間相関分析に基づく遠隔音声認識のフロントエンド

合澤 隆拓[1,2]　坂東 宜昭[2]　糸山 克寿[1]　西田 健次[1]　中臺 一博[1,3]
[1]東京工業大学　[2]産業技術総合研究所　[3](株) ホンダ・リサーチ・インスティチュート・ジャパン

## 1. はじめに

混合音から個別の音声を抽出する音声強調や音源分離は，雑音や他の話者の音声が混入する遠隔音声認識のフロントエンドとして不可欠である [1]．スマートスピーカに代表されるように，単一話者の遠隔音声認識は高い性能を達成しているが，複数人が参加する会話の音声認識は未だ多くの課題が残されている [2]．ホームパーティーのような複数人の自由な会話を認識できれば，スマートホームや家庭用ロボットとの自然な対話や，人語を解する補聴器・拡張現実の実現が期待できる．

遠隔音声認識には，未知環境でも頑健に動作するフロントエンドが不可欠である．例えば，混合複素角度中心ガウスモデル (cACGMM) [3] などのブラインド音源音源分離 (BSS) [1,3,4] は，観測信号のみから適応的に音声を強調・分離できる．しかし従来の BSS の多くは，線形の生成モデルに基づくため性能に限界があった．そこで，非線形生成モデルに基づく深層フルランク空間相関分析 (Neural FCA) [5] が提案されている．この手法は，使用環境の混合音を用いて非線形生成モデルを教師なし学習し，混合音から音声を精緻に分離できる．Neural FCA は，数値混合音による性能評価のみ報告されているが，実収録音でも高い性能を発揮すると期待できる．

本研究では，Neural FCA に基づく遠隔音声認識のフロントエンドについて報告する．従来の Neural FCA は，混合音に含まれる音源数が既知と仮定しており，音源数が常に変動する日常会話の認識には適さない．そこで，他の手法で得た発話区間情報を生成モデルに導入し，音源数の変動に対応した弱教師あり学習へ拡張する．提案法は，ホームパーティでの会話を収録した CHiME-6 データセット [2] を用いて評価した．

## 2. 弱教師あり深層フルランク空間相関分析

Neural FCA [5] に発話区間変数を補足情報として導入した弱教師あり学習について説明する．

### 2.1 問題設定

本手法では，ホームパーティのような $N$ 人が会話している状況で，以下の問題設定により音源分離を行う．

**入力**: $M$ チャネル混合音 $\mathbf{x}_{ft} \in \mathbb{C}^M$ と話者 $n = 1, \ldots, N$ の時間フレーム $t$ での発話有無 $u_{nt} \in \{0, 1\}$
**出力**: 話者 $n$ の分離音 $\hat{s}_{nft} \in \mathbb{C}$

ここで，$f = 1, \ldots, F$ および $t = 1, \ldots, T$ はそれぞれ，周波数および時間インデックスを表す．

A Front-End of Distant Speech Recognition Based on Neural Full-Rank Spatial Covariance Analysis: *T. Aizawa, Y. Bando, K. Itoyama, K. Nishida, K. Nakadai*

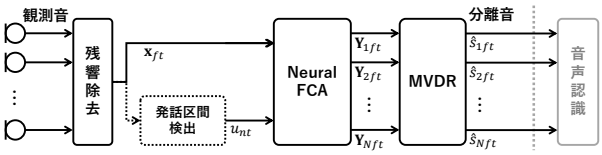


図 1: 弱教師あり Neural FCA に基づくフロントエンド

### 2.2 生成モデル

提案法では，観測混合音 $\mathbf{x}_{ft}$ を以下のように $N+1$ 個の音源信号 $s_{nft} \in \mathbb{C}$ の和で表す．

$$\mathbf{x}_{ft} = \sum_{n \in \mathfrak{N}_t} \mathbf{a}_{nf} s_{nft} \tag{1}$$

ただし，$\mathfrak{N}_t = \{0\} \cup \{n | u_{nt} = 1\}$ は時間 $t$ に存在する音源の集合，$\mathbf{a}_{nf} \in \mathbb{C}^M$ は音源 $n$ のステアリングベクトルであり，$n = 0$ は雑音を表す．音源信号のパワースペクトル密度 (PSD) $s_{nft}$ は，次式のように音源の特徴を表す潜在ベクトル $\mathbf{z}_{nt} \in \mathbb{R}^D$ を用いて表現する．

$$s_{nft} \sim \mathcal{N}_{\mathbb{C}}\left(0, g_{\theta,f}(\mathbf{z}_{nt})\right), \quad z_{ntd} \sim \mathcal{N}\left(0, 1\right) \tag{2}$$

