Interspeech 2024
1-5 September 2024, Kos, Greece

Towards Explainable Monaural Speaker Separation with Auditory-based Training

Hassan Taherian[1], *Vahid Ahmadi Kalkhorani*[1], *Ashutosh Pandey*[2], *Daniel Wong*[2], *Buye Xu*[2], *DeLiang Wang*[1]

[1]The Ohio State University, USA
[2]Meta Reality Labs Research, USA

taherian.1@osu.edu

Abstract

Permutation ambiguity is a major challenge in training monaural talker-independent speaker separation. While permutation invariant training (PIT) is a widely used technique, it functions as a 'black box', providing little insight into which auditory cues lead to successful training. We introduce a new approach to speaker separation by leveraging differences in pitch and onset, which are both prominent cues for auditory scene analysis. We propose pitch-based and onset-based training to resolve permutation ambiguity, assigning speakers by their pitch frequencies and onset times, respectively. This approach offers a more explainable training strategy than PIT. We also propose a hybrid criterion combining these cues to improve separation performance in challenging conditions such as same-gender speakers or close utterance onsets. Evaluation results show that pitch and onset criteria each perform competitively to PIT and the hybrid criterion surpasses PIT in separating two-speaker mixtures.

Index Terms: explainable speaker separation, pitch-based training, onset-based training, permutation invariant training.

1. Introduction

Human listeners have a remarkable ability to segregate multiple sound sources, including multi-talker speech mixtures. Perceptual research in auditory scene analysis (ASA) reveals various grouping or segregation cues, such as pitch, amplitude modulation, and location [1, 2]. Traditionally, computational auditory scene analysis (CASA) [3] performs speech separation based on ASA principles. CASA organizes sound sources through simultaneous and sequential grouping by leveraging auditory cues including common periodicity and common onset [4, 5].

Recently, the deep learning based approach has been firmly established as the mainstream methodology for sound separation [6], resulting in dramatic performance improvement over the traditional CASA and speech enhancement [7] approaches. For monaural speaker separation, a DNN (deep neural network) is trained so that its output layers are assigned to distinct speakers in a multi-talker mixture. For a DNN model to effectively separate untrained speakers, it must be talker-independent. A major challenge in achieving talker independency is how to assign output layers to the underlying speakers during training with mixtures of a large number of speakers. With incorrect output-speaker assignment, DNN training would not converge due to conflicting gradients. This is known as the permutation ambiguity problem [8, 9]. A widely adopted solution to this problem is utterance-wise permutation invariant training (PIT) [9]. PIT tackles permutation ambiguity by evaluating the losses for every potential output-speaker assignment, and uses the one with the lowest loss to train DNN models.

Since the introduction PIT for addressing permutation ambiguity, the focus of research in speaker separation has largely been on improving DNN architectures, with impressive performance gains [10, 11, 12, 13, 14, 15]. Despite its effectiveness, PIT relies on spectrogram comparisons in the loss calculation and offers little insight into the auditory cues that are learned during training and employed during testing. Its 'black box' nature makes it difficult to explain why it works, and how to link to the large body of perceptual research in ASA. The lack of explainability also hinders the prediction of model behavior in untrained complex acoustic environments.

As an effort towards explainable speaker separation, this paper proposes novel training criteria that align more closely with auditory grouping cues. Motivated by the importance of pitch and onset in ASA, particularly in speaker separation, we investigate pitch-based and onset-based training methods; more specifically, we propose to assign speaker labels consistently according to their average fundamental frequencies[1] ($F0$) and onset times, respectively. In pitch-based training, a DNN model focuses on pitch differences for speaker separation, effectively learning to differentiate talkers based on distinct pitches. Onset-based training separates speakers by leveraging the different starting times of their utterances. Compared to PIT, these criteria are more explainable, allowing for a closer link to psychoacoustics. By incorporating pitch and onset in the training phase, we can analyze model performance in terms of pitch differences or onset time differences, and draw comparisons with human performance in similar conditions. This approach also helps to access the potential limitations of auditory cues in source separation and provides insights into mechanisms for further performance improvement.

Additionally, we propose a hybrid criterion that combines pitch and onset training in challenging scenarios where a single cue is not discriminative enough. Evaluations on the standard WSJ2-MIX dataset [8] show that pitch-based and onset-based models perform competitively to those trained with PIT, while the model trained with the hybrid criterion surpasses PIT in separating two-speaker mixtures.

