論文 / 著書情報
Article / Book Information

論題	単一チャンネル音声分離のためのマルチチャンネルモデルを用いた知識蒸留手法
Title	Multi-Channel Knowledge Distillation for Single-Channel Speech Separation
著者	二通大地, Roland Hartanto, 篠田浩一
Authors	Daichi Nitsu, Roland Hartanto, Koichi Shinoda
出典	電子情報通信学会技術研究報告, Vol. 125, no. 74, pp. 10-15
Citation	IEICE technical report, Vol. 125, no. 74, pp. 10-15
発行日 / Pub. date	2025, 6
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright(c) 2025 IEICE

一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2025-3(2025-06)

単一チャンネル音声分離のための マルチチャンネルモデルを用いた知識蒸留手法

二通　大地†　　ローランド　ハルタント†　　篠田　浩一†

† 東京科学大学　〒152–8552 東京都目黒区大岡山 2–12–1
E-mail: †{nitsu@ks.c,roland@ks.c,shinoda@c}.titech.ac.jp

あらまし　音声分離は，雑音環境下での音声処理において重要な技術であり，補聴器，音声認識システム，スマートスピーカーなど幅広い分野で応用されているが，単一チャンネルでは空間情報が利用できず，性能が劣る．本研究では，マルチチャンネルモデルから単一チャンネルモデルへの知識蒸留手法 MCKD-SS を提案する．提案手法では，音声分離と音源到来方向（DOA）推定を同時に行うマルチタスクモデルから中間層の空間情報を抽出し，単一チャンネルモデルに蒸留する．SMS-WSJ や WHAMR!で性能向上を確認し，無響環境では効果が限定的だったが，雑音・残響環境での有効性を示した．

キーワード　音声分離, 知識蒸留, 深層学習

Multi-channel Models Knowledge Distillation for Single-Channel Speech Separation

Daichi NITSU†, Roland HARTANTO†, and Koichi SHINODA†

† Institute of Science Tokyo　2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8552 Japan
E-mail: †{nitsu@ks.c,roland@ks.c,shinoda@c}.titech.ac.jp

Abstract　Speech separation is an essential technology for processing speech in noisy environments, with applications in hearing aids, speech recognition systems, and smart speakers. However, single-channel models cannot utilize spatial information and tend to have lower performance. This study proposes MCKD-SS (Multi-Channel Knowledge Distillation for Single-Channel Speech Separation), which distills spatial information from the intermediate layers of a multi-task model that simultaneously performs speech separation and direction of arrival (DOA) estimation. Experiments on the SMS-WSJ and WHAMR! datasets confirmed performance improvements, demonstrating effectiveness in noisy and reverberant environments, although the effect was limited in clean conditions.

Key words　Speech Separation, Knowledge Distillation, Deep Learning

1. はじめに

近年，音声認識システムの発展により，人間の音声を正確に処理する技術が実用化されている．しかし，多くのシステムは単一話者を前提に設計されており，複数話者が同時に発話する状況では性能が大きく低下する．この問題に対処するには，混合音声から個々の音源を分離する音声分離技術の開発が不可欠である．

音源分離技術は古くから研究が行われており，従来は信号処理技術に基づく手法が主流であった．しかし，近年の深層学習技術の飛躍的な進展に伴い，ディープニューラルネットワークを用いた手法が注目されている [1]．最近の研究では，リカレントニューラルネットワーク（RNN）[2] や自己注意（self-attention）[3] 機構を活用した新たなアーキテクチャの導入により，性能向上が図られている [4]〜[8]．これらの手法は，従来手法と比較して，混合音声から各音源を高精度で分離する能力を有しており，多くの音響環境で優れた性能を示している．

音声分離技術は，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利用することで空間的情報（例：到来方向，位相差など）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的情報を利用できず，時間・周波数領域の特徴に依存するために性能が劣るが，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を持つ．