ここで，$g_{\theta,f} : \mathbb{R}^D \to \mathbb{R}_+$ は，$\mathbf{z}_{nt}$ から PSD を出力するパラメータ $\theta$ を持つ深層ニューラルネットワーク (DNN) であり，次節の通り事前に収集した多チャネル混合音から学習する．以上より，観測混合音 $\mathbf{x}_{ft}$ は，以下の多変量複素ガウス分布に従う．

$$\mathbf{x}_{ft} \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sum_{n \in \mathfrak{N}_t} g_{\theta,f}(\mathbf{z}_{nt}) \mathbf{H}_{nf}\right) \tag{3}$$

ただし，$\mathbf{H}_{nf} = \mathbf{a}_{nf}\mathbf{a}_{nf}^{\mathsf{H}} \in \mathbb{S}_+^{M \times M}$ は音源 $n$ の空間相関行列 (SCM) である．本稿では，SCM をフルランクに緩和し，$\mathbf{a}_{nf}$ の比較的小さな変動を許容する．

### 2.3 弱教師あり事前学習

本学習では，多チャネル混合音と発話区間から，対数周辺尤度 $\log p_\theta(\mathbf{X}|\mathbf{H}, \mathbf{U})$ を最大にするよう音源モデル $g_{\theta,f}$ を弱教師あり学習する．この対数周辺尤度は直接計算困難なので，以下の推論モデル $q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U})$ を導入した変分償却推論 [6] を行う．

$$q_\phi(\mathbf{Z}|\mathbf{X}, \mathbf{U}) = \prod_{n,t,d} \mathcal{N}_{\mathbb{C}}\left(z_{ntd} \mid \mu_{\phi,ntd}(\mathbf{C}), \sigma^2_{\phi,ntd}(\mathbf{C})\right) \tag{4}$$

ここで $\mu_{\phi,ntd}(\mathbf{C}) \in \mathbb{R}$ と $\sigma^2_{\phi,ntd}(\mathbf{C}) \in \mathbb{R}_+$ は，特徴量 $\mathbf{C}$ を入力とするパラメータ $\phi$ を持つ DNN の出力である．特徴量 $\mathbf{C}$ は $\mathbf{X}$ と $\mathbf{U}$ から計算されるが，詳細は 3.2 節で述べる．変分償却推論では，学習データに対する以下の変分下限 $\mathcal{L}$ を最大化するように，DNN のパラメータ $\theta$

と $\phi$，SCM $\mathbf{H}_{nf}$ を一挙に最適化する．

$$\mathcal{L}=\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{X}|\mathbf{Z},\mathbf{H},\mathbf{U})]-\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})|p(\mathbf{Z})] \quad (5)$$

この最大化により，パラメータ $\theta$ と SCM $\mathbf{H}_{nf}$ は $\log p_\theta(\mathbf{X}|\mathbf{H},\mathbf{U})$ を最大化するように，パラメータ $\phi$ は，$\mathcal{D}_{\mathrm{KL}}[q_\phi(\mathbf{Z}|\mathbf{X},\mathbf{U})|p_\theta(\mathbf{Z}|\mathbf{X},\mathbf{H},\mathbf{U})]$ を最小化するように学習される．各 DNN のパラメータは変分自己符号化器 [6] と同様に，誤差逆伝播法により最適化する．SCM $\mathbf{H}_{nf}$ は，以下の更新則 [7] を繰り返して最適化する．

$$\mathbf{H}_{nf} \leftarrow \frac{1}{\sum_{t=1}^{T} u_{nt}} \sum_{t\in\{t|u_{nt}=1\}} \frac{1}{g_{\theta,f}(\mathbf{z}^*_{nt})}\hat{\mathbf{X}}_{nft} \quad (6)$$

$$\hat{\mathbf{X}}_{nft}=\mathbf{Y}_{nft}+\mathbf{Y}_{nft}(\mathbf{Y}^{-1}_{:ft}\mathbf{x}_{ft}\mathbf{x}^{\mathsf{H}}_{ft}\mathbf{Y}^{-1}_{:ft}-\mathbf{Y}^{-1}_{:ft})\mathbf{Y}_{nft} \quad (7)$$

ただし，$\mathbf{Y}_{:ft}=\sum_{n=1}^{N}\mathbf{Y}_{nft}\in\mathbb{S}_+^M$ は，各音源ごとの $\mathbf{Y}_{nft}=g_{\theta,f}(\mathbf{z}^*_{nt})\mathbf{H}_{nf}\in\mathbb{S}_+$ の和であり，$\mathbf{z}^*_{nt}\sim q_\phi(\mathbf{Z}|\mathbf{X})$ は潜在ベクトルのサンプルである．本稿では，$\theta$ と $\phi$ を 1 回更新するごとに $\mathbf{H}$ を 5 回更新する．

### 2.4 推論方法

学習した DNN を用いて，未知の混合音を分離できる．具体的には，混合音 $\mathbf{X}$ に対して，$z_{ntd}\leftarrow\mu_{\phi,ntd}(\mathbf{C})$) を初期値として，$\log p_\theta(\mathbf{X}|\mathbf{H},\mathbf{U},\mathbf{Z})$ を最大にするよう $\mathbf{z}_{nt}$ と $\mathbf{H}_{nf}$ を推定する [5]．分離音 $\hat{s}_{nft}$ は，得られた PSD $g_{\theta,f}(\mathbf{z}_{nt})$ と SCM $\mathbf{H}_{nf}$ から，最小分散無歪 (MVDR) ビームフォーマ を用いて得る．

