



Audio Enhancement from Multiple Crowdsourced Recordings: A Simple and Effective Baseline

Shiran Aziz, Yossi Adi, Shmuel Peleg

School of Computer Science and Engineering
The Hebrew University of Jerusalem, Israel

shiran.aziz@mail.huji.ac.il

Abstract

With the popularity of cellular phones, events are often recorded by multiple devices from different locations and shared on social media. Several different recordings could be found for many events. Such recordings are usually noisy, where noise for each device is local and unrelated to others. This case of multiple microphones at unknown locations, capturing local, uncorrelated noise, was rarely treated in the literature. In this work we propose a simple and effective crowdsourced audio enhancement method to remove local noises at each input audio signal. Then, averaging all cleaned source signals gives an improved audio of the event. We demonstrate the effectiveness of our method using synthetic audio signals, together with real-world recordings. This simple approach can set a new baseline for crowdsourced audio enhancement for more sophisticated methods which we hope will be developed by the research community. Code, dataset, and models are available.

Index Terms: Audio enhancement, Time-frequency filtering, Crowdsourced denoising, User-Generated recordings

1. Introduction

Cellular phones are powerful multimedia devices, capable of quality recording of events around us. In particular, public events are often captured by multiple people from different locations. Many such user-generated recordings are also uploaded to social media, where several different recordings could be found for each event. In most cases user recordings have noisy audio signals, where noises are local to each device, and unrelated to each other due to the distance between users. Crowdsourced audio enhancement aims to use all available audio signals of an event, creating an audio signal that excludes the local noises at each input signal.

Unlike more traditional single-channel and multi-channel denoising approaches [1, 2, 3, 4, 5, 6], in crowdsourced audio enhancement there is no prior definition of noise. Instead, noise is defined as a sound that is not common to most input audio. Hence, while local sounds will be removed, any global sounds that is present in all input signals will remain. For instance, consider several people shooting with their cell phones videos of a musical concert from different locations in the hall. The music coming from the main stage will be captured in all recordings, however the background noise will be unique to each of the recordings.

This work presents a straightforward method for crowdsourced audio enhancement. The method is based on filtering noisy space-time outliers from the input spectrograms considering both upper and lower thresholds. Specifically, we start by computing the Short-Time Fourier Transform (STFT) of all input signals. For each Time-Frequency (TF) cell we exam-

ine the magnitude values given to it by each input signal, and outlier values in each cell are removed. We define outliers as values which are substantially higher than the median magnitude value of the corresponding TF cell. The enhanced signal is constructed by averaging all amplitudes in each TF cell that are not considered as outliers. We evaluated the proposed method considering both synthetic and in-the-wild recordings. Results suggest that the proposed method significantly outperforms the baseline methods considering a diverse set of sources and background noises. The proposed method is simple and straightforward, requires no training, hence can serve as a foundational baseline for comparison with more sophisticated statistical techniques.

2. Related Work

While much work has been done on audio enhancement using multi-channel microphone arrays [7], most papers are *position-aware* and address the case where the properties of the microphones and the relationship between them are known and constant [8, 9, 10].

Combining user recordings should be *position-agnostic*, as we do not have any prior information on the relative position of the microphones. The authors in [11] were the first, to the best of our knowledge, to address crowdsourced audio enhancement from unrelated recordings. They proposed creating an improved audio signal, where the possible corruptions in each input signal can be missing frequencies or missing time periods.

Another relevant line of work is *scene-agnostic* multi-microphone speech processing. The authors in [12] proposed a deep learning based solution for speech dereverberation considering a varying number of microphone array at different positions. Some papers [13, 14] are focused on a setup where the target speaker is always closest to the microphone array. Unlike this approach we only have a single clean source, and we can not assume that one microphone has the cleanest recording of this source. Recently, [15] showed, in parallel to our work, a flexible multichannel speech enhancement, for a varying number of microphones at random positions inside a room. Though they show impressive results they focus on indoor speech recordings with relatively small distances between the microphones in the array. Unlike the crowdsourced speech enhancement task, these lines of work assume that all sources are captured by all microphones. Similarly, in Independent Component Analysis (ICA) [16] multi-channel speech separation is done by finding a linear representation of non-Gaussian data so that the components are as statistically independent as possible. Notice, ICA considers equal number of sources and microphones. Following such line of research the authors in [17] proposed the *Full-rank spatial Covariance Analysis* (FCA) method, while recently the authors in [18] proposed the *fastFCA*, which extends such re-

search direction and proposed a method for source separation for the undetermined case of more sources than microphones.