本研究では，単一チャンネルモデルの性能向上を目的に，マ

ルチチャンネルモデルからの知識蒸留を用いた新たな学習方法 MCKD-SS（**M**ulti-**C**hannel **K**nowledge **D**istillation for **S**ingle-Channel Speech **S**eparation）を提案する．提案手法では，音声分離と音源到来方向（Direction of Arrival: DOA）推定を同時に行うマルチタスクモデルから，中間層の空間情報を抽出し，単一チャンネルモデルに蒸留する．本手法により，単一チャンネルモデルの高性能化と広範な実用性の両立を図る．

2 節では，音声分離の従来手法と知識蒸留に関する背景を述べる．3 節では，提案手法である MCKD-SS の詳細を説明する．4 節では，実験の設定について，使用するデータセットや評価指標を含めて記述する．5 節では，実験結果の分析および考察を行う．6 節では，本研究の結論と今後の展望について述べる．

2. 従 来 研 究

2.1 深層学習に基づく単一チャンネル音声分離

深層学習に基づく単一チャンネル音声分離は，単一のマイクロフォンから収録された混合音声信号から個々の音源を分離する技術であり，空間情報が得られないため，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 [9] 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 [10] が広く用いられている．また，TF-GridNet [11] 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であり，エンコーダ・デコーダ構造に加え，複数の TF-GridNet Block からなる特徴抽出部を持つ．各ブロックは，フレーム内スペクトル特徴を扱う Intra-Frame Full-Band Module，帯域内時間変化を学習する Sub-Band Temporal Module，フレーム間の長距離依存性を扱う Cross-Frame Self-Attention Module から構成されており，時間・周波数の両方向の情報を精緻に捉えることができる．さらに，Transformer ベースの時間次元と周波数次元の両方に自己注意を適用する構造を持つ Sepformer [6] や，短時間変化も補足可能な FSMN（Feedforward Sequential Memory Network）を統合した Mossformer2 [7] なども登場しており，長短期依存性の統合処理によって高い分離性能を実現している．

2.2 深層学習に基づくマルチチャンネル音声分離

深層学習に基づく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，より高精度な分離が可能である [1]．近年では，空間情報と時間-周波数情報を統合的に学習する手法が提案されている．

たとえば，TF-GridNet のマルチチャンネル拡張版 [5] のほか，音源方向やマイクロフォン間の空間的関係性を自己注意機構によって学習する SpatialNet [8]，および SpatialNet を基に，帯域構造に応じたモジュールを組み合わせて設計された CrossNet [12] などがあり，いずれも実環境下で高い分離性能を示している．

音声分離における課題の一つに「順序が不定な出力と正解の対応関係をどう扱うか」という順列曖昧性問題（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）[13] や，マルチチャンネルであれば DOA（Direction of Arrival）に基づいて順序を定める Location-Based Training（LBT）[14] を使うことができる．

MSDET（Multitask Speaker Separation and DOA Estimation Training）[15] 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も向上し，特に音源間の方向差が大きい環境で高い効果を示すことが報告されている．

2.3 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 [16], [17]．出力分布（Soft Target）や中間特徴を活用して，生徒モデルの性能向上を図る．

音声処理分野では，マルチチャンネルモデルの空間情報を単一チャンネルモデルに蒸留する研究が進んでいる．Horiguchi ら [18] は，マルチチャンネルモデルと単一チャンネルモデルの相互学習を導入し，両者の出力を同時に最適化することで話者ダイアライゼーションの性能を改善した．

また，Xu ら [19] は，単一チャンネル音声を仮想的にバイノーラル音声に変換する手法を提案し，教師モデルのエンコーダから得られる空間的特徴を生徒モデルに伝達している．具体的には，単一チャンネル信号を左右チャンネルの疑似信号に変換し，それをバイノーラルモデルに入力，得られた中間特徴をヒントとして用いることで，空間的知識を単一チャンネルモデルに模倣させている．

本研究では，MSDET を教師モデルとし，その中間層で学習された空間的特徴を TF-GridNet ベースの単一チャンネルモデルに蒸留することで，空間情報の利用が困難な環境でも高精度な音声分離を実現することを目指す．

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，マルチチャンネルモデルが持つ空間情報を知識蒸留により単一チャンネルモデルに学習させる新しい手法 MCKD-SS（**M**ulti-**C**hannel **K**nowledge **D**istillation for **S**ingle-Channel Speech **S**eparation）を提案する．

