ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746022

SELECTIVE MUTUAL LEARNING: AN EFFICIENT APPROACH FOR SINGLE CHANNEL SPEECH SEPARATION

Ha Minh Tan⋆ *Duc-Quang Vu*⋆ *Chung-Ting Lee* *Yung-Hui Li* *Jia-Ching Wang*

Dept. of Computer Science and Information Engineering, National Central University, Taiwan

ABSTRACT

Mutual learning, the related idea to knowledge distillation, is a group of untrained lightweight networks, which simultaneously learn and share knowledge to perform tasks together during training. In this paper, we propose a novel mutual learning approach, namely selective mutual learning. This is the simple yet effective approach to boost the performance of the networks for speech separation. There are two networks in the selective mutual learning method, they are like a pair of friends learning and sharing knowledge with each other. Especially, the high-confidence predictions are used to guide the remaining network while the low-confidence predictions are ignored. This helps to remove poor predictions of the two networks during sharing knowledge. The experimental results have shown that our proposed selective mutual learning method significantly improves the separation performance compared to existing training strategies including independently training, knowledge distillation, and mutual learning with the same network architecture.

Index Terms— Supervised speech separation, monophonic source separation, time domain audio separation.

1. INTRODUCTION

In recent years, deep learning-based speech separation has achieved impressive performance. The speech separation methods include two main categories such as time-frequency (TF) domain-based methods and time-domain-based methods. In the TF domain, the speech separation methods aim to approximate the estimated spectrum with the clean spectrum of the speakers [1, 2]. Moreover, several methods use the TF mask as a training target, so they have achieved significant performance by improving the accuracy for the estimated mask [3, 4, 5, 6, 7]. The short-time Fourier transform (STFT) is an important technique adopted to create a TF representation of the mixed signal. These TF representations are separated into individual sources which are used to reconstruct the source waveform by the inverse STFT (iSTFT). There are several limitations that arise in the TF domain-based methods. For example, the TF representation in a complex domain contains both the magnitude and phase of the signal, however, addressing the phase problem is very difficult, so most of the proposed methods only solve the magnitude of the mixture and have an upper bound on separation performance. Although several approaches use the phase information to design the reference masks, e.g., the phase-sensitive mask [8] and complex ratio mask [9], the separation performance still exists in the upper bound. Moreover, source separation in the TF domain for effective performance requires high-frequency resolution e.g., the window length of 32 ms for speech separation [3, 4] and 90 ms for music separation [10]. Therefore, these approaches are limited in applications for very short latency systems or real-time systems.

The recent time-domain models overcome the TF domain's limitations, the time-domain-based methods directly separate the mixed waveform, and have achieved significant progress, e.g., the time-domain audio separation networks (TasNet) [11], fully-convolutional time-domain audio separation network (Conv-TasNet) [12], and the time convolutional networks (TCNs) [13, 14, 15]. Very recently, the dual-path recurrent neural network (DPRNN) [16] is proven to be promising for speech separation, which is capable of modeling extremely long-time sequences with state-of-the-art performance. The DPRNN first divides the long input sequence into shorter segments and length-fixed segments and then it performs local and global processing using an intra-segment RNN and an inter-segment RNN iteratively on segments. The interleaving architecture allows the inter-segment RNN to process information across segments and the intra-segment RNN processes the local segments independently. Therefore, the DPRNN architecture is a promising choice for long sequences. The intra-segment and inter-segment in DPRNN's architecture are the powerful techniques that have been used in recent advanced models such as DPTNet [17] and SepFormer [18]. Besides, various lightweight methods based on DPRNN have been proposed such as GroupComm-DPRNN [19], GC3-DPRNN [20], Sandglasset [21], etc.

To boost the performance for lightweight models, knowledge distillation (KD) [22] is one of the common approaches because of its ability to transfer a big pre-trained network's interpretation capability to other smaller networks without reducing the performance capability of these small networks (see Fig. 1 (a)). However, this approach requires a teacher

⋆ The first two authors contributed equally.

