論文 / 著書情報
Article / Book Information

論題(和文)	マルチチャンネルモデルを用いた知識蒸留による単一チャンネル音声分離手法
Title(English)	Single-Channel Speech Separation Using Knowledge Distillation from Multichannel Models
著者(和文)	二通大地, HartantoRoland, 篠田浩一
Authors(English)	Daichi Nitsu, Roland Hartanto, Koichi Shinoda
出典(和文)	日本音響学会 第154回（2025年秋季）研究発表会講演論文集, , , pp. 329–330
Citation(English)	, , , pp. 329–330
発行日 / Pub. date	2025, 9

マルチチャンネルモデルを用いた知識蒸留による単一チャンネル音声分離手法*

☆二通大地，ローランドハルタント，篠田浩一 (東京科学大)

1　はじめに

音声分離技術は，音声認識や音声処理のフロントエンドとして利用されるほか，会議の議事録作成，聴覚障害者支援，スマートスピーカーなど多様な応用が可能な重要な技術である。近年，深層学習の発展により，音声分離技術は大きく進歩している。

音声分離は，マイクロフォンの数に応じてマルチチャンネルモデルと単一チャンネルモデルに分類される。マルチチャンネルモデルは複数のマイクロフォンを用いることで，音源到来方向（DOA; Direction of Arrival）や位相差などの空間情報を活用でき，高い分離性能を実現する。一方，単一チャンネルモデルは空間情報を利用できず，時間・周波数領域の特徴に依存するため，性能は劣るが，マイク 1 つで運用できる利便性から実社会での応用範囲が広いという利点がある。

本研究では，単一チャンネルモデルの性能向上を目的として，マルチタスク学習で訓練されたマルチチャンネルモデルから知識を蒸留する手法を提案する。

2　従来研究

2.1　深層学習に基づく音声分離

単一チャンネル音声分離は，空間情報を利用できないため，主に音響的特徴（時間-周波数領域）に依存した手法が採用されている。時間領域に基づく手法としては，TasNet [1] およびその改良版である畳み込みニューラルネットワーク（CNN）を基盤にした Conv-TasNet [2] が広く用いられている。また，TF-GridNet [3] は時間-周波数領域の統合的な特徴処理を実現するモデルや，Transformer を用いた Sepformer [4] なども登場している。

マルチチャンネル音声分離は，複数のマイクロフォンから得られる空間情報（DOA，位相差など）を活用することで，単一チャンネル音声分離より高精度な分離が可能である。例えば，TF-GridNet のマルチチャンネル拡張や，空間的な関係性を自己注意機構によって学習する SpatialNet [5] が提案されている。

音声分離における課題の一つに，出力と正解の順序が不定である順列曖昧性問題（Permutation Ambiguity）がある。この問題に対しては，出力と正解のすべての組み合わせを評価して最適な順列を選ぶ Permutation Invariant Training（PIT） [6] や，マルチチャンネル環境であれば DOA に基づいて順序を定める Location-Based Training（LBT） [7] が利用される。

MSDET（Multitask Speaker Separation and DOA Estimation Training） [8] は，音声分離と DOA 推定を同時に行うマルチタスク学習を採用しており，空間情報をより効果的に活用する。これにより，推定された DOA 情報を利用して順列曖昧性の解消に加え，音源分離そのものの精度の向上も達成した。

2.2　音声処理における知識蒸留

知識蒸留（Knowledge Distillation）は，大規模な教師モデルの知識を軽量な生徒モデルに転移することで，モデル圧縮や高速化を実現する手法である。出力分布（Soft Target）や中間特徴を活用することで，生徒モデルの性能向上を図る。音声処理分野においても，マルチチャンネルモデルの知識を単一チャンネルモデルに蒸留する研究が進んでおり，性能向上が報告されている [9] [10]。

3　提案手法

本研究では，単一チャンネル音声分離モデルの性能向上を目的として，マルチタスク学習によって訓練されたマルチチャンネルモデルから空間情報を知識蒸留する手法を提案する。提案手法では，教師モデルに音声分離と DOA 推定を同時に行うマルチタスクモデル MSDET を採用する。MSDET は TF-GridNet を分離モデルとして用い，高精度な音声分離を実現する。一方，生徒モデルは単一チャンネル入力の TF-GridNet であり，MSDET と同じ構造を持つが，入力チャネル数と出力処理が異なる。教師モデルの各 TF-GridNet ブロックの中間出力を抽出し，それを生徒モデルの対応するブロックと一致させるように学習する。具体的には，中間特徴の L1 ノルムの差を蒸留損失として定義する。この蒸留により，生徒モデルは空間情報を間接的に獲得できる。これにより，単一チャンネル環境でも分離性能の向上が期待される。

提案手法の全体構成を Fig. 1 に示す。

---

*Single-Channel Speech Separation Using Knowledge Distillation from Multichannel Models by Daichi NITSU, Roland HARTANTO, and Koichi SHINODA (Institute of Science Tokyo)

