Target Sound Signal Processing with Weakly Supervised Learning

Helin Wang
Peking University
E-mail: wanghl15@pku.edu.cn

Abstract

Target sound signal processing aims to mimic human's selective auditory attention by detecting, extracting and removing a target sound source from a multi-source environment. Existing methods rely on simulated pairwise data that is hard to collect and has an obvious gap with real-world applications. To address this issue, in this research program, I propose weakly supervised learning methods for target sound signal processing to utilize large-scale real data. In addition, a curriculum learning approach is proposed to help with the training of target sound signal processing by a easy-to-hard learning manner. Furthermore, I discuss other potential research including network structure and loss function, using additional information and using self-supervised learning.

I. Introduction and Motivation

Sounds carry a great deal of critical information about our surroundings, from individual physical events to sound scenes as a whole [1]. The human brain is able to focus auditory attention on a particular sound in a multi-source environment, such as human voice, music and acoustic events. In this way, people can easily listen to the parts of their interests and mask out the the parts they are not interested in, referred to as the cocktail party problem [2]. However, research in the design of computational systems for target sound signal processing has been left largely unexplored. Target sound signal processing tasks have a lot of potential applications, such as multimedia information indexing [3], bio-acoustic species and migration monitoring [4], noise monitoring for smart cities [5] and speech enhancement [6], [7]. Our daily lives could be greatly improved if we could develop hearing devices that select the sound that we want to listen to depending on the situation.

There are three main tasks in target sound signal processing, *i.e.* detection, extraction and removal. Given a target sound's reference audio or label (*e.g.* 'babycry'), target sound detection (TSD) aims to detect the target sound signal from a mixture audio, which outputs the onset and offset times of target sound. Target sound extraction (TSE) aims at extracting the target sound signal from the mixture audio, while target sound removal (TSR) focuses on the suppression of the target sound signal. This research project mainly studies TSE and TSR tasks, which are more challenging than TSD task. Based on different target sound sources, TSE and TSR can be extended to speaker extraction (using target speakers as the reference) [8], [9], [10], instrument extraction (using music instruments as the reference) [11], [12], acoustic events selection (using target acoustic events as the reference) and so on.

So far, thanks to the development of deep learning, these tasks have been widely studied with supervised learning methods. Such methods always relies on pairwise synthetic data (called strongly-labelled data in this research project), which contains both the mixture sound signal and the target sound signal. However, reliance on this synthetic training data is problematic because good performance depends upon the degree of match between the training data and real-world audio, especially in terms of the acoustic conditions and distribution of sources. It is also hard to collect clean pairs for some rare sound types like bio-acoustic species and rare instruments. Thus, building a system to detect and extract such sounds is more challenging.

Based on the above challenges, I propose to solve the target sound signal processing tasks with weakly supervised learning. More specifically, instead of training with the strongly-labelled data, a large-scaled weakly-labelled data can be used, in which we only have the labels of what sound types happen in the

audio without the clean parts of the target sound signal. Such weakly-labelled data is much easier to obtain on the Internet platforms (such as *freesound*[1] and Youtube videos[2]) and real recordings. In order to leverage such data, this research project proposes three approaches towards different weakly supervised settings, and the contributions and aims can be summarized as follows:

1) I propose to use the result of sound classification (or detection) system to guide the training of the TSE (or TSR) networks. The objective function requires the classifier (or detector) applied to a separated source to output weak labels for the class corresponding to that source and zeros for all other classes. The objective function also enforces that the separated sources sum to the mixture.
2) I propose a reference invariant training (RIT) method, in which the model separates the mixture sound signal into a variable number of latent sources with the corresponding references, so that the separated sources can be remixed to approximate the original mixtures.
3) As the model is hard to converge trained by weakly-labelled data, a curriculum learning-based training method can be further designed to train the neural networks by a easy-to-hard manner.
4) I also study more effective supervised methods, including the network structure, the loss function, using additional information and employing self-supervised learning.

II. Related Works

In this section, I will introduce some of the related works, including sound source separation, target sound extraction, weakly supervised learning and curriculum learning.

