深層学習に基づく仮想マイク生成技術の劣決定音源分離問題への適用の検討 *

☆瀬川華子（筑波大），落合翼，マーク・デルクロア，中谷智広，
池下林太郎，荒木章子（NTT），山田武志（筑波大），牧野昭二（早稲田大/筑波大）

1 はじめに

マイク数の制限によるアレイ信号処理の性能低下を緩和するためのアプローチとして，実マイク信号から仮想的にマイク数を増加させる仮想マイク生成の技術 [1] が検討されている．近年我々は，深層学習に基づく仮想マイク生成技術 (neural network based virtual microphone estimator, NN-VME) [2] を提案した．NN-VME では，時間領域の信号予測ニューラルネットワーク (NN) を使用することで，時間領域信号として仮想マイク信号の振幅と位相を同時に推定する．NN-VME は音響的な仮定を陽に置くことなく，教師あり学習の枠組みで NN に仮想マイクの推定規則を学習データからデータドリブンに獲得できることが期待される．

先行研究 [2] では，単一話者の条件下におけるビームフォーミングへの適用のみが評価された．しかし，NN-VME の枠組みは，特定の音響条件やアレイ信号処理技術に依存するものではなく，複数人話者の同時発話といった音響条件にも適用可能なものであると期待される．しかし，単一話者の場合では，観測される実マイク信号から，1 人分の話者の空間情報 (e.g. 話者の位置やインパルス応答) を推定できれば良いが，複数の話者が存在する場合，各話者についての空間情報を同時に推定することが必要になるため，推定の難易度が格段に高い問題設定となる．NN-VME が少数の実マイク信号から，各話者についての空間情報を保存した仮想マイク信号を十分な精度で推定可能であるかは必ずしも明らかではない．また，音源分離タスクにおいて，時間領域の信号予測モデルは残響がある状況においてその推定精度が低くなる問題が報告されており，残響条件下において NN-VME が十分な精度を達成できるかは必ずしも明らかではない．先行研究 [2] では，NN-VME のこのような条件における有効性について十分な検討がされていなかった．

そうした背景のもと，本稿では，NN-VME の有効性について検証するため，より多様な以下のような実験設定のもとで評価を行った．(1) NN-VME を，単一話者と比較してより複雑な複数話者における劣決定の音響条件に適用した．(2) 残響時間が NN-VME の仮想マイクロホン推定精度に与える影響を調べるため，残響時間がそれぞれ 0 ms，100 ms，200 ms，300 ms の評価用発話データを作成し，残響時間に関する包括的な実験評価を行った．(3) NN-VME が生成する仮想マイク信号の，後段の処理であるアレイ処理技術からの独立性を確認するために，ビームフォーミングに加えて，代表的なブラインド音源分離手法（BSS）(e.g. IVA [3] と ILRMA [4]）に対する NN-VME の有効性を評価した．

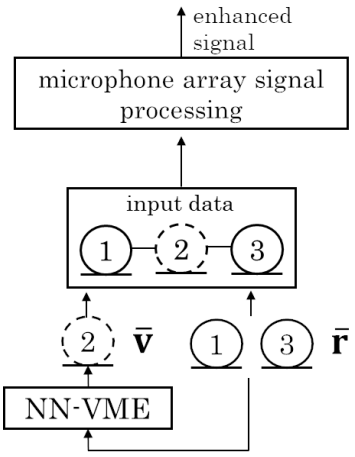

Fig. 1 Overview of array signal processing with NN-VME

2 深層学習に基づく仮想マイク生成技術

2.1 全体の構成

$\bar{\mathbf{r}}_c \in \mathbb{R}^{\mathcal{T}}$ は c 番目の実マイクにて録音された時間領域の観測信号を，$\bar{\mathbf{v}}_{c'} \in \mathbb{R}^{\mathcal{T}}$ は c' 番目の仮想マイクに対応する時間領域の推定信号を表すものとする．ここで，$\mathcal{T}$ は波形の長さ (サンプル数) を表す．仮想マイク技術とは，実マイク信号 $\bar{\mathbf{r}} = \{\bar{\mathbf{r}}_{c=1}, \ldots, \bar{\mathbf{r}}_{c=C_r}\} \in \mathbb{R}^{\mathcal{T}\times C_r}$ を入力として，仮想マイク信号 $\bar{\mathbf{v}} = \{\bar{\mathbf{v}}_{c'=1}, \ldots, \bar{\mathbf{v}}_{c'=C_v}\} \in \mathbb{R}^{\mathcal{T}\times C_v}$ を推定する技術である．ここで，C_r, C_v は入力として用いる実マイク数と推定する仮想マイク数をそれぞれ表す．また，本稿では，オーバーラインの付いたシンボル (e.g. $\bar{\mathbf{r}}$) は，時間領域信号であることを表すものとする。