3.1 教師モデルと生徒モデル

図 1 は，通常の単一チャンネルモデル（TF-GridNet）の学習プロセスを示している．このプロセスでは，入力音声を時間周波数（Time-Frequency: TF）表現に変換し，TF-GridNet ブロックを通じて特徴を抽出および処理が行われる．最終的にモデルは分離結果を出力し，ターゲット信号との誤差を最小化するように学習される．この従来手法は構造が単純である一方，単一チャンネルの特性上，マルチチャンネルデータの持つ情報を活用することはできないという制約がある．

図 2 に示す本研究の提案手法 MCKD-SS は，単一チャンネルモデル（生徒モデル）を基盤とし，マルチチャンネルモデル（教師モデル）からの知識蒸留によって分離性能の向上を図るものである．

教師モデルにはマルチチャンネル音声分離と DOA 推定を同

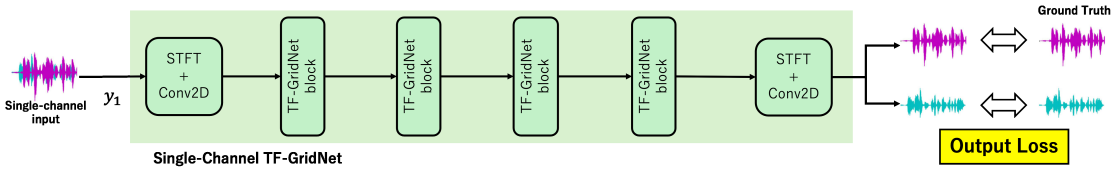

図 1　従来の単一チャンネル音声分離手法

Fig. 1　Conventional single-channel speech separation

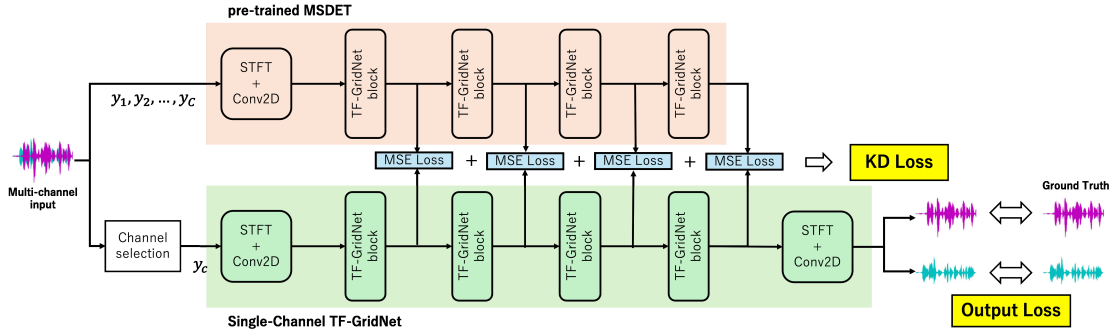

図 2　本研究の提案手法（MCKD-SS）

Fig. 2　Proposed method (MCKD-SS)

時に行うマルチタスクモデル MSDET [15] を採用する．MSDET は，音声分離タスクと DOA 推定タスクを統合的に学習することで，マルチチャンネル情報を効果的に活用し，高精度な音声分離性能を実現する．本研究では，TF-GridNet を分離モデルとして用いた MSDET の事前学習済み（pre-trained）モデルを使用した．一方，生徒モデルには単一チャンネルの TF-GridNet モデルを採用する．生徒モデルの目的は，マルチチャンネルモデルの空間情報を学習し，単一チャンネル環境でも高い音声分離性能を発揮することである．教師モデルと生徒モデルの構造を比較すると，主な違いは入力処理層と出力処理層にあるが，両モデルのコアとなる TF-GridNet ブロックの構造とサイズは一致している．

本研究では，教師モデルの持つ知識を効率的に蒸留するために，各 TF-GridNet ブロックの中間出力を活用する知識蒸留手法を導入する．具体的には，これらの中間出力を蒸留プロセスに利用することで，教師モデルが持つ空間的情報を生徒モデルに反映させる．この手法により，単一チャンネルモデルの性能を向上させることが期待される．

