# Speech Enhancement with Weakly Labelled Data from AudioSet

*Qiuqiang Kong, Haohe Liu, Xingjian Du, Li Chen, Rui Xia, Yuxuan Wang*

ByteDance, Shanghai, China

{kongqiuqiang, liuhaohe.7, duxingjian.real, chenli.cloud,
rui.xia, wangyuxuan.11}@bytedance.com

## Abstract

Speech enhancement is a task to improve the intelligibility and perceptual quality of degraded speech signals. Recently, neural network-based methods have been applied to speech enhancement. However, many neural network-based methods require users to collect clean speech and background noise for training, which can be time-consuming. In addition, speech enhancement systems trained on particular types of background noise may not generalize well to a wide range of noise. To tackle those problems, we propose a speech enhancement framework trained on weakly labelled data. We first apply a pretrained sound event detection system to detect anchor segments that contain sound events in audio clips. Then, we randomly mix two detected anchor segments as a mixture. We build a conditional source separation network using the mixture and a conditional vector as input. The conditional vector is obtained from the audio tagging predictions on the anchor segments. In inference, we input a noisy speech signal with the one-hot encoding of "Speech" as a condition to the trained system to predict enhanced speech. Our system achieves a PESQ of 2.28 and an SSNR of 8.75 dB on the VoiceBank-DEMAND dataset, outperforming the previous SEGAN system of 2.16 and 7.73 dB respectively.

**Index Terms**: Speech enhancement, weakly labelled data, AudioSet.

## 1. Introduction

Speech enhancement (SE) is a task to improve the intelligibility and perceptual quality of degraded speech signals. Speech enhancement has many applications, such as teleconference, mobile phone calls, automatic speech recognition and hearing aids [1]. Early works of speech enhancement include signal processing methods such as [2] and non-negative matrix factorization (NMF) [3] methods. Those conventional methods perform well under stationary noise but have limited performance under non-stationary noise or in low signal-to-noise ratio (SNR) environments. Recently, neural network-based methods have been proposed for speech enhancement, such as denoising autoencoder [4], fully connected neural networks [5], recurrent neural networks (RNNs) [6], convolutional neural networks [7, 8] (CNNs), time-domain CNNs [9, 10, 11] and generative adversarial networks (GANs) [12, 13] and a sound selector was proposed in [14]. Those neural network-based speech enhancement methods require clean speech and background noise for training.

However, there are several limitations of previous speech enhancement methods. First, previous neural network-based speech enhancement methods require clean speech and background noise for training, while collecting clean speech and background noise can be difficult and time-consuming. Users need to collect different datasets for building speech enhance- ment systems for different scenarios. Second, there is a domain mismatch between datasets collected in laboratories and in the real world. The types of background sounds recorded in laboratories [15] are limited, and do not cover a wide range of background sounds in the world. In addition, speech datasets such as TIMIT [16] and VoiceBank [17] mostly contain neutral emotion speech, while there can be various emotions of speech in our real life.

Recently, there are several source separation and speech enhancement systems that do not require clean sources for training. For example, universal source separation systems [18, 19] were proposed to separate sources in an unsupervised way. The advantage of unsupervised learning is that there are much large-scale unlabelled audio data on the internet. Another way of building source separation systems is to use weakly labelled data [20, 21]. Weakly labelled data only contain audio tags of audio clips, without extra information. In [22], pretrained sound event detection systems and audio tagging systems are used to build the source separation systems trained on the weakly labelled AudioSet [23]. AudioSet contains hundreds of different sound classes from YouTube and provides a larger variety of sounds than previous speech and noise datasets. In [24], the authors proposed an audio filtering system trained on weakly labelled data. However, there is a lack of work using weakly labelled data to build speech enhancement systems.

