SPAIN-NET: SPATIALLY-INFORMED STEREOPHONIC MUSIC SOURCE SEPARATION

Darius Petermann and Minje Kim

Indiana University, Department of Intelligent Systems Engineering, Bloomington, IN, USA 47408

ABSTRACT

With the recent advancements of data driven approaches using deep neural networks, music source separation has been formulated as an instrument-specific supervised problem. While existing deep learning models implicitly absorb the spatial information conveyed by the multi-channel input signals, we argue that a more explicit and active use of spatial information could not only improve the separation process but also provide an entry-point for many user-interaction based tools. To this end, we introduce a control method based on the stereophonic location of the sources of interest, expressed as the panning angle. We present various conditioning mechanisms, including the use of raw angle and its derived feature representations, and show that spatial information helps. Our proposed approaches improve the separation performance compared to location agnostic architectures by 1.8 dB SI-SDR in our Slakh-based simulated experiments. Furthermore, the proposed methods allow for the disentanglement of same-class instruments, for example, in mixtures containing two guitar tracks. Finally, we also demonstrate that our approach is robust to incorrect source panning information, which can be incurred by our proposed user interaction.

Index Terms— music source separation, positional encoding, panning, conditioning, neural networks

1. INTRODUCTION

Musical source separation (MSS), a task consisting in isolating various musical constituents from a given music mixture, has been an active research area for decades now. The problem is challenging due to the typical *underdetermined* nature of musical signals (i.e., lesser number of channels than sources), hence it has been addressed via machine learning, e.g., spectrogram decomposition [1, 2]. Recently, deep learning and data driven approaches have advanced this field of study significantly. A typical deep learning-based MSS system can be trained in a supervised fashion by comparing the model's output to the ground-truth source signals. It is also common to employ the concept of masking in the feature space, such as ideal ratio masking (IRM) [3] on the coefficients of the short-time Fourier transform (STFT) [4], while a direct waveform estimation is also common, such as seen in Wave-U-Net or Demucs [5, 6, 7].

In this paper, we focus on the stereophonic mixtures. In music especially, stereo channel settings are a widely popular format and usually preferred over monophonic mixtures, since it conveys a larger spatial field for a more enjoyable listening experience. Discussing the professional stereophonic mixing process is out of the scope of this paper as it is artistic and complicated. It is however important to note that each music source tends to have unique stereophonic characteristics, such as a panning location in the stereophonic panorama. For example, Fig. 1 portrays what a typical panning configuration for Pop music could look like. These typical configurations can however change depending on the music genre, instrumentation, and the mixing engineer's creative freedom over the process, making supervised learning challenging.

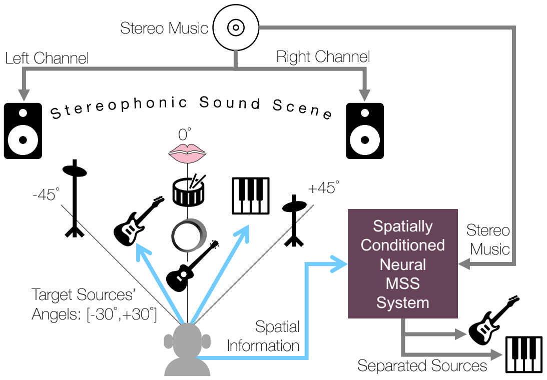

Fig. 1: Diagram of the overall proposed system. Notice that the stereo field location is addressed in degrees.

Indeed, stereophonic MSS has added another dimension to the MSS problem. There are models assuming a source-specific spatial panning position and disjoint orthogonality among sources in the time-frequency domain, such as DUET [8], ADRess [9, 10], and PROJECT [11]. While these models are strictly instrument-agnostic, it is known that the source and spatial modeling approaches can be combined together as in multichannel nonnegative matrix factorization [12] and the separation of the main-versus-accompaniment using the source-filter model [13]. Likewise, knowing of or assuming about the source locations in the stereophonic audio scene can help improve source separation.