3.2　損 失 関 数

出力損失関数は，生徒モデルが推定した音声とクリーンなターゲット音声との誤差を最小化することを目的とする．ターゲット音声は，話者ごとのクリーン音声で構成されており，モデルは Permutation Invariant Training（PIT）[13] を用いて学習を行う．PIT は，推定された音源とターゲット音源の最適なペアリングを動的に探索し，話者の順序に依存しない音源分離を可能にする手法である．

この損失関数は，従来の Scale-Invariant Signal-to-Distortion Ratio（SI-SDR）[20] 損失を基盤としている．SI-SDR は，音声信号間の類似性をスケールに依存しない形で評価する指標であり，音源分離タスクにおける損失関数や性能評価によく用いられる．さらに，本研究では，[11] を参考に SI-SDR 損失に加え Mixture Constraint (MC) 損失項を導入している．MC 損失は，混合信号全体の整合性を維持することを目的としており，出力損失関数は次式で定式化される：

$$\mathcal{L}_{\text{Output}} = \mathcal{L}_{\text{SI-SDR}} + \frac{1}{N}\left\| \sum_{c=1}^{C} \hat{\alpha}_q^{(c)} \hat{s}_q^{(c)} - \sum_{c=1}^{C} s_q^{(c)} \right\|_1,$$

ここで，$\|\cdot\|_1$ は L1 ノルム，N はサンプル数，C は話者数を表す．$\hat{\alpha}_q^{(c)}$ はスケーリング係数，$\hat{s}_q^{(c)}$ は推定音源，$s_q^{(c)}$ はターゲット音源を表す．

教師-生徒フレームワークにおける知識蒸留の損失関数は，教師モデルと生徒モデルの中間層出力間の類似性を評価し，空間情報の伝達を行う．具体的には，対応する TF-GridNet ブロックの出力の差を L1 ノルムで測定し，次式で定式化される：

$$\mathcal{L}_{\text{KD}} = \frac{1}{BM}\sum_{i=1}^{B}\left\| S^i - T^i \right\|_1,$$

ここで，S^i と T^i はそれぞれ生徒モデルと教師モデルの i 番目の TF-GridNet ブロックの出力を示し，M は S^i, T^i の要素数を表し，B は TF-GridNet ブロックの数を表す．

最終的な損失関数は，出力損失と知識蒸留損失の加重和として定義され，以下のように表される：

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Output}} + w\mathcal{L}_{\text{KD}},$$

ここで，w は蒸留損失に対する重みを表し，本研究では $w = 0.01$ に設定した．この設定により，出力損失を主に最適化しつつ，知識蒸留による特徴伝達を補助的に活用している．

4.　実 験 条 件

4.1　データセット

本研究では，音声分離モデルの学習には SMS-WSJ [21] を用

い，評価には SMS-WSJ，WSJ0-2mix [22]，WHAMR! [23] の 3 種のデータセットを使用した．これらのデータセットは異なる音響条件を反映しており，提案手法の性能を多角的に検証する目的で選定した．

SMS-WSJ [21] は，反響条件下での 2 話者音声分離の性能評価に広く利用されているベンチマークデータセットである．WSJ0 および WSJ1 コーパスに含まれる音声から構成され，トレーニング 87.4 時間，検証 2.5 時間，テスト 3.4 時間のデータが用意される．話者セットは重複せず，サンプリングレートは 8 kHz，直径 20 cm の円形 6ch マイクアレイで収録されている．音源距離は 1.0〜2.0 m，残響時間（T60）は 0.2〜0.5 秒でランダムに設定され，20〜30 dB のホワイトノイズが加えられている．本研究では，単一チャンネルモデルの学習には 1 チャンネル目を，教師モデルにも同データセットを使用した．

WSJ0-2mix [22] は，無響環境における単一チャンネル音声分離の性能評価に広く利用されているベンチマークデータセットである．WSJ0 コーパスのクリーン音声を使用し，2 話者の混合音声を提供する．テストデータは 4.8 時間（3,000 個）で，話者間エネルギー差は − 5〜5 dB に設定され，サンプリングレートは 8 kHz である．