In this work, we adapt the weakly labelled source separation system [20] to build our speech enhancement system. The contribution of this work includes 1) We propose a speech enhancement system trained on weakly labelled data only. This allows us to use the large-scale AudioSet containing hundreds of sound classes to train the speech enhancement system, rather than using datasets with limited sizes. 2) Our proposed speech enhancement system can be trained without clean speech data. 3) We propose a data mining strategy for selecting anchor segments from AudioSet, where the selected anchor segment pairs have disjoint audio tags. This reduces label ambiguity of conditional vectors and improves enhancement quality. 4) We simplify and propose a revised training function to train the source separation system compared to [20]. 5) Our proposed speech enhancement system outperforms many previous speech enhancement systems trained on clean speech and background noise data. 6) For the speech enhancement task, we evaluate various metrics such as PESQ, CSIG that were not discussed in [20]. We also extend the loss function calculated on spectrogram [20] to a loss function calculated in the waveform domain. As far as we know, our proposed system is the first attempt to use weakly labelled data for speech enhancement.

This paper is organized as follows: Section 2 introduces our speech enhancement system trained on weakly labelled data. Section 3 shows the experiment results. Section 4 concludes this work.

## 2. Speech Enhancement with Weakly Labelled Data

### 2.1. Neural Network-Based Speech Enhancement

Recently, neural network-based methods have been applied to speech enhancement, and have outperformed conventional speech enhancement methods [5]. The neural network-based speech enhancement methods require pairs of noisy speech and clean speech for training. We denote a noisy speech as $x \in \mathbb{R}^L$, and its corresponding clean speech as $s \in \mathbb{R}^L$, where $L$ is the number of samples in an audio clip. Then, a neural network learns a mapping: $f : x \mapsto s$, where $f$ can be modeled by a neural network with learnable parameters, such as fully connected neural networks [5], RNNs [6], CNNs [7, 8] or time-domain CNNs [9, 10, 11]. we denote the enhanced speech as $\hat{s} = f(x)$. In training, the parameters of $f$ can be optimized by minimizing a loss function $l(\hat{s}, s)$, such as a mean absolute error (MAE) loss:

$$l_{\text{MAE}} = \|\hat{s} - s\|_1, \tag{1}$$

where $\| \cdot \|_1$ is an $l_1$ norm. In inference, the enhanced speech $\hat{s}$ can be calculated by $\hat{s} = f(x)$, where $x$ is a noisy speech. This speech enhancement method has been widely adopted in conventional speech enhancement methods [5, 6, 7, 8, 9, 10, 11]. However, one disadvantage of those methods is that clean speech and noisy speech pairs are required for training. For real applications, users need to collect a large amount of clean speech and background noise data, which can be time-consuming. To address this problem, we propose a speech enhancement framework that can be trained on weakly labelled data.

### 2.2. Speech Enhancement with Weakly Labelled Data

Our proposed speech enhancement system is trained on a large-scale weakly labelled AudioSet [23] dataset containing 527 kinds of sound classes. Most audio clips have durations of 10 seconds. AudioSet is weakly labelled, that is, each audio clip is only labelled with tags, but without onset and offset times of sound events. Also, AudioSet does not indicate clean speech, where speech is usually mixed with other sounds under unknown SNR. Previous work [20] proposed to build source separation systems from weakly labelled data. We show that the data mining strategy in [20] is not suitable for speech enhancement. One reason is that speech is the largest sound class in AudioSet. When randomly selecting two anchor segments from AudioSet, anchor segments $s_1$ and $s_2$ usually have overlapped audio tags. In this case, the separation system will be hard to train. To solve this problem, we propose an anchor segment mining algorithm described in Section 2.5.

To begin with, we denote two anchor segments containing different sound classes as $s_1$ and $s_2$ respectively. The anchor segments $s_1$ and $s_2$ are detected from two audio clips that most likely to contain sound events. The anchor segments $s_1$ and $s_2$ are mined from a mini-batch of anchor segments to have disjoint audio tags. The disjoint tags are important to train the speech enhancement system. In training, we build a neural network to learn a mapping:

$$f(s_1 + s_2, c_1) \mapsto s_1, \tag{2}$$

where $c_1 \in [0, 1]^K$ is a conditional vector that controls what source to separate, and $K$ is the number of sound classes in AudioSet. Equation (2) simplifies the training function in [20], and performs well in the speech enhancement task. In training, there is no need for $s_1$ or $s_2$ to be clean. The conditional vector $c_1$ is calculated by using an audio tagging system applied on $s_1$. To explain, if $s_1$ contains both "Speech" and "Water". When conditioning on the audio tagging probability $c_1$, the system (2) will separate both "Speech" and "Water". In inference, the enhanced speech $\hat{s}$ can be obtained by input a noisy speech $x$, and by setting the conditional vector $c$ as the one-hot encoding of "Speech":

$$\hat{s} = f(x, c). \tag{3}$$

To explain, the training of the speech enhancement system described in (2) does not require clean speech. In inference, we can predict clean speech from input noisy speech using (3).

### 2.3. Sound Event Detection for Selecting Anchor Segments

To begin with, we randomly select two sound classes from AudioSet. For each sound class, we randomly select an audio clip in AudioSet. However, there is no information on when sound events occur in audio clips. Therefore, we apply a sound event detection (SED) system [22] to predict the frame-wise presence probability of the sound class to detect anchor segments. The anchor segments $s_1$ and $s_2$ are 2-second audio segments used to constitute a mixture as input. The SED system is a DecisionLevelMax system from PANNs [22], which applies log mel spectrogram as input feature, and uses a 14-layer CNN as a classification model. Each convolutional layer has a kernel size of $3 \times 3$. The convolutional layers are followed by a time-distributed fully connected layer with $K$ outputs to predict the frame-wise presence probability of sound classes. The frame-wise predictions are max pooled along the time axis to obtain clip-wise predictions. We denote the weak labels of an audio clip as $y \in \{0, 1\}^K$, and its clip-wise prediction as $\hat{y} \in [0, 1]^K$. The SED system is trained by minimizing a binary cross-entropy loss [22] between predicted and target weakly labelled tags:

$$\text{loss} = -\sum_{k=1}^{K} y_k \ln \hat{y}_k + (1 - y_k) \ln(1 - \hat{y}_k). \tag{4}$$

In inference, the frame-wise prediction $p_{\text{SED}}(t, k)$ can be obtained by using the SED system. For example, if there are $T$ frames in the audio clip, then the shape of SED prediction $p_{\text{SED}}(t, k)$ will be $T \times K$. The first row of Fig. 1 shows the log mel spectrogram of a 10-second audio clip from AudioSet containing "Speech" and other sound classes. The second row shows the frame-wise SED prediction of "Speech". To detect the anchor segment that most likely to contain the $k$-th sound event, we define the area of probability in an anchor segment as:

$$q_k(t) = \sum_{t - \tau/2}^{t + \tau/2} p_{\text{SED}}(t, k), \tag{5}$$

where $\tau$ is the duration of anchor segments. Then, the time stamp of the anchor segment $t_{\text{anchor}}$ is obtained by:

$$t_{\text{anchor}} = \underset{t}{\arg\max} \, q_k(t), \tag{6}$$

where $t_{\text{anchor}}$ is the centre time of the anchor segment $s_1$. That is, we detect anchor segment $s_1$ that most likely to contain sound events, as shown in the red block in Fig. 1. Similarly, we select anchor segment $s_2$ from another audio clip. Then, we mix $s_1 + s_2$ as input to (2).
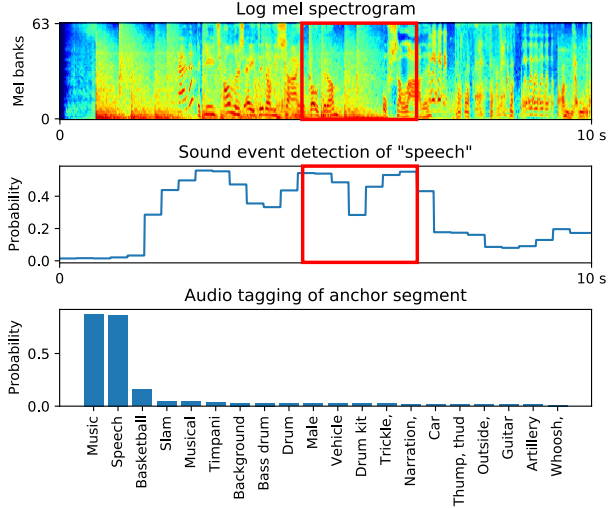
Figure 1: *Top: log mel spectrogram of a 10-second audio clip from AudioSet; Middle: predicted SED probability of "Speech", where red block shows the selected anchor segment; Bottom: predicted audio tagging probabilities of the anchor segment.*