In this work, we address the case where each audio signal has an independently added noise. Similarly to our setup, the authors in [19] propose the *Max-elimination* method, which removes at each time-frequency cell the signal having a maximal amplitude.

3. Model

We address an audio source S , recorded by m independent microphones at unknown locations. It is assumed that microphones are far from each other such that all m noises are different and uncorrelated. Let A_1, \dots, A_m be the input signals from each of the microphones, where each signal A_i is composed of the source signal S together with added noise N_i . Our proposed method starts by temporally aligning all input signals and normalize their magnitude. Then, we denoise the input signals using time-frequency filtering.

We start by describing the alignment and normalization method. We temporally align the audio signals with the method proposed by [20], using time-frequency magnitude peaks. Alignment is done by finding correspondences between frequencies and time differences of detected pairs of peaks¹. After temporal alignment, the corresponding peaks used in the alignment process can be assumed to belong to the clean audio source, and the amplitude at the corresponding peaks are normalized accordingly. We normalize by estimating the multiplicative constant between all corresponding pairs in the log-spectrogram: we select the signal with the maximum number of matched peaks as anchor signal, and normalize the others multiplying by the corresponding α value for each pair. Formally, given matched spectral peaks $\{(|P_n^A|, |P_n^B|)\}_{n=1}^N$ where $|P_n^A|$ is the amplitude of the frequency peak of signal A in the n^{th} matched pair with signal B . Using the log spectrogram amplitude, we estimate the coefficient using the mean of all the pairs as $\alpha^{AB} = (\sum_{n=1}^N |P_n^A|) / \sum_{n=1}^N (|P_n^B|)$.

Once all signals are aligned and normalized, we propose to estimate the source signal S from the m noisy input signals $\{A_i\}_{i=1}^m$ by removing outliers from TF cells. Formally, for all input signals A_i , we start by extracting the STFTs Y_i , using 2048 FFT coefficients, window size of 2048, and overlap ratio of 0.25. Next, for each (t, f) cell in the magnitude spectrogram we perform the following: (i) given all input magnitude spectrograms $|Y_i|$ compute the median amplitude in each (t, f) cell, denoted as $C(t, f)$; (ii) define as outliers those signals whose spectrogram magnitudes are above or below given thresholds, that depend of the above computed median $C(t, f)$. Formally, $|Y_i(t, f)| > \lambda_1 C(t, f)$ or $|Y_i(t, f)| < \lambda_2 C(t, f)$, where λ_1 and λ_2 are hyper-parameters calibrated on the held-out set; (iii) The denoised magnitude spectrogram $G(t, f)$ is constructed by averaging the values of all signals that were not detected as outliers. Intuitively, when $|Y_i(t, f)|$ is substantially larger than the median or substantially lower than the median of all input signals, it is considered as noise. Next, we relax the prior constraint using γ . We look at TF cells in the neighborhood of a removed cell, and also remove those that fulfills a relaxed constraint. If all signals in a cell are removed by the above process, only the upper bound is used. Lastly, we convert $G(t, f)$ back to a time-domain signal by applying inverse STFT using the mean phase of all signals. In all experiments we use $\lambda_1 = 1.15$, $\gamma = 1.1$ and $\lambda_2 = 0.01$. A pseudo-code of the