WHAMR! [23] は，WSJ0-2mix に雑音と残響を加えた拡張版である．カフェ，街頭，公共交通機関など日常環境由来の非定常な雑音と残響（T60 が 0.2 から 1.0 秒）が加えられており，信号対雑音比（SNR）は − 6 から 3 dB の範囲で設定され，話者とマイクの距離は 0.66 から 2.0 m である．

4.2 学 習 設 定

本研究では，音声分離モデルの実装に ESPnet ツールキット [24] を使用し，教師モデルには，TF-GridNet をセパレーターとする MSDET の事前学習済みモデルを用いた．一方，生徒モデルは，TF-GridNet に関する先行研究 [5] 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．学習率は 0.001 から開始し，トレーニング中に勾配が発散する問題が確認されたことから，発散を防ぎつつ安定した学習を実現するために，学習率を段階的に半分に減らすスケジュールを適用した．具体的には，トレーニング中の損失が収束せず，勾配が不安定になる兆候（勾配が Inf になる現象）が観測されたタイミングで学習率を半分に調整した．この勾配の発散は知識蒸留モデルおよび非知識蒸留モデルの双方で確認されたため，発散は知識蒸留特有の課題ではない．そのため，この学習率調整が両モデル間の比較における公平性を損なうことはなく，性能評価の妥当性を保つ設定であると判断した．加えて，最適化には Adam を採用し，勾配クリッピングの L2 ノルムは 1.0 に設定した．

4.3 評 価 指 標

音声分離性能を評価するため, 客観的な指標として ESTOI [25], SI-SDR [20] を用いた．ここで，以下の定義に従い，クリーンな音声データを s，処理された音声データを $\hat{s}$，環境雑音を n とする（$s+n$ が音声強調の処理対象となる音声データを示す）．

ESTOI (Extended Short-Time Objective Intelligibility) は，音声信号の明瞭度を評価する指標である．元の STOI (Short-Time Objective Intelligibility) は，短時間ごとの音声フレームの相関を数値化することで，目標音声と処理後の音声の知覚的な類似度を評価する．ESTOI はこれを拡張し，フレーム単位での評価精度を高めることで，音声の自然さや明瞭度をより詳細に捉える．

SI-SDR (Scale-Invariant Signal-to-Distortion Ratio) は，音声のスケールに依存せず，分離された音声と元音声との類似度を評価する指標である．歪み成分や混合成分の残留を定量的に評価することで，分離精度を示す．計算式は以下の通りである：

$$\mathrm{SI\text{-}SDR}(s,\hat{s}) = 10\log_{10}\left(\frac{\|s\|^2}{\|s-\beta\hat{s}\|^2}\right),$$
$$\text{where}\quad \beta = \arg\min_{\beta}\|s-\beta\hat{s}\|^2$$

これらの指標はいずれも，値が大きいほど高い音質・明瞭度を示す．

表 1　SMS-WSJ における音声分離性能の比較
Table 1　Comparison of speech separation performance on SMS-WSJ

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**

5. 実 験 結 果

5.1 SMS-WSJ における音声分離性能の評価

表 1 は，TF-GridNet における音声分離性能を ESTOI と SI-SDR の 2 つの指標で評価した結果を示している．本研究では，オリジナル論文 [5] の結果を基準に再現実験を行い，提案手法 MCKD-SS の性能を検証し，既存の手法との比較を行った．

再現実験では，オリジナル論文の結果と比較して ESTOI が 92.40 から 92.99 へと向上し，SI-SDR も 16.20dB から 16.99dB へと改善した．この性能向上は，再現実験時に学習率を調整することでモデルの勾配発散を抑え，学習を安定化させたことが影響している可能性がある．この結果はオリジナル論文の結果と完全に一致していないことを意味し，何らかの実験条件が異なっている可能性がある．この点を踏まえ，再現実験の結果を基準としながら提案手法の有効性を検証した．

提案手法である MCKD-SS は，再現実験モデルと比較して ESTOI が 92.99 から 93.04 へと向上し，SI-SDR も 16.99dB から 17.12dB へと改善した．

