Improved separation of closely-spaced speakers by exploiting auxiliary direction of arrival information within a U-Net architecture

Stijn Kindt, Alexander Bohlender, Nilesh Madhu *
IDLab, Department of Electronics and Information Systems, Ghent University
imec, Ghent, Belgium
stijn.kindt@ugent.be, alexander.bohlender@ugent.be, nilesh.madhu@ugent.be

Abstract

Microphone arrays use spatial diversity for separating concurrent audio sources. Source signals from different directions of arrival (DOAs) are captured with DOA-dependent time-delays between the microphones. These can be exploited in the short-time Fourier transform domain to yield time-frequency masks that extract a target signal while suppressing unwanted components. Using deep neural networks (DNNs) for mask estimation has drastically improved separation performance. However, separation of closely spaced sources remains difficult due to their similar inter-microphone time delays. We propose using auxiliary information on source DOAs within the DNN to improve the separation. This can be encoded by the expected phase differences between the microphones. Alternatively, the DNN can learn a suitable input representation on its own when provided with a multi-hot encoding of the DOAs. Experimental results demonstrate the benefit of this information for separating closely spaced sources.

1. Introduction

Speaker separation is the extraction of individual speech signals from a mixture of multiple overlapping talkers and additive noise. This has several use cases, such as automatic speech recognition (ASR) and transcription, telecommunication devices, and hearing aids [1]. Typically, separation is done in the short-time Fourier transform (STFT) domain. Because of the spectro-temporal sparsity and approximate disjointness of speech [21] in the STFT representation, a target speaker can be extracted from the mixture by selecting the time-frequency (TF) bins which are dominated by that speaker. This corresponds to applying a mask to the STFT representation of the microphone signal, where the mask has values close to 1 when the target source is dominant at the TF bin, and 0 when an interferer or noise is dominant. As an alternative to the direct application of the masks to extract the target signal, the masks can be integrated into a spatial filtering framework, where they can control the updates of the different components of adaptive beamformers. For example, in [12, 18] the masks guide the adaptation of the spatial statistics required for minimum variance distortionless response (MVDR) beamforming or the multichannel Wiener filter (MWF). For a good source separation, therefore, estimation of robust time-frequency masks for each source is the key.

If there is only one active speaker at one time, a single microphone may be sufficient to separate speech from background noise. The structure of speech is then used to detect the target signal in the noisy mixture. However, when multiple speakers are active simultaneously, they cannot be separated based on generic speech structure alone. Then additional information is needed about the specific speaker characteristics, such as the gender of the target speaker [8] or some latent space embedding of the speaker characteristics [10, 24]. With a compact microphone array, however, multiple overlapping speakers can be separated without the the need for prior knowledge on the speaker characteristics - as long as they are not co-located in space. The extra information for the source separation then comes from the spatial diversity: the time difference of arrival (TDOA) of the signals at the different microphones is dependent on the speaker locations. This information can be exploited to define appropriate time-frequency masks for the separation.

However, the spatial diversity is limited when the sources are closely spaced. Indeed, when the sources get closer, they generate increasingly similar TDOAs. Thus, separating such closely spaced sources becomes difficult. In this work, we investigate the possible advantages of adding auxiliary information, in the form of direction of arrival (DOA) information, to a deep neural network (DNN)-based mask estimation framework. Two different techniques of embedding this information are studied. The first approach uses hand-crafted features: the expected phase difference at the microphones corresponding to DOAs where active

*This work is supported by the Research Foundation - Flanders (FWO) under grant numbers G081420N and 11G0721N

speakers are located. The second approach lets the DNN derive a suitable representation from a multi-hot encoded vector representing active speaker DOAs.

In order to generate these features, we will assume the target locations to be perfectly known. Of course, extracting this information from the microphone signals is also challenging if the sources are closely spaced but this is outside the scope of this paper. We note, however, that additional sensors, such as a camera, can help in this regard.

In terms of DNNs, architectures based on convolutional and recurrent neural layers are efficient and have been shown to perform well for speech processing, see, e.g., [3–5, 15, 26]. Our baseline, therefore, is a straightforward multi-channel extension of a state-of-the-art convolutional recurrent U-net architecture for speech enhancement (CRUSE), originally proposed in [26] (and optimised in [4]) for single-microphone noise suppression.