Here, we claim that MSS could further benefit from some additional, high-level, spatial information, if it is provided more directly. In that regard, our approach can be seen as a variant of *informed* source separation [14], where aligned scores [15], the user's query [16], and even the user's scribble on the spectrograms [17] can serve as the auxiliary information. Similarly, we envision that the sources' spatial locations can be used as the auxiliary input to a machine learning-based MSS system as shown in [18] for speech separation. We postulate that spatial information would be useful when the other features, such as spectral, timbral, or temporal characteristics, are not discriminative enough, e.g., in unseen instruments or artificially synthesized sound.

To this end, we propose to *condition* a deep neural network (DNN) using the spatial information of the sources of interest, which we call spatially informed networks (SpaIn-Net). Injecting prior-knowledge into deep learning has been well investigated for MSS applications, for example, the target source's label [19, 20], a query audio signal that describes the target source [16], etc. To the best of our knowledge, the proposed model is the first attempt in the informed MSS literature to condition a DNN using spatial information of the sources.

The proposed model applies the conditioning idea to one of the state-of-the-art MSS systems, called Open-Unmix + CrossNet (XUMX) [21]. We investigate various conditioning mechanisms and show that they overall improve the MSS performance compared to the baseline unconditioned XUMX model. Note that the system also adds an interactive interface entry point, allowing for an *inaccurate* user input that still helps MSS, opening up a new direction to user-centered applications. The robustness to the noisy user input differentiates SpaIn-Net from the setup in [18].

2. METHODOLOGY

2.1. Baseline Model

Our baseline model, the XUMX architecture, was introduced as part of the *Music Demixing Challenge 2021* [22] as an extension of Open-Unmix (UMX) [23]. The XUMX model's superiority comes from its advanced loss functions. First, the *multi-domain* loss function computes the source reconstruction loss both in the frequency and time domain, for the former mean-squared loss compares the magnitudes of source and reconstruction, while the latter employs weighted signal-to-distortion ratio (wSDR) on the time-domain signals, directly. Second, the model also employs a *combination* loss that examines all partial mixtures and their reconstruction, e.g., the mixture of guitar and bass versus the mixture of the estimated guitar and bass, and so on. In this work, we opt to use of the multi-domain loss as the sole loss function. The concept of combining sources is also used within the model where the source-specific features are averaged up across the original UMX network's source-specific extraction streams. We inherit the XUMX model to construct our baseline and the proposed systems, although we opted out of the combination loss which degrades the separation performance in our same-source separation task.

2.2. Spatial embeddings

Since the conditioning process combines heterogeneous data types, which in our case consist of stereo audio signals and the sources' spatial information, it needs a careful design to benefit from both modalities. First, it is reasonable to assume that the audio signals are in high dimensional space. In our XUMX baseline, for example, the input signal goes through STFT, resulting in an F dimensional input vector at t-th time step, where F is defined by the frame size. Meanwhile, as for the spatial conditions, we opt to use the angle of the source instrument's panning location in the stereophonic sound field as illustrated in Fig. 1. For example, if the user wants to separate guitar and piano, the corresponding panning location will be -30° and $+30^\circ$, respectively. These scalars are obviously not descriptive enough when it comes to professionally engineered music, where the instruments can have ambient effects that disperse the perceived panning location of the source. However, considering the potential user interface that may benefit from its simplicity, we employ the scalar angle value to inform the MSS system.

One obvious approach to combine these two types of information is to concatenate the angle value to the spectrum, e.g., by appending each of the K angle values of K sources to each of the corresponding XUMX source-specific inference streams, forming an $F+1$ dimensional vector per inference stream.