¹We use the implementation from <https://github.com/worldveil/dejavu>

Algorithm 1

```

1: for  $i = 1, 2, \dots, m$  do
2:    $Y_i = STFT(A_i)$ 
3:    $M_i \leftarrow$  ones of the shape  $Y_i$ 
4: end for
5:  $C \leftarrow \text{median}(\{|Y_1|, \dots, |Y_m|\})$ 
6: for  $i = 1, 2, \dots, m$  do
7:   for  $(t, f)$  in  $Y_i$  do:
8:     if  $|Y_i(t, f)| > \lambda_1 C(f, t)$  or
        $|Y_i(t, f)| < \lambda_2 C(f, t)$  then
9:        $M_i(f, t) \leftarrow 0$ 
10:    end if
11:  end for
12:   $G_i^0 = Y_i$ ,  $G_i^1 \leftarrow M_i \odot Y_i$ 
13:  while  $G_i^{j-1} \neq G_i^j$  do
14:    for  $(t, f)$  which  $M_i(t, f) = 0$  do
15:      for  $t-1 \leq s \leq t+1$  and
         $f-1 \leq g \leq f+1$  do
16:        if  $|Y_i(s, g)| > \gamma C(s, g)$  then
17:           $M_i(s, g) \leftarrow 0$ 
18:        end if
19:        update  $G_i^j \leftarrow M_i \odot Y_i$ 
20:      end for
21:    end while
22:   $G \leftarrow (\sum_{i=1}^m G_i^{final}) / (\sum_{i=1}^m M_i)$ 
23: return  $ISTFT(G)$ 

```

proposed algorithm can be found at Algorithm 1.

4. Experiments

We evaluate the proposed approach considering two different setups: (i) a synthetically generated dataset; and (ii) a dataset of real-world, user recordings collected from the web. The use of synthetic dataset allows us to evaluate the proposed approach in a controlled setting, exploring different noise levels and different types of noises. We also demonstrate that the proposed approach can generalize to user recording obtained from YouTube.

4.1. Datasets

Synthetic Recordings. We artificially generated noisy inputs by mixing source and noise signals. As common audio source we use either music from the MUSDB18 benchmark [21] or speech from the LibriSpeech corpus [22]. All audio samples were resampled to 16kHz. Each common audio source signal S is duplicate to k channels, while for each channel we add an independent noise. Each noise signal is multiplied by a different constant, a_i , which reflects the desired Signal-to-Noise Ratio (SNR) of the input signals. Formally, let N_1, \dots, N_k be the noises added to each channel, the a_i coefficients are computed as follows,

$$a_i = \sqrt{P(S) / (10^{SNR_{ab}/10} \cdot P(N_i))}, \quad (1)$$

$$P(S) = \sum_{n=1}^{\frac{\text{len}(S)}{\tau}} \left(\max_{t \in (n\tau, (n+1)\tau)} S(t) \right)^2,$$

Where τ is a time interval, which was set to be 1 second. We consider different types of noises such as speech, environmental

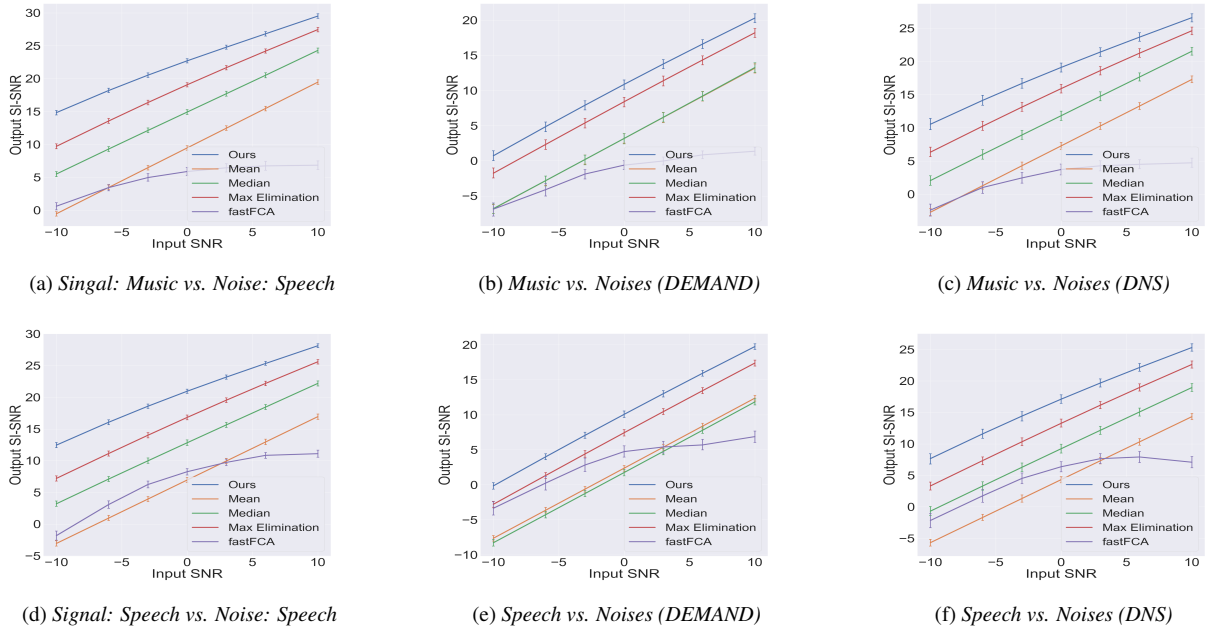


Figure 1: Combining 5 synthetic noisy audio signals: Average SI-SNR of enhanced signal as a function of the SNR of the input signals, and 95% confidence interval on 100 experiments. Methods compared: (i) Mean: Using the mean of all signals. (ii) Median: Replacing the mean magnitude with the median magnitude in each TF cell. (iii) Max Elimination [19]: Removing the maximal magnitude in each TF cell. (iv) fastFCA: model the contribution of each source as a complex Gaussian distribution with zero mean. (v) Our Crowdsourced Enhancement, consistently having the best results.

Table 1: signal: speech vs. Noise: speech / signal: Speech vs. Noises (DEMAND) / Music vs. Noise: Speech. Same as Fig. 1, but with PESQ and STOI as evaluation metric. The 95% confidence interval ranges between 0.05 – 0.18 and 0.01 – 0.02 respectively

SNR	PESQ					STOI				
	MEAN	MEDIAN	FASTFCA	MAX ELIMI	OURS	MEAN	MEDIAN	FASTFCA	MAX ELIMI	OURS
-10	1.07 / 1.05 / 1.26	1.16 / 1.06 / 1.34	1.10 / 1.18 / 1.20	1.25 / 1.12 / 1.59	1.61 / 1.22 / 2.18	0.54 / 0.60 / 0.39	0.69 / 0.64 / 0.59	0.60 / 0.73 / 0.45	0.77 / 0.73 / 0.70	0.86 / 0.78 / 0.82
-6	1.10 / 1.09 / 1.21	1.29 / 1.11 / 1.62	1.23 / 1.33 / 1.31	1.46 / 1.26 / 1.97	1.98 / 1.44 / 2.69	0.64 / 0.69 / 0.52	0.78 / 0.73 / 0.70	0.71 / 0.80 / 0.57	0.84 / 0.81 / 0.79	0.91 / 0.85 / 0.87
-3	1.14 / 1.14 / 1.30	1.44 / 1.19 / 1.91	1.39 / 1.42 / 1.41	1.69 / 1.44 / 2.32	2.32 / 1.69 / 3.04	0.70 / 0.76 / 0.61	0.83 / 0.79 / 0.77	0.78 / 0.84 / 0.65	0.88 / 0.85 / 0.84	0.95 / 0.89 / 0.90
0	1.23 / 1.24 / 1.51	1.64 / 1.33 / 2.26	1.58 / 1.56 / 1.59	1.98 / 1.68 / 2.72	2.68 / 2.00 / 3.34	0.77 / 0.81 / 0.69	0.87 / 0.84 / 0.82	0.84 / 0.86 / 0.71	0.91 / 0.89 / 0.88	0.95 / 0.92 / 0.92
3	1.35 / 1.41 / 1.80	1.91 / 1.53 / 2.64	1.78 / 1.64 / 1.78	2.32 / 1.98 / 3.06	3.02 / 2.35 / 3.50	0.82 / 0.86 / 0.76	0.90 / 0.88 / 0.86	0.87 / 0.86 / 0.75	0.93 / 0.92 / 0.91	0.96 / 0.94 / 0.94
6	1.54 / 2.32 / 2.15	2.22 / 1.79 / 2.98	2.00 / 1.70 / 1.96	2.68 / 2.32 / 3.36	3.33 / 2.71 / 3.79	0.86 / 0.89 / 0.83	0.93 / 0.91 / 0.90	0.90 / 0.85 / 0.78	0.95 / 0.94 / 0.93	0.97 / 0.95 / 0.96
10	1.90 / 2.04 / 2.66	2.68 / 2.24 / 3.41	2.20 / 1.88 / 1.97	3.14 / 2.82 / 3.68	3.69 / 3.21 / 4.01	0.91 / 0.93 / 0.88	0.90 / 0.94 / 0.92	0.92 / 0.87 / 0.80	0.97 / 0.96 / 0.94	0.98 / 0.97 / 0.96

noises, hammering, keyboard typing, dogs barking, etc. Speech data were obtained from the LibriSpeech corpus, while other types of noises were extracted from either DEMAND [23] or AudioSet [24].

Real-world Recordings. We have collected real world user recordings of live music shows from YouTube. Multiple different clips were collected for each covered performance. As the clips were taken by independent users, we align and normalize all these recordings before processing. Overall, we collected ~300 video recordings from 4 different music shows.

4.2. Baselines

We evaluate the proposed method against four baselines. The first one, denoted as MEAN, is constructed by taking the average of all input audio signals. The second baseline, denoted as MEDIAN, is constructed by computing the STFT of all signals and of the average signal, and replacing the magnitude of the average signal in each TF cell with the median magnitude of all input signals in that TF cell [11]. Another baseline is the FASTFCA [18]. In the time-frequency domain, each source contribution is modeled as a zero-mean Gaussian random variable whose covariance represents the source’s spatial properties.

We used the implementation described in [18]. The last baseline is the Maximum Component Elimination [19], in which the magnitude of the average signal at each TF cell replaced by the average magnitude of all input signals in that TF cell after removing the maximal magnitude.

4.3. Model Evaluation

To assess the quality of the reconstructed audio in relation to the reference signal the Invariant Signal-to-Noise Ratio (SI-SNR) [25], PESQ [26], [27], and STOI [28] were used as an objective methods, while we use the MULTI Stimulus test with Hidden Reference and Anchor (MUSHRA) [29] test as a subjective one. We conducted a human listening test using a web platform [30], asking participants to rate the quality of recordings on a scale of 0 to 100 [31].

4.4. Results

Results for the synthetic data can be seen on Figure 1 and Table 1 considering either music or speech as the source signal with various types of noises and SNR values. In all experiments we use $k = 5$ sources. Notice, as this is a synthetic dataset, we have the perfect alignment, hence we skip the alignment process in

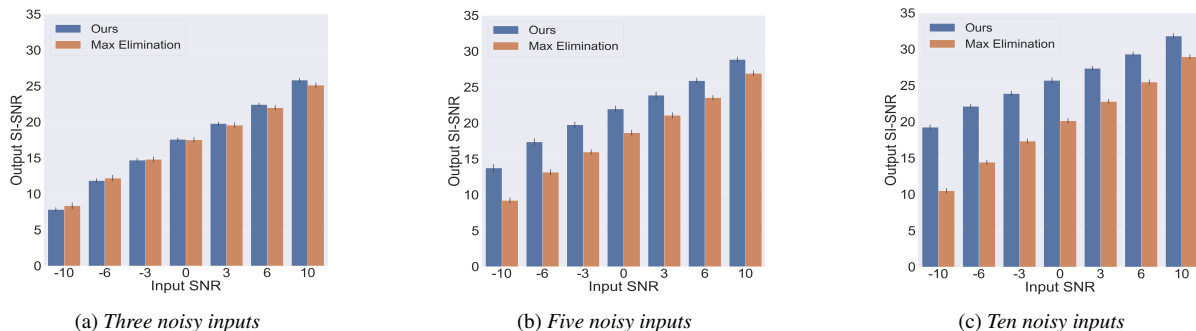


Figure 2: Average SI-SNR of enhanced signal, combining 3, 5, and 10 synthetic noisy audio signals. Source signal is music, and noise is speech. Max elimination [19], the best baseline, is compared with our Crowdsourced Enhancement. As expected, the benefit of our method over the baseline increases as more noisy signals are combined together.

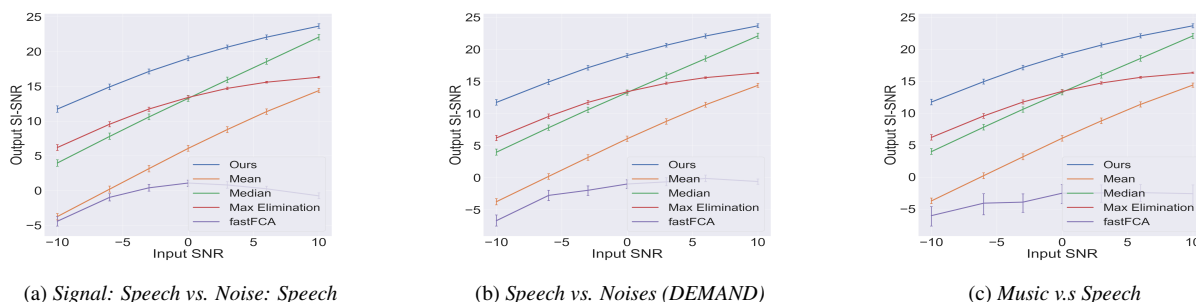


Figure 3: Same as Fig. 1, but with simulated packet loss, where each noisy input signals also has a randomly placed one second of silence. Max Elimination, the best baseline under additive noise, fails in this case.

this setting. We report the SI-SNR, STOI and PESQ metric between each of the methods against the clean target signal. Under each of the evaluated setups we extracted the enhanced signal using five synthesized noisy signals. Results suggest that the proposed method is significantly better than the evaluated baselines. This is more noticeable at low SNR values (e.g., -5, -10). Interestingly, when considering environmental noises from DEMAND, the gap between the proposed method and the evaluated baselines is smaller. In Figure 2 we compare our method to Max Elimination [19], considering different number of sources. Notice, the proposed method is superior to the Max Elimination method with an exception of three sources considering low SNR values. This implies that the proposed method can benefit from a large number of input sources.

Next, we experiment with a packet loss setting, where we assume random parts of each input signals may be missing. We inject a low energy white Gaussian noise in the missing periods, to prevent numerical issues with fastFCA. To simulate that, we randomly erase one second from each input signal independently. Results are presented in Figure 3. Results suggest the proposed method is superior to the evaluated baselines under this setting as well. Interestingly, as we go to higher SNR values, the Max Elimination method converges towards the mean. This can be explained as the Max Elimination considers one element less than the mean method and for high SNR values it is often not a noisy element. Notice that the median method is not affected by the packet loss as it will ignore it anyway.

We perform subjective tests following the MUSHRA protocol [31], asking participants to rate the quality of recordings on a scale of 0 to 100. Obtained ratings: Max elimination [19],

the closest prior art, got 48.4 ± 2.9 ; our method got 67.4 ± 2.6 , (mean \pm 95% confidence interval). This suggests that the proposed method is superior to the evaluated baselines also considering subjective metrics. Code, datasets, models and audio examples are available at the following link:

https://shiranziz.github.io/crowdsourced_audio_enhancement/

4.5. Recording from a Live Performance

Finally, we evaluate the proposed approach on real recordings of live music shows collected from YouTube. As no ground truth is given when we enhance the crowdsourced recordings, the results can be examined on the website.

5. Conclusions

We presented a simple and effective method for noise removal from crowdsourced recordings. The method examines individual time-frequency cells, and removes noisy input signals whose magnitude are outliers. The method can handle additive noise by removing outliers that are higher than the median signal, and can also handle silent moments (e.g., packet loss) by removing outliers lower than the median. We believe the development of simple and competitive baselines are crucial for constructing efficient solutions for real-world tasks. Although being simple, the proposed method improves over prior work, hence can be served as a new baseline for more complicated statistical methods which will be developed by the community in future work.

6. References

- [1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [2] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1652–1664, 2016.
- [3] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.
- [4] A. Defossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.
- [5] S. Araki, N. Ono, K. Kinoshita, and M. Delcroix, "Projection back onto filtered observations for speech separation with distributed microphone array," in *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. IEEE, 2019, pp. 291–295.
- [6] S. E. Chazan, L. Wolf, E. Nachmani, and Y. Adi, "Single channel voice separation for unknown number of speakers under reverberant and noisy settings," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3730–3734.
- [7] J. Benesty, I. Cohen, and J. Chen, *Fundamentals of signal enhancement and array signal processing*. John Wiley & Sons, 2017.
- [8] E. Vincent, T. Virtanen, and S. Gannot, *Audio source separation and speech enhancement*. John Wiley & Sons, 2018.
- [9] P. Tzirakis, A. Kumar, and J. Donley, "Multi-channel speech enhancement using graph neural networks," in *ICASSP, 2021*, pp. 3415–3419.
- [10] A. Pandey, B. Xu, A. Kumar, J. Donley, P. Calamia, and D. Wang, "Multichannel speech enhancement without beamforming," in *ICASSP, 2022*, pp. 6502–6506.
- [11] M. Kim and P. Smaragdus, "Collaborative audio enhancement using probabilistic latent component sharing," in *ICASSP, 2013*, pp. 896–900.
- [12] Y. Yemini, E. Fetaya, H. Maron, and S. Gannot, "Scene-agnostic multi-microphone speech dereverberation," in *Interspeech, 2021*, pp. 1129–1133.
- [13] T. Yoshioka, X. Wang, and D. Wang, "Picknet: Real-time channel selection for ad hoc microphone arrays," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 921–925.
- [14] H. Taherian, S. E. Eskimez, T. Yoshioka, H. Wang, Z. Chen, and X. Huang, "One model to enhance them all: array geometry agnostic multi-channel personalized speech enhancement," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 271–275.
- [15] A. Jukić, J. Balam, and B. Ginsburg, "Flexible multichannel speech enhancement for noise-robust frontend," in *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2023, pp. 1–5.
- [16] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.
- [17] N. Q. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1830–1840, 2010.
- [18] N. Ito, R. Ikeshita, H. Sawada, and T. Nakatani, "A joint diagonalization based efficient approach to underdetermined blind audio source separation using the multichannel wiener filter," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1950–1965, 2021.
- [19] N. Stefanakis and A. Mouchtaris, "Maximum component elimination in mixing of user generated audio recordings," in *IEEE Int. Workshop on Multimedia Signal Processing (MMSP)*, 2017, pp. 1–6.
- [20] A. Wang, "An industrial strength audio search algorithm," in *IS-MIR*, 2003, pp. 7–13.
- [21] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: <https://doi.org/10.5281/zenodo.1117372>
- [22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *ICASSP, 2015*, pp. 5206–5210.
- [23] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," in *Int. Congress on Acoustics (ICA)*, 2013.
- [24] J. F. Gemmeke, D. P. 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 *ICASSP, 2017*, pp. 776–780.
- [25] Y. Luo and N. Mesgarani, "Tasnet: time-domain audio separation network for real-time, single-channel speech separation," in *ICASSP, 2018*, pp. 696–700.
- [26] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [27] M. Wang, C. Boeddeker, R. G. Dantas, and ananda seelan, "ludlows/python-pesq: supporting for multiprocessing features," may 2022. [Online]. Available: <https://doi.org/10.5281/zenodo.6549559>
- [28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.
- [29] R. B. ITU, "Method for the subjective assessment of intermediate quality level of coding systems," *Recommendation ITU-R BS. 1534*, 2001.
- [30] M. Schoeffler, S. Bartoschek, F.-R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra—a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, no. 1, 2018.
- [31] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.