Fig. 1 は, 仮想マイクの生成と生成された仮想マイク信号のアレイ信号処理への適用の流れを示す．Fig. 1 では，簡単のために，入力として扱う実マイクのチャネル数を $C_r = 2$，出力である仮想マイクのチャネル数を $C_v = 1$ とした場合を想定し，チャネル 1, 3 の実マイク信号からチャネル 2 の位置に存在する信号を仮想マイク信号として生成する様子を示している．

---

* Application of Neural Network-based Virtual Microphone Estimator to Source Separation in Underdetermined Situation by Hanako Segawa (Univ. of Tsukuba) , Tsubasa Ochiai, Marc Delcroix, Tomohiro Nakatani, Rintaro Ikeshita, Shoko Araki (NTT) , Takeshi Yamada (Univ. of Tsukuba) , Shoji Makino (Waseda Univ./Univ. of Tsukuba)

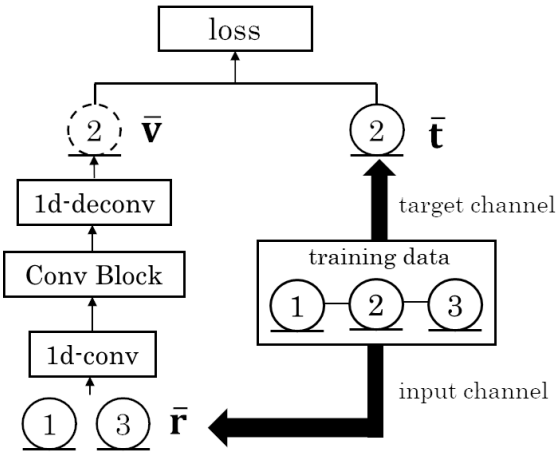

Fig. 2　Network architecture and overview of training of neural network-based virtual microphone estimator

推定された仮想マイク信号を実マイク信号と結合することで，仮想的にマイク数が拡張された観測マイクアレイ信号 $\bar{\mathbf{x}} = [\bar{\mathbf{r}}, \bar{\mathbf{v}}] \in \mathbb{R}^{\mathcal{T}\times(C_r+C_v)}$ を得る．こうして仮想的にマイクの数を増やしたマイクアレイ信号 $\bar{\mathbf{x}}$ に対し，アレイ信号処理技術を適用することで，アレイ信号処理技術の性能や適用範囲の向上が可能になると期待される．

2.2　ネットワーク構造

NN-VME [2] では，NN を使用し， 実マイク信号 $\bar{\mathbf{r}}$ を入力として，仮想マイク信号 $\bar{\mathbf{v}}$ を予測する:

$$\bar{\mathbf{v}} = \text{NN-VME}(\bar{\mathbf{r}}) \tag{1}$$

ここで，NN-VME(·) は，仮想マイクを推定するための NN モデルを表す．

NN-VME のネットワーク構造には，高い精度で時間領域の信号を推定することが可能な Conv-Tasnet [5] に基づいたものを採用している．Conv-TasNet は Fig. 2 の左側に示すように, 時間領域の信号を encoder で直接的に受け取り，internal convolution block で処理した後，decoder layer によって時間領域信号を直接的に出力する．こうした convolution-based encoder/decoder architecture を採用することで NN-VME は仮想マイク推定に必要な振幅と位相を，時間領域信号として同時に推定することが可能になっている．