While appending the scalar to the input vector might be a valid way, we investigate more elaborated methods to carefully examine the impact of spatial information on MSS. We observe that the main issue might be that the two dimensions are very different, e.g., $F \gg 1$. Out of various other ways to condition a neural network, such as FiLM [24], in this paper, we adopted the positional encoding method proposed in the Transformer model [25] that expands the scalar variable's dimension using sinusoids. The original positional encoding scheme converts a nonnegative integer value (e.g., the word order index within the input sentence) into a sinusoidal function, represented in a D-dimensional vector. The shape of the output positional embedding vector differs based on the scalar input for discrimination. However, the original formulation is defined only for nonnegative integers, thus necessitating a variant to cover negative numbers, i.e., source positions on the left channel.

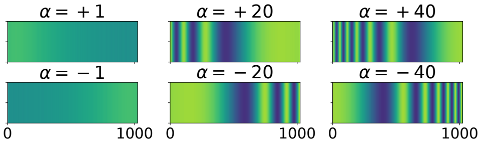

Fig. 2: Examples of positional encoding embeddings for various angular degree values. Note that the further apart from the center the values are, the more disparity is reflected in their resulting positional encoding maps.

Hence, our proposed positional encoding is designed to create the vector version of both positive and negative scalars. First, the positive side is defined similarly to the Transformer's. $\mathcal{P}$ is a function of the angle value in degree $0 \le \alpha \le +45$ and the dimension index i that varies from 0 to $D/2$, where D is the target dimension:

$$\mathcal{P}(2i,\alpha) = \sin\left(\frac{\alpha}{45^{\frac{2i}{D}}}\right), \quad \mathcal{P}(2i+1,\alpha) = \cos\left(\frac{\alpha}{45^{\frac{2i}{D}}}\right). \quad (1)$$

Here, $\mathcal{P}$ is defined by alternating sine and cosine functions. For a given input scalar α, the sinusoidal function "slows down" its frequency exponentially as the dimension i increases. The result is a sinusoidal function that gradually decreases its frequency in the higher dimension (Fig 2, the first row). α contributes to the overall frequency of this resulting sinusoidal function: the smaller α is, the more it reduces the overall frequency and vice versa. For the negative angles, we flip these sinusoids in the left-right direction, so that the ripple area (defined here as the faster changing frequency portion of the positional encoding vector) appears on the opposite side:

$$\mathcal{N}(2i,\alpha) = \sin\left(\frac{\alpha}{45^{\frac{D-2i}{D}}}\right), \mathcal{N}(2i+1,\alpha) = \cos\left(\frac{\alpha}{45^{\frac{D-2i}{D}}}\right). \quad (2)$$

Note that in Fig 2 $D = 1024$ is a dimension chosen empirically among other options (e.g., 512, 2048, etc.).

2.3. Conditioning mechanisms

We condition the XUMX baseline by combining the spatial information with the spectrum. Let $\boldsymbol{a}_t^{(k)} \in \mathbb{R}^D$ hold the positional embedding representation of the k-th target source's panning location at the given time frame t, which is the output of the function $\mathcal{P}$ or $\mathcal{N}$ depending on the sign of the angle α. In this paper we limit our discussion to the static source cases, so we drop the time index t from $\boldsymbol{a}^{(k)}$. Also, note that $D = 1$ denotes the case where we do not apply the positional encoding and just use the raw angle values, directly.

While $\boldsymbol{a}^{(k)}$ denotes the "ground-truth" panning angle, we also take the incorrect user input into account. To that end, we employ another notation, a noise-injected angle $\bar{\boldsymbol{a}}^{(k)} = \boldsymbol{a}^{(k)} + \epsilon$, where ϵ is a random deviation amount sampled from a uniform distribution defined between $[-\delta, \delta]$: $\epsilon \sim \mathcal{U}(-\delta, \delta)$. We will revisit the difference between $\boldsymbol{a}^{(k)}$ and $\bar{\boldsymbol{a}}^{(k)}$ in the experiments.