A. Sound Source Separation

Sound source separation refers to the technology of extracting a single sound source signal from a given mixed sound signal [13]. In this case, the number of mixed channels used is usually one, two, or more, however, mixed channels of more than five are rare [14]. There should be no less than two sources of audio for the dataset to be considered as mixed, and the general range goes from two to ten. The concept of the signal source is somewhat vague. For example, a cello, a viola or a violin can be considered as a separate sound source, but they can also be a part of the sound source of orchestral instruments.

In the development of sound source separation, several models and methods have been proposed which can be generally divided into supervised and unsupervised learning algorithms. A supervised learning algorithm is a machine learning algorithm that makes predictions using a training dataset consisting of input values and corresponding output values. The size of the training dataset determines the predictive ability of the model, that is, the accuracy of sound source separation. Among these methods, the typical ones are the Support Vector Machine (SVM), the Gaussian Mixture Model (GMM), and the Denoising Autoencoder (DAE) [15]. Unsupervised learning algorithm [16] is also a machine learning algorithm that can find some hidden relationships in the unmarked data to get the correct separation of sound signals. Among these methods, the typical ones are Independent Component Analysis (ICA), Sparse Coding, and Non-negative matrix factorization (NMF) [17]. Non-negative matrix factorization and Denoising Autoencoder are popular methods [18], [19] for the separation of sound signals, however, the latter requires a large set of source signal training to ensure its correct performance which, in many cases, is difficult to achieve. Furthermore, the rise of deep learning technology has made great progresses on the study of supervised sound source separation. Recent sound source separation approaches have been formulated in either the time-frequency (T-F, or spectrogram) representation of the mixture signal, which is estimated from the waveform using the short-time Fourier transform (STFT) [20], or directly the waveform domain such as Conv-TasNet [21], Dual-path RNN [22] and Sepformer [23]. All these methods usually rely on simulated paired data for training.

[1]https://freesound.org/
[2]https://www.youtube.com/

B. Target Sound Extraction

Human beings can perceive a target sound that we are interested in from a multi-source environment by the selective auditory attention [24], [25] and such applications have been widely explored in machine hearing. Such target sound extraction tasks can be seen as an extension of sound source separation, where the difference is target sound extraction needs to attend to the reference information instead of separating all sound types. Compared with general sound source separation, target sound extraction has its own advantages and values for studying. For one thing, general sound source separation often require knowing or estimating the number of sound sources in the mixture in advance. However, the number of sound sources cannot always be known in advance in real world applications [10]. For another, sound source separation methods may suffer from what is called global permutation ambiguity, where the separated sound for the same sound source may not stick to the same output stream when crossing long pauses or duration [21]. Target sound extraction can naturally solve these problems because it only focuses on the sound sources we are interested in and other sound sources become background noise or interference.

Recent popular methods for target sound extraction are SpEx [10], speaker beam [8] and Voicefilter [26], in which the basic network structures are inspired by the methods in sound source separation. Similar to sound source separation, the performance of extraction highly relies on the synthetic paired data.

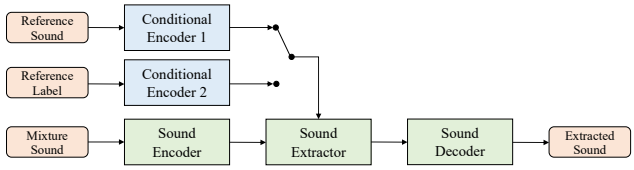
C. Weakly Supervised Learning and Unsupervised Learning

Beside supervised learning methods, there is a type of method called weakly supervised learning algorithm [27]. This algorithm does not require a highly trained set of labels and is more specific than abstract unsupervised learning. As for sound areas, weakly supervised learning has been successfully used in classification tasks like automated speech recognition [28] and audio classification [29]. In this methods, the training data only provides labels, without the onset and offset timing information, which is more challenging than supervised learning. Other studies explored unsupervised learning algorithm for sound applications. A popular one is Autoencoder (AE), which can learn the efficient representation of input dataset through unsupervised learning, along with an extension model called Variational Autoencoder (VAE) [30].