2. Mask-based source separation

2.1. Signal model

We assume that a mixture of the target speech and interference speech is captured by an M-element microphone array in a reverberant and noisy room. Thus, each captured speech signal can be modelled by convolving the dry signal at the source location with the speaker-location dependent room impulse response (RIR). So we may model the mixture with J speakers at microphone m as:

$$y_m(n) = \sum_{j=1}^{J} h_{m,j}(n) * s_j(n) + v_m^{\text{add}}(n)\,, \qquad (1)$$

where $s_j(n)$ is the (dry) speech signal of source j, $h_{m,j}(n)$ is the RIR modelling the direct path ($h_{m,j}^{\text{dir}}(n)$) and reflections ($h_{m,j}^{\text{ref}}(n)$) from the location of source j to microphone m, $*$ is the convolution operator and $v_m^{\text{add}}(n)$ is the additive noise at microphone m. Using the STFT representation, $x_{m,j}(n) = h_{m,j}^{\text{dir}}(n) * s_j(n)$ and $v_m(n) = h_{m,j}^{\text{ref}}(n) * s_j(n) + v_m^{\text{add}}(n)$, we can write (1) as:

$$Y_m(l,k) = \sum_{j=1}^{J} X_{m,j}(l,k) + V_m(l,k)\,, \qquad (2)$$

where k is the frequency index and l the frame index of the STFT. We assume that the speakers do not move during utterances, which is indeed a valid assumption in many situations *e.g.* people sitting around a table for a meeting, or at the bar. By stacking the individual microphone signals into a column vector, a more compact representation is obtained as:

$$\mathbf{Y}(l,k) = \sum_{j=1}^{J} \mathbf{X}_j(l,k) + \mathbf{V}(l,k)\,, \qquad (3)$$

where $\mathbf{X}_j(l,k) = [X_{1,j}(l,k), \cdots, X_{M,j}(l,k)]^T$ and $\mathbf{V}(l,k) = [V_1(l,k), \cdots, V_M(l,k)]^T$.

2.2. Separation by time-frequency masks

The well-known properties of sparsity and disjointness of speech signals in their STFT representation [21] imply that each TF bin is typically dominated by one source. Thus, by identifying and preserving the TF bins dominated by a target speaker j and suppressing the remaining TF bins, an estimate $\widehat{X}_j(l,k)$ of the target speaker signal can be obtained. In effect, this corresponds to generating a speaker specific mask $\mathcal{M}_j(l,k)$, which has values close to 1 for TF bins (l,k) dominated by $X_j(l,k)$ and values close to 0 otherwise, and applying it to the STFT spectrum of the chosen reference microphone as:

$$\widehat{X}_j(l,k) = \mathcal{M}_j(l,k) Y_{\text{ref}}(l,k)\,. \qquad (4)$$

The separation mask can be defined in a wide variety of ways (see, e.g. [28]). Here, without loss of generality, we choose the (bounded) spectral magnitude mask (SMM):

$$\mathcal{M}_j(l,k) = \min\left(\left(\frac{|X_{\text{ref},j}(l,k)|^2}{|Y_{\text{ref}}(l,k)|^2}\right)^{\beta}; 1\right)\,, \qquad (5)$$

where β is a parameter that controls the trade-off between speech distortion and suppression of the interference and noise. We set $\beta = 1$, which suppresses the interferer(s) and noise more aggressively compared to the typical choice of $\beta = 0.5$, at the cost of a slightly increased distortion of the target signal. Further, as it is easier to learn a bounded target, we clip the SMM at 1. Also, without loss of generality, we assume microphone 1 is chosen as the reference.

3. DNN-based mask estimation

The masks needed for the separation are typically estimated using a DNN, e. g., a fully-convolutional network such as Conv-TasNet [17], which performs an end-to-end separation, or a convolutional recurrent neural network such as [5], which operates in the STFT domain. To guarantee a high speech quality when an STFT-based masking is performed, it is particularly important that the TF mask captures the *local* structure of the target. One way to accomplish this is to process frequency subbands separately, as done in [3, 15]. A more computionally efficient solution is given by encoder-decoder architectures, where local information can be preserved by means of skip connections between encoder and decoder. In recent years, many TF mask estimation approaches of this type have been proposed, e. g., [6, 14]. In this work, we consider the optimised convolutional recurrent U-net for speech enhancement (CRUSE) structure from [4], and extend it to the multi-channel case.

3.1. Extended CRUSE for multichannel separation

The extended CRUSE architecture is depicted in Fig. 1. The convolutional layers in the first part of the U-Net (the encoder) have a kernel size of $(2,3)$ and a stride $(1,2)$ along