Meanwhile, the input mixture signal goes through the first feature extraction step, which is STFT in our XUMX setup. The left and right channel signals go through STFT individually, resulting in a stacked magnitude spectrogram $|\boldsymbol{X}| \in \mathbb{R}_+^{2F\times T}$, whose upper and bottom halves are the left and right channel spectrograms, respectively. Once again, based on the static source assumption, we repeatedly concatenate the spatial embedding $\boldsymbol{a}^{(k)}$ to all T spectra. Given that we can have up to K such embedding vectors, the final conditioned input vector at time t is $\left[|\boldsymbol{X}_{:,t}|^\top, \boldsymbol{a}^{(k)\top}\right]^\top \in \mathbb{R}^{2F+D}$ for the k-th XMUX inference stream, which estimates the k-th source.

Adding the two vectors is also a popular option as in the Transformer model. To that end, the system must make sure that $D = 2F$, so that the addition operation holds: $|\boldsymbol{X}_{:,t}| + \boldsymbol{a}^{(k)} \in \mathbb{R}^F$.

Finally, we also try adaptive instance normalization (AdaIN), which was originally proposed in the context of image style-transfer [26] with the aim to statistically align a given set of content feature to some target style feature. In our case, the style and feature contents denote two different modalities: the spatial information as the content feature $\boldsymbol{a}^{(k)}$ and the spectra as the style feature $|\boldsymbol{X}_{:,t}|$. AdaIN's goal is to align their mean and standard deviation as follows:

$$\text{AdaIN}(|\boldsymbol{X}_{:,t}|, \boldsymbol{a}^{(k)}) = \sigma(\boldsymbol{a}^{(k)}) \left(\frac{|\boldsymbol{X}_{:,t}| - \mu(|\boldsymbol{X}_{:,t}|)}{\sigma(|\boldsymbol{X}_{:,t}|)} \right) + \mu(\boldsymbol{a}^k), \tag{3}$$

Here, for every frame t we align the mean and variance of the positional encoding $\boldsymbol{a}^k$ with those of the input spectrogram $|\boldsymbol{X}_{:,t}|$.

3. DATASET AND EXPERIMENTAL SETUP

3.1. The dataset

Since we seek supervised MSS, access to the isolated ground-truth sources is necessary during training. In this view, we opt to work with the Slakh dataset [27], which comprises 2,100 songs and 34 instrument categories, for a total of 145 hours of audio data in mono format at a sampling rate of 44.1kHz. Slakh allows a full control of the originally monophonic sources—we freely relocate their stereophonic panning locations using constant power panning laws (See eq. (4)). Compared to other alternative choices, such as MUSDB [28] or MedleyDB [29], the use of Slakh avoids unnecessary downmixing of stereophonic original sources, which then have to be upmixed for stereo panning. The downside is that Slakh's stem tracks are originally MIDI sources rendered by virtual instruments. It is also true that our constant power panning may not represent the real-world professional mixing process. We follow Slakh's original split schemes. However, due to the four-source separation setup which omits other source categories (Sec. 3.3), the size of each of the training sets naturally reduces to 120 hours.

3.2. Mixing procedure

We use constant power panning (CPP) laws to assign each of the target sources a randomly chosen panning angle. We ensure that the sources' relative levels in the resulting stereo mix remain the same by using the CPP laws. For an amplitude of the monophonic stem signal $m(n)$ at time index n, the CPP law defines the gain of left and right channels as follows:

$$\begin{aligned} x_\text{L}(n) &= (\sqrt{2}/2)(\cos\alpha + \sin\alpha)m(n) \\ x_\text{R}(n) &= (\sqrt{2}/2)(\cos\alpha - \sin\alpha)m(n), \end{aligned} \tag{4}$$

which are then multiplied to $m(n)$ to construct the stereo channels x_L and x_R. The resulting stereo signal $\boldsymbol{x}$ should convey a perceived panning location that matches the target angle α.

3.3. The proposed experiments

In order to assess the validity of our approach and showcase that the conditioning spatial information benefits the separation task, we design multiple experimental setups.