## 3. 実験・評価

CHiME-6 データセットで提供されている実収録音を用いて提案法を評価した．

### 3.1 データセット

CHiME-6 データセットでは，kitchen, dining, living からなる室内で収録されたディナーパーティを認識する．収録には，各パーティごとに 5 または 6 台の 4 チャネルマイクアレイ (Microsoft Kinect v2) が使用され，一つのエリアに少なくとも 2 つのマイクアレイが設置されている．各録音は 16 kHz で収録され，train, dev および eval セットに分割されている．

### 3.2 実験設定

提案法である弱教師あり Neural FCA は，CHiME-6 Challenge のベースライン音声認識器 [2] を用いて単語誤り率 (WER) を評価した．Neural FCA の DNN は，文献 [5] を参考に構築した．推論モデルは 16 層の 1 次元畳み込み層から，音源モデルは 3 層の 1 次元畳み込み層からなる．推論モデルの入力は，基準マイクロホンと他のマイクロホン間のチャネル間位相差，cACGMM の分離マスクのいずれか 1 つおよび混合音のスペクトログラムを入力した．潜在変数 $D$ の次元は 20 とした．スペクトログラムは，窓長 1024 サンプル，ホップ長 256 サンプルの短時間フーリエ変換で得た．音源モデルを精緻に学習させるため，変分下限 $\mathcal{L}$ の Kalback-Leibler (KL) 項の重みを周期的に変動させる KL アニーリング [5] を行った．また，計算機のメモリの制約上，各アレイの両端のマイクから 8 本をパワーの大きい順に選択して使用した．提案法は，CHiME-6 Challenge のベースラインシステムであった，cACGMM に基づく枠組みと比較した．また，文献 [5] で用いられていた多チャネル Wiener フィルタ (MWF) による分離も比較した．

表 1: CHiME-6 での音声認識性能 (WER)

| 手法 | | dev | eval |
|---|---|---|---|
| ベースライン (cACGMM) | | 51.8 | 51.3 |
| 弱教師あり Neural FCA | | | |
| モノラル混合音+位相差 | MWF | 74.7 | 70.9 |
| | MVDR | 55.3 | 52.6 |
| モノラル混合音+分離マスク | MWF | 68.7 | 68.9 |
| | MVDR | 51.5 | 51.3 |

### 3.3 実験結果

表 1 に音声認識性能を WER で示す．まず，推論モデルの入力に cACGMM の分離マスクを用いたほうがチャネル間位相差を用いる場合比べて性能が向上した．これは実混合音のみからの学習は難しく，分離の補助情報を与えた方が音源分離を学習しやすくなったためと考えられる．ビームフォーマには，MVDR を用いた場合が MWF より高い性能を示した．これは，MVDR は線形時不変フィルタのため，分離歪みが抑制され，性能向上につながったと考えられる．ベースラインである cACGMM との比較では，cACGMM の分離マスクを用いて MVDR で分離した場合に dev セットでは 0.3 pt 改善し，eval セットでは同程度の性能となった．

## 4. おわりに

本稿では，Neural FCA の弱教師あり学習に基づく，遠隔音声認識のフロントエンドを開発した．推論モデルの入力に cACGMM の分離マスクを用いた場合に，チャネル間位相差と比較して性能が向上することを確認した．今後は，学習の安定化により認識性能の向上を目指すとともに，より高性能な認識器での評価を行う．

**謝辞**: 本研究の一部は，JST ACT-X JPMJAX200N および NEDO の支援を受けた．

## 参考文献

[1] K. Shimada et al. Unsupervised speech enhancement based on multichannel NMF-informed beamforming for noise-robust automatic speech recognition. *IEEE/ACM TASLP*, 27(5):960–971, 2019.
[2] S. Watanabe and other. CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *CHiME 2020 Workshop*, 1–7, 2020.
[3] C. Boeddeker *et al.* Front-end processing for the CHiME-5 dinner party scenario. In *CHiME5 Workshop*, 1–6, 2018.
[4] K. Sekiguchi et al. Fast multichannel nonnegative matrix factorization with directivity-aware jointly-diagonalizable spatial covariance matrices for blind source separation. *IEEE/ACM TASLP*, 28:2610–2625, 2020.
[5] Y. Bando et al. Neural full-rank spatial covariance analysis for blind source separation. *IEEE SPL*, 28:1670–1674, 2021.
[6] D. P. Kingma et al. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.
[7] H. Sawada et al. Experimental analysis of EM and MU algorithms for optimizing full-rank spatial covariance model. In *EUSIPCO*, 885–889, 2021.