
Learning to Separate Voices by Spatial Regions

Zhongweiyang Xu¹ Romit Roy Choudhury¹

Abstract

We consider the problem of audio voice separation for binaural applications, such as earphones and hearing aids. While today’s neural networks perform remarkably well (separating 4+ sources with 2 microphones) they assume a known or fixed maximum number of sources, K . Moreover, today’s models are trained in a supervised manner, using training data synthesized from generic sources, environments, and human head shapes.

This paper intends to relax both these constraints at the expense of a slight alteration in the problem definition. We observe that, when a received mixture contains too many sources, it is still helpful to separate them by region, i.e., isolating signal mixtures from each conical sector around the user’s head. This requires learning the fine-grained spatial properties of each region, including the signal distortions imposed by a person’s head. We propose a two-stage self-supervised framework in which overheard voices from earphones are pre-processed to extract relatively clean *personalized* signals, which are then used to train a region-wise separation model. Results show promising performance, underscoring the importance of personalization over a generic supervised approach. (audio samples available at our project website¹). We believe this result could help real-world applications in selective hearing, noise cancellation, and audio augmented reality.

1. Introduction

Audio source separation research (Luo et al., 2019; Gu et al., 2019; 2020; Jenrungrot et al., 2020) has focused extensively on separating sources from a generic microphone array (e.g.,

¹Department of Electrical and Computer Engineering, University of Illinois Urbana-Champaign, Illinois, US. Correspondence to: Zhongweiyang Xu <zx21@illinois.edu>, Romit Roy Choudhury <croy@illinois.edu>.

Proceedings of the 39th International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

¹<https://uiuc-earable-computing.github.io/binaural/>

a table-top teleconference system, robots, cars, etc.). When these microphones are on “earable” devices, such as hearing aids and earphones, new opportunities emerge. In particular, the human face/ears/head alter the arriving audio signals in sophisticated ways, ultimately helping the brain infer important spatial attributes of the signal (Blauert, 1996). Importantly, this *head-related transfer function* (HRTF) is different across users, and harnessing this personalized filter remains a rich area of exploration in various fields of science (Zhang et al., 2021; Yang & Roy Choudhury, 2021).

This paper aims to explore the potential benefits of personal HRTFs in binaural source separation, such as for hearing aids, earphones, or glasses. Our hypothesis is that the personal HRTF encodes considerable spatial diversity and this diversity can aid source separation compared to a baseline with generic HRTFs. Of course, the spatial diversity may not be enough to separate two sources arriving from nearby angles; in fact even typical humans cannot achieve more than 20° angular resolution (Yang & Roy Choudhury, 2021). However, if the diversity can separate sources by broad angular regions (e.g., isolate one mixture per region shown in Figure 1), various applications may benefit from it. A user’s hearing aid, for example, could help isolate a target voice in front of him from all other voices in other regions. Even if two voices arrive from the front, separating that two-voice mixture from all other mixtures could still be beneficial to noise cancellation, augmented reality, and other applications.

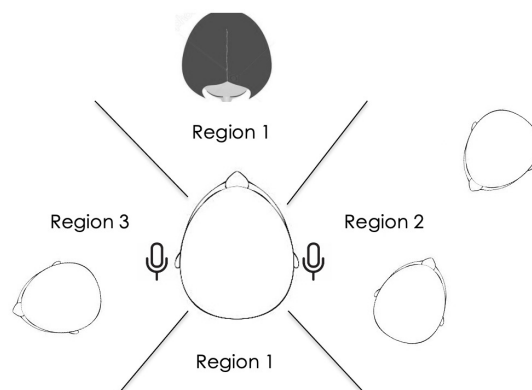


Figure 1. $K=4$ sources in $R=3$ spatial regions (or cones). The front and back cones identified as the same region due to spatial aliasing (i.e., signals arriving from the front and back cones can produce the same time difference at the ears).

Separating voices by region has a second benefit. If the voices can be separated into R mixtures (with one mixture per region), the separation algorithm does not need to know the number of sources, K , in the recorded signal. This offers an important relaxation from past work (Luo & Mesgarani, 2019; Luo et al., 2020; Tan et al., 2021; Subakan et al., 2021). Specifically, although recent deep learning models have performed remarkably well in source separation (separating $K=4+$ sources from 2 microphone recordings), they have assumed a known K or at least a fixed maximum K (Nachmani et al., 2020). Region-based separation obviates the need to pre-specify K since it would always output at most R mixtures. Of course, the tradeoff is that the output signals may not be isolated voices, rather they would be mixtures of voices if more than one are located in the same region. This paper makes this compromise to gain from HRTF-based personalization and immunity from K .

The difficulty in harnessing the personalized HRTF lies in extracting clean binaural signals for any given individual. Clearly, a user Alice has abundant opportunities to record ambient signals on her own earphones, but any such binaural recording would likely be polluted and this polluted data would need separation. Thus, recent work (Han et al., 2020; Tan et al., 2021) have used generic data (i.e., anechoic source signals filtered with HRTF databases) and used this clean dataset to train a general binaural separation model (a supervised approach).

To gain from HRTF personalization, one potential approach would be to require the user to record sounds in a quiet room, from many different directions. This would help train a personalized supervised model, although at the expense of significant user effort.

Our idea is to utilize Alice’s own recordings (from everyday scenarios) and *opportunistically* extract out relatively clean signals (whenever possible). We propose a pre-processing module that uses spectral and spatial techniques to identify *when reliable separation is viable*; otherwise, the signal segment is discarded. Our algorithm relies on a Gaussian mixture model (GMM) to identify when a signal is relatively clean (i.e., high SINR), versus situations where signal-clusters have merged deceptively to appear as one signal. Moreover, given earphone microphones are separated by a relatively large distance (diameter of human head), the algorithm must also cope with heavy spatial aliasing at higher frequencies.

The output of our pre-processing module is expected to yield relatively clean sources that embed the user’s personalized HRTF. Using these sources, we synthesize region-wise voice mixtures, and then train a neural network-based separation model. Since the reference signals are synthesized from Alice’s own recorded signals, *the training is self-supervised*. We show that even though the reference

signals are not perfectly clean, the region-wise separation model can still learn to separate source mixtures effectively. Results from our self-trained model outperforms supervised models trained with generic HRTFs by $2+dB$. Our model is not data-hungry and can achieve voice separation without requiring any knowledge of K .

Our contributions are: (1) recognizing the combined gain from personalization, self-supervision, and relaxed assumptions on K in exchange for region-wise source separation. (2) a proposed pre-processing module and a network architecture that realize this gain, and (3) extensive comparisons that show how spatial cues (embedded in personalized HRTFs) can play a crucial role in source separation.