2.3　学習方法

NN-VME は，教師あり学習のフレームワークを採用している．NN-VME では，システムの開発時には実際のシステムの運用時よりもマイク数の制限が少ないことを想定し，教師あり学習に必要な入力信号と目的信号の組である $\{\bar{\mathbf{r}}, \bar{\mathbf{t}}\}$ を用意する．ここで，$\bar{\mathbf{t}} = \{\bar{\mathbf{t}}_{c'=1}, \ldots, \bar{\mathbf{t}}_{c'=C_v}\} \in \mathbb{R}^{\mathcal{T}\times C_v}$ を表し，また $\bar{\mathbf{t}}_{c'} \in \mathbb{R}^{\mathcal{T}}$ は c' 番目の仮想マイクに対応する時間領域での目的信号を表す．Fig. 2 の右側はこの準備過程について示しており，学習時に利用できる実マイク信号集合の内，一部の集合を $\bar{\mathbf{r}}$ として入力信号に，残りの集合を $\bar{\mathbf{t}}$ として目的信号に割り当てる様子を図示している．

損失関数としては，次のような仮想マイクの位置において実際に観測される観測信号 (i.e. target signal $\bar{\mathbf{t}}$) と NN によって推定された仮想マイク信号 $\bar{\mathbf{v}}$ = NN-VME($\bar{\mathbf{r}}$) の間の scale-dependent signal-to-noise ratio を採用した．

$$\mathcal{L}_{\text{VM}} = \sum_{c'=1}^{C_v} 10\log_{10}\frac{||\bar{\mathbf{t}}_{c'}||^2}{||\bar{\mathbf{t}}_{c'} - \bar{\mathbf{v}}_{c'}||^2} \tag{2}$$

仮想マイクの位置で観測される信号 (振幅や位相) は，音源とマイクの位置関係や残響条件等，様々な条件に依存する．多様な状況下での仮想マイク推定を可能とするため，NN-VME の学習では，音源位置やマイク位置，雑音条件や残響条件の異なる多様なデータを用いて学習を行う．これにより，音響条件に頑強な仮想マイクの推定が可能になると期待される．

3　評価実験

3.1　実験条件

3.1.1　評価データ

本実験では，Wall Street Jounal (WSJ) corpus に収録されている音声データと CHiME-4 corpus に収録されている雑音データを用いて，3 名の話者による同時発話を模擬したシミュレーション混合音声データセットを作成した．インパルス応答は, image method を用いて作成した．学習データの作成時には残響時間 (T_{60}) を 0 ms から 300 ms の範囲でランダムに選択し，評価用データとしては，残響時間が 0 ms, 100 ms, 200 ms, 300 ms にそれぞれ固定した 4 種類の条件でインパルス応答を作成した．各発話の生成において，マイクと各話者の位置はランダムに配置するものとした．話者間の SIR は，1 人目の話者を基準として，−3 dB ∼ 3 dB の範囲で設定し，拡散性雑音の SNR は 20 dB に設定した．上記の条件を満たすように, 学習用データとして 30,000 混合音声, 検証用データを 5,000 混合音声，評価用データとして 5,000 混合音声をそれぞれ生成した．

使用した CHiME-4 corpus におけるマイクの配置を Fig. 3 に示す．本研究の実験では，合計 6 チャネルある信号のうち，4, 5, 6 の 3 チャネルの信号を使用した．

3.1.2　評価に用いたアレイ処理手法

本稿では，劣決定条件かつ残響のある音響条件における NN-VME の有効性を検証するため，先行研究 [2] で用いた NN マスクベースの MVDR ビームフォーマ (NN-BF) [6] に加えて, 代表的なブラインド音源分離手法である独立ベクトル分析（Independent Vector Analysis, IVA）[3] と独立低ランク行列分析 (Independent Low-rank Matrix Analysis, ILRMA) [4] を採用し評価を行った．

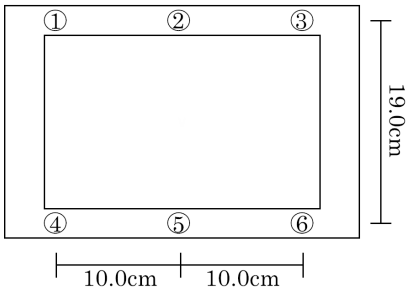

Fig. 3　Microphone arrangement of CHiME4 corpus

MVDR を含む指向性の null を形成するビームフォーマでは，マイク数 – 1 個の null を生成するため，劣決定条件下では妨害話者の発話をすべて消すことは原理的に不可能であり，マイク数によって達成できる強調性能が制約される．また，IVA や ILRMA に代表される分離行列を推定するフレームワーク では，劣決定条件下 (マイク数 < 話者数) では音源の解を一意に決定することができず，マイク数によって手法が適用できない状況が存在し得る．よって，仮想マイク生成によってマイク数を増加することで，劣決定条件におけるビームフォーミングの性能向上や，ブラインド音源分離を劣決定条件に適用可能になることが期待される．

3.1.3　評価システムの設定

本研究では，NN-VME のネットワーク構造として Conv-Tasnet をベースとしたものを採用した．[5] の表記に従って，ハイパーパラメータを $N = 256$, $L = 20$, $B = 256$, $H = 512$, $P = 3$, $X = 8$, $R = 4$ と設定した．また，モデルの最適化には初期学習率 0.0001 とした Adam アルゴリズムと勾配クリッピングを採用し，200 エポックまで学習を行った．Conv-Tasnet の実装には，[7] の Github リポジトリで提供されるソフトウェアを利用した．

NN-BF で使用する時間周波数マスクを推定するため，permutation invariant training に基づく音源分離モデルを採用した．ネットワーク構造等の基本的な設定は，NN-VME のものと同一である．[8] に従って，NN によって推定された (音源分離された) 各話者の時間領域信号に対し STFT を 適用し，観測信号の振幅との比を取ることで時間周波数マスクを獲得した．また，IVA と ILRMA には，オープンソースプラットフォームである pyroomacoustics [9] の実装を使用した．分離行列を推定するために，IVA では 50 回，ILRMA では 100 回の反復を行った。また，ILRMA における NMF の基底数は 2 と設定した．

NN-BF，IVA，ILRMA それぞれにおける短時間フーリエ変換の窓長は，32 ms, 64 ms, 128 ms, 256 ms, 512 ms の中から各残響時間ごとに最も SDR が高かったものを採用し，シフト長は窓長の 1/4 とした．

3.1.4　評価尺度

本実験では，評価尺度として BSSEval version 4 の signal-to-distortion ratio (SDR) を採用した．

時間領域の推定信号 $\bar{\mathbf{s}}_{\mathrm{est}} \in \mathbb{R}^{\mathcal{T}}$ と参照信号 $\bar{\mathbf{s}}_{\mathrm{ref}} \in \mathbb{R}^{\mathcal{T}}$ が与えられたもとで，BSSEval version 4 の SDR は次のように定義される:

$$\mathrm{SDR}(\bar{\mathbf{s}}_{\mathrm{ref}}, \bar{\mathbf{s}}_{\mathrm{est}}) = 10 \log_{10} \frac{||\bar{\mathbf{s}}_{\mathrm{ref}}||^2}{||\bar{\mathbf{s}}_{\mathrm{ref}} - \bar{\mathbf{s}}_{\mathrm{est}}||^2} \quad (3)$$

本実験では，仮想マイク単体での推定精度を評価するための評価尺度として，$\mathrm{SDR}_{\mathrm{VM}} = \mathrm{SDR}(\bar{\mathbf{t}}, \bar{\mathbf{v}})$ を使用した．ここで，2.1.2 節より $\bar{\mathbf{v}}$ は NN-VME によって推定された仮想マイク信号を，$\bar{\mathbf{t}}$ はその参照信号となる仮想マイクの位置に存在する実マイク信号をそれぞれ表す．また，仮想マイク信号をアレイ信号処理 (Array Processing) に適用した際の有効性を評価するための評価尺度として，$\mathrm{SDR}_{\mathrm{AP}} = \mathrm{SDR}(\bar{\mathbf{s}}, \bar{\mathbf{y}})$ を使用した．ここで，$\bar{\mathbf{y}} \in \mathbb{R}^{\mathcal{T}}$ は時間領域での分離信号を，$\bar{\mathbf{s}} \in \mathbb{R}^{\mathcal{T}}$ はその対応する残響ありクリーン音声 (i.e. spatial image) を表す．評価の際，リファレンスマイクとして 4 ch 目の信号を使用した．最終的な音源分離性能は，3 人の話者の $\mathrm{SDR}_{\mathrm{AP}}$ の平均値として計算した．このとき，各推定信号と各参照信号の permutation は BSSEval の実装に基づき SIR のスコアに基づいて決定された．

Table 1　SDR of virtual microphone signal [dB]

			T_{60}[ms]			
mictype	eval ch	ref ch	0	100	200	300
RM	4	5	2.9	3.7	2.6	2.2
RM	6	5	4.6	4.4	3.6	2.7
VM	5 (4,6)	5	17.8	15.9	12.5	9.9

3.2　結果と考察

3.2.1　仮想マイク信号レベルでの評価

表 1 は仮想マイクの推定精度を評価する $\mathrm{SDR}_{\mathrm{VM}}$ (3.1.4 節参照) のスコアを示す．表中において，mictype の RM は実マイク信号を表し，VM は NN-VME で推定された仮想マイク信号を表す．eval ch の列は，SDR を計算する際に推定信号として使用される仮想マイクや実マイクのチャネルインデックスを示し，また ref ch の列は，SDR を計算する際に参照信号として使用される実マイクのチャネルインデックスを示す．VM の項において，5 (4, 6) のように表記された場合には，チャネル 4,6 の実マイク信号を入力として推定されたチャネル 5 の位置のマイク信号であることを示す．ここでは，仮想マイクの $\mathrm{SDR}_{\mathrm{VM}}$ に加えて，評価のベースラインとして仮想マイクの位置に隣接する実マイクの $\mathrm{SDR}_{\mathrm{VM}}$ を計算した．

表 1 から，NN-VME によって推定された仮想マイク信号 (i.e. 5 (4, 6)) が，隣接するマイクで録音された実マイク信号 (i.e. 4 or 6) と比較しても十分に高い $\mathrm{SDR}_{\mathrm{VM}}$ を達成することが確認された．この結果は，NN-VME (i.e. 教師あり学習に基づく仮想マイク生成) が，複数人が同時に発話する劣決定条件下においても，観測された少数の実マイク信号から，それぞれの話者の空間情報を推測し，仮想マイク信号を生成するポテンシャルを持つことを示すものと考えられる．

また表から，残響が少ない環境下 (i.e. $T_{60} = 0$ms or $T_{60} = 100$ ms) において特に高い $\mathrm{SDR_{VM}}$ を示す一方で，残響が長くなるにつれて (隣接する実マイクと比較すると十分に高い $\mathrm{SDR_{VM}}$ ではあるもののの) $\mathrm{SDR_{VM}}$ が低くなっていく傾向も確認された．これらの結果は，NN-VME による仮想マイクの推定の際に，残響の影響を低減する方法 を検討する必要があることを示唆するものと考えられる．

3.2.2 アレイ信号処理レベルでの評価

表 2 は, 評価に用いたアレイ信号処理手法 (i.e. NN-BF, IVA, ILRMA) の音源分離性能を評価する $\mathrm{SDR_{AP}}$ (3.1.4 節参照) のスコアを示す．ここで，Method は評価に用いたアレイ信号処理手法を示す．used ch は処理に利用したチャネルインデックスを示している．ただし，RM，VM の列は使用した実マイクのチャネルと使用した仮想マイクのチャネルをそれぞれ表す．例えば, 2 行目の“ (2) NN-BF”は, 2 チャネル分の実マイク (i.e. RM = 4,6) と 1 チャネル分の仮想マイク (i.e. VM = 5 (4, 6)) を用いて構築された NN-BF の結果を示す．また，“ (3) NN-BF”，“ (6) IVA”，“ (9) ILRMA”の結果は，3 チャネル分の実マイクを利用した際のそれぞれの手法の結果を示すものであり，仮想マイク生成によって達成し得る上限性能 (NN-VME が仮想マイクの位置にある実マイクを完全に予測した場合に達成される) に対応するものである．また，表中の – は，劣決定条件下 (i.e. RM = 4,6) では，IVA と ILRMA の適用が不可能であることを示すものである．

表から，仮想マイクを利用した“ (2) NN-BF”は，仮想マイクを除いて同数の実マイクを用いた “ (1) NN-BF” と比較して，高い $\mathrm{SDR_{AP}}$ を達成していることが確認される．この結果は，表 1 で検証された仮想マイクの推定精度が，後段のアレイ信号処理技術へと応用された際に，その性能向上に貢献するレベルを達成していることを示すものと考えられる．

また，仮想マイクを利用した“ (5) IVA”や“ (8) ILRMA”は，特に残響が少ない環境下 ($T_{60} = 0$ ms または $T60 = 100$ ms) において，未処理の音声と比較して $\mathrm{SDR_{AP}}$ の向上を達成していることが確認される．この結果は，NN-VME が劣決定条件という本来手法を適用できない条件下 (i.e. “ (4) IVA” や “ (7) ILRMA”) において，IVA や ILRMA に代表される BSS の技術を適用可能にするポテンシャルを持つこと示すものと考えられる．

表から，NN-BF が残響時間によらず，仮想マイクの利用による一定の精度向上を示している一方で，仮想マイクを利用した “ (5) IVA”や “ (8) ILRMA”の性能が, 残響時間が多い場合に (e.g. $T_{60} = 300$ ms) , 特に低下している様子が確認される．これは，表 1 で確認された残響による仮想マイクの推定精度の劣化の影響に加え，“ (6) IVA” や “ (9) ILRMA”の結果に示されているような IVA や ILRMA そのものの性能劣化が要因であると考えられる．

Table 2　SDR improvement of enhanced signal [dB]

		used ch		$\mathrm{T_{60}}$[ms]			
	Method	RM	VM	0	100	200	300
(1)	NN-BF	4,6	None	5.8	6.1	6.2	5.8
(2)	〃	4,6	5 (4, 6)	8.1	8.3	8.2	7.1
(3)	〃	4,6,5	None	10.1	10.6	10.6	9.1
(4)	IVA	4,6	None	–	–	–	–
(5)	〃	4,6	5 (4, 6)	6.1	5.2	4.4	2.6
(6)	〃	4,6,5	None	7.7	7.8	6.4	4.2
(7)	ILRMA	4,6	None	–	–	–	–
(8)	〃	4,6	5 (4, 6)	4.6	4.0	2.8	1.2
(9)	〃	4,6,5	None	6.9	7.1	6.0	4.1

4　おわりに

本稿では，NN-VME の有効性を検証するために，様々な音響条件 (i.e.　複数話者, 残響条件) やアレイ信号処理 (i.e.　ニューラルネットワークベースのビームフォーマ，ブラインド音源分離手法) に対して NN-VME を適用する実験を行った。実験結果から，NN-VME が複数話者の発話においても仮想マイクの推定が可能であることや，NN-VME の適用によって劣決定条件下では本来適用できないブラインド音源分離手法 (i.e.　IVA，ILRMA) の適用が可能になることが確認された．一方で，残響による仮想マイクの推定精度の低下も確認された．今後の課題として，NN-VME のフレームワークに残響除去手法を組み合わせるなど，こうした残響の影響を低減するための検討が重要になってくると考えられる．

謝辞　本研究の一部は科研費 19H04131 の助成を受けた．

参考文献

[1] H. Katahira et al., “Nonlinear speech enhancement by virtual increase of channels and maximum SNR beamformer,” EURASIP JASP, Vol. 2016, No. 1, pp. 1-8, 2016.

[2] T. Ochiai et al. ”Neural network-based virtual microphone estimator,” in Proc. ICASSP, pp. 6114-6118, 2021.

[3] T. Kim, “Real-time independent vector analysis for convolutive blind source separation,” IEEE Trans. on Circuit and Systems, vol. 57, no. 7, pp. 1431-1438, 2010.

[4] D. Kitamura et al, “Determined blind source separation unifying independent vector analysis and nonnegative matrix factorization,” IEEE/ACM Trans. ASLP, vol. 24, no. 9, pp. 1626-1641, 2016.

[5] Y. Luo et al, “Conv-TasNet: Surpassing ideal time – frequency magnitude masking for speech separation,” IEEE/ACM Trans. ASLP, vol. 27, no. 8, pp. 1256-1266, 2019.

[6] J. Heymann et al, “Neural network based spectral mask estimation for acoustic beamforming,” in Proc. ICASSP, pp. 196-200, 2016.

[7] “https://github.com/funcwj/conv-tasnet,” .

[8] T. Ochiai et al. ”Beam-TasNet: Time-domain audio separation network meets frequency-domain beamformer,” In Proc. ICASSP, pp. 6384-6388, 2020.

[9] “https://github.com/LCAV/pyroomacoustics,” .