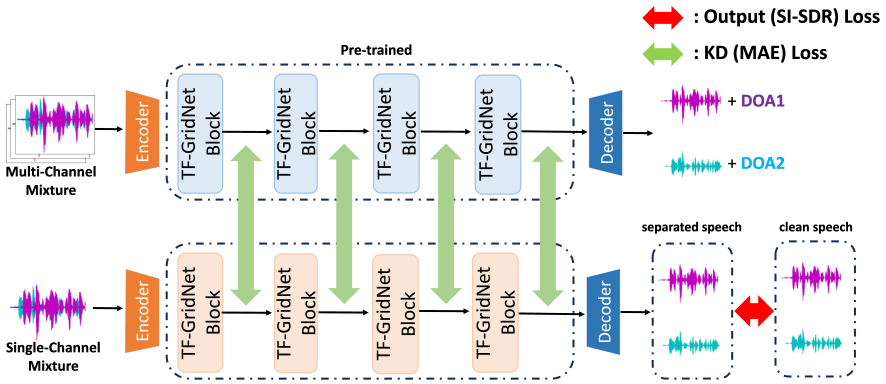

Fig. 1　提案手法の概略図

3.1　損失関数

最終的な損失関数は，出力損失 L_{Output} と蒸留損失 L_{KD} の加重和で表される。

$$L_{\text{Total}} = L_{\text{Output}} + w \cdot L_{\text{KD}}, \tag{1}$$

ここで，L_{Output} は SI-SDR 損失と Mixture Constraint（MC）損失の和であり，w は蒸留損失に対する重みで，本研究では $w = 0.01$ に設定した。また，PIT を用いて学習を行っている。

4　実験

4.1　実験設定

本研究では，単一チャンネル音声分離モデルの学習に SMS-WSJ データセットを用い，SMS-WSJ，WSJ0-2mix，WHAMR! の各データセットで性能を評価した。SMS-WSJ は，WSJ0 コーパスの音声を基に作成された残響環境下の 2 話者混合音声であり，ホワイトノイズが付加されている。WSJ0-2mix も WSJ0 コーパスを基にしたデータセットで，無響環境を想定している。一方，WHAMR! は WSJ0-2mix にカフェや路上など実環境の雑音と残響を加えたデータセットである。

モデルの実装には ESPnet を用い，教師モデルには TF-GridNet をセパレータとする MSDET の事前学習済みモデルを使用した。生徒モデルは，TF-GridNet に関する先行研究で使用されたモデルサイズおよび学習設定に従っている。

評価指標には，音声の明瞭度を評価する ESTOI と，分離性能を示す SI-SDR を用いた。

4.2　実験結果

SMS-WSJ データセットにおける提案手法の評価結果を Table 1 に示す。従来の TF-GridNet（再現実験）と比較して，提案手法は ESTOI で 0.05 %，SI-SDR で 0.13 dB の性能向上が確認された。

次に，学習に使用していないデータセットでの性能評価結果を Table 2 に示す。無響環境の WSJ0-2mix では提案手法による改善は見られなかったが，雑音や残響を含む WHAMR! においては SI-SDR が 0.3 dB 向上するなど，わずかな性能向上が確認された。これらの結果から，提案手法は特に雑音・残響環境下において有効であることが示された。

Table 1　SMS-WSJ における音声分離性能の比較

モデル	ESTOI ↑ (%)	SI-SDR ↑ (dB)
TF-GridNet（論文）	92.40	16.20
TF-GridNet（再現実験）	92.99	16.99
MCKD-SS（提案手法）	**93.04**	**17.12**

Table 2　WSJ0-2mix および WHAMR! における音声分離性能の比較

Dataset	System	ESTOI ↑ (%)	SI-SDRi ↑ (dB)
WSJ0-2Mix	TF-GridNet (再現実験)	**76.68**	**9.52**
	TF-GridNet + 提案手法	69.51	5.36
WHAMR!	TF-GridNet (再現実験)	47.91	4.29
	TF-GridNet + 提案手法	**47.92**	**4.59**

5　おわりに

本研究では，単一チャンネル音声分離の性能向上を目的に，マルチチャンネルモデルの中間層情報を活用する知識蒸留手法を提案した。SMS-WSJ において従来手法を上回る性能を示し，特に雑音や残響環境で有効であることが確認された。

謝辞　本研究は JSPS 科研費 JP23H00490 の助成を受けたものです。

参考文献

[1] Luo, Mesgarani, ICASSP, 696-700, 2018
[2] Wu et al., PSGEC, 965-969, 2023
[3] Wang et al., TASLP, vol.31, 3221-3236, 2023
[4] Subakan et al., ICASSP, 21-25, 2021
[5] Quan, Li, TASLP, vol.32, 1310-1323, 2024
[6] Kolbæk et al., TASLP, vol.25, 1901-1913, 2017
[7] Taherian et al., TASLP, vol.30, 2791-2800, 2022
[8] Hartanto et al., Interspeech, 2170-2174, 2024
[9] Horiguchi et al., SLT, 620-625, 2023
[10] Xu, LSP, vol.31, 386-390, 2024