2. Formulation and Baseline

2.1. Problem Formulation

Consider two microphones at the human ears that hear multiple ambient voices from all around the head. Assume space is partitioned into R regions, and the signals from each region i form a mixture y_i . These per-region mixtures can be modeled at the left and right microphones as:

$$y_i^l = \sum_{j=1}^{N_i} h_{ij}^l * s_{ij} \quad y_i^r = \sum_{j=1}^{N_i} h_{ij}^r * s_{ij} \quad (1)$$

Here $s_{ij} \in \mathbb{R}^{1 \times T}$ is the j^{th} signal in the i^{th} region; h_{ij}^l and h_{ij}^r are the corresponding head-related impulse response (HRIR) for the direction from which source s_{ij} arrives. The $*$ denotes the convolution operation and N_i denotes the number of sources in the i^{th} region. Then, the recorded mixture at each microphone would be a summation over all region-based mixtures as follows:

$$m^l = \sum_{i=1}^R y_i^l \quad m^r = \sum_{i=1}^R y_i^r \quad (2)$$

Here m^l and m^r are the left and right microphone recordings. The goal of region-based separation is to estimate y_i^l and y_i^r from m^l and m^r , for all $i \in [1, R]$.

2.2. Supervised Separation as Baseline

For a supervised approach, the binaural mixtures m_l and m_r are fed into a separation model f_θ with parameters θ . The model predicts R binaural sounds which corresponds to R regions:

$$\hat{y}_1^l, \hat{y}_1^r, \hat{y}_2^l, \hat{y}_2^r, \dots, \hat{y}_R^l, \hat{y}_R^r = f_\theta(m_l, m_r)$$

To optimize the model parameters θ , the loss function of our system contains two parts: *active loss* and *inactive loss*. The active loss is for the regions that have voices — we want

each region’s output to contain all the sources inside that region. The inactive loss is for the regions those contain no active voices — we want these regions to output an empty source. Thus we adopt the loss function in (Wisdom et al., 2020a) but without permutation invariant training (PIT). Specifically, assume reference region mixtures we want to learn are $y_i^l, y_i^r, i \in [1, R]$, ordered so that first M reference region mixtures are active. Then, the loss function is:

$$\begin{aligned} Loss = & \sum_{i=1}^M [L_{\text{SNR}}(y_i^l, \hat{y}_i^l) + L_{\text{SNR}}(y_i^r, \hat{y}_i^r)] \\ & + \sum_{i=M+1}^R [L_{\text{inactive}}(m^l, \hat{y}_i^l) + L_{\text{inactive}}(m^r, \hat{y}_i^r)] \end{aligned}$$

The L_{SNR} loss in the equation below is the negative SNR loss with a constant $\tau=10^{-\text{SNR}_{\text{max}}/10}$, $\text{SNR}_{\text{max}}=30\text{dB}$.

$$L_{\text{SNR}}(y, \hat{y}) = 10\log_{10}(\|y - \hat{y}\|^2 + \tau\|y\|^2)$$

The constant τ is to assign a 30dB maximum SNR to prevent the network from optimizing for one single source. The detailed reasoning is explained clearly in (Wisdom et al., 2020b;a).

The L_{inactive} is to enforce the network to output empty source for inactive regions. Thus the inactive loss is:

$$L_{\text{inactive}}(x, \hat{y}) = 10\log_{10}(\|\hat{y}\|^2 + \tau\|x\|^2)$$

For supervised training, we generate training data by convolving voice sources with human head-related impulse responses (HRIR)². We consider two cases: First, we assume we know the person’s HRTF; we use this filter to create the dataset and train the separation model. Assume the separation performance is P_{personal} for this case. Second, assume we don’t know the person’s HRTF and we train the separation model using a generic HRTF database. Say the separation performance for this person is P_{general} . Obviously, $P_{\text{personal}} > P_{\text{general}}$ because the first case is learning the person’s personalized HRTF. However, P_{personal} is not achievable since the personal HRTF of a given person is not known in practice. Our goal is to opportunistically learn the personalized HRTF at a region-wise granularity — we expect that these personal spatial cues, even though self-supervised (hence imperfect), will help outperform P_{general} and take us close to P_{personal} .

3. Two Stage Model

The first stage aims to accept binaural recordings from a user’s ear-device and output relatively clean voice sources along with their directions of arrival (DoA). The output

²HRIR is the time domain representation of an HRTF. The HRIR varies as a function of the signal’s direction of arrival.

sources should not be contaminated too much so that they preserve the personal HRTF (naturally embedded in the signals). The challenge lies in identifying and eliminating deceptive mixtures that appear as single or two sources, or when two sources appear separable but have corrupt spatial cues.

In stage 2, we use these sources and their DoAs to synthesize larger mixtures of many sources *per region*. This region-wise mixture-dataset is then used to train our separation model. The output sources from stage 1 serve as reference signals for our loss function, thereby self-training the model. We elaborate on the two stages next.

3.1. Stage 1: Reliable Source Extraction

We intend to spatially cluster the two-microphone recordings, but such techniques are not without limitations. Today, state-of-the-art spatial clustering exploits the inter-microphone time difference (ITD) and inter-microphone level difference (ILD) as the key spatial features for clustering (Yilmaz & Rickard, 2004; Mandel et al., 2007; Mandel & Ellis, 2007; Weiss et al., 2008; Mandel et al., 2009). Briefly, if $M^l(t, f)$ and $M^r(t, f)$ denote the STFT of m^l and m^r , we can calculate the inter-microphone phase difference $\Delta\phi(t, f)$ for each time–frequency (t – f) bin:

$$\Delta\phi(t, f) = \angle \frac{M^l(t, f)}{M^r(t, f)} \quad (3)$$

Assuming the two microphones are sufficiently close to avoid spatial aliasing, the ITD for each t – f bin can be directly estimated from $\Delta\phi(t, f)$ as $\text{ITD}(t, f) = \frac{\Delta\phi(t, f)}{2\pi f}$. Figure 2 visualizes this where $\Delta\phi(t, f)$ is computed from two low frequency t – f bins, and mapped to the ITD axis.

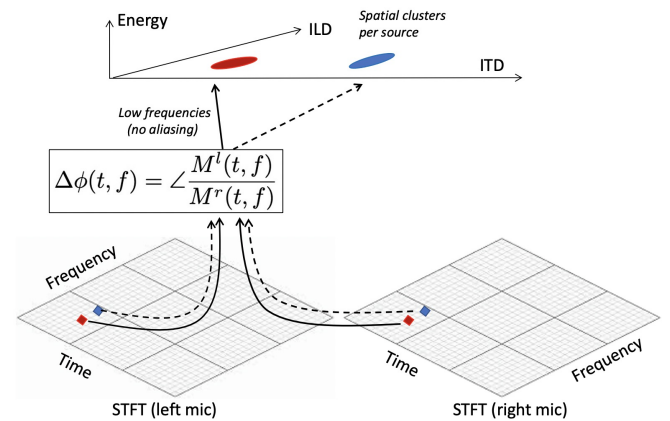


Figure 2. From STFT to clustering on the ITD–ILD space.

Similarly, the amplitude-level difference ILD can be computed as:

$$\text{ILD}(t, f) = 20\log_{10} \frac{|M^l(t, f)|}{|M^r(t, f)|} \quad (4)$$

Assuming the ILD only varies with direction (and not across frequency), it is possible to spatially cluster on the ITD+ILD dimensions. Each cluster maps back to the t - f masks on the STFT, ultimately achieving decent spatial separation of the two (red and blue) sources.

Unfortunately, limitations emerge even if all the above assumptions hold: (1) Clustering assumes each t - f bin could only contain one single source – (Yilmaz & Rickard, 2004) calls this assumption W-Disjoint Orthogonality (W-DO). This approximately holds with $K = 2$ or 3 voice sources; with more sources, signals “collide” in t - f bins. (2) The second problem is that if two sources arrive from nearby angles, their spatial features blend into a single cluster, making separation impossible.

In non-ideal cases, i.e., when the assumptions do not hold, additional issues emerge. (3) Since human heads are relatively large in comparison to the wavelength at higher voice frequencies, spatial aliasing becomes a problem. If $\Delta\tau_{\max}$ is the maximum inter-microphone time difference (between the two ears), the minimum frequency for possible aliasing is: $f_{\text{aliasing}} = \frac{1}{2 \times \Delta\tau_{\max}}$. For average human head shapes, $f_{\text{aliasing}} \approx 1000\text{Hz}$, implying that more than 87% of the frequencies get aliased (Figure 3 shows the multiple ambiguous ITDs due to spatial aliasing from the high frequency t - f bins). (4) Finally, the human HRTF is frequency selective, hence the ILD also varies per-frequency; modeling this variation is hard since it is unique to each individual. To mitigate these 4 problems, we design a *selective spatial clustering* algorithm, discussed next.

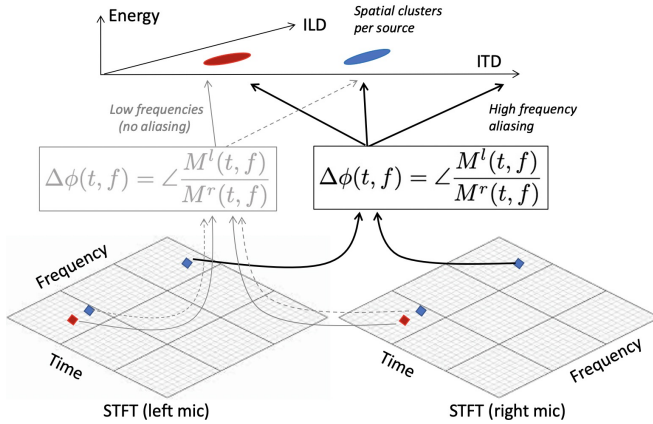


Figure 3. Higher frequencies produce spatial aliasing (i.e., due to smaller wavelengths, the measured phase difference can translate to many possible time differences or ITDs). The example blue signal in the high-frequency t - f bin shows 3 possible ITDs, which affects the separation of red and blue source signals.

Selective Spatial Clustering

Algorithm 1 presents the pseudo code; we explain the key steps below.

Step 1: We conservatively estimate f_{aliasing} (based on maximum possible human head size) and use the unaliased frequency bins to estimate ITD. We cluster on ITD and look for 1 peak or 2 adequately separated peaks.

Step 2: A single peak indicates either a single source, or multiple (angularly) nearby sources that have merged (in ITD) to become a single peak. We fit a Gaussian on this peak and accept the peak if the estimated variance is less than a threshold σ_{th}^2 .

We show that the standard deviation (STD) of the ITD distribution is a robust indicator of whether voice directions are angularly nearby or separated. Figure 4 shows a clear STD gap between a single source and 2-source mixtures when the 2 sources are 20° apart. This guides our choice of σ_{th}^2 .

The time-frequency mask corresponding to this ITD peak gives us one source or one mixture, from a specific region. We add this source to our database.

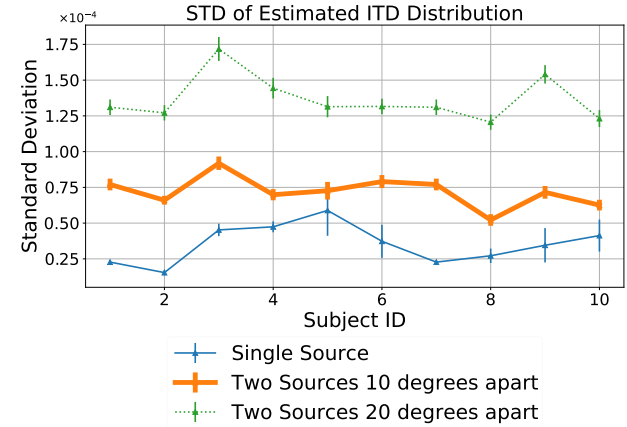


Figure 4. Standard deviation of estimated ITDs for single clean source and 2-source mixtures with 10 and 20 degree apart. The wide gap between 10 and 20 degrees allows for reliable separation.

Step 3: Similarly, 2 peaks indicate 2 sources or 2 mixtures, but we need them to be sufficiently separated in ITD to gain confidence that they have not mutually contaminated each other. For this, we fit the ITDs to a 2-component *gaussian mixture model* (GMM) and check if the variances are less than σ_{th}^2 , and their means differ more than $\Delta\tau_{min}$. If the fitted Gaussians satisfy none of these (conservative) conditions, we deem the sound segment unsuitable for separation and discard it. Otherwise, we proceed to further separation.

Step 4: Given audio frequencies far exceed f_{aliasing} , we need to cluster on high frequency bins (and cope with ITD aliasing). Past work (Mandel et al., 2007; Mandel & Ellis, 2007; Weiss et al., 2008; Mandel et al., 2009) shows that inter-microphone phase differences (IPD) are highly noisy at these frequencies, but ILDs are helpful due to greater frequency-sensitivity. Motivated by this, we aim to estimate ILDs for the two spatial sources.

Algorithm 1 Selective Spatial Separation

Input:

Real world binaural recording m^l, m^r
 Aliasing frequency threshold $f_{aliasing}$
 Variance threshold for peak detection σ_{th}^2
 Dual source ITD difference threshold $\Delta\tau_{min}$
 Time Domain Dominating Factor α

Step 1: Use unaliasing t-f bins to find ITD distribution

$$M^l(t, f) \leftarrow STFT(m^l) \quad M^r(t, f) \leftarrow STFT(m^r)$$

$$\Delta\phi(t, f \in (0 : f_{aliasing})) \leftarrow \angle \frac{M^l(t, f \in (0 : f_{aliasing}))}{M^r(t, f \in (0 : f_{aliasing}))}$$

$$ITDs(t, f \in (0 : f_{aliasing})) \leftarrow \frac{\Delta\phi(t, f \in (0 : f_{aliasing}))}{2\pi f}$$
Step 2: Output the binaural recording and the estimated ITD when the ITDs show one single obvious peak

$$\mu^*, \sigma^* \leftarrow \arg \max \mathcal{N}(ITDs; \mu, \sigma^2)$$
if $\sigma^* < \sigma_{th}$ **then**

 Return $m^l, m^r, ITD = \mu^*$
end if
Step 3: Discard the recordings when the ITDs does not show two obvious peaks, or the two peaks are too close

Use 2-component Gaussian Mixture Model to fit ITDs

 Let $\mu_1^*, \mu_2^*, \sigma_1^*, \sigma_2^*$ be the optimized means and standard deviations of the two components

if $\sigma_1^* > \sigma_{th}$ or $\sigma_2^* > \sigma_{th}$ or $|\mu_1^* - \mu_2^*| < \Delta\tau_{min}$ **then**

 Discard m^l, m^r , Return

end if
Step 4: Cluster the non-aliasing frequency bins and generate non-aliasing t-f bin masks

 Cluster the t-f bins using the GMM for $f < f_{aliasing}$, get two unaliasing t-f bin masks $mask_1^u(t, f), mask_2^u(t, f)$ for the two separable spatial sounds

Step 5: Use each source's dominating time bins to estimate the source's ILD(f) for aliasing frequencies.

$$E_1(t) \leftarrow \sum_f (mask_1^u \times (|M^l(t, f)|^2 + |M^r(t, f)|^2))$$

$$E_2(t) \leftarrow \sum_f (mask_2^u \times (|M^l(t, f)|^2 + |M^r(t, f)|^2))$$
repeat
 $T_1 \leftarrow \{t | E_1(t) > \alpha \times E_2(t)\}$
 $T_2 \leftarrow \{t | E_2(t) > \alpha \times E_1(t)\}$
 $\alpha \leftarrow 0.9 \times \alpha$
until None of T_1 and T_2 is empty

$$ILDs(t, f) \leftarrow 20 \log_{10} \frac{|M^l(t, f)|}{|M^r(t, f)|}; \quad f \geq f_{aliasing}$$

$$ILD_1(f) \leftarrow \text{mean}(\{ILDs(t, f) | t \in T_1, f \geq f_{aliasing}\})$$

$$ILD_2(f) \leftarrow \text{mean}(\{ILDs(t, f) | t \in T_2, f \geq f_{aliasing}\})$$

$$ILD_{threshold}(f) \leftarrow \frac{ILD_1(f) + ILD_2(f)}{2}$$

 Use the $ILD_{threshold}$ to get masks for aliasing frequency t-f masks $mask_1^a, mask_2^a$
Step 7: Apply both masks and return the separated spatial sounds with ITD labels

$$mask_1(t, f) \leftarrow \text{concatenate}(mask_1^u(t, f), mask_1^a(t, f))$$

$$mask_2(t, f) \leftarrow \text{concatenate}(mask_2^u(t, f), mask_2^a(t, f))$$

$$s_1^{l,r} \leftarrow iSTFT(mask_1(t, f) \times M^{l,r})$$

$$s_2^{l,r} \leftarrow iSTFT(mask_2(t, f) \times M^{l,r})$$

 Return $s_1^{l,r}, s_2^{l,r}, ITD_1 = \mu_1^*, ITD_2 = \mu_2^*$

If only one source was active at time t , per-frequency ILD estimation would be easy — we would record the ILDs for each high frequency bin. With mixtures of signals, this is problematic. However, given source signals are mostly uncorrelated, we expect to find time bins in which only one of the sources dominate. How can we tell when one source dominates?

Step 5: We compute the time–frequency masks estimated from the lower (unaliased) frequencies, compute the energy corresponding to each mask, and test if energy $E_1(t)$ exceeds $E_2(t)$ by a factor of α .

If source 1 dominates at certain time instants T_i , we compute the mean per-frequency ILD from those time instances. We perform the same for source 2. This yields the per-frequency ILD for each source.

Step 6: For all time bins where no source dominates, we compute the ILD for each high frequency bin and compare against the mean ILDs recorded in Step 5. That frequency bin is assigned to the source whose ILD matches better. At this point, every t - f bin has been assigned a mask.

Step 7: The masks are applied and after an inverse STFT, the two signals (or mixtures) are extracted. The ITDs corresponding to the signals/mixtures are recorded – this gives us the region from which the signal arrived. These separated signals/mixtures and their associated regions are entered into a region-wise source database.

3.2. Stage 2: Self-trained Region-wise Separation Model

Figure 5 shows our model pipeline – the output from stage 1 is a database of relatively clean sources (and their DoAs). Stage 2 uses these sources to synthesize region-wise mixtures and then *mixes these mixtures* to create the binaural recordings. This dataset trains our separation model.

Our neural network model for region-based voice separation is the feature concatenation TasNet, derived from (Han et al., 2020). Single-channel TasNet contains three modules: encoder, Temporal Convolutional Network (TCN), and decoder. The linear encoder is a list of kernels to transform the time domain signal to an STFT-like 2-d representation. This representation is fed into the TCN module to predict a real mask for each source. After the masks are applied to the representation, the linear decoder transforms the masked representations back to time domain.

The feature concatenation TasNet uses $\cos(\text{ITD})$, $\sin(\text{ITD})$, and ILD of all time-frequency bins as co-channel features. The co-channel features are then concatenated with the encoder output of left or right recordings (in the channel dimension) to generate a new representation for both the left and right channel's mixture. Then representations are fed into TCN to yield left/right separation masks for all sources, re-

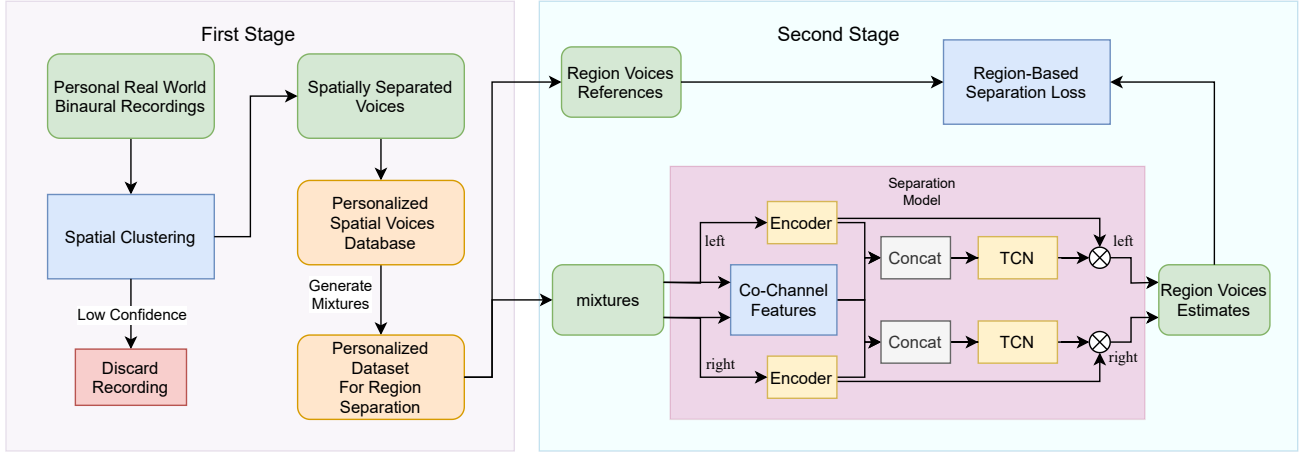


Figure 5. Two stage pipeline for self-supervised region-based separation using real-world personal recordings. The first stage collects binaural sounds and uses spatial clustering to create a personal database. The second stage uses this database for self-supervised training.

spectively. The remainder is same as single-channel TasNet, except that our architecture is designed to output sources for both left and right channels.

We expect our model to learn the spatial cues (from the personalized HRTFs) embedded in the binaural recordings. We expect that advantages from the personalized spatial information will outweigh the disadvantage of *partially-clean* reference signals, outperforming supervised training models that use generic HRTFs. Finally, our model makes no assumptions on the number of sources, K .

4. Experiments and Evaluation

To configure the feature concatenation TasNet, we set $N = 512$, $L = 32$, $B = 128$, $Sc = 128$, $P = 3$, $X = 8$, $R = 3$, following the convention in (Luo & Mesgarani, 2019). To calculate the co-channel features $\cos(\text{ITD})$, $\sin(\text{ITD})$, and ILD , we use 256-bin STFT with hop size 16 to make sure the STFT can be aligned with the encoder output. Hanning window is applied when calculating the STFT. The model is trained on 4 1080ti GPUs using the ADAM optimizer with batch size 4. The learning rate is set to be 10^{-3} .

4.1. Region-Based Supervised Training

HRTF dataset: For supervised region-based separation, we use the CIPIC HRTF database (Algazi et al., 2001). The CIPIC HRTF database contains real-world recorded Head Related Impulse Responses (HRIR) for 45 subjects, with 50 different azimuths and 25 different elevations, at roughly 5 degrees of angular resolution. For our experiments, we only use the horizontal plane with 50 different azimuth angles, divided into three regions, as shown in Figure 1. The front and back cones of region 1 add up to $90 + 90 = 180^\circ$, while regions 2 and 3 are 90° each.

Voice source and mixture dataset: We use the LibriMix dataset (Cosentino et al., 2020), sampled at 16KHz, without considering noise and reverberation. With the script used in (Dovrat et al., 2021), Libri5Mix is used for training and validation, while Libri2Mix, Libri3Mix, Libri4Mix, and Libri5Mix are used for testing.

Creating binaural mixtures: To form a binaural mixture, we first assign a voice source to a randomly chosen region, and then select a random angle θ from within that region. The voice source is then convolved with the corresponding $\text{HRIR}(\theta)$ – this forms one of the components of the mixture. To create a mixture of K sources, we randomly choose K from $[2, 5]$, and repeat the same procedure. With K HRIR-convolved sources, we sum them to generate the mixture. Observe that the HRIR is distinct for left and right ears, so we obtain a pair of mixtures – called the binaural mixture.

Generic vs. personalized training: To characterize the gap between generic and personal HRTF, we train two models: (1) The training sources are all convolved with the test subject’s personal HRIR – as discussed earlier, this gives the upperbound on performance. (2) For the generic model, the training sources are convolved with a random person’s HRIR, chosen randomly from a database of 27 people’s HRIRs. During test, the model is tested with the test subject’s HRIR.

Total users and models: For both cases, there are 3 sets of testing data, corresponding to 3 subjects. Thus, overall we have 4 supervised training models, i.e., 1 generic HRTF-based model, and 3 personalized models for the 3 testing sets (from each subject).

Basic metric: We use signal-to-noise-ratio (SNR) to assess separation quality. Observe that our model outputs binaural sounds which needs to preserve the ILD, hence the com-

monly used SI-SDR metric (Roux et al., 2018) does not apply. Thus, we compute SNR between a reference signal $x \in R^{1 \times T}$ and the estimated signal $\hat{x} \in R^{1 \times T}$ as:

$$\text{SNR}(x, \hat{x}) = 10 \log_{10} \left(\frac{\|x\|_2^2}{\|x - \hat{x}\|_2^2} \right)$$

Extending metric to region-wise mixtures: For region-based separation, the notion of a voice source gets extended to a region-wise mixture. So the reference signal in this case is the true mixture from that region, while the estimated signal is the estimated mixture from the same region. In the special case where all sources are from the same region (and the other regions have no active sources), we simply use the same SNR equation from above — we term this single-region SNR or “S-SNR”. However, when multiple regions are active, we modify the metric to SNR improvement (SNRi). Assuming that the mixture of region-wise mixtures is denoted as $m \in R^{1 \times T}$, we define SNRi as:

$$\text{SNRi}(x, \hat{x}, m) = \text{SNR}(x, \hat{x}) - \text{SNR}(x, m)$$

When 2 regions are active, we average the two SNRi and report them as 2-SNRi. Since this is a binaural estimation, we average over the left/right microphones as well. Similarly, for samples from 3 active regions, we report 3-SNRi.

Results: Table 1 reports the main performance results averaged over 3 testing subjects for LibriKMix, where K is the number of sources between 2 and 5. All results are in dB scale and the model names are summarized below:

- **GENERAL**: training with generic HRTFs.
- **PERSONALIZED**: trained with subject’s personal HRTF.
- **SELF**: self-supervised training from stage 1 outputs.
- **SEMI**: semi-supervised training with some clean sources.
- **FEW-SEMI**: semi-supervised with less than 1hr dataset.

Note that GENERAL and PERSONALIZED are supervised, while SELF, SEMI, and FEW-SEMI are self-supervised. *Evident from the table, the SNR gap between GENERAL and PERSONALIZED is significant, characterizing the room for improvement available to self-supervised and personalization-based approaches.*

4.2. Self-supervised Training

Creating the “dirty” source dataset: For fair comparison, the training data for our self-supervised model is drawn from the same audio/HRTF dataset, except that they are deliberately mixed with interfering sources (other binaural speeches) and then fed to our *Selective Spatial Filter* in stage 1. The output of stage 1, which still contains interference (hence called a “dirty” signal) is then used as training and reference signals in stage 2. Specifically, for each clean source $s^{r,l}$ in Libri5Mix, we convolve with a randomly chosen HRIR(θ) to create a binaural source. Then we mix this

Table 1. Separation SNR compared between supervised and self-supervised models for increasing number of sources. The * indicates upper-bound performance with personal, clean sources.

MODEL	K	S-SNR	2-SNRi	3-SNRi
GENERAL	2	21.0	12.0	N/A
PERSONALIZED*	2	36.5	16.7	N/A
SELF	2	31.4	13.9	N/A
SEMI	2	33.1	15.1	N/A
FEW-SEMI	2	31.0	15.0	N/A
GENERAL	3	20.9	11.2	13.1
PERSONALIZED*	3	36.5	15.3	16.8
SELF	3	32.2	12.8	14.9
SEMI	3	33.5	14.0	15.7
FEW-SEMI	3	31.1	13.8	15.5
GENERAL	4	20.5	10.9	12.7
PERSONALIZED*	4	36.3	14.6	16.0
SELF	4	33.9	12.5	14.4
SEMI	4	33.8	13.4	15.0
FEW-SEMI	4	31.0	13.2	14.8
GENERAL	5	21.4	10.2	12.2
PERSONALIZED*	5	35.8	13.7	15.2
SELF	5	34.3	11.8	13.8
SEMI	5	33.8	12.6	14.3
FEW-SEMI	5	31.0	12.4	14.1

binaural source with other sources that are also convolved with the same HRIR but a different random angle ω . Of course, the mixing is done per channel to yield binaural mixtures. When these mixtures are fed to our stage 1, the “dirty” output $\hat{s}^{l,r}$ serves as our self-training data.

Configuring stage 1 model: For stage 1’s spatial clustering model, we use 1024-point STFT with 512 overlap using Hanning window. We set $f_{aliasing} = 562Hz$, which is about the 36th bin in the FFT. We set $\alpha = 5$, $\sigma_{th} = 0.00007$ second — this value was set empirically based on our discussion on Figure 4.

Result: Figure 6 shows the performance of our spatial clustering algorithm for 5 subject’s HRTFs. The separation improves as the angular separation increases between the sources in the mixture (recall that the filter only accepts 1 or 2 sources and discards 3+ source mixtures).

Creating “dirty” mixtures: Using the dirty sources $\hat{s}^{l,r}$ we generate the dirty mixtures $\hat{y}^{l,r}, \hat{m}^{l,r}$. While supervised training trains on $y^{l,r}$, self-supervised training — using only 2-source mixtures — trains on $\hat{y}^{l,r}, \hat{m}^{l,r}$. The performance is tested on many sources (Libri2Mix, Libri3Mix, Libri4Mix, and Libri5Mix).

Results: Table 1 shows the results. Evidently, even if the training references are “dirty”, they can still guide the model to outperform the general supervised model (trained on clean sources). SELF outperforms GENERAL by 10+dB in terms of S-SNR and more than 1.5dB in terms of 2-

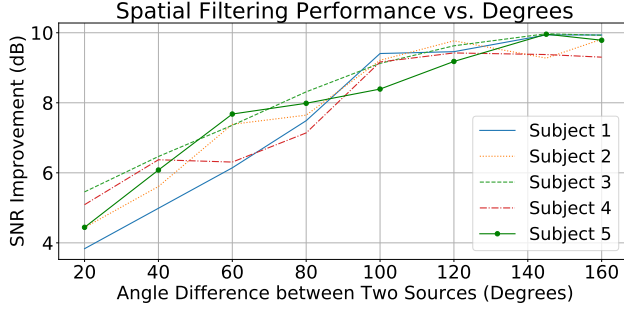


Figure 6. Stage 1’s ability to separate 2-source mixtures as a function of the angular separation between sources.

SNRi and 3-SNRi. Further, the performance improvement is higher with more sources.

4.3. Semi-supervised and Few-shot training

Semi-supervised (SEMI): Our dirty sources were all derived from mixtures. In practice, when a user wears her earphone/hearing-aid or glasses in everyday life, their earphone is likely to record single sources as well — in fact, they should be quite common. To avail this benefit, we create the training samples with 50% of clean sources and apply the same method for mixture generation. We call this the semi-supervised model (SEMI) and add to Table 1 as another point of comparison.

Few-shot (FEW-SEMI): We further consider the case when there are limited amount of real-world recordings. Given that Libri5Mix offers 56 hours of training mixtures, such a dataset may consume several days if a user must collect them in the real-world. Thus, we only use 56 minutes of personal data to fine-tune the GENERAL model, and then test if personalization can still outperform the 56-hour trained generic-HRTF model.

Results: Table 1 shows favorable results, indicating that FEW-SEMI can learn the personal HRTF even from limited data, thus preserving most of the gains over GENERAL. Further, the small gap between FEW-SEMI and SEMI suggests that region-based separation generalizes well (instead of over-fitting).

4.4. Classical Vs. Region-based Source Separation

New metric: It is difficult to compare region-based separation with classical each-source-separation using permutation invariant training (PIT). Hence we consider one special case of target speech extraction where the target speaker is alone in the front-back region and all other speakers (interferers) are in other regions. The goal is to compare the SNR of only the separated target speech.

Dataset: We train the neural network with identical settings

to perform classical source separation using PIT with SNR loss. We train on Libri4Mix, and Libri5Mix separately, to obtain 2 models. *This means the classical source separation model is assuming the number of sources is known, while the proposed region-based model does not.* We use Libri4Mix, and Libri5Mix as the test set for this experiment. To synthesize binaural sounds, we use a randomly selected HRIR from region 1 (front-back region) for the target voice, and then randomly select HRIRs from other two regions for the interfering voices.

Results: Figure 7 plots the results. Evidently, even though the classical source separation model assumes the correct number of sources, its performance for separating the target speech is substantively worse than the proposed region-based model. This result is strongly suggestive that spatial information may be far more valuable than spectral information, when it comes to separating multi-channel binaural voice recordings.

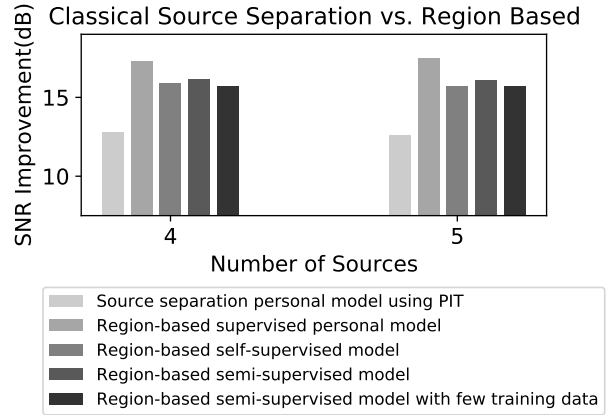


Figure 7. Classical versus region-based separation: The target speaker is in region 1 while all interference speakers are in other regions. SNR improvement reported for target speech extraction.

4.5. Points of Discussion

Why not increase the number of regions? Recall from Figure 1 that the front and back cones together form a single region (Region 1). This is because we want signals producing the same ITDs to be located in the same region. It is possible to increase the number of regions while still satisfying this property — Figure 8 shows an example with 5 regions. However, such designs are not free of tradeoffs. Specifically, at $16kHz$, the typical time difference of arrival (TDoA) is around $0.3ms$ which translates to 5 audio samples. This means 5 samples need to embed the spatial signature of any given region, a fundamentally difficult proposition even for deep neural networks. As compute power and number of microphones increase in earables, separating voices into more regions will become easier.

For our region setup, what if a source lies near the boundary between two regions? To cope with this, it is possible to

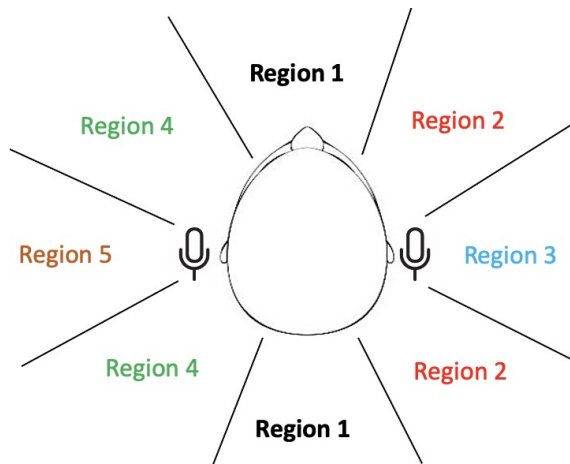


Figure 8. Example of 5 regions such that sources with same ITDs fall in the same region.

define soft region boundaries, i.e., assign the source to the neighbor region if the source is very close to the boundary ($\approx 10^\circ$). However, in our experiments, we did not need to tackle this issue because we were limited by the resolution of the HRTF database. In other words, since HRTFs are available at a granularity of 10° , we assigned nearby sources to the closest HRTF angle, which automatically made the region assignment.

5. Related Work

Single-Channel Speech Separation. Single channel source separation methods like (Luo & Mesgarani, 2019; Luo et al., 2020; Subakan et al., 2021) are able to separate speech sources successfully using only single channel input. These models are all trained with permutation invariant training (Yu et al., 2017; Kolbæk et al., 2017). Deep clustering (Hershey et al., 2015; Wang et al., 2018; Luo et al., 2018; Chen et al., 2017) is another approach to the permutation problem. Current solutions to the unknown number of sources problem are from 2 strategies. (Chazan et al., 2020; Nachmani et al., 2020) tries to solve the problem of variable sources by assuming a maximum K . (Takahashi et al., 2019) tries to solve this problem by decoding sources in a recursive manner until no sources are left.

Neural Binaural Speech Separation. Binaural recordings have also been used for neural speech separation. With multiple microphones, spatial information offers another cue for source separation. (Gu et al., 2019; 2020) tries to learn inter-channel features for multi-channel speech separation. (Han et al., 2020; Tan et al., 2021) uses parallel shared encoders for binaural speech separation, while preserving the interaural cues. (Jenrungrot et al., 2020) proposes a binary search algorithm to continue searching for active sources. This work is similar to ours in the sense that it also uses region-wise separation to solve the unknown K

problem. Since they use 4 microphones in their experiments, their model can localize sound sources with fine granularity. However, in the binaural case (e.g., earphones and hearing aids) front back confusion limits the approaches in literature.

Classical Binaural Speech Separation. Binaural speech separation without neural models is a well studied topic. These methods are essentially aims to cluster the T-F bins of the mixture based on interaural cues. DUET (Yilmaz & Rickard, 2004) clusters using 2 microphone recordings assuming no spatial aliasing. EM based methods (Mandel et al., 2007; Mandel & Ellis, 2007; Weiss et al., 2008; Mandel et al., 2009) attempt to exploit binaural cues like interaural time difference (ITD) and Interaural level difference (ILD) to cluster the T-F bins in STFT. To avoid spatial aliasing, they employ graphical models to model each bin’s ILD and IPD distributions. These methods also assume approximate W-disjoint Orthogonality (Yilmaz & Rickard, 2004) which are violated with many source mixtures. However, these methods achieve reliable performance with few sources, especially when they are not close to each other.

Self-Supervised Neural Speech Separation. (Maciejewski et al., 2018) shows that supervised speech separation model’s performance degrades when channels mismatch between training and testing data. Certain self-supervised and unsupervised models are specifically designed to mitigate this problem. (Drude et al., 2019; Tzinis et al., 2019; Seetharaman et al., 2018) uses spatial clustering to guide deep clustering. A limitation with these methods is that spatial clustering might generate clusters that contain several very close sources, which cannot guide source separation models to separate all the sources.

6. Conclusion

The importance of spatial cues in voice separation has been studied extensively. However, the gap between generic and personalized spatial cues has been relatively less explored. This paper finds that the human’s head-related filter embeds valuable spatial signatures that can be learnt at coarse granularity (i.e., region-wise). The performance gains are robust, and importantly, can be achieved in a self-supervised manner. Moreover, such region-wise voice separation also obviates the need to know the number of sources, thus relaxing an important assumption in practice. We believe the findings could aid important applications for hearing aids and earphones, such as selective hearing, noise cancellation, and audio-based augmented reality.

7. Acknowledgments

We thank NSF (award numbers: 1918531, 1910933, 1909568, and 2008338, MRI-2018966) and NIH (award number: 1R34DA050262-01) for partially funding this research.

References

- Algazi, V., Duda, R., Thompson, D., and Avendano, C. The cipic hrtf database. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No.01TH8575)*, pp. 99–102, 2001. doi: 10.1109/ASPAA.2001.969552.
- Blauert, J. *Spatial Hearing: The Psychophysics of Human Sound Localization*. The MIT Press, 10 1996. ISBN 9780262268684. doi: 10.7551/mitpress/6391.001.0001. URL <https://doi.org/10.7551/mitpress/6391.001.0001>.
- Chazan, S. E., Wolf, L., Nachmani, E., and Adi, Y. Single channel voice separation for unknown number of speakers under reverberant and noisy settings. *CoRR*, abs/2011.02329, 2020. URL <https://arxiv.org/abs/2011.02329>.
- Chen, Z., Luo, Y., and Mesgarani, N. Deep attractor network for single-microphone speaker separation. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar 2017. doi: 10.1109/icassp.2017.7952155. URL <http://dx.doi.org/10.1109/ICASSP.2017.7952155>.
- Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., and Vincent, E. Librimix: An open-source dataset for generalizable speech separation, 2020.
- Dovrat, S., Nachmani, E., and Wolf, L. Many-speakers single channel speech separation with optimal permutation training, 2021.
- Drude, L., Hasenklever, D., and Haeb-Umbach, R. Unsupervised training of a deep clustering model for multichannel blind source separation, 2019.
- Gu, R., Wu, J., Zhang, S.-X., Chen, L., Xu, Y., Yu, M., Su, D., Zou, Y., and Yu, D. End-to-end multi-channel speech separation, 2019.
- Gu, R., Zhang, S.-X., Chen, L., Xu, Y., Yu, M., Su, D., Zou, Y., and Yu, D. Enhancing end-to-end multi-channel speech separation via spatial feature learning, 2020.
- Han, C., Luo, Y., and Mesgarani, N. Real-time binaural speech separation with preserved spatial cues, 2020.
- Hershey, J. R., Chen, Z., Roux, J. L., and Watanabe, S. Deep clustering: Discriminative embeddings for segmentation and separation, 2015.
- Jenrungrot, T., Jayaram, V., Seitz, S., and Kemelmacher-Shlizerman, I. The cone of silence: Speech separation by localization, 2020.
- Kolbæk, M., Yu, D., Tan, Z.-H., and Jensen, J. Multi-talker speech separation with utterance-level permutation invariant training of deep recurrent neural networks, 2017.
- Luo, Y. and Mesgarani, N. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, Aug 2019. ISSN 2329-9304. doi: 10.1109/taslp.2019.2915167. URL <http://dx.doi.org/10.1109/TASLP.2019.2915167>.
- Luo, Y., Chen, Z., and Mesgarani, N. Speaker-independent speech separation with deep attractor network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(4):787–796, Apr 2018. ISSN 2329-9304. doi: 10.1109/taslp.2018.2795749. URL <http://dx.doi.org/10.1109/TASLP.2018.2795749>.
- Luo, Y., Ceolini, E., Han, C., Liu, S.-C., and Mesgarani, N. Fasnnet: Low-latency adaptive beamforming for multi-microphone audio processing, 2019.
- Luo, Y., Chen, Z., and Yoshioka, T. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation, 2020.
- Maciejewski, M., Sell, G., García-Perera, L. P., Watanabe, S., and Khudanpur, S. Building corpora for single-channel speech separation across multiple domains. *CoRR*, abs/1811.02641, 2018. URL <http://arxiv.org/abs/1811.02641>.
- Mandel, M. I. and Ellis, D. P. Em localization and separation using interaural level and phase cues. In *2007 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 275–278. IEEE, 2007.
- Mandel, M. I., Ellis, D. P., and Jebara, T. An em algorithm for localizing multiple sound sources in reverberant environments. In *Advances in neural information processing systems*, pp. 953–960, 2007.
- Mandel, M. I., Weiss, R. J., and Ellis, D. P. Model-based expectation-maximization source separation and localization. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(2):382–394, 2009.
- Nachmani, E., Adi, Y., and Wolf, L. Voice separation with an unknown number of multiple speakers, 2020.
- Roux, J. L., Wisdom, S., Erdogan, H., and Hershey, J. R. Sdr - half-baked or well done?, 2018.
- Seetharaman, P., Wichern, G., Roux, J. L., and Pardo, B. Bootstrapping single-channel source separation via unsupervised spatial clustering on stereo mixtures, 2018.

- Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., and Zhong, J. Attention is all you need in speech separation, 2021. *Speech and Signal Processing (ICASSP)*, pp. 451–455, 2021. doi: 10.1109/ICASSP39728.2021.9414448.
- Takahashi, N., Parthasaarathy, S., Goswami, N., and Mitsu-fuji, Y. Recursive speech separation for unknown number of speakers, 2019.
- Tan, K., Xu, B., Kumar, A., Nachmani, E., and Adi, Y. Sagnn: Self-attentive gated rnn for binaural speaker separation with interaural cue preservation. *IEEE Sig-nal Processing Letters*, 28:26–30, 2021. ISSN 1558-2361. doi: 10.1109/lsp.2020.3043977. URL <http://dx.doi.org/10.1109/LSP.2020.3043977>.
- Tzinis, E., Venkataramani, S., and Smaragdis, P. Unsuper-vised deep clustering for source separation: Direct learn-ing from mixtures using spatial information. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019. doi: 10.1109/icassp.2019.8683201. URL <http://dx.doi.org/10.1109/ICASSP.2019.8683201>.
- Wang, Z.-Q., Roux, J. L., and Hershey, J. R. Alterna-tive objective functions for deep clustering. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 686–690, 2018. doi: 10.1109/ICASSP.2018.8462507.
- Weiss, R., Mandel, M., and Ellis, D. Source separation based on binaural cues and source model constraints. pp. 419–422, 09 2008. doi: 10.21437/Interspeech.2008-51.
- Wisdom, S., Erdogan, H., Ellis, D., Serizel, R., Turpault, N., Fonseca, E., Salamon, J., Seetharaman, P., and Hershey, J. What’s all the fuss about free universal sound separation data?, 2020a.
- Wisdom, S., Tzinis, E., Erdogan, H., Weiss, R. J., Wilson, K., and Hershey, J. R. Unsupervised sound separation using mixture invariant training, 2020b.
- Yang, Z. and Roy Choudhury, R. Personalizing head related transfer functions for earables. pp. 137–150, 08 2021. doi: 10.1145/3452296.3472907.
- Yilmaz, O. and Rickard, S. Blind separation of speech mixtures via time-frequency masking. *Signal Processing, IEEE Transactions on*, 52:1830 – 1847, 08 2004. doi: 10.1109/TSP.2004.828896.
- Yu, D., Kolbæk, M., Tan, Z.-H., and Jensen, J. Permutation invariant training of deep models for speaker-independent multi-talker speech separation, 2017.
- Zhang, M., Wang, J.-H., and James, D. L. Personalized hrtf modeling using dnn-augmented bem. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics,*