5.2 音源間の到来方向差ごとの SI-SDR 比較

表 2 は，SMS-WSJ データセットを用いて音源間の到来方向差（DOA 差）$\triangle d$ ごとに SI-SDR を評価した結果を示している．提案手法である MCKD-SS は，全ての音源間の到来方向差において再現実験モデルを上回る性能を示している．音源がほぼ正面に位置している $0° \le d < 10°$ の範囲では，MCKD-SS が再現実験モデルを 0.14 dB 上回る結果を示した．次に，$10° \le d < 30°$ および $30° \le d < 60°$ の範囲では，それぞれ 0.15 dB の向上が確認された．最後に，$60° \le d \le 180°$ の範囲では，MCKD-SS が 0.19 dB の向上を示し，DOA 差が大きいほど性能向上が顕著であった．これらの結果から，MCKD-SS（提案手法）は，DOA 差が小さい状況でも一定の有効性を示す一方で，DOA 差が大き

表 2　音源間の到来方向差ごとの SI-SDR 比較
Table 2　SI-SDR comparison by inter-source DOA difference

モデル	$0° \leq \triangle d < 10°$	$10° \leq \triangle d < 30°$	$30° \leq \triangle d < 60°$	$60° \leq \triangle d \leq 180°$
TF-GridNet（再現実験）	16.79	17.19	17.18	17.41
MCKD-SS（提案手法）	**16.93**	**17.34**	**17.33**	**17.60**

表 3　WSJ0-2mix における音声分離性能の比較
Table 3　Comparison of speech separation performance on WSJ0-2mix

モデル	ESTOI ↑ (%)	SI-SDR ↑ (dB)
TF-GridNet（再現実験）	**76.68**	**9.52**
MCKD-SS（提案手法）	69.51	5.36

表 4　WHAMR!における音声分離性能の比較
Table 4　Comparison of speech separation performance on WHAMR!

モデル	ESTOI ↑ (%)	SI-SDR ↑ (dB)
TF-GridNet（再現実験）	47.91	-1.84
MCKD-SS（提案手法）	**47.92**	**-1.54**

い状況では性能向上がより目立つ傾向があることが示された．

5.3　異なるデータセットにおける音声分離性能の評価

SMS-WSJ でトレーニングされたモデルの一般化性能を評価するために，WSJ0-2mix および WHAMR!データセットを用いて検証を行い，トレーニングデータとは異なる環境での性能を明らかにした．

表 3 に示される WSJ0-2mix での評価結果によると，再現実験モデルは ESTOI が 76.68，SI-SDR が 9.52 dB であった．一方，提案手法である MCKD-SS は，ESTOI が 69.51，SI-SDR が 5.36dB と，再現実験モデルに比べて低い性能を示した．一方，表 4 に示される WHAMR!での評価結果では，再現実験モデルが ESTOI が 47.91，SI-SDR が-1.84dB を記録したのに対し，MCKD-SS は ESTOI が 47.92，SI-SDR が-1.54dB とわずかに性能が向上している．全体として，MCKD-SS は，雑音と残響を含む環境下では一定の効果を発揮する一方で，無響環境では性能が低下する傾向が確認された．

5.4　考　　察

SMS-WSJ での評価により，提案手法 MCKD-SS は ESTOI および SI-SDR の両指標で再現実験モデルを上回る性能を示した．音源間の到来方向差（DOA 差）ごとの分析では，DOA 差が大きいほど性能向上が顕著であり，これはマルチチャンネルモデルの性能向上傾向 [15] と一致していた．この結果は，知識蒸留が単一チャンネル音声分離モデルにおいて空間情報の活用を促進し，性能向上に寄与している可能性を示唆している．

また，一般化性能の評価では，無響環境の WSJ0-2mix においては効果が限定的だった一方，雑音・残響を含む WHAMR!では SMS-WSJ と同様，性能向上が確認された．これは，残響を手がかりとした空間情報の間接的な利用が寄与した結果と考えられる．

音声分離モデルは，話者の音程や声質といった話者特徴などの音響情報と，音源の位置や方向に関連する空間的情報を組み合わせて処理する．一般的に，単一チャンネルモデルでは直接的な空間的情報を得ることが難しく，主に音響的情報に基づいて分離を行う．一方，知識蒸留を導入した提案手法では，単一チャンネルモデルにおいても空間的情報がより活用されていることが確認された．この点は，マルチチャンネルモデルから学習した空間的情報の処理能力が，知識蒸留を通じて単一チャンネルモデルに伝達された結果と考えられる．特に残響や雑音が存在する環境下で効果的に機能する点は，単一チャンネル音声分離モデルの新たな可能性を広げる重要な成果といえる．

6.　ま　と　め

本研究では，単一チャンネル音声分離性能の向上を目的に，マルチチャンネルモデルの知識を蒸留する手法 MCKD-SS を提案した．教師モデルには，音声分離と音源到来方向（DOA）推定を同時に行うマルチタスクモデル MSDET を用い，その中間層の出力を生徒モデル（TF-GridNet）に転移することで，空間情報の活用を可能にした．

SMS-WSJ における評価では，従来手法を上回る ESTOI および SI-SDR を達成した．特に DOA 差が大きい場合により性能向上が見られた．また，WSJ0-2mix と WHAMR! による検証では，無響環境では効果が限定的だった一方，雑音・残響環境では提案手法が従来手法を上回る性能を示した．

これらの結果から，MCKD-SS は単一チャンネルモデルに空間的な分離能力を付与できる有効な手法であり，実環境下での活用に向けた有望なアプローチであると確認された．

今後は，さらなる一般化性能の検証が課題である．特に，異なる残響・雑音条件や実環境データにおける性能の確認が必要である．また，蒸留損失と出力損失の重み付けや損失関数の改良（コサイン類似度や KL ダイバージェンスの導入）による性能向上も期待される．加えて，他のモデル構造への展開や，知識蒸留に代わるファインチューニングによる性能改善も有効なアプローチと考えられる．

謝　　辞

本研究は JSPS 科研費 JP23H00490 の助成を受けた．

文　　献

[1] Wang, D. and Chen, J.: Supervised Speech Separation Based on Deep Learning: An Overview, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 26, No. 10, pp. 1702–1726 (online), DOI: 10.1109/TASLP.2018.2842159 (2018).

[2] Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, *Neural Comput.*, Vol. 9, No. 8, pp. 1735–1780 (online), DOI: 10.1162/neco.1997.9.8.1735 (1997).

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I.: Attention Is All You Need (2017).

[4] Luo, Y., Chen, Z. and Yoshioka, T.: Dual-Path RNN: Efficient Long Sequence Modeling for Time-Domain Single-Channel Speech Separation, *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 46–50 (online), DOI: 10.1109/ICASSP40776.2020.9054266 (2020).

[5] Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y. and Watanabe, S.: TF-GridNet: Integrating Full- and Sub-Band Modeling for Speech Separation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 31, pp. 3221–3236 (online), DOI: 10.1109/TASLP.2023.3304482 (2023).

[6] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M. and Zhong, J.: Attention Is All You Need in Speech Separation, *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25 (online), DOI: 10.1109/ICASSP39728.2021.9413901 (2021).

[7] Zhao, S., Ma, Y., Ni, C., Zhang, C., Wang, H., Nguyen, T. H., Zhou, K., Yip, J. Q., Ng, D. and Ma, B.: MossFormer2: Combining Transformer and RNN-Free Recurrent Network for Enhanced Time-Domain Monaural Speech Separation, *Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10356–10360 (online), DOI: 10.1109/ICASSP48485.2024.10445985 (2024).

[8] Quan, C. and Li, X.: SpatialNet: Extensively Learning Spatial Information for Multichannel Joint Speech Separation, Denoising and Dereverberation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 1310–1323 (online), DOI: 10.1109/TASLP.2024.3357036 (2024).

[9] Luo, Y. and Mesgarani, N.: TasNet: Time-Domain Audio Separation Network for Real-Time, Single-Channel Speech Separation, *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700 (online), DOI: 10.1109/ICASSP.2018.8462116 (2018).

[10] Wu, H., Lu, X., Feng, Y., Zeng, Z., Huang, Z., Hu, Z. and Li, Z.: ConvTasNet Based Transformer Sound Noise Reduction and Condition Recognition Network, *Proceedings of the 2023 Power System and Green Energy Conference (PSGEC)*, pp. 965–969 (online), DOI: 10.1109/PSGEC58411.2023.10255967 (2023).

[11] Wang, Z.-Q., Cornell, S., Choi, S., Lee, Y., Kim, B.-Y. and Watanabe, S.: TF-GRIDNET: Making Time-Frequency Domain Models Great Again for Monaural Speaker Separation, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5 (online), DOI: 10.1109/ICASSP49357.2023.10094992 (2023).

[12] Lin, Y., Zhang, T., Mao, Y. and Zhong, S.: CrossNet: A Low-Latency MLaaS Framework for Privacy-Preserving Neural Network Inference on Resource-Limited Devices, *IEEE Transactions on Dependable and Secure Computing*, pp. 1–17 (online), DOI: 10.1109/TDSC.2024.3431590 (2024).

[13] Kolbæ k, M., Yu, D., Tan, Z.-H. and Jensen, J.: Multitalker Speech Separation With Utterance-Level Permutation Invariant Training of Deep Recurrent Neural Networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 25, No. 10, pp. 1901–1913 (online), DOI: 10.1109/TASLP.2017.2726762 (2017).

[14] Taherian, H., Tan, K. and Wang, D.: Multi-Channel Talker-Independent Speaker Separation Through Location-Based Training, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 30, pp. 2791–2800 (online), DOI: 10.1109/TASLP.2022.3202129 (2022).

[15] Hartanto, R., Sakti, S. and Shinoda, K.: MSDET: Multitask Speaker Separation and Direction-of-Arrival Estimation Training, *Proceedings of Interspeech 2024*, pp. 2170–2174 (online), DOI: 10.21437/Interspeech.2024-2537 (2024).

[16] Buciluã, C., Caruana, R. and Niculescu-Mizil, A.: Model Compression, *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, ACM, pp. 535–541 (2006).

[17] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network (2015).

[18] Horiguchi, S., Takashima, Y., Watanabe, S. and Garcia, P.: Mutual Learning of Single- and Multi-Channel End-to-End Neural Diarization, *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 620–625 (online), DOI: 10.1109/SLT54892.2023.10023388 (2023).

[19] Xu, X.: Improving Monaural Speech Enhancement by Mapping to Fixed Simulation Space With Knowledge Distillation, *IEEE Signal Processing Letters*, Vol. 31, pp. 386–390 (online), DOI: 10.1109/LSP.2024.3355746 (2024).

[20] Le Roux, J., Wisdom, S., Erdogan, H. and Hershey, J. R.: SDR – Half-Baked or Well Done?, *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630 (online), DOI: 10.1109/ICASSP.2019.8683855 (2019).

[21] Drude, L., Heitkaemper, J., Boeddeker, C. and Haeb-Umbach, R.: SMS-WSJ: Database, Performance Measures, and Baseline Recipe for Multi-Channel Source Separation and Recognition (2019).

[22] Hershey, J. R., Chen, Z., Le Roux, J. and Watanabe, S.: Deep Clustering: Discriminative Embeddings for Segmentation and Separation, *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 31–35 (online), DOI: 10.1109/ICASSP.2016.7471631 (2016).

[23] Maciejewski, M., Wichern, G., McQuinn, E. and Le Roux, J.: WHAMR!: Noisy and Reverberant Single-Channel Speech Separation, *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700 (online), DOI: 10.1109/ICASSP40776.2020.9053327 (2020).

[24] Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A. and Ochiai, T.: ESPnet: End-to-End Speech Processing Toolkit, *Proceedings of Interspeech 2018*, pp. 2207–2211 (online), DOI: 10.21437/Interspeech.2018-1456 (2018).

[25] Jensen, J. and Taal, C. H.: An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 24, No. 11, pp. 2009–2022 (online), DOI: 10.1109/TASLP.2016.2585878 (2016).