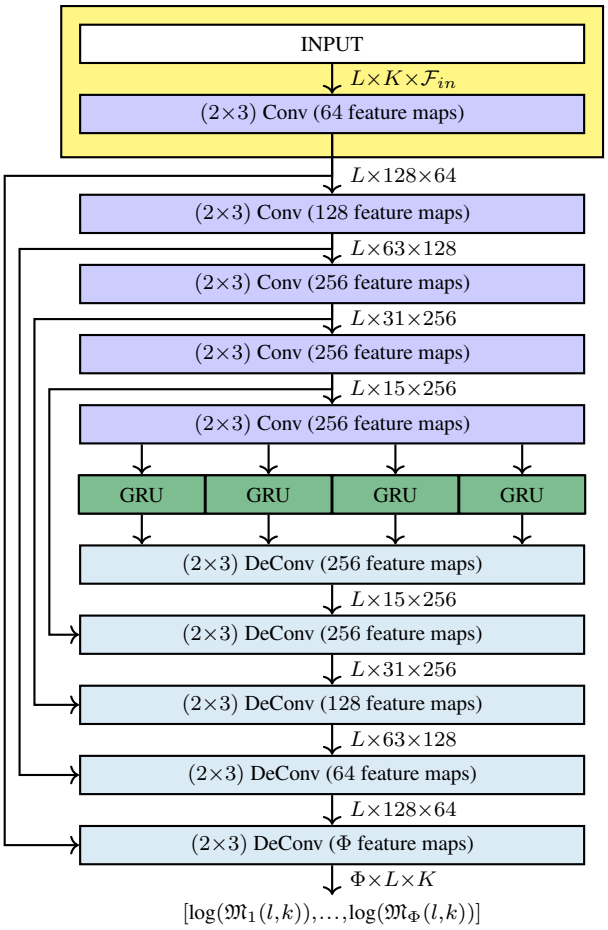


Figure 1: U-Net structure (assuming $K = 257$ discrete frequencies at the input). The bigger yellow box at the top will be defined by the selection of the input features. This choice of input features will also dictate the dimension ($\mathcal{F}_{in}$) of input features.

the time and frequency dimension respectively. Thus, each encoder layer successively reduces the frequency dimension by half, while the feature dimension increases as depicted.

This is repeated 5 times until we get a latent space representation. The feature and frequency dimensions are then flattened to form the feature dimension for the gated recurrent unit (GRU) [7]. To reduce the complexity of the model, the features are divided into 4 groups, which are processed by 4 GRU layers of smaller size in parallel [4]. The outputs of the GRUs are 'unflattened' into the frequency and feature dimensions. Deconvolutional layers in the decoder are then used to reverse the dimensionality reduction of convolutional layers in the encoder. Additive skip connections with a learnable scaling and bias are inserted between encoder and decoder [4]. These propagate information throughout the network, and make it easier for the network to learn via back-propagation.

After each convolutional and deconvolutional layer, batch normalisation is applied. All layers, except for the output layer, use the Rectified Linear Unit (ReLU) activation function.

The yellow box in Fig. 1 will change depending on the input features: it will either be the baseline inputs, discussed in Sec. 3.2, or one of the novel input features, incorporating auxiliary information, discussed in Sec. 4.

3.2. Input features

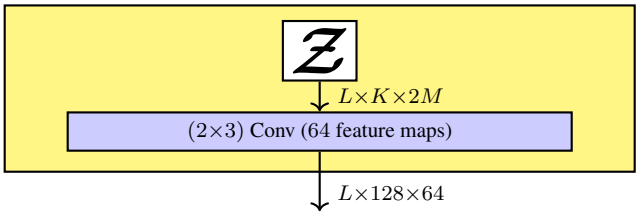


Figure 2: Baseline input features: $\boldsymbol{\mathcal{Z}}$ is a $L\times K\times2M$ tensor where the third dimension is given by the vector $\boldsymbol{\mathcal{Z}}(l,k)$ defined in (8). The elements of $\boldsymbol{\mathcal{Z}}(l,k)$ consists of the real and imaginary part of the normalised amplitudes from the microphones signals. $\mathcal{F}_{in} = 2M$.

As input features for the reference (baseline) model, we straightforwardly change the single channel inputs from [4] to integrate the spatial diversity available through the microphone array. We take the real and imaginary parts of the normalised amplitude $\frac{Y_m(l,k)}{||\mathbf{Y}(l,k)||_2}$ at each microphone [30]:

$$\mathcal{Z}_m^R(l,k) = \Re\left\{\frac{Y_m(l,k)}{||\mathbf{Y}(l,k)||_2}\right\} \tag{6}$$

and

$$\mathcal{Z}_m^I(l,k) = \Im\left\{\frac{Y_m(l,k)}{||\mathbf{Y}(l,k)||_2}\right\} \tag{7}$$

where $\|.\|_2$ is the ℓ_2 norm of a vector.

We use the following short hand notation in Fig. 2:

$$\boldsymbol{\mathcal{Z}}(l,k) = [\mathcal{Z}_1^R(l,k), \mathcal{Z}_1^I(l,k), \cdots, \mathcal{Z}_M^R(l,k), \mathcal{Z}_M^I(l,k)] \tag{8}$$

where $\boldsymbol{\mathcal{Z}}(l,k)$ is a $2M$ vector, to form the third dimension of the $L \times K \times 2M$ tensor $\boldsymbol{\mathcal{Z}}$.

Since the spatial information is essentially present in the phase, the chosen representation encodes this information well. However, compared to using the phase ($\angle Y_m(l,k)$) directly, the above representation is advantageous as it avoids the 2π phase wrapping problem.

While there is also some spatial information, like room reverberation, contained in the amplitude, normalising the amplitude across the microphones delivers (in our experience) a better generalisation to scale, speakers and also to different signal types. With this set of features, we obtain an input dimension of $\mathcal{F}_{in} = 2M$ for each TF bin.