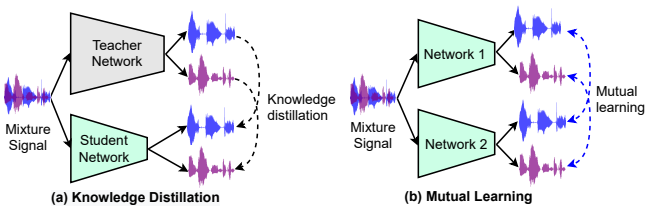

Fig. 1. **Comparison between ML and KD for speech separation.** The black line is the forward path; the black dashed line indicates the feature distillation from the pre-trained teacher to the student meanwhile the blue dashed line illustrates the knowledge sharing between two networks in the ML approach.

network to guide the student network. Moreover, the distillation process usually is performed one way, i.e., from teacher to student. In contrast, mutual learning (ML) approaches use a group of students to learn and share knowledge simultaneously with each other during the training phase (see Fig. 1 (b)). ML has become a promising approach to increase performance on many different tasks such as image classification [23], audio tagging [24], speech recognition [25], etc.

In this paper, we propose a simple yet robust approach to boost the performance of the deep models for speech separation, namely the selective mutual learning (SML) method. Inspired by conventional ML, there are two networks in our method, and these networks are trained simultaneously to solve the task together. Moreover, they are to be like a couple of friends, learning and sharing knowledge with each other. Different from existing ML methods, which always teach each other throughout the training process, in our approach, only high-confidence predictions are used to guide the remaining network while the low-confidence predictions are ignored. Specifically, the predictions with high values of the scale-invariant source-to-noise ratio (SI-SNR) of network 1 are used to guide the training of network 2 and otherwise at each epoch. This helps our model avoids using poor predictions to train and promotes networks towards highly accurate predictions. Our contribution is summarized as follows:

(1) We present a new approach based on ML for speech separation. To our best knowledge, this is the first ML method that is built for speech separation.

(2) We propose a novel SML strategy. It allows removing poor predictions between two networks during the sharing knowledge process.

(3) Experimental results show that our SML approach outperforms other training mechanisms such as independently training, KD, and ML with the same network architecture. The details are discussed in Section 3.

2. METHODOLOGY

2.1. Problem Setup

Given a mixed speech signal x from the independent speech signals, the goal of the speech separation task is to separate the component speech signals from x. In this paper, we focus on x as the mixed signal from two single signals (assume g and h). Our goal is how to extract g and h separately from x. Let $X = \{x^{(1)}, x^{(2)}, ..., x^{(N)}\}$ denote the set of the mixed speeches, $S = \{s^{(1)}, s^{(2)}, ..., s^{(N)}\}$ is the set of reference clean sources and $\hat{S} = \{\hat{s}^{(1)}, \hat{s}^{(2)}, ..., \hat{s}^{(N)}\}$ is the set of estimated sources where $s^{(i)} = [g^{(i)}, h^{(i)}]$ and $\hat{s}^{(i)} = [\hat{g}^{(i)}, \hat{h}^{(i)}]$. N illustrates the numbers of training samples. In time-domain speech separation frameworks, SI-SNR [26] usually is used as the training target (as a loss function) to compare the similarity between two signals (e.g., a reference clean source s and an estimated source $\hat{s}$). The SI-SNR is defined as follows:

$$P\left(s, \hat{s}\right) = 10\log_{10}\frac{||\alpha \cdot s||^2}{||\hat{s} - \alpha \cdot s||^2} \quad (1)$$

where $\alpha = \frac{\langle \hat{s}, s \rangle s}{||s||^2}$. The $s \in \mathbb{R}^{1\times t}$ and $\hat{s} \in \mathbb{R}^{1\times t}$ where t denotes the length of the signals. Scale invariance is ensured when s and $\hat{s}$ are normalized to zero-mean. In the training process, we utilize the utterance-level permutation invariant training (uPIT) [4] to address the source permutation issue.

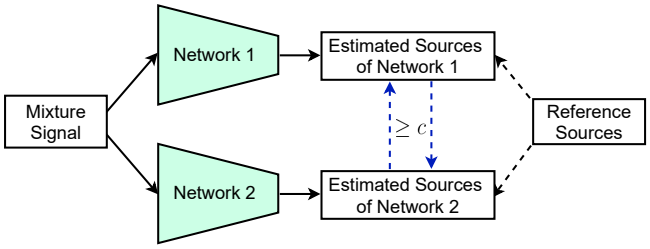

Fig. 2. **Overview of our proposed SML approach for speech separation.** In which, the black dashed line indicates the supervised learning loss and the blue dashed line illustrates the knowledge sharing between two networks in the SML framework. The hyper-parameter c denotes the confidence factor.