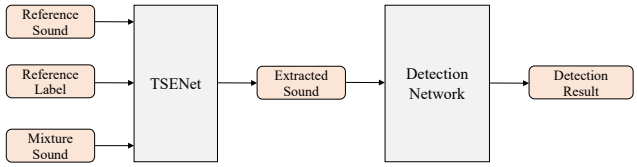
Although several methods have been proposed to use weakly supervised learning or unsupervised learning for sound source separation tasks [31], [32], [33], there is not any attempt yet for target sound extraction. As additional reference information should be attended to in target sound extraction tasks, effective ways to use weakly supervised learning need to be explored.

D. Curriculum Learning

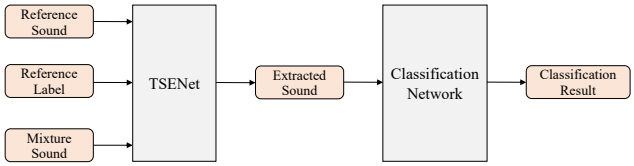
In most deep neural network (DNN)-based models, examples are considered in a random order during training, which is different from the learning process by human brain. Humans learn the basic (easy) concepts sooner and the advanced (hard) concepts later, which is basically reflected in all the curricula taught in schooling systems around the world, as humans learn much better when the examples are not randomly presented but are organized in a meaningful order [34]. Curriculum Learning is a similar strategy for training a machine learning model, which can achieve two important benefits: (i) an increase of the convergence speed of the training process and (ii) a better accuracy. Bengio *et al.* [35] are the first to formalize the easy-to-hard training strategies in the context of machine learning, proposing the curriculum learning (CL) paradigm. This seminal work inspired many researchers to pursue curriculum learning strategies in various application domains, such as weakly supervised object localization [36], object detection [37] and neural machine translation [38] among many others. In this research project, I plan to introduce the curriculum learning to target sound signal processing tasks to solve the problem of hard training for weakly supervised learning.



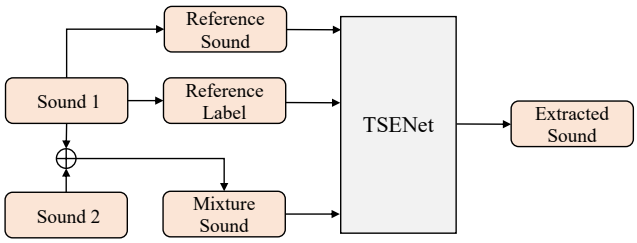
(a) Supervised Method.



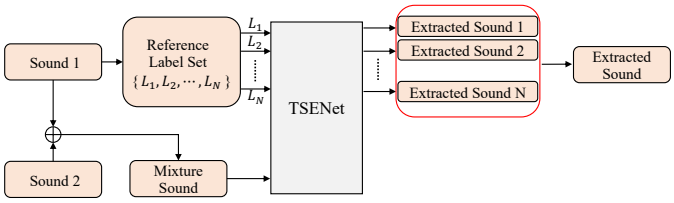
(b) Weakly Supervised Method with Sound Detection.



(c) Weakly Supervised Method with Sound Classification.



(d) Weakly Supervised Method with Mixture Training



(e) Weakly Supervised Method with Reference Invariant Training

Fig. 1. Frameworks with supervised and weakly supervised methods for TSE.

III. Methodology

In this section, I will firstly introduce the supervised TSE methods, followed by proposed weakly supervised methods with different settings. Then I will introduce a curriculum learning approach for TSE. Finally, I will discuss some potential studies.