## 2.4. Audio Tagging for Constructing Conditional Vector

In the source separation system in (2), the conditional vector $c_1$ controls what sources to separate from $s_1 + s_2$. However, the challenging problem is that there are no ground truth tags of $s_1$, so $c_1$ is unknown. Using the tags of the audio clip as the tags of $s_1$ is incorrect, because $s_1$ is only a short segment of the audio clip. In addition, there can be multiple sound events in $s_1$, but they are not labelled in the tags of the AudioSet dataset analyzed by [25]. To address this problem, we apply an audio tagging system on $s$: $c = g_{AT}(s)$ to estimate the conditional vector $c$. The audio tagging system provides a better tag estimation than the SED system. The audio tagging system $g_{AT}$ is a 14-layer CNN of PANNs [22]. The 14-layer CNN consists of several convolutional layers. Then, global average and max pooling are applied to summarize the feature maps into a fixed dimension embedding vector. Finally, a fully connected layer is applied to the embedding vector to predict the presence probability of sound events. The training of the audio tagging system applies the binary cross-entropy described in (4). The advantage of using audio tagging prediction rather than one-hot encoding of labels to build $c_1$ is that $g_{AT}(s_1)$ provides a better estimation of sound events probability in $s_1$ than weak labels. From the top to bottom in Fig. 1 shows the log mel spectrogram of a 10-second audio clip, the SED result of the audio clip, and the audio tagging probabilities of the selected anchor segment. For example, the predominant sound events in $s_1$ are "Music" and "Speech". Other sound classes in $s_1$ include "Basketball" and "Slam", etc.

## 2.5. Anchor Segment Mining

Previous works randomly select anchor segments to train general source separation systems [20]. We show that anchor segment mining is important for speech enhancement. One reason is that around 50% of audio clips in AudioSet are labelled as speech. Therefore, randomly selecting anchor segments $s_1$ and $s_2$ will often lead to $s_1$ and $s_2$ have overlap tags. In this case,

---

**Algorithm 1** Anchor segment mining.

1: Mini-batch of anchor segments: $S = \{s_1, ..., s_B\}$, and their predicted tags: $R = \{r_1, ..., r_B\}$.
2: **for** $r_1 \in R$ **do**
3:     **for** $r_2 \in R$ **do**
4:         **if** $r_1 \cap r_2 = \emptyset$ **then**
5:             Collect anchor segments of $r_1$ and $r_2$ to constitute $s_1$ and $s_2$.
6:             Remove $r_1$ and $r_2$ from $R$.
7:         **end if**
8:     **end for**
9: **end for**

---

the separation result of equation (2) can be incorrect. For example, if both $s_1$ and $s_2$ contain "Speech", when the conditional vector $c_1$ is the one-hot encoding of "Speech", the system in (2) will only separate "Speech" from $s_1$, but not "Speech" from $s_2$. However, we aim to separate all "Speech" from both $s_1$ and $s_2$. To address this problem, we propose an anchor segment mining method to mine $s_1$ and $s_2$ to have disjoint conditional vectors.

Our proposed anchor segment mining algorithm is described as follows. In training, we denote the mini-batch size as $B$. For each mini-batch, we randomly select $B$ sound classes from $K$ sound classes from AudioSet without replacement. Then, we detect anchor segments as described in Section 2.3. We denote the detected mini-batch of anchor segments as $\{s_1, ..., s_B\}$. Then, we calculate the conditional vectors $\{c_1, ..., c_B\}$ by $g_{AT}(\cdot)$. When selecting pairs of anchor segments $s_1$ and $s_2$, we need to ensure their conditional vectors $c_1$ and $c_2$ are disjoint. The conditional vectors $c_b$ have continuous values. Therefore, we apply thresholds to $c_b$ to predict their present tags $r_b$. The thresholds are calculated from PANNs [22] with equal precision and recall for each sound class. Then, we propose a mining algorithm described in Algorithm 1 to select pairs of anchor segments to have disjoint predicted tags from the mini-batch to constitute $s_1$ and $s_2$.

## 2.6. Separation Model

We propose to use convolutional UNets [26, 20] on the spectrogram of the mixture $s_1 + s_2$ to build separation systems. To begin with, the waveform of a mixture is transformed into a spectrogram. A UNet consists of an encoder and a decoder. The encoder consists of 12 convolutional layers with kernel sizes of $3 \times 3$ to extract high-level representations. Downsampling layers with sizes of $2 \times 2$ are applied to every two convolutional layers. The decoder is symmetric to the encoder with 12 convolutional layers. Transposed convolutional layers are used to upsample feature maps after every two convolutional layers. Shortcut connections are added between encoder and decoder layers with the same hierarchies. In each convolutional layer, the conditional vector $c_1$ is multiplied with a learnable matrix $\mathbf{V}$, and is added to the feature maps as bias. This bias information controls what sound classes to separate from a mixture. We denote a conditional layer as:

$$y = \text{relu}(\text{bn}(\mathbf{W} * x) + \mathbf{V}c), \quad (7)$$

where $\mathbf{W}$ is the convolutional kernel that is used to be convolved with $x$, and $\text{relu}(\cdot)$ is a ReLU non-linearity [] and $\text{bn}(\cdot)$ is a batch normalization [27] layer. The decoder outputs a spectrogram mask with values between 0 and 1, and is multiplied to the mixture spectrogram to obtain the separated spectrogram

of $s_1$. Then, an inverse short time Fourier transform (ISTFT) is applied on the separated spectrogram using the phase of the mixture to recover $\hat{s}_1$. The separation system is trained by minimizing the loss function (1).

## 3. Experiments

Our speech enhancement system is trained on the balanced subset of the weakly labelled AudioSet [23] containing 20,550 audio clips with 527 sound classes. The audio clips have durations of 10-second. Audio clips are weakly labelled, and there can be multiple sound events in an audio clip. There are 5,251 audio clips containing "Speech". To begin with, we resample all audio clips to 32 kHz to be consistent with the configuration of PANNs [22]. The sound event detection and audio tagging systems from PANNs are used to detect anchor segments as described in Section 2.5. To build the separation system, we extract spectrograms of mixtures using short-time Fourier transform (STFT) with a window size 1024 and a hop size 320. All anchor segments have durations of 2 seconds. We set the mini-batch size to 24. Adam optimizer [28] is used for training. We trained the system for 1 million iterations using a single Tesla-V100-SXM2-32GB GPU card in one week.

We evaluate our proposed speech enhancement system directly on the test set of the VoiceBank [17] and DEMAND [15] datasets without using the training data of VoiceBank and DE-MAND. Different from previous speech enhancement methods that are trained on the VoiceBank and DEMAND dataset, our proposed speech enhancement system is only trained on the AudioSet dataset while tested on the voiceBANK-DEMAND dataset. Therefore, the evaluation of our system is an out-of-domain speech enhancement system evaluation. There are 824 paired noisy and clean speech for testing in the VoiceBank-DEMAND dataset. Each audio clip has a sample rate of 48 kHz. The noisy speech have four SDR settings of 15, 10, 5 and 0 dB. There are 10 types of noise, including 2 types of synthetic noise and 8 types of noise from DEMAND. There are 28 speakers from VoiceBank.

Following previous works of speech enhancement [29, 12, 30], we apply Perceptual evaluation of speech quality (PESQ) [31], Mean opinion score (MOS) predictor of signal distortion (CSIG), MOS predictor of background-noise intrusiveness (CBAK), MOS predictor of overall signal quality (COVL) [32] and segmental signal-to-ratio noise (SSNR) [33] to evaluate the speech enhancement performance. Table 1 shows that noisy speech without enhancement achieves PESQ, CSIG, CBAK, COVL, SSNR of 1.97, 3.35, 2.44, 2.63 and 1.68 dB respectively. Our proposed speech enhancement system achieves a PESQ of 2.28, outperforming the Wiener [29] and SEGAN [12] systems. Our system achieves a CBAK of 2.96 and an SSNR of 8.75 dB, outperforming the Wiener and SEGAN systems of 2.68 and 5.07 dB, indicating the effectiveness of training speech enhancement with weakly labelled data. On the other hand, our system achieves a CSIG of 2.43 and COVL of 2.30, lower than other systems, indicating that our speech enhancement system may lose details of speech, especially the high-frequency component of speech as shown in Fig. 2.

The left and right columns of Fig. 2 visualizes two speech enhancement examples of our proposed system. From top to bottom rows show the log mel spectrogram of noisy speeches, target clean speeches and enhanced speeches respectively. Our speech enhancement trained on weakly labelled data is successful in enhancing speech from noisy signals. Considering that our system is trained on weakly labelled data only, and does not
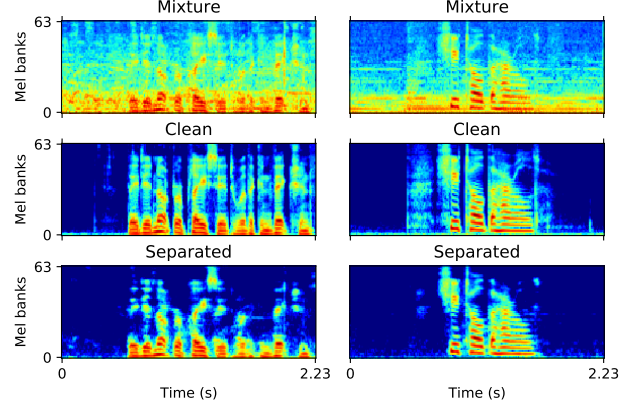


Figure 2: *The left and right columns show the speech enhancement result of two examples. Top: log mel spectrogram of noisy speech; Middle: ground truth clean speech; Bottom: enhanced speech.*

Table 1: *Speech enhancement results*

|              | PESQ | CSIG | CBAK | COVL | SSNR |
|--------------|------|------|------|------|------|
| Noisy        | 1.97 | 3.35 | 2.44 | 2.63 | 1.68 |
| Wiener [29]  | 2.22 | 3.23 | 2.68 | 2.67 | 5.07 |
| SEGAN [12]   | 2.16 | 3.48 | 2.94 | 2.80 | 7.73 |
| Wave-U-Net [30] | 2.40 | 3.52 | 3.24 | 2.96 | 9.97 |
| Proposed     | 2.28 | 2.43 | 2.96 | 2.30 | 8.75 |

use any training data from VoiceBank-DEMAND. We show that training a speech enhancement system from weakly labelled data is possible. We provide our speech enhancement demos in the following links[1][2].

## 4. Conclusion

In this work, we propose a speech enhancement framework trained on weakly labelled data from AudioSet. Our proposed speech enhancement system does not require clean speech or background noise data, which can be time-consuming to collect in the real world. We show that speech enhancement with weakly labelled data provides promising results when evaluated on the out-of-domain speech enhancement task. We propose to use sound event detection and audio tagging systems from pre-trained audio neural networks (PANNs) to detect anchor segments. We propose an anchor segment mining algorithm to mine anchor segments with disjoint tags for training. We build conditional UNet sound separation systems for speech enhancement, where the conditional vector controls what sound events to separate. In inference, enhanced speech can be obtained by input noisy speech and the one-hot encoding of "Speech" to the trained system. Our proposed system outperforms the Wiener and SEGAN systems evaluated on the VoiceBank-DEMAND dataset in the PESQ, CBAK, and SSNR metrics without using any training data from VoiceBank-DEMAND. In the future, we will continue to improve the speech enhancement performance trained on weakly labelled data.

---

[1] https://www.youtube.com/watch?v=q3hVnpNcpBI
[2] https://www.youtube.com/watch?v=DzQvn820u8E

# 5. References

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice.* CRC press, 2013.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing (TASLP)*, vol. 32, no. 6, pp. 1109–1121, 1984.

[3] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 21, no. 10, pp. 2140–2151, 2013.

[4] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *INTERSPEECH*, 2013, pp. 436–440.

[5] Y. Xu, J. Du, L. Dai, and C. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, vol. 23, no. 1, pp. 7–19, 2014.

[6] F. Weninger, H. Erdogan, S. Watanabe, E. Vincent, J. Le Roux, J. R. Hershey, and B. Schuller, "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," in *International Conference on Latent Variable Analysis and Signal Separation (LVA-ICA)*. Springer, 2015, pp. 91–99.

[7] S. R. Park and J. Lee, "A fully convolutional neural network for speech enhancement," in *INTERSPEECH*, 2017.

[8] S. Fu, Y. Tsao, and X. Lu, "SNR-Aware Convolutional Neural Network Modeling for Speech Enhancement." in *INTERSPEECH*, 2016, pp. 3768–3772.

[9] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] D. Rethage, J. Pons, and X. Serra, "A Wavenet for speech denoising," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5069–5073.

[11] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6875–6879.

[12] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *INTERSPEECH*, 2017.

[13] C. Donahue, B. Li, and R. Prabhavalkar, "Exploring speech enhancement with generative adversarial networks for robust speech recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5024–5028.

[14] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," 2020.

[15] J. Thiemann, N. Ito, and E. Vincent, "DEMAND: a collection of multi-channel recordings of acoustic noise in diverse environments," in *Proceedings of Meetings on Acoustics*, 2013.

[16] J. S. Garofolo, "TIMIT acoustic phonetic continuous speech corpus," *Linguistic Data Consortium*, 1993.

[17] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *International Conference Oriental COCOSDA with Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013.

[18] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[19] Wisdom, S. and Tzinis, E. and Erdogan, H. and Weiss, R. J. and Wilson, K. and Hershey, J. R., "Unsupervised speech separation using mixtures of mixtures," in *Workshop on International Conference on Machine Learning (ICML)*, 2020.

[20] Q. Kong, Y. Wang, X. Song, Y. Cao, W. Wang, and M. D. Plumbley, "Source separation with weakly labelled data: An approach to computational auditory scene analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 101–105.

[21] F. Pishdadian, G. Wichern, and J. Le Roux, "Finding strength in weakness: Learning to separate sounds with weak supervision," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2386–2399, 2020.

[22] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[23] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776–780.

[24] B. Gfeller, D. Roblek, and M. Tagliasacchi, "One-shot conditional audio filtering of arbitrary sounds," *arXiv preprint arXiv:2011.02421*, 2020.

[25] E. Fonseca, S. Hershey, M. Plakal, D. P. Ellis, A. Jansen, and R. C. Moore, "Addressing missing labels in large-scale sound event recognition using a teacher-student framework with loss masking," *IEEE Signal Processing Letters*, vol. 27, pp. 1235–1239, 2020.

[26] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," in *International Society for Music Information Retrieval (ISMIR)*, 2017, pp. 745–751.

[27] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[29] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1996, pp. 629–632.

[30] C. Macartney and T. Weyde, "Improved speech enhancement with the wave-u-net," *arXiv preprint arXiv:1811.11307*, 2018.

[31] ITU-T, "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.

[32] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE/ACM Transactions on audio, speech, and language processing (TASLP)*, vol. 16, no. 1, pp. 229–238, 2007.

[33] S. R. Quackenbush, T. P. Barnwell, and M. A. Clements, *Objective measures of speech quality.* Prentice Hall, 1988.