The rest of the paper is organized as follows. Section 2 reviews related works. In Section 3, we describe a baseline separation model, pitch- and onset-based training, and the hybrid

This research was supported in part by an National Science Foundation grant (ECCS-2125074), a research contract from Meta Reality Labs, the Ohio Supercomputer Center, the Delta Supercomputer Center (OCI 2005572), and the Pittsburgh Supercomputer Center (NSF ACI-1928147).

[1] Pitch is a perceived attribute, while fundamental frequency is a physical property of sound. For simplicity, we will use the two terms interchangeably.

10.21437/Interspeech.2024-530

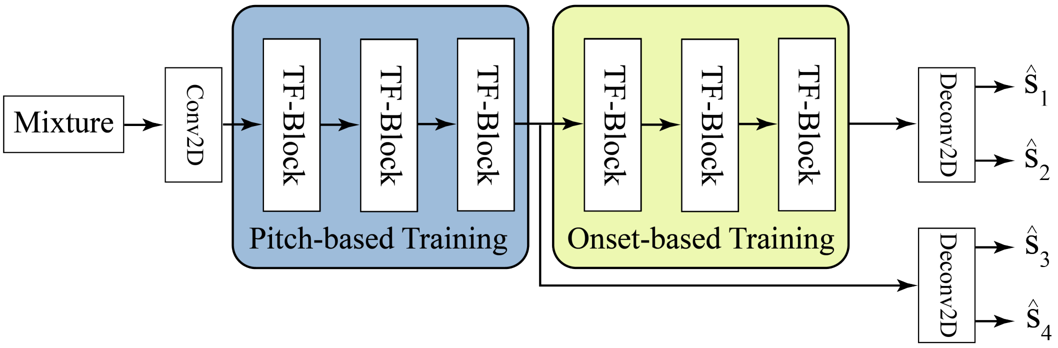

Figure 1: *Diagram of the proposed hybrid criterion based on the TF-GridNet architecture, divided into two distinct parts for pitch and onset training. The outputs from each part use a shared Deconv2D layer, generating four outputs ($\hat{s}_1$ and $\hat{s}_2$ for onset, $\hat{s}_3$ and $\hat{s}_4$ for pitch) for a two-speaker mixture. During inference, only the outputs from the onset part are utilized.*

criterion. We then present the experimental setup and the evaluation results in Section 4 and Section 5, respectively. Finally, we conclude the paper in Section 6.

2. Related Works

Several studies have explored the utilization of pitch and onset cues for speaker separation and automatic speech recognition tasks [16, 17]. Wang et al. [18] combined a separation model with a PIT-based pitch-tracking network for multi-speaker pitch tracking, where the estimated pitch is fed to another separation model as an additional input for further improvement. [19] proposed a method that integrates pitch tracking and speaker separation, consisting of two stages: pitch extraction from mixtures and speaker separation conditioned on the extracted pitch, utilizing a conditional generative adversarial network. These approaches, however, perform separation in distinct stages. Unlike these methods, our work integrates pitch as a direct training criterion within a single separation model to tackle the permutation ambiguity problem more efficiently. Furthermore, Pandey et al. [20] introduced attentive training for extracting the target speaker based on the earliest speaker onset, focusing on a target speaker extraction. In contrast, our approach aims to extract all speakers at once.

3. Algorithm Description

3.1. Baseline Separation Model

In this study, we employ the TF-GridNet [15] architecture as our baseline separation model. TF-GridNet achieved state-of-the-art performance across various speaker separation benchmarks. The TF-GridNet model processes time-frequency (TF) units in a grid-like pattern and is composed of multiple blocks. The input to the TF-GridNet is a stack of real and imaginary (RI) components of the mixture short-time Fourier transform (STFT). First, the model uses a two-dimensional convolution (Conv2D) layer to compute a embedding for each TF unit. The embeddings are then processed by a series of blocks. Each block features three components: the first two utilize bidirectional Long Short-Term Memory (BLSTM) networks to process full-band spectral features within individual frames and temporal information across frequencies. The final component employs a self-attention module designed to process information across frames to capture long-range contexts. Finally, the output of the last block is processed by a two-dimensional deconvolution (Deconv2D) to obtain the predicted RI components for all speakers. Similar to many existing supervised separation models, TF-GridNet employs the utterance-level PIT criterion for training. PIT uses fixed output-speaker pairings for an entire utterance and selects the pairing that minimizes the overall loss across all possible permutations of speaker outputs [9]:

$$\mathcal{L}_{\text{PIT}} = \min_{\phi \in \Phi} \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{\phi(n)}), \tag{1}$$

where $\hat{S}_n$ and $S_{\phi(n)}$ represents the estimated and clean speech signals for speaker n in the time domain, and $\mathcal{L}$ is the loss function, here employing the widely adopted scale-invariant signal-to-distortion ratio (SI-SDR) loss [21]. Symbol Φ denotes the set of all permutations of N speakers, with ϕ representing a specific permutation.

3.2. Auditory-based Training

To address the permutation ambiguity problem, we propose leveraging auditory cues of speakers for talker-independent speaker separation. This study introduces two novel training criteria based on speaker pitch and onset time. Assuming $F0_1$, $F0_2, \ldots, F0_N$ represent the fundamental frequencies of the N speakers averaged over the utterance, we define the pitch-based training loss function as:

$$\mathcal{L}_{\text{Pitch}} = \frac{1}{N} \sum_{n=1}^{N} \mathcal{L}(\hat{S}_n, S_{F0_n}). \tag{2}$$

$F0$ estimation is only required during training. For $F0$ estimation, we employ a pitch tracking algorithm, specifically RAPT [22]. RAPT, a time-domain pitch tracking method, extracts $F0$ information by estimating signal periodicity through the normalized cross-correlation function. For each mixture, we extract the $F0$ for each speaker's clean speech and average the $F0$ over the entire utterance, excluding unvoiced frames. In pitch-based training, output-speaker assignments follow the average $F0$ order, with the first output corresponding to the speaker with the lowest average $F0$ and the last output to the speaker with the highest average $F0$.

Similarly, the permutation ambiguity problem can be effectively resolved by utilizing onset time information of speakers, motivated by the observation that speakers rarely start talking simultaneously in real-life conversations. We define onset-based training with the following loss function:

$$\mathcal{L}_{\text{Onset}} = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}(\hat{S}_n, S_{t_n}), \tag{3}$$

where $t_1, t_2, \ldots, t_N$ are speaker indices arranged in ascending order according to their onset time. In onset-based training, output-speaker assignments are determined by onset order, assigning the speaker with the earliest onset to the first DNN output, and so on.

3.3. Combination of Pitch and Onset Cues

Employing either the proposed pitch or onset criterion, our model is designed to distinguish speakers using a single auditory cue. However, it is anticipated that this approach might underperform under scenarios where speakers' pitches are closely matched — such as those of the same gender when utilizing pitch-based training — or when their onset times are close, which could challenge onset-based training. To address these limitations, integrating pitch and onset cues can improve separation performance. We hypothesize that a model incorporating both cues will consistently surpass models trained exclusively on either pitch or onset. This combined approach exploits both pitch and onset information, enabling the model to separate speakers more effectively, especially in situations where relying on a single cue is insufficient.

To achieve this, we introduce a hybrid model that combines pitch-based and onset-based training through a unified training strategy. As illustrated in Fig. 1, the TF-GridNet architecture, which comprises B blocks, is divided into two parts. The first part, comprising the initial $B/2$ blocks, is optimized using the pitch-based criterion, while the second part, consisting of the remaining $B/2$ blocks, is trained by the onset-based criterion. The rationale for this sequential arrangement is motivated by the real-world observation that speakers in conversations rarely start speaking at exactly the same moment. Consequently, the onset-based training, applied in the latter part, acts as a decisive cue for speaker separation.

During training, the output from the final block of the pitch-focused part is passed through a shared Deconv2D layer, which generates the predicted STFT components for all speakers. This hybrid model is trained using the following loss function:

$$\mathcal{L}_{\text{Hybrid}} = \mathcal{L}_{\text{Pitch}} + \mathcal{L}_{\text{Onset}}. \tag{4}$$

4. Experimental Setup

We evaluated our proposed training criteria using the WSJ0-2mix and WSJ0-3mix datasets [8], both of which are established benchmarks for assessing the performance of monaural speaker separation. These datasets include a 30-hour training set and a 10-hour validation set, generated by selecting random speakers from the Wall Street Journal (WSJ0) training set and mixing their speeches at various signal-to-noise ratios (SNRs) ranging from 0 dB to 5 dB. To examine the performance of onset-based training models, we introduced four additional test sets derived from the WSJ0-2mix test data by incorporating speaker onset differences of 0.25s, 0.50s, 0.75s, and 1.0s. This was done by shifting the onset of one speaker in each test utterance by the specified durations. All audio samples were processed at a sampling rate of 8 kHz. We did not use any dynamic mixing techniques. Performance metrics reported include the signal-to-distortion ratio improvement (ΔSDR) [23], scale-invariant signal-to-noise ratio improvement (ΔSI-SNR) [21], perceptual evaluation of speech quality (PESQ), and extended short-time objective intelligibility (ESTOI) [24], to measure separation performance, speech quality and speech intelligibility, respectively.

Table 1: *Average ΔSDR (dB), ΔSI-SDR (dB), PSEQ and ESTOI (%) of different training criteria on WSJ0-2MIX and WSJ0-3MIX.*

Method	ΔSDR	ΔSI-SDR	PESQ	ESTOI
WSJ0-2MIX				
Unprocessed	0.00	0.00	1.68	56.10
PIT	23.71	23.58	4.07	97.18
Pitch Based Training	23.25	23.11	4.04	96.88
WSJ0-3MIX				
Unprocessed	0.00	0.00	1.42	38.51
PIT	20.15	19.99	3.40	90.93
Pitch Based Training	19.44	19.30	3.30	89.44

Our preprocessing steps involved normalizing the sample variance of each mixture segment to 1.0, followed by adjusting the clean target sources accordingly. We utilized the STFT with a window length of 32 ms, hop length of 8 ms, and employing a square-root Hann window. Our model, TF-GridNet, comprised 14.5 million parameters including $B = 6$ blocks, a BLSTM with 256 units, and 64 channels, with a kernel size of 4 and stride of 1.

For training, 4-second segments were sampled from each mixture, and the RAPT [22] algorithm was employed to extract the pitch of clean speech signals for each speaker within the frequency range of 60-404 Hz. During training, we apply an energy-based voice activity detector [1] to remove silent frames at the beginning of each utterance. We introduced variable onset differences during training by randomly shifting the onset of one speaker in each sample by 0.2 to 1.2 seconds, resulting training segments with an overlap ratio of 70% to 95%. The models were optimized using the Adam optimizer, with gradient norms capped at 1.0. We adopted a learning schedule with a ramp-up phase, peaking at a learning rate of 0.001 after a warm-up period of 100K steps, as proposed in [25]. The training was conducted on two NVIDIA A100 GPUs for approximately 8 days.

5. Evaluation Results

Table 1 compares the performance of pitch-based training with PIT on the WSJ0-2MIX and WSJ0-3MIX datasets. For 2-speaker mixtures, we observe that pitch-based training produces competitive results compared to PIT, with a ΔSI-SDR of 23.11 dB, only 0.47 dB lower than PIT. For 3-speaker mixtures, the pattern is similar, with pitch-based training yielding a ΔSI-SDR of 19.30 dB, 0.69 dB lower than PIT.

We speculate that the performance of pitch-based training is influenced by the pitch difference between speakers. Specifically, it tends to underperform in scenarios where the average pitch difference is small, such as in same-gender mixtures. To further investigate this, we analyzed the results for two-speaker mixtures, categorizing them based on the pitch difference between speakers and their genders. Figure 2 illustrates a scatter plot comparing the ΔSI-SDR values for both PIT and pitch-based training on the WSJ0-2MIX dataset for male-male,

[1] Available at: https://github.com/wiseman/py-webrtcvad

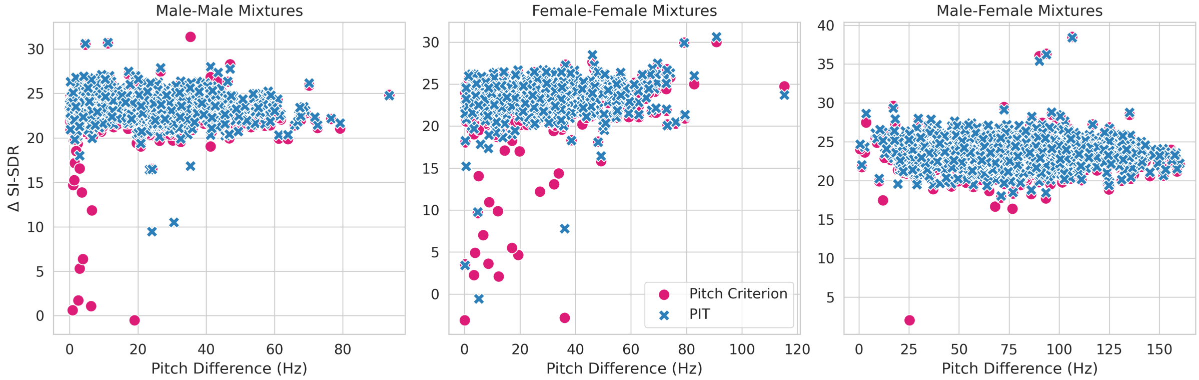

Figure 2: *Comparison of* Δ*SI-SDR between pitch-based training and PIT on the WSJ0-2MIX dataset for male-male, female-female, and male-female mixtures.*

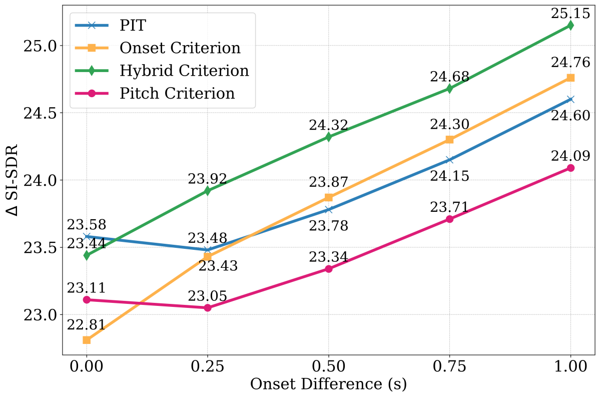

Figure 3: Δ*SI-SDR of different training criteria based on speaker onset differences on WSJ0-2MIX. 0.0s onset difference indicates the original WSJ0-2MIX test set.*

female-female, and male-female mixtures. We observe that pitch-based training is on par with PIT for male-female mixtures, irrespective of the average pitch difference. However, in same-gender mixtures, pitch-based training falls short of PIT's performance when the pitch difference is below 20 Hz. This indicates that pitch-based training remains effective for same-gender mixtures as well, only underperforming in scenarios where the average pitch difference between speakers is small.

To evaluate the performance of onset-based training, we utilized the WSJ0-2MIX dataset with different speaker onset differences. Figure 3 displays the ΔSI-SDR results for different training criteria with respect to the speaker onset differences. Onset-based training achieves a 22.81 dB SI-SDR improvement with a 0.0s onset difference (the original WSJ0-2MIX test set). This performance can be attributed to the fact that utterances in the WSJ0-2MIX mixtures do not start exactly at the same time, typically featuring some initial silence. As the onset difference increases, even with a minimal increase to 0.25s, the performance of onset-based training improves significantly. It surpasses PIT in scenarios where the onset difference exceeds 0.50s.

When we combine pitch and onset-based training, it becomes evident that the hybrid model outperforms both the pitch-based and onset-based models on their own. This integration makes the model more robust and enhances its performance compared to when only a single criterion is used, especially in cases where one criterion alone is insufficient for separation. The hybrid model provides comparable results to PIT in 0.0 onset difference. However, it surpasses PIT's performance when the onset difference is greater than 0.0 seconds. Figure 4 compares the performance of the combined criterion to pitch-based training on the WSJ0-2MIX dataset with a 1.0s onset difference. The plot indicates that the combined criterion is more effective than pitch-based training alone, particularly when the average pitch difference between speakers is small. With the hybrid model, we exploit the explainability of each criterion within the model to enhance the overall performance.

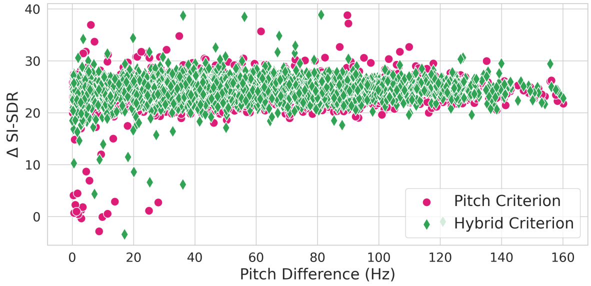

Figure 4: Δ*SI-SDR scatter plot based on speaker pitch difference for pitch-onset training and pitch-based training on WSJ0-2MIX with 1.0s onset difference.*

6. Conclusions

In this study, we have introduced a new approach to address permutation ambiguity in speaker separation by leveraging auditory cues, specifically pitch frequencies and onset times. By combining pitch-based and onset-based training, our method enhances the separation performance in challenging scenarios. This study represents an initial step towards explainable speaker separation. The proposed training criteria may be used to address other tasks susceptible to permutation ambiguity, such as DNN-based speaker diarization [26]. Future work will explore additional grouping cues, such as vocal-tract length [2], and integrate spatial locations [27].

7. References

[1] A. S. Bregman, *Auditory scene analysis: The perceptual organization of sound*. MIT press, 1994.

[2] C. J. Darwin, D. S. Brungart, and B. D. Simpson, “Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers,” *The Journal of the Acoustical Society of America*, vol. 114, pp. 2913–2922, 2003.

[3] D. L. Wang and G. J. Brown, Eds., *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[4] M. Cooke and D. P. Ellis, “The auditory organization of speech and other sources in listeners and computational models,” *Speech communication*, vol. 35, pp. 141–177, 2001.

[5] G. Hu and D. L. Wang, “Monaural speech segregation based on pitch tracking and amplitude modulation,” *IEEE Transactions on Neural Networks*, vol. 15, pp. 1135–1150, 2004.

[6] D. L. Wang and J. Chen, “Supervised speech separation based on deep learning: An overview,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, pp. 1702–1726, 2018.

[7] P. C. Loizou, *Speech enhancement: theory and practice*. CRC Press, 2007.

[8] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, “Deep clustering: Discriminative embeddings for segmentation and separation,” in *Proc. ICASSP*, 2016, pp. 31–35.

[9] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, “Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 1901–1913, 2017.

[10] Y. Luo and N. Mesgarani, “Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 1256–1266, 2019.

[11] Y. Liu and D. L. Wang, “Divide and conquer: A deep CASA approach to talker-independent monaural speaker separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, pp. 2092–2102, 2019.

[12] Y. Luo, Z. Chen, and T. Yoshioka, “Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation,” in *Proc. ICASSP*, 2020, pp. 46–50.

[13] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, “Attention is all you need in speech separation,” in *Proc. ICASSP*, 2021, pp. 21–25.

[14] J. Rixen and M. Renz, “QDPN - quasi-dual-path network for single-channel speech separation,” in *Proc. Interspeech*, 2022, pp. 5353–5357.

[15] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, “TF-GridNet: Integrating full-and sub-band modeling for speech separation,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 31, pp. 3221–3236, 2023.

[16] C. Weng, D. Yu, M. L. Seltzer, and J. Droppo, “Single-channel mixed speech recognition using deep neural networks,” in *Proc. ICASSP*, 2014, pp. 5632–5636.

[17] ——, “Deep neural networks for single-channel multi-talker speech recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1670–1679, 2015.

[18] K. Wang, F. Soong, and L. Xie, “A pitch-aware approach to single-channel speech separation,” in *Proc. ICASSP*, 2019, pp. 296–300.

[19] X. Li, Y. Wang, Y. Sun, X. Wu, and J. Chen, “PGSS: pitch-guided speech separation,” in *Proc. AAAI*, vol. 37, 2023, pp. 13 130–13 138.

[20] A. Pandey and D. L. Wang, “Attentive training: A new training framework for speech enhancement,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 1360–1370, 2023.

[21] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR–half-baked or well done?” in *Proc. ICASSP*, 2019, pp. 626–630.

[22] D. Talkin and W. B. Kleijn, “A robust algorithm for pitch tracking (RAPT),” *Speech coding and synthesis*, vol. 495, p. 518, 1995.

[23] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 1462–1469, 2006.

[24] J. Jensen and C. H. Taal, “An algorithm for predicting the intelligibility of speech masked by modulated noise maskers,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 2009–2022, 2016.

[25] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, “Specaugment: A simple data augmentation method for automatic speech recognition,” *arXiv:1904.08779*, 2019.

[26] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. García, “Encoder-decoder based attractors for end-to-end neural diarization,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 1493–1507, 2022.

[27] H. Taherian, K. Tan, and D. L. Wang, “Multi-channel talker-independent speaker separation through location-based training,” *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 30, pp. 2791–2800, 2022.