2.2. Selective Mutual Learning

As shown in Fig. 2, there are two networks in the SML framework, namely network 1 and network 2. Let denote $\mathcal{F}_1$ and $\mathcal{F}_2$ correspond to the network 1 and network 2, respectively. Let their corresponding parameters be θ_1 and θ_2. We denote $\hat{s}_1^{(i)}$ and $\hat{s}_2^{(i)}$ as the estimated output sources of the $\mathcal{F}_1$ and $\mathcal{F}_2$ with the input $x^{(i)}$ i.e., $\hat{s}_1^{(i)} = \mathcal{F}_1(x^{(i)}; \theta_1)$ and $\hat{s}_2^{(i)} = \mathcal{F}_2(x^{(i)}; \theta_2)$. There are two main loss terms in the loss function of the SML

framework, including the loss function for $\mathcal{F}_1$ and the other for $\mathcal{F}_2$. In particular, the loss of each network is calculated as:

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{sup}(s^{(i)}, \hat{s}_1^{(i)}) + \lambda\mathcal{L}_{sml}(s^{(i)}, \hat{s}_2^{(i)}, \hat{s}_1^{(i)})\Big) \quad (2)$$

$$\mathcal{L}_2 = \frac{1}{N}\sum_{i=1}^{N}\Big(\mathcal{L}_{sup}(s^{(i)}, \hat{s}_2^{(i)}) + \lambda\mathcal{L}_{sml}(s^{(i)}, \hat{s}_1^{(i)}, \hat{s}_2^{(i)})\Big) \quad (3)$$

where $\mathcal{L}_1$ and $\mathcal{L}_2$ correspond to the overall loss functions of $\mathcal{F}_1$ and $\mathcal{F}_2$. The hyper-parameter λ is a hyper-parameter representing the weight of $\mathcal{L}_{sml}$ and $\mathcal{L}_{sup}$ denotes the supervised loss between the reference clean source and the estimated source. $\mathcal{L}_{sup}$ is calculated by:

$$\mathcal{L}_{sup}(s^{(i)}, \hat{s}_1^{(i)}) = -P(s^{(i)}, \hat{s}_1^{(i)}) \quad (4)$$

$$\mathcal{L}_{sup}(s^{(i)}, \hat{s}_2^{(i)}) = -P(s^{(i)}, \hat{s}_2^{(i)}) \quad (5)$$

Besides, $\mathcal{L}_{sml}$ is the SML loss function between two networks. Specifically, we use the estimated source of network 1 as a reference pseudo-source to guide the training of network 2 and likewise for the training of network 1. Moreover, we only retain the high-quality estimated sources and ignore the low-quality estimated sources during the mutual learning phase. To evaluate the quality of the estimated sources, we use a hyper-parameter c, called the confidence factor. The detail of the $\mathcal{L}_{sml}$ loss is illustrated as follows:

$$\mathcal{L}_{sml}(s^{(i)}, \hat{s}_2^{(i)}, \hat{s}_1^{(i)}; c) = -\mathbb{1}\Big(P(s^{(i)}, \hat{s}_2^{(i)}) \geq c\Big)P(\hat{s}_2, \hat{s}_1) \quad (6)$$

$$\mathcal{L}_{sml}(s^{(i)}, \hat{s}_1^{(i)}, \hat{s}_2^{(i)}; c) = -\mathbb{1}\Big(P(s^{(i)}, \hat{s}_1^{(i)}) \geq c\Big)P(\hat{s}_1, \hat{s}_2) \quad (7)$$

In Eq. 6, $-\mathbb{1}$ is an indicator function that evaluates to -1 if $P(s^{(i)}, \hat{s}_2^{(i)}) \geq c$, i.e, only estimated sources $\hat{s}_2^{(i)}$, which are compared to the reference clean sources $s^{(i)}$ and greater than the confidence factor c, will be utilized as the reference pseudo-sources. The low-quality estimated sources (i.e., $-\mathbb{1}$ is 0 if $P(s^{(i)}, \hat{s}_2^{(i)}) < c$) are removed. Likewise with Eq. 7. The detailed procedure of the SML framework in the training process is illustrated in Algorithm 1.