A. Supervised Framework for TSE

The framework of supervised learning method for TSE is shown in Fig. 1 (a). Given the mixture sound $\mathbf{x}$ and the reference information (can either be a reference sound $\mathbf{r_S}$ and/or a reference label $\mathbf{r_L}$), the TSE model aims to estimate the target sound $\mathbf{y}$,

$$\hat{\mathbf{y}} = f_{TSE}(\mathbf{x}, \mathbf{r_S}, \mathbf{r_L}; \theta_{TSE}) \tag{1}$$

where $\hat{\mathbf{y}}$ is the estimated target sound, $f_{TSE}(\cdot)$ denotes the TSE model and θ_{TSE} denotes the corresponding parameters. More specifically, the reference sound and reference label are firstly fed into two different conditional encoders, which output two conditional embeddings that represent the reference information.

$$\mathbf{e_S} = f_{ConE1}(\mathbf{r_S}; \theta_{ConE1}) \tag{2}$$

$$\mathbf{e_L} = f_{ConE2}(\mathbf{r_L}; \theta_{ConE2}) \tag{3}$$

where $f_{ConE1}(\cdot)$ and $f_{ConE2}(\cdot)$ denote the conditional encoders for the reference sound and reference label respectively and θ_{ConE1} and θ_{ConE2} denote the corresponding parameters. The extraction network contains three main parts: (i) a sound encoder, which extracts acoustic feature from the mixture sound; (ii) a sound extractor, which inputs the acoustic feature and the conditional embeddings and outputs the extracted intermediate feature of the target sound (usually a mask [10]); (iii) a sound decoder, which is used to

transform the estimated intermediate feature into the extracted waveform $\hat{\mathbf{y}}$. A very popular time-domain loss function in this task is the multi-scale scale-invariant signal-to-distortion ratio (SI-SDR) loss [39], denoted as $\mathcal{L}_{sisdr}$.

$$\mathcal{L}_{sisdr}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = 10\log_{10}\left(\frac{\left\|\frac{\langle\hat{\mathbf{y}},\mathbf{y}\rangle}{\langle\mathbf{y},\mathbf{y}\rangle}\mathbf{y}\right\|^2}{\left\|\frac{\langle\hat{\mathbf{y}},\mathbf{y}\rangle}{\langle\mathbf{y},\mathbf{y}\rangle}\mathbf{y} - \hat{\mathbf{y}}\right\|^2}\right) \tag{4}$$

where $\langle\cdot,\cdot\rangle$ is the inner product. To ensure scale invariance, the signals $\hat{\mathbf{y}}$ and $\mathbf{y}$ are normalized to zero-mean prior to the SI-SDR calculation.

B. Weakly Supervised Framework for TSE

As described above, supervised methods need pairwise training data, *i.e.* $\mathbf{x}$, $\mathbf{y}$, $\mathbf{r_S}$ and $\mathbf{r_L}$, which are quite difficult to obtain in real-world recordings. Therefore, previous methods mainly rely on simulated data that has a large gap with real applications. To address this issue, I propose to utilize the weakly labelled data from real world, which is much easier to obtain but more challenging. Based on different weakly-labelled settings, I propose four methods for TSE.

1) Weakly Supervised Method with Sound Detection: For a mixture sound which we do not have the clean target sound but we have the labels of sounds happened and their onset & offset times, a sound detection network can be used to guide the learning of TSE network. As shown in Fig. 1 (b), we firstly pre-train a detection network which can classify different sound types and give their onset & offset times. Next, we set the detection network as the back-end of the TSE network and the extracted sound is fed into the detection network to carry out detection. Finally, the whole system is trained by the ground truth of target sound detection. In this way, the objective function requires the detector applied to the extracted sound to output weak labels in the regions of the target sound source and zeros for all other sources and regions. Let $\hat{\mathbf{z}}$ and $\mathbf{z}$ be the estimated target sound detection result and the corresponding ground truth, the network is trained with a binary cross-entropy loss,

$$\mathcal{L}_{detect}\left(\mathbf{z}, \hat{\mathbf{z}}\right) = \sum_{i=1}^{n}\left(-\hat{z}_i\log(z_i) - (1-\hat{z}_i)\log(1-z_i)\right) \tag{5}$$

where n denotes the number of sound types.

2) Weakly Supervised Method with Sound Classification: Similar to the weakly supervised method with sound detection, if there is only labels of sounds provided without their timing information, a sound classification network can be used to guide the learning of TSE network. As shown in Fig. 1 (c), the training strategy is similar to weakly supervised method with sound detection, and the only difference is changing the back-end from a detection network to a classification network. Theoretically, weakly supervised method with sound classification is more hard to train than that with sound detection.