- 4S: The first MSS task involves four distinct musical sources, namely guitar, strings, bass, and piano.
- 4S2G: A more challenging four-source separation task that contains two guitar sources (with no strings).
- D0, D1, DF, and $\texttt{DF}_\texttt{AdaIN}$: To validate the impact of different choices of spatial information dimension D, we investigate two options $D = 1$ and $D = F$. Note that D0 stands for the XUMX baseline where no spatial conditioning is used, while $\texttt{DF}_\texttt{AdaIN}$ is the $D = F$ case where AdaIN is applied.
- CAT vs. ADD: CAT indicates the combination option that concatenates $|\boldsymbol{X}_{:,t}|$ and $\boldsymbol{a}^{(k)}$. ADD, however, denotes the case when the two are added together. Once again, D0 ignores this option.
- $\bar{\alpha}_\text{Tr}$ vs. α_Tr and $\bar{\alpha}_\text{Te}$ vs. α_Te: We distinguish the two training cases depending on the type of auxiliary input, i.e., whether the angle is contaminated by the noise ($\bar{\alpha}_\text{Tr}$) or not (α_Tr). Note that when D0, this training option is turned off and disregarded, as the baseline does not use spatial information. We sample ϵ from a uniform distribution defined over a range of $[-8, +8]$. There are two types of test experiments defined similarly: $\bar{\alpha}_\text{Te}$ and α_Te. Our goal is to make sure the system works robustly even on a noisy test signal $\bar{\alpha}_\text{Te}$.

For example, $\texttt{4S2G-D1-CAT-}\alpha_\texttt{Tr}\texttt{-}\bar{\alpha}_\texttt{Te}$ indicates a model trained and tested on the two-guitar mixture using the raw source angle added to the spectra as the conditioning mechanism. Here, the auxiliary input is noisy to reflect users' incorrect estimation of the source locations during the test time. However, the model is trained on exact source locations without any noisy angle involved. Meanwhile, 4S-D0 means the XUMX baseline tested on the default four-source separation experiment with no spatial information involved (or ignored if there is any).

4. EXPERIMENTAL RESULTS AND DISCUSSIONS

To evaluate the performance of the various models involved, we consider the following well-established metrics: signal-to-distortion ratio (SDR), source-to-interference ratio (SIR), source-to-artifacts ratio (SAR), and, additionally, source image to spatial distortion ratio (ISR) to properly measure the spatial reconstruction quality in our stereophonic setup [30].

Table 1 presents the results on our first task (4S). We first observe a considerable improvement of at least 11 dB in terms of SDR coming from all of the systems, including our baseline model, over the input mixture. More importantly, we note that our first proposed model $\texttt{4S-D1-CAT-}\alpha_\texttt{Tr}$ outperforms the baseline by 1.4 dB on average. Although this scalar raw angle value is imbalanced compared to the high-dimensional spectrum vector, its efficacy signifies the importance of spatial information in MSS. Furthermore, the proposed positional encoding-based conditioning method successfully brings an additional improvement (0.2 dB) as shown in our $\texttt{4S-DF-ADD-}\alpha_\texttt{Tr}$ models, although the improvement is not too significant. Due to the space limitation, we exclude DF-CAT results that are not too different from D1-CAT, while still being worse.

The injection of noise into α *during training* does not seem to consistently improve the performance if we compare the $\bar{\alpha}_\text{Tr}$ and α_Tr

Table 1: BSS Eval improvements observed on **Task 4S** for the CrossNet baseline model and our proposed models: 4S-D1-CAT-$\alpha_{\mathtt{Tr}}$ taking the raw angle scalar and 4S-DF-ADD-$\alpha_{\mathtt{Tr}}$. Note that 4S-D1-CAT-$\bar{\alpha}_{\mathtt{Tr}}$ is trained on noisy angle.