3.3. Network output

For the output, we adopt the approach of [3]. The potential target locations are divided into Φ different angular sections, each corresponding to one DOA class. For each section ϕ, the network generates a mask $\mathfrak{M}_\phi(l,k)$ that can be used to extract a speaker from that direction. The correct mask for any speaker is then selected based on their (known or estimated) location: $\mathcal{M}_j(l,k) = \mathfrak{M}_\phi(l,k)$ if source j is located in angular section ϕ at time frame l (later written as $\mathbb{1}(\phi_j(l) = \phi)$). This mask is then applied as in (4).

The advantage of the chosen output representation, where different outputs correspond to different directions, is the implicit resolution of permutation. Thus, additional measures to resolve the permutation problem, *e.g.* permutation invariant training (PIT) [29], are not needed.

The outputs of the DNN are set to estimate the log-masks $\log(\mathcal{M}_j(l,k))$. In this manner the dynamic range of the mask values is better utilised and a more accurate estimation of lower values is obtained. However, the log-masks have no lower bound, which is undesirable for a training target. Thus, a mingain g_{min} is imposed to limit the suppression. This is achieved by setting the output activation function to be a clipped linear function between g_{min} and 1. Another benefit of the mingain is that it can also reduce artifacts such as musical tones.

During training, we consequently minimise the mean squared error (MSE) loss between the estimated log-mask $\log \widehat{\mathcal{M}}_j(l,k)$ and the desired log-mask $\log \mathcal{M}_j(l,k)$ over all active sources, as in [3]:

$$\mathcal{L} = \sum_{l,k,j} \left(\log \widehat{\mathcal{M}}_j(l,k) - \log \mathcal{M}_j(l,k) \right)^2 . \qquad (9)$$

Masks for directions without active speakers are treated as *don't cares* and do not contribute to the loss function.

4. Incorporation of auxiliary DOA information

We will show in the evaluation in Sec. 5 that this baseline system is good in separating multiple sources in general. In contrast however, the separation of closely spaces sources leaves some room for improvement. To improve upon these situations, we propose to add extra DOA information to the network. With this extra information, the network should be able to extract more useful information from the very first network layer. Additionally, there should be less confusion between closely spaced sources, since the network already knows that these sources exist and are in close proximity.

We present the two different options: the first one is the use of hand crafted features, while the second one uses a multi-hot encoding of the DOA, allowing the network to learn its own representation.

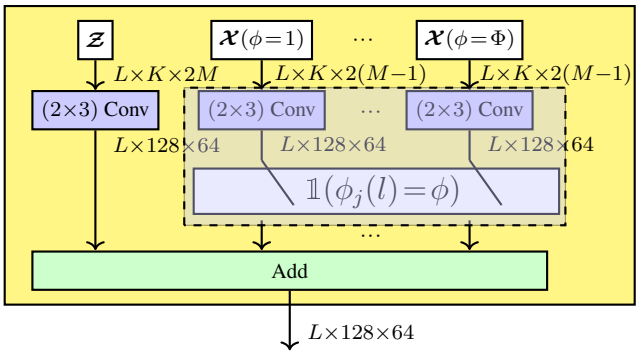


Figure 3: Baseline features with auxiliary *expected phase difference* features $\boldsymbol{\mathcal{X}}(\phi)$, as defined in (13). Additionally, the weights of the first convolutional layer are dependent on the target DOA: $\mathbb{1}(\phi_j(l) = \phi)$, indicated in the dotted box. $\mathcal{F}_{in} = 2M + 2J(M-1)$.

4.1. Expected phase differences

A representation of the DOAs that permits their inclusion within the input to the neural network is given by the corresponding *expected phase difference* between the microphones of the array, as used in *e.g.* [19]. To avoid unneeded redundancy, we only take the expected phase difference between the reference (first) and mth microphone:

$$2\pi f_k \Delta\tau_m(\phi) = 2\pi f_k \tau_m(\phi) - 2\pi f_k \tau_1(\phi) , \qquad (10)$$

where f_k is the central frequency at the frequency bin k and $\Delta\tau_m(\phi)$ is the time delay at microphone m, of a plane wave originating from the direction corresponding to DOA index ϕ, measured with respect to the array reference. Mathematically, $\Delta\tau_m(\phi) = [r_m^x, r_m^y][cos(\phi), sin(\phi)]^T/c$, with r_m^x and r_m^y the x- and y-coordinate of the mth channel with respect to the microphone reference microphone, and c the speed of sound.

To match the real and imaginary inputs at (6) and (7), we take the cosine and sine of the phases at (10) respectively:

$$\mathcal{X}_m^C(\phi,k) = \cos\left(2\pi f_k \Delta\tau_m(\phi)\right) , \qquad (11)$$

$$\mathcal{X}_m^S(\phi,k) = \sin\left(2\pi f_k \Delta\tau_m(\phi)\right) . \qquad (12)$$

We make Φ tensors $\boldsymbol{\mathcal{X}}(\phi)$ of size $L \times K \times 2(M-1)$, where the third dimension is given by the $2(M-1)$ vector:

$$\boldsymbol{\mathcal{X}}(\phi,k) = [\mathcal{X}_2^C(\phi,k), \mathcal{X}_2^S(\phi,k), \cdots, \mathcal{X}_M^C(\phi,k), \mathcal{X}_M^S(\phi,k)] . \qquad (13)$$

The same elements are repeated over the frame dimension L so the input size corresponds to $\boldsymbol{\mathcal{Z}}$. Note: as the expected phase difference for the reference microphone is always 0, it conveys no additional information and is thus not included.

This manner of including auxiliary DOA information is depicted in Fig. 3. In order to give the network maximum flexibility, the weights of the first convolutional layer

(for each $\boldsymbol{\mathcal{X}}(\phi)$) is dependent on the DOA and the corresponding expected phase differences. This means that we have Φ different sets of weights. Further, only the inputs corresponding to an (estimated) active target speaker at time frame l are passed though: $\mathbb{1}(\phi_j(l) = \phi)$. Others are multiplied with zeros as to have no influence on the mask estimation. When all J speakers are active, this yields an additional $2J(M-1)$ features per TF bin: $\mathcal{F}_{in} = 2M + 2J(M-1)$.

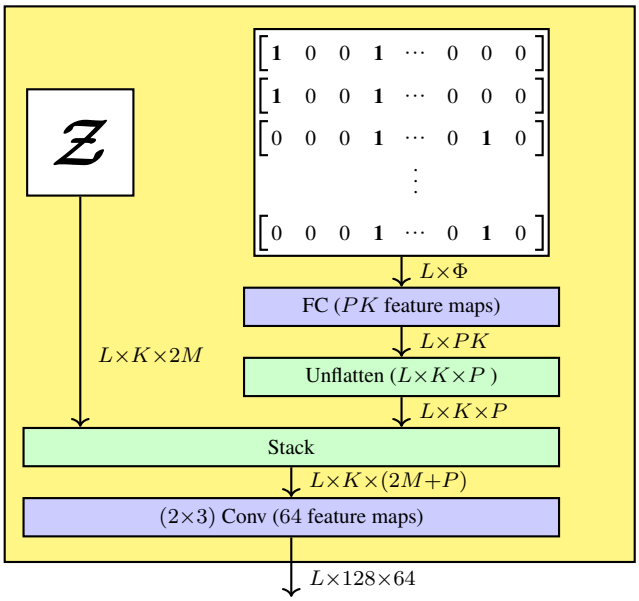


Figure 4: Baseline features with auxiliary features obtained from *multi-hot encoding inputs*. $\mathcal{F}_{in} = 2M + P$.

4.2. Multi-hot encoding

Alternatively, we can let the network determine a suitable representation on its own via a multi-hot input vector: we supply a Φ sized input vector for each time frame which indicates in what angular sector a speaker is active. A 1 is assigned to the ϕth element when a source is active at the location with index ϕ. This is then used as input for a fully connected (FC) layer with PK output features. The same FC layer is reused for all time frames. The encoding thus has no temporal context. We concatenate the newly generated features with the input of the baseline method by setting the convolutional layer to have an output size of PK. Here P is the number of additional input features for each TF bin. This is depicted in Fig. 4

We also tried to increase the representation power to possibly better exploit information about which combinations of sources are concurrently active. This was done by adding additional layers between the multi-hot input and the stacking operation in Fig. 4. However, empirically, it was found to not improve the method.

The multi-hot generated features contribute an additional P features, thus $\mathcal{F}_{in} = 2M + P$. We choose $P = 2JM$, which yields a similar number of features as for the expected phase differences.

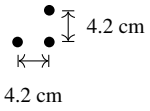


Figure 5: The 3-element microphone array used for the experiments.

5. Experiments

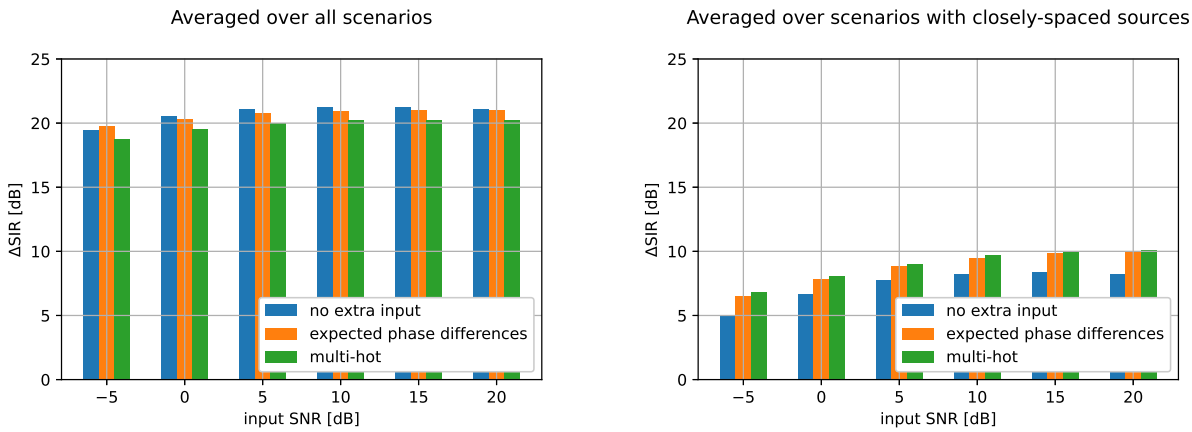
For the experiments, we use a planar array of $M = 3$ microphones. These are placed in an isoceles right-angled triangle configuration, where the lengths of the catheti are $4.2\,\mathrm{cm}$, as depicted in Fig. 5. This array geometry is used in both training and evaluation.

The sampling frequency is $16\,\mathrm{kHz}$. An STFT frame length of 512 samples (with 50% overlap between frames) is chosen, resulting in $K = 257$ frequency bins (positive frequency spectrum). The g_{min} is set to $-40\,\mathrm{dB}$.

5.1. Training

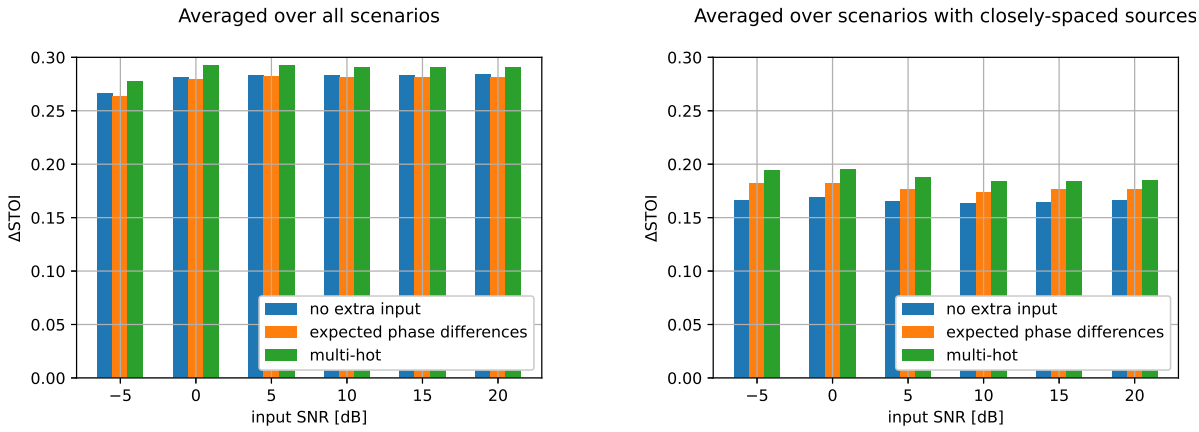
For training, we used the TIMIT [9] and PTDB-TUG speech datasets [20]. During the training, different scenarios are simulated where either one or two sources are concurrently active, similar to [2]. Thus, we set maximum number of active sources J to 2. The location of each speaker is constant until the source becomes inactive (silent). When the source becomes active again, a new location is randomly assigned to the source. Source activity and inactivity are modelled as two states of a Markov chain, where a transition between the two states occurs once every $1.5\,\mathrm{s}$ on average. See [2] for more details on the training setup.

The source signals are convolved with RIRs simulated using [11]. There are 10 different rooms with reverberation times ranging from $RT_{60} = 0.2\,\mathrm{s}$ to $0.8\,\mathrm{s}$. Further, for simplicity, we consider that the speakers are present only in the $180°$ angular space around the front of the array. We divide this region into angular sections of $5°$ width, resulting in $\Phi = 37$ different sections and, consequently, 37 different masks at the output. Different sets of RIRs are produced for training and validation. For the additive noise, we simulate temporally uncorrelated diffuse noise, as described by [13], with input SNRs ranging from $0\,\mathrm{dB}$ to $30\,\mathrm{dB}$. We stress that the network is not specifically trained to separate only closely spaced sources, since the locations are chosen arbitrarily, and there are also cases where only one speaker is active. Thereby, we can ensure that an improved separation of closely spaced sources *does not come at the cost of a reduced usefulness of the system in other scenarios*.



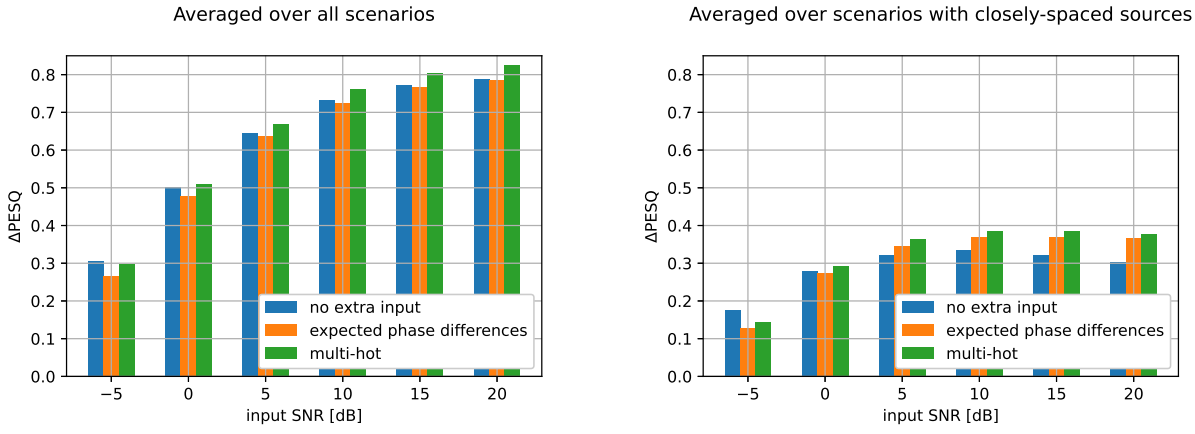
(a) (b)

Figure 6: The ΔSIR metrics (as a function of different input SNRs) for all simulated cases on the left, and for the subset where sources are separated by only 20 degrees or less on the right.



(a) (b)

Figure 7: The ΔSTOI metrics for all simulated cases (left) and the subset with only closely spaced sources (right).



(a) (b)

Figure 8: The ΔPESQ metrics for all simulated cases (left) and the subset with only closely spaced sources (right).

5.2. Evaluation

The test RIRs for the evaluation are simulated via a different generator: pyroomacoustics [23]. We generated 327 random scenarios. Each scenario has a random room dimension between $\{4, 4, 2\}$ and $\{8, 8, 4\}$ m. Since the focus of this work is on source separation, we only consider cases where two speakers are active. The locations of both speakers are fixed during one simulation. We make sure to simulate approximately one third of the cases where the sources are closely spaced ($\leq$ 20 degrees apart), in order to have a representative sample size. Each source signal consists of 5 speech utterances. The utterances are taken from the TSP speech database [16]. For each room, 6 different input SNRs are generated between $-5\,\mathrm{dB}$ and $20\,\mathrm{dB}$, where the noise is again temporally uncorrelated and spatially diffuse.

Evaluation metrics are computed for two sets of scenarios: a first set where all scenarios are included, and a second set consisting only of cases where the sources are separated by 20 degrees or less. This second set consists of 114 scenarios.

5.3. Metrics

We consider three metrics: the first is the source-to-interference ratio (SIR), as defined by [27]. This is an important metric for source separation since it indicates how much the interferer is suppressed relative to the target.

The other metrics focus on perceptual quality (PESQ: perceptual evaluation of speech quality [22]) and intelligibility (STOI: short-time objective intelligibility [25]). These metrics offer important, complementary information on the separation performance, since the SIR alone can be misleading: a decent SIR can be achieved by suppressing all of the interfering source, while only keeping a small portion of the target speech. This would however lead to unintelligible, poor quality speech.

5.4. Results and discussion

In Fig. 6, the ΔSIR metric is plotted, *i.e.*, the gain with respect to the input signal. The baseline system, without extra DOA information, yields a slightly better performance for almost every input SNR when the results for all spacings are averaged (Fig. 6a) . However, for all three variants, the ΔSIR is very high, so that the minor difference is insignificant. In general, we can conclude that the extra DOA information does not have an influence on the performance of the CRUSE architecture. The network can infer the spatial information on its own.

In contrast, the advantage of the extra DOA information is clearly visible when only closely spaced sources are considered (Fig. 6b). In these scenarios, incorporating auxiliary information yields a consistent gain of 2 to $2.5\,\mathrm{dB}$ over all input SNRs. However, the ΔSIR is, in general, less than when the sources are farther apart. This is not surprising because of the difficulty of separating sources with the considered compact 3-microphone array when their angles of arrival are similar.

Comparing the hand crafted expected phase difference features to the multi-hot encoding, the multi-hot encoding comes out on top for almost every case. This leads us to conclude that the network can learn a better representation than the expected phase differences to encode the DOA information. Either way, we would expect the multi-hot encoding to perform at least as well, since it could generate a representation equal to the expected phase differences.

The improved interferer suppression obtained by incorporating the auxiliary information is evident when listening to examples with a spacing of 20 degrees between the two speakers. Some samples can be found as Supplementary material at the AVSS site (will be moved to a website after paper acceptance).

The STOI and PESQ graphs from Fig. 7 and Fig. 8 validate these informal perceptual observations, and are largely in line with what we observed in the SIR graphs: averaging over all inter source distances does not show a *significant* benefit of the additional DOA input features (even though the PESQ and STOI metrics favour these systems), but when looking at the performance for closely spaced sources only, the benefit becomes clear.

There is one outlier: the PESQ score for closely spaced sources at $-5\,\mathrm{dB}$. Here, the original input features do seem the have an edge. However, this is not in line with our observations when listening to the examples ourselves. This is likely because PESQ is less reliable at low input SNRs.

6. Conclusions

We incorporated additional DOA information at the input of a recurrent convolutional U-net in order to improve the separation of closely spaced sources with a compact microphone array.

Two representations of DOA information were considered: expected phase differences and multi-hot encoding. For sources that are farther apart, the additional inputs did not have significant impact. This shows that, generally, the network can separate sources effectively without requiring knowledge on the exact target locations.

In situations where the sources are closely spaced, on the other hand, both proposed methods were found to improve the separation. Of the two, the multi-hot encoder slightly outperformed the handcrafted expected phase difference features, indicating that the network is able to generate a superior representation.

For this work, we assumed the DOAs to be known. Future work will investigate the influence of DOA errors, resulting from estimation of the DOAs.

References

[1] A. Bertrand. Applications and trends in wireless acoustic sensor networks: A signal processing perspective. In *2011 18th IEEE symposium on communications and vehicular technology in the Benelux (SCVT)*, pages 1–6. IEEE, 2011. 1

[2] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu. Exploiting temporal context in CNN based multisource doa estimation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1594–1608, 2021. 5

[3] A. Bohlender, A. Spriet, W. Tirry, and N. Madhu. Neural networks using full-band and subband spatial features for mask based source separation. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 346–350. IEEE, 2021. 2, 4

[4] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev. Towards efficient models for real-time deep noise suppression. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 656–660. IEEE, 2021. 2, 3

[5] S. Chakrabarty and E. A. P. Habets. Time–frequency masking based online multi-channel speech enhancement with convolutional recurrent neural networks. *IEEE Journal of Selected Topics in Signal Processing*, 13(4):787–799, 2019. 2

[6] S. E. Chazan, H. Hammer, G. Hazan, J. Goldberger, and S. Gannot. Multi-microphone speaker separation based on deep DOA estimation. In *Proc. 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019. 2

[7] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. In *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, 2014. 3

[8] D. Ditter and T. Gerkmann. Influence of speaker-specific parameters on speech separation systems. In *INTERSPEECH*, pages 4584–4588, 2019. 1

[9] J. S. Garofolo. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993*, 1993. 5

[10] M. Ge, C. Xu, L. Wang, E. S. Chng, J. Dang, and H. Li. Spex+: A complete time domain speaker extraction network. *Proc. Interspeech 2020*, pages 1406–1410, 2020. 1

[11] E. A. Habets. Room impulse response generator. *Technische Universiteit Eindhoven, Tech. Rep*, 2(2.4):1, 2006. 5

[12] E. A. Habets, J. Benesty, S. Gannot, and I. Cohen. The MVDR beamformer for speech enhancement. In *Speech processing in modern communication*, pages 225–254. Springer, 2010. 1

[13] E. A. Habets and S. Gannot. Generating sensor signals in isotropic noise fields. *The Journal of the Acoustical Society of America*, 122(6):3464–3470, 2007. 5

[14] M. M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann. Combining adaptive filtering and complex-valued deep postfiltering for acoustic echo cancellation. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 121–125, 2021. 2

[15] X. Hao, X. Su, R. Horaud, and X. Li. Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6633–6637, 2021. 2

[16] P. Kabal. TSP speech database. *McGill University, Database Version*, 1(0):09–02, 2002. 7

[17] Y. Luo and N. Mesgarani. Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019. 2

[18] N. Madhu and R. Martin. A versatile framework for speaker separation using a model-based speaker localization approach. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):1900–1912, 2011. 1

[19] P. Pertilä and J. Nikunen. Distant speech separation using predicted time–frequency masks from spatial features. *Speech communication*, 68:97–106, 2015. 4

[20] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf. A pitch tracking corpus with evaluation on multipitch tracking scenario. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011. 5

[21] S. Rickard and O. Yilmaz. On the approximate W-disjoint orthogonality of speech. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–529. IEEE, 2002. 1, 2

[22] 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 *IEEE Intl. Conf. on acoustics, speech, and signal processing.*, volume 2, pages 749–752, 2001. 7

[23] R. Scheibler, E. Bezzam, and I. Dokmanić. Pyroomacoustics: A python package for audio room simulation and array processing algorithms. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355. IEEE, 2018. 7

[24] R. Sinha, M. Tammen, C. Rollwage, and S. Doclo. Speaker-conditioned target speaker extraction based on customized lstm cells. In *Speech Communication; 14th ITG Conference*, pages 1–5. VDE, 2021. 1

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE Intl. Conf. on acoustics, speech and signal processing*, pages 4214–4217, 2010. 7

[26] K. Tan and D. Wang. A convolutional recurrent neural network for real-time speech enhancement. In *Interspeech*, volume 2018, pages 3229–3233, 2018. 2

[27] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE transactions on audio, speech, and language processing*, 14(4):1462–1469, 2006. 7

[28] D. Wang and J. Chen. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1702–1726, 2018. 2

[29] D. Yu, M. Kolbæk, Z.-H. 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)*, pages 241–245. IEEE, 2017. 4

[30] Y. Yu, W. Wang, and P. Han. Localization based stereo speech source separation using probabilistic time-frequency masking and deep neural networks. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1):1–18, 2016. 3