3) Weakly Supervised Method with Mixture Training: The above two methods use additional task (*i.e.* detection and classification) to help with the TSE task, however, the TSE network is not directly optimized. There are two potential drawbacks, (i) the performance is highly determined by the accuracy of detection and classification, and (ii) the quality of the extracted sound may be poor as the extraction network only needs to fit the back-end networks without any other restriction. To overcome these problems, I propose a mixture training method which simulates training pairs with mixture sounds to train weakly supervised TSE. To be more specific, as shown in Fig. 1 (d), a new mixture sound can be simulated with two original mixture sounds that includes different sound types. We can use one of the original mixture sounds as a target sound and consider the other as the interference. After that, the training pipeline is the same as the supervised method, for which the aim of the TSE network is to extract the target sound from the simulated mixture sound.

4) Weakly Supervised Method with Reference Invariant Training: The mixture learning method uses the whole mixture sound or all of the labels of the mixture sound as reference, however, in real-world applications, we often need to extract only one sound source from the mixture. Therefore, the best way to eliminate this gap is considering single sound source in the training stage, which is actually hard to accomplish because we do not have single sound source data. Inspired by the permutation invariant training (PIT) [40] which solves the output permutation problem caused by the lack of a unique source class for each output, I extend the mixture training method to a reference invariant training method. The key idea is estimating each sound source with each reference label and using the summation of these sources as the extracted target sound. As shown in Fig. 1 (e), consider two mixture sounds $\mathbf{x_1}$ and $\mathbf{x_2}$, we use $\mathbf{x_1}$ as the target sound. Let $\{L_1, L_2, \cdots, L_N\}$ be the sound labels within $\mathbf{x_1}$, where N denotes the number of labels, the TSE network inputs each label and outputs the corresponding extracted sound $\{\hat{\mathbf{y}}_\mathbf{1}, \hat{\mathbf{y}}_\mathbf{2}, \cdots, \hat{\mathbf{y}}_\mathbf{N}\}$. The final extracted target sound is calculated by the summation of them,

$$\hat{\mathbf{y}} = \sum_{i=1}^{N} \hat{\mathbf{y}}_\mathbf{i} \tag{6}$$

C. Curriculum Learning Approach for TSE

To design a curriculum learning strategy for training neural networks, which settings should be used to define hard samples and easy samples is very important. For TSE tasks, these settings could be the number of labels within the mixture, the length of audio, the signal-to-distortion rate and so on. We could further explore different setting to better train the TSE models with weakly supervised learning.

D. Other Potential Studies for TSE

Apart from the learning and training strategies, we could further make other potential research, such as the network and loss function, using additional information and using self-supervised pre-training.

(i) Network and loss function: Current existing popular network structures are Conv-TasNet [21], Dual-path RNN [22], and DPTNet [41]. However, the computational complexity is quite high especially with the increase of audio length. In addition, existing methods often need a long context to make inferences and are poor at streaming applications. Thus, more effective, efficient and economic network structure should be explored. Furthermore, new loss functions could be studied to better achieve the extraction, *e.g.* giving more weights the loss of overlapping regions to enhance the separation of overlapping regions.

(ii) Using additional information: We could use the timing information (*i.e.* onset and offset times) to guide the extraction, and we could fuse visual information to form a audio-visual TSE task.

(iii) Using self-supervised pre-training: Motivated by the success of large-scale self-supervised pre-training in natural language processing [42] and computer vision [43] areas, we could further explore self-supervised pre-training models to help with downstream separation and extraction tasks. More specifically, the learned representation from the self-supervised pre-training can replace the sound encoder module in Fig. 1 (a).

IV. Research Schedule

To achieve my potential research contributions, I divided the research program into the following distinct tasks:

(i) Stage 1: Preparation (Month 0-12)

Conduct thorough literature survey on the methodology of sound source separation, target sound extraction, weakly supervised learning, unsupervised learning and self-supervised learning.

(ii) Stage 2: Collect Data (Month 12-18)

Collect and preprocessing data.

(iii) Stage 3: Basic Development (Month 18-21)

Construct a preliminary target sound extraction system with supervised learning (baseline), including

different network structures and implementation details; Evaluate the performance of baseline system and analyze the results.

(iv) Stage 4: Construct Novel Methods (Month 21-31)

Integrate all the achievements and build novel systems for target sound extraction with weakly supervised learning; Evaluate the performance of proposed methods in different experimental setups; Apply the TSE system into a practical APP or devices.

(v) Stage 5: Documentation (Month 31-36)

Complete data analysis and thesis writing.

REFERENCES

[1] T. Virtanen, M. D. Plumbley, and D. Ellis, *Computational analysis of sound scenes and events*. Springer, 2018.

[2] S. Haykin and Z. Chen, "The cocktail party problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005. [Online]. Available: https://doi.org/10.1162/0899766054322964

[3] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 131–135. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7952132

[4] D. Stowell and D. Clayton, "Acoustic event detection for multiple overlapping similar sources," in *2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2015, New Paltz, NY, USA, October 18-21, 2015*. IEEE, 2015, pp. 1–5. [Online]. Available: https://doi.org/10.1109/WASPAA.2015.7336885

[5] J. P. Bello, C. T. Silva, O. Nov, R. L. DuBois, A. Arora, J. Salamon, C. Mydlarz, and H. Doraiswamy, "SONYC: a system for monitoring, analyzing, and mitigating urban noise pollution," *Commun. ACM*, vol. 62, no. 2, pp. 68–77, 2019. [Online]. Available: https://doi.org/10.1145/3224204

[6] J. Benesty, S. Makino, and J. Chen, *Speech enhancement*. Springer Science & Business Media, 2006.

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

[8] M. Delcroix, K. Zmolíková, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 5554–5558. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8462661

[9] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 86–90. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8682245

[10] C. Xu, W. Rao, E. S. Chng, and H. Li, "Spex: Multi-scale time domain speaker extraction network," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 1370–1384, 2020. [Online]. Available: https://doi.org/10.1109/TASLP.2020.2987429

[11] S. Uhlich, F. Giron, and Y. Mitsufuji, "Deep neural network based instrument extraction from music," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*. IEEE, 2015, pp. 2135–2139. [Online]. Available: https://doi.org/10.1109/ICASSP.2015.7178348

[12] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications, and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, 2014. [Online]. Available: https://doi.org/10.1109/MSP.2013.2271648

[13] F. Weninger, J. Geiger, M. Wöllmer, B. Schuller, and G. Rigoll, "The munich 2011 chime challenge contribution: Nmf-blstm speech enhancement and recognition for reverberated multisource environments," in *Proc. Machine Listening in Multisource Environments (CHiME 2011), satellite workshop of Interspeech 2011, ISCA, Florence, Italy*, 2011.

[14] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Speech Audio Process.*, vol. 14, no. 4, pp. 1462–1469, 2006. [Online]. Available: https://doi.org/10.1109/TSA.2005.858005

[15] M. M. R. Pimpale, S. Therese, and V. Shinde, "A survey on: sound source separation methods [j]," *International Journal*, vol. 3, no. 11, pp. 580–584, 2016.

[16] T. Virtanen, "Unsupervised learning methods for source separation in monaural music signals," in *Signal Processing Methods for Music Transcription*. Springer, 2006, pp. 267–296.

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13, Papers from Neural Information Processing Systems (NIPS) 2000, Denver, CO, USA*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds. MIT Press, 2000, pp. 556–562. [Online]. Available: https://proceedings.neurips.cc/paper/2000/hash/f9d1152547c0bde01830b7e8bd60024c-Abstract.html

[18] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *2003 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (IEEE Cat. No. 03TH8684)*. IEEE, 2003, pp. 177–180.

[19] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *2017 IEEE Global Conference on Signal and Information Processing, GlobalSIP 2017, Montreal, QC, Canada, November 14-16, 2017*. IEEE, 2017, pp. 1265–1269. [Online]. Available: https://doi.org/10.1109/GlobalSIP.2017.8309164

[20] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018. [Online]. Available: https://doi.org/10.1109/TASLP.2018.2842159

[21] 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, no. 8, pp. 1256–1266, 2019. [Online]. Available: https://doi.org/10.1109/TASLP.2019.2915167

[22] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 46–50. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9054266

[23] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, "Attention is all you need in speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2021, Toronto, ON, Canada, June 6-11, 2021*. IEEE, 2021, pp. 21–25. [Online]. Available: https://doi.org/10.1109/ICASSP39728.2021.9413901

[24] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the acoustical society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[25] S. Getzmann, J. Jasny, and M. Falkenstein, "Switching of auditory attention in "cocktail-party" listening: Erp evidence of cueing effects in younger and older adults," *Brain and cognition*, vol. 111, pp. 1–12, 2017.

[26] Q. Wang, H. Muckenhirn, K. W. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. Lopez-Moreno, "Voicefilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 2728–2732. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1101

[27] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *National science review*, vol. 5, no. 1, pp. 44–53, 2018.

[28] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, ser. ACM International Conference Proceeding Series, W. W. Cohen and A. W. Moore, Eds., vol. 148. ACM, 2006, pp. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[29] Y. Xu, Q. Kong, W. Wang, and M. D. Plumbley, "Large-scale weakly supervised audio classification using gated convolutional neural network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018*. IEEE, 2018, pp. 121–125. [Online]. Available: https://doi.org/10.1109/ICASSP.2018.8461975

[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014. [Online]. Available: http://arxiv.org/abs/1312.6114

[31] 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 *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 101–105. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053396

[32] F. Pishdadian, G. Wichern, and J. L. Roux, "Learning to separate sounds from weakly labeled scenes," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 91–95. [Online]. Available: https://doi.org/10.1109/ICASSP40776.2020.9053055

[33] S. Wisdom, E. Tzinis, H. Erdogan, R. J. Weiss, K. W. Wilson, and J. R. Hershey, "Unsupervised sound separation using mixture invariant training," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/28538c394c36e4d5ea8ff5ad60562a93-Abstract.html

[34] P. Soviany, R. T. Ionescu, P. Rota, and N. Sebe, "Curriculum learning: A survey," *CoRR*, vol. abs/2101.10382, 2021. [Online]. Available: https://arxiv.org/abs/2101.10382

[35] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, ser. ACM International Conference Proceeding Series, A. P. Danyluk, L. Bottou, and M. L. Littman, Eds., vol. 382. ACM, 2009, pp. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380

[36] R. T. Ionescu, B. Alexe, M. Leordeanu, M. Popescu, D. P. Papadopoulos, and V. Ferrari, "How hard can it be? estimating the difficulty of visual search in an image," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2157–2166. [Online]. Available: https://doi.org/10.1109/CVPR.2016.237

[37] M. Shi and V. Ferrari, "Weakly supervised object localization using size estimates," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9909. Springer, 2016, pp. 105–121. [Online]. Available: https://doi.org/10.1007/978-3-319-46454-1_7

[38] T. Kocmi and O. Bojar, "Curriculum learning and minibatch bucketing in neural machine translation," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, R. Mitkov and G. Angelova, Eds. INCOMA Ltd., 2017, pp. 379–386. [Online]. Available: https://doi.org/10.26615/978-954-452-049-6_050

[39] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*. IEEE, 2019, pp. 626–630. [Online]. Available: https://doi.org/10.1109/ICASSP.2019.8683855

[40] D. Yu, M. Kolbæk, Z. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 241–245. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7952154

[41] J. Chen, Q. Mao, and D. Liu, "Dual-path transformer network: Direct context-aware modeling for end-to-end monaural speech separation," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 2642–2646. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-2205

[42] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*,

J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[43] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "Videobert: A joint model for video and language representation learning," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 7463–7472. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00756