	Models	4S-D0					4S-D1-CAT-$\alpha_{\mathtt{Tr}}$					4S-D1-CAT-$\bar{\alpha}_{\mathtt{Tr}}$					4S-DF-ADD-$\alpha_{\mathtt{Tr}}$				
	Instruments	Gtr.	Str.	Pia.	Bas.	Avg.	Gtr.	Str.	Pia.	Bas.	Avg.	Gtr.	Str.	Pia.	Bas.	Avg.	Gtr.	Str.	Pia.	Bas.	Avg.
	Mixture SDR	−12.3	−22.7	−3.5	−10.2	−12.2	−12.3	−22.7	−3.5	−10.2	−12.2	−12.3	−22.7	−3.5	−10.2	−12.2	−12.3	−22.7	−3.5	−10.2	−12.2
	Δ SDR	10.9	15.6	8.1	9.9	11.1	**12.4**	17.6	8.1	11.3	12.3	12.2	18.1	**8.5**	11.2	12.5	12.0	**18.4**	**8.5**	**11.7**	**12.7**
4S	ISR	1.5	0.2	5.6	3.6	2.7	2.7	1.5	5.5	4.1	3.4	**2.8**	1.6	**6.0**	4.1	**3.5**	2.3	**1.7**	5.9	**4.3**	**3.5**
-$\alpha_{\mathtt{Te}}$	SAR	3.1	3.5	8.6	5.6	5.2	**5.3**	4.2	9.4	**6.8**	6.4	**5.3**	**4.8**	**9.6**	6.7	**6.6**	**5.3**	4.7	9.5	**6.8**	**6.6**
	SIR	5.3	0.5	12.6	4.0	5.6	**9.0**	1.1	**14.4**	7.8	8.1	8.7	**2.6**	14.2	**8.3**	**8.4**	8.9	1.7	13.9	7.7	8.1
	Δ SDR	10.9	15.6	8.1	9.9	11.1	**12.3**	17.5	8.0	11.3	12.3	12.1	18.1	**8.5**	11.2	**12.5**	11.7	**18.2**	**8.5**	**11.7**	**12.5**
4S	ISR	1.5	0.2	5.6	3.6	2.7	**2.7**	1.5	5.4	4.1	3.4	**2.7**	1.6	**5.9**	4.0	**3.6**	2.1	**1.7**	**5.9**	**4.3**	3.5
-$\bar{\alpha}_{\mathtt{Te}}$	SAR	3.1	3.5	8.6	5.6	5.2	5.3	4.2	9.4	**6.7**	6.4	**5.4**	4.8	**9.6**	6.6	6.6	5.3	**5.2**	**9.6**	**6.7**	**6.7**
	SIR	5.3	0.5	12.6	4.0	5.6	8.8	0.8	**14.5**	7.7	8.0	8.6	**2.3**	14.0	**8.3**	**8.3**	**9.2**	1.8	14.1	7.5	8.1

Table 2: BSS Eval improvements observed on **Task 4S2G** for the CrossNet baseline model and our proposed models: 4S2G-D1-CAT-$\alpha_{\mathtt{Tr}}$ taking the raw angle scalar and 4S2G-DF-ADD-$\alpha_{\mathtt{Tr}}$. Note that 4S2G-D1-CAT-$\bar{\alpha}_{\mathtt{Tr}}$ is trained on noisy angle.