Algorithm 1: SML for speech separation

Input: Training mixed set X, the reference set S.
$\mathcal{F}_1, \mathcal{F}_2$: The network 1 and network 2.
λ, c: Loss weight and confidence factors.
Initialize: θ_1 and θ_2 for $\mathcal{F}_1$ and $\mathcal{F}_2$, respectively.

repeat
  $x, s = get\_batch(X, S)$
  $\hat{s}_1 = \mathcal{F}_1(x; \theta_1)$
  $\hat{s}_2 = \mathcal{F}_2(x; \theta_2)$
  Calculate $\mathcal{L}_1 = \mathcal{L}_{sup}(s, \hat{s}_1) + \lambda\mathcal{L}_{sml}(s, \hat{s}_2, \hat{s}_1; c)$
  Calculate $\mathcal{L}_2 = \mathcal{L}_{sup}(s, \hat{s}_2) + \lambda\mathcal{L}_{sml}(s, \hat{s}_1, \hat{s}_2; c)$
  Update parameters θ_1 and θ_2
until *convergence*
return $(\mathcal{F}_1, \theta_1)$ and $(\mathcal{F}_2, \theta_2)$

3. EXPERIMENTS

3.1. Dataset and Evaluation Metrics

Dataset. We evaluated the SML approach on the speaker-independent speech separation task using the WSJ0-2mix dataset. This is the common two-speaker benchmark used for speech separation in recent years. The WSJ0-2mix dataset includes a 30-hour training dataset, a 10-hour validation dataset, and a 5-hour evaluation dataset. The signal mixtures between -5 and 5 dB signal-to-noise ratio (SNR) are artificially generated by selecting random female and male speakers in the WSJ0 dataset. The training and validation datasets have been generated from speakers in `si_tr_s` from the WSJ0 dataset while the evaluation dataset is generated using by utterances with 16 speakers in the `si_dt_05` and `si_et_05` from the WSJ0 dataset, respectively. All the mixed waveforms are sampled down from 16 kHz to 8kHz for reducing the computational consumption.

Metrics. The separation performance of the proposed method is evaluated with scale-invariant signal-to-noise ratio improvement (SI-SNRi) [26], source-to-distortion ratio improvement (SDRi), short-time objective intelligibility (STOI), and perceptual evaluation of speech quality (PESQ). The SI-SNRi and SDRi metrics used in [27, 4, 11]. STOI [28] varies from 0 to 1, and reflects a correlation of voice intelligibility in hearing tests. PESQ [29] values range from -0.5 to 4.5, and are widely used to evaluate separation performance. All the larger values of the metrics lead to a better separation system.

3.2. Implementation Details

The SML approach includes a pair of networks. We use the DPRNN [16] backbone with 3-blocks (instead of 6-blocks as in the official version) for both networks in the SML framework. Each DPRNN block contains two BLSTM layers for local and global processing. We use the Adam optimizer [30], the initial learning rate of $1e^{-4}$ and a decaying rate of 0.98 for every two epochs. The confidence factor c is initialized to 15 and increases by +1 after 10 epochs. The maximum value of c is set to 20. The λ is set to 0.001. The training process was considered converged and automatically stops when the validation loss of both networks don't decrease in 15 consecutive epochs. All the networks are trained with uPIT to maximize SI-SNR and 4-second long segments. A gradient clipping method with a maximum L2-norm of 5 is used dur-

ing training and the batch size of 4.

3.3. Performance Comparison

Comparison with other mechanisms. To examine the effectiveness of our proposed SML, we have compared the separated performance of the SML method to other training mechanisms including the baseline (independently training), KD, and ML. We use the online training in the KD method i.e., both the teacher and student networks are simultaneously updated during the training process. The process of knowledge transfer takes only one way in the training process i.e., student network learns from reference sources and is imparted knowledge from teacher network. The ML method uses all predictions of the two networks for sharing knowledge while SML only retains the high-quality predictions for sharing knowledge. All methods are trained with the same network architecture i.e., DPRNN 3-blocks. The teacher network of the KD approach is the DPRNN 6-blocks. As shown in Table 1, the performance of the KD approach insignificantly improves compared to the baseline method (14.2 dB vs 14.1 dB SI-SNRi). R. Aihara et al., [31] have shown that there was no difference in the student network's performance with and without the teacher. The reason behind this issue is the huge difference between the teacher network's output and that of the students. In [32, 33, 34], the authors have demonstrated that the capacity gap between the large teacher and the small student is one of the notable limitations of the KD approach. As a result, the student network cannot effectively exploit the knowledge transferred from the teacher.

Table 1. Comparison of the performance of the SML approach with other mechanisms on the WSJ0-2mix dataset.

Mechanisms	SI-SNRi	SDRi	STOI	PESQ
Baseline	14.1	14.2	0.92	2.82
KD	14.2	14.3	0.92	2.83
ML (network 1)	14.4	14.6	0.92	2.86
ML (network 2)	14.5	14.8	0.93	2.88
SML (network 1)	**15.1**	**15.3**	**0.94**	**2.93**
SML (network 2)	**14.8**	**15.0**	**0.94**	**2.92**

Different from KD, there is no capacity gap in the ML approach, thus the performance of the students in the ML approach improved but not significantly compared to the baseline method (14.5 vs 14.1 dB SI-SNRi). With both KD and ML approaches, all output predictions from the teacher network or friend network are utilized to guide the training of the remaining network. However, in many cases, the outputs of these networks are not good predictions (incorrect knowledge), thus training with both correct and incorrect knowledge causes a significant decrease in the performance of the network. Our SML approach can address the above limitation via a mechanism of selectively sharing knowledge with each other. As a result, SML performance outperforms both the baseline, KD, and ML approaches on both all metrics.

Comparison with the state-of-the-art methods. In this part, we present our experimental studies of the SML approach against the state-of-the-art methods. As shown in Table. 2, the proposed SML method has obtained higher SI-SNRi and SDRi while utilizing a smaller model size. In particular, our performance is only lower than 3.7 dB SI-SNRi compared to DPRNN [16] while the model size is much smaller than DPRNN (0.9M vs 2.6M). Comparison with the same model size i.e., DPRNN 3-blocks with 0.9M params, DPRNN 3-blocks only achieves 14.1 dB SI-SNRi and 14.2 dB SDRi, meanwhile, our method outperforms DPRNN 3-blocks by 1.0 dB SI-SNRi and 1.1 dB SDRi. It implies that the SML approach drives each network to learn more useful information from sharing/learning knowledge with each other. As a result, our proposed SML method has better generality and extensibility for any network backbone.

Table 2. Comparison of the SML approach to state-of-the-art methods on the WSJ0-2mix dataset. In which, the model size is the total of parameters used in each model.

Method	Model Size	SI-SNRi	SDRi
DPCL++ [27]	13.6M	10.8	-
ADANet [5]	9.1M	10.4	10.8
uPIT-BLSTM [4]	92.7M	-	10.0
TasNet-BLSTM [11]	-	10.8	11.1
Conv-TasNet [12]	5.1M	15.3	15.6
FurcaNeXt [35]	51.4M	-	18.4
DPRNN 6-blocks [16]	2.6M	**18.8**	**19.0**
DPRNN 3-blocks [16]	0.9M	14.1	14.2
SML (DPRNN 3-blocks)	**0.9M**	**15.1**	**15.3**

4. CONCLUSION

In this work, we have proposed a novel SML approach for speech separation, to improve the performance of each peer network in a cohort. In our proposed SML method, each network is trained with supervised learning and knowledge distillation selectively from the other network. The exchange and sharing of knowledge between two networks help to enhance the generalization performance. The experimental results have shown that the SML approach outperforms other training mechanisms such as independently training, KD, and ML. In addition, only with 0.9M params, the SML approach achieves competitive performance compared to state-of-the-art heavy networks.

5. REFERENCES

[1] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder.," in *Interspeech*, 2013, vol. 2013, pp. 436–440.

[2] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM TASLP*, vol. 23, no. 1, pp. 7–19, 2014.

[3] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Interspeech*, 2016.

[4] Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM TASLP*, vol. 25, no. 10, pp. 1901–1913, 2017.

[5] Yi Luo, Zhuo Chen, and Nima Mesgarani, "Speaker-independent speech separation with deep attractor network," *IEEE/ACM TASLP*, vol. 26, no. 4, pp. 787–796, 2018.

[6] Zhong-Qiu Wang, Jonathan Le Roux, and John R Hershey, "Alternative objective functions for deep clustering," in *ICASSP*. IEEE, 2018, pp. 686–690.

[7] Ha Minh Tan and Jia-Ching Wang, "Single channel speech separation using enhanced learning on embedding features," in *GCCE*. IEEE, 2021, pp. 430–431.

[8] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *ICASSP*. IEEE, 2015, pp. 708–712.

[9] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM TASLP*, vol. 24, no. 3, pp. 483–492, 2015.

[10] Y. Luo, Z. Chen, J. R Hershey, J. Le Roux, and N. Mesgarani, "Deep clustering and conventional networks for music separation: Stronger together," in *ICASSP*. IEEE, 2017, pp. 61–65.

[11] Yi Luo and Nima Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP*. IEEE, 2018, pp. 696–700.

[12] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM TASLP*, vol. 27, no. 8, pp. 1256–1266, 2019.

[13] J. Wang, Max W. Y. Lam, D. Su, and D. Yu, "Tune-in: Training under negative environments with interference for attention networks simulating cocktail party effect," in *AAAI*, 2021.

[14] Shaojie Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *ArXiv*, vol. abs/1803.01271, 2018.

[15] M. W. Y. Lam, J. Wang, D. Su, and D. Yu, "Mixup-breakdown: a consistency training method for improving generalization of speech separation models," in *ICASSP*. IEEE, 2020.

[16] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*. IEEE, 2020, pp. 46–50.

[17] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," *arXiv preprint arXiv:2007.13975*, 2020.

[18] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong, "Attention is all you need in speech separation," in *ICASSP*. IEEE, 2021, pp. 21–25.

[19] Y. Luo, C. Han, and N. Mesgarani, "Ultra-lightweight speech separation via group communication," in *ICASSP*. IEEE, 2021.

[20] Yi Luo, Cong Han, and Nima Mesgarani, "Group communication with context codec for lightweight source separation," *IEEE/ACM TASLP*, vol. 29, pp. 1752–1761, 2021.

[21] M. WY Lam, J. Wang, D. Su, and D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation," in *ICASSP*. IEEE, 2021, pp. 5759–5763.

[22] G. E. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *ArXiv*, vol. abs/1503.02531, 2015.

[23] Y. Zhang, T. Xiang, Timothy M Hospedales, and H. Lu, "Deep mutual learning," in *IEEE CVPR*, 2018, pp. 4320–4328.

[24] Yifang Yin, Harsh Shrivastava, Ying Zhang, Zhenguang Liu, Rajiv Ratn Shah, and Roger Zimmermann, "Enhanced audio tagging via multi-to single-modal teacher-student mutual learning," in *AAAI*, 2021, vol. 35, pp. 10709–10717.

[25] R. Masumura, M. Ihori, A. Takashima, T. Tanaka, and T. Ashihara, "End-to-end automatic speech recognition with deep mutual learning," in *APSIPA*. IEEE, 2020, pp. 632–637.

[26] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP*. IEEE, 2019, pp. 626–630.

[27] Y. Isik, J. L. Roux, Z. Chen, S. Watanabe, and J. R Hershey, "Single-channel multi-speaker separation using deep clustering," *INTERSPEECH*, 2016.

[28] C. H Taal, R. C Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *TASLP*, vol. 19, pp. 2125–2136, 2011.

[29] ITU-T Recommendation, "Perceptual evaluation of speech quality (pesq): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[30] D. P Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[31] Ryo Aihara, Toshiyuki Hanazawa, Yohei Okato, Gordon Wichern, and Jonathan Le Roux, "Teacher-student deep clustering for low-delay single channel speech separation," in *ICASSP*. IEEE, 2019, pp. 690–694.

[32] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, "Knowledge distillation: A survey," *IJCV*, vol. 129, no. 6, pp. 1789–1819, 2021.

[33] Duc-Quang Vu, Ngan Le, and Jia-Ching Wang, "Teaching yourself: A self-knowledge distillation approach to action recognition," *IEEE Access*, vol. 9, pp. 105711–105723, 2021.

[34] Duc-Quang Vu, Jia-Ching Wang, et al., "A novel self-knowledge distillation approach with siamese representation learning for action recognition," in *VCIP*. IEEE, 2021, pp. 1–5.

[35] L. Zhang, Z. Shi, J. Han, A. Shi, and D. Ma, "Furcanext: End-to-end monaural speech separation with dynamic gated dilated temporal convolutional networks," in *MMM*. Springer, 2020, pp. 653–665.