	Models	4S2G-D0					4S2G-D1-CAT-$\alpha_{\mathtt{Tr}}$					4S2G-D1-CAT-$\bar{\alpha}_{\mathtt{Tr}}$					4S2G-DF-ADD-$\alpha_{\mathtt{Tr}}$				
	Instruments	Gtr1	Gtr2	Pia.	Bas.	Avg.	Gtr1	Gtr2	Pia.	Bas.	Avg.	Gtr1	Gtr2	Pia.	Bas.	Avg.	Gtr1	Gtr2	Pia.	Bas.	Avg.
	Mixture SDR	−15.2	−16.5	−2.3	−15.0	−12.2	−15.2	−16.5	−2.3	−15.0	−12.2	−15.2	−16.5	−2.3	−15.0	−12.2	−15.2	−16.5	−2.3	−15.0	−12.2
	Δ SDR	9.5	10.5	7.5	12.7	10.0	**12.5**	13.1	7.7	14.9	**12.1**	12.1	**13.2**	**7.8**	**15.1**	**12.1**	12.3	12.9	7.7	14.0	11.7
4S2G	ISR	−1.2	−0.7	6.3	3.0	1.9	**1.4**	1.2	6.3	3.7	**3.1**	1.2	**1.3**	6.3	3.7	3.1	1.1	1.1	**6.4**	**3.8**	**3.1**
-$\alpha_{\mathtt{Te}}$	SAR	5.9	5.9	9.3	5.1	6.5	5.9	6.2	**9.9**	6.6	7.1	**6.5**	**6.6**	**9.9**	6.6	**7.4**	5.7	**6.6**	9.8	**6.7**	7.2
	SIR	−2.3	−3.2	12.9	2.1	2.4	**4.8**	**4.5**	15.8	**7.0**	**8.0**	4.4	4.0	**16.1**	**7.0**	7.9	4.0	3.2	14.0	4.9	6.5
	Δ SDR	9.5	10.5	7.5	12.7	10.0	**12.1**	12.9	7.7	14.8	**11.9**	11.7	**13.0**	**7.8**	**15.1**	**11.9**	12.0	12.7	7.7	13.7	11.5
4S2G	ISR	−1.2	−0.7	6.3	3.0	1.9	**0.9**	1.3	6.2	3.6	3.0	**0.9**	**1.4**	6.3	**3.7**	**3.0**	0.8	1.2	**6.4**	3.6	**3.0**
-$\bar{\alpha}_{\mathtt{Te}}$	SAR	5.9	5.9	9.3	5.1	6.5	5.8	6.4	9.8	6.5	7.1	**6.5**	6.5	9.9	6.5	**7.4**	5.3	**6.8**	**10.0**	**6.6**	7.2
	SIR	−2.3	−3.2	12.9	2.1	2.4	**4.8**	**3.5**	15.6	**7.1**	**7.7**	4.1	**3.5**	**16.3**	6.9	**7.7**	3.7	2.9	14.2	4.7	6.4

Table 3: SI-SDR improvements averaged over all four sources on **Task 4S** for various conditioning approaches. The clean source angles are used for the test signals (α_{Te}) and the models are trained from accurate source angles (α_{Tr}).

	D0	DF$_{\mathtt{AdaIN}}$	D16-CAT	D32-CAT	D64-CAT
Mixture SDR	−12.2	−12.2	−12.2	−12.2	−12.2
Average SDR	11.1	11.6	12.3	11.7	12.5

models' performance on the noise injected test experiments $\bar{\alpha}_{\text{Te}}$. Essentially, it means that the model trained from the accurate source location can still generalize to the test-time inaccurate conditioning. We believe this robustness comes from the fact that (a) the model performs non-spatial source separation anyway (b) the model implicitly extracts and uses the spatial information from the input stereo signal at least to some degree. Meanwhile, 4S-D1-CAT-$\bar{\alpha}_{\mathtt{Tr}}$ does not significantly deteriorate the separation performance on the test set with clean spatial information α_{Te}.

Table 2 presents the results from our more challenging second task 4S2G due to the two overlapping guitar sources that share similar spectral and timbral characteristics. This second task promotes our approach more rightfully as their potentially different spatial positions can dissociate the confusingly overlapping sources. We observe a more substantial improvement from our models over the baseline once again, especially for Gtr1 and Gtr2, of over 3 and 2.6 dB, respectively, on the test signals with accurate angles. The improvement is still substantial when the test source angles are not accurate: 2.6 and 2.4 dB. This jump in performance is also clearly reflected in the SIR scores; particularly for the two guitars where the baseline's SIR is substantially low (-2.3 and -3.2 dB) while our method showcases a clear merit (7.1 and 7.7 dB improvement). This demonstrates how an uninformed system may poorly manage to dissociate between identical instruments while ours may succeed. Once again, due to space-constraints, we opt to exclude DF-CAT, which did not perform nearly as well as DF-ADD. With that in mind, this points us to conclude that the better performance of DF-ADD does not necessarily lie in the size of D but in the conditioning approach.

In Table 3 we share additional insight over the different choices of D when they are concatenated to the spectrum as well as the use of the AdaIN option. We found that most of these choices consistently improve the baseline unconditioned model D0, while the simplest D1 option shows the best performance.

5. CONCLUSIONS

In this paper we presented SpaIn-Net, that incorporated a conditioning mechanism for musical source separation, by making use of spatial information. The network was informed of the position of each target source, which could be provided by the user during inference. We proved the benefit of our approach by leading a set of experiments involving diverse musical instrument stems drawn from the Slakh dataset and by exploring different conditioning methods. The outcome of our experiments showed a clear separation improvement and robustness toward incorrect user input on challenging stereo mixtures, both in favor of our method. In addition, we showcased a difficult mixing scenario involving multiple instruments of the same class and demonstrated that our approach improved the separation by 2.8 dB on average. While SpaIn-Net showed great promises coupled with a XMUX baseline, we point out that it presents new doors to a relatively unexplored field and that it can serve as a preliminary base for many potential user-centered applications in the future. Source codes and sound examples can be found: https://saige.sice.indiana.edu/research-projects/spain-net

6. REFERENCES

[1] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, pp. 529–540, 2008.

[2] N. Ono, K. Miyamoto, J. Le Roux, H. Kameoka, and S. Sagayama, "Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, 2008.

[3] A. Narayanan and D. L. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[4] P.-S. Huang, M. Kim, M. Hasegawa-Johnson, and P. Smaragdis, "Singing-voice separation from monaural recordings using deep recurrent neural networks," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[5] F. Lluís, J. Pons, and X. Serra, "End-to-End Music Source Separation: Is it Possible in the Waveform Domain?" in *Proc. of the Annual Conference of the International Speech Communication Association (Interspeech)*, 2019.

[6] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2018, pp. 334–340.

[7] Défossez *et al.*, "Music Source Separation in the Waveform Domain," *arXiv preprint arXiv:1911.13254*, 2019.

[8] S. Rickard, *The DUET Blind Source Separation Algorithm*. Dordrecht: Springer Netherlands, 2007, pp. 217–241.

[9] D. Barry, E. Coyle, and B. Lawlor, "Real-time sound source separation: Azimuth discrimination and resynthesis," in *Audio Engineering Society Convention 117*, 2004.

[10] S. Sofianos, A. Ariyaeeinia, and R. Polfremann, "Towards effective singing voice extraction from stereophonic recordings," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[11] D. Fitzgerald, A. Liutkus, and R. Badeau, "Projection-based demixing of spatial audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1560–1572, May 2016.

[12] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[13] J.-L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *Proc. of the European Signal Processing Conference (EUSIPCO)*, 2009.

[14] A. Liutkus, J.-L. Durrieu, L. Daudet, and G. Richard, "An overview of informed audio source separation," in *2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013.

[15] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1205–1215, 2011.

[16] J.-H. Lee, H.-S. Choi, and K. Lee, "Audio query-based music source separation," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[17] N. J. Bryan and G. J. Mysore, "An efficient posterior regularized latent variable model for interactive sound source separation," in *Proc. of the International Conference on Machine Learning (ICML)*, 2013.

[18] Z. Chen *et al.*, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *IEEE Spoken Language Technology Workshop (SLT)*, 2018.

[19] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, "Class-conditional embeddings for music source separation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019.

[20] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a Control Mechanism in the U-Net for Multiple Source Separations," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2019.

[21] R. Sawata, S. Uhlich, S. Takahashi, and Y. Mitsufuji, "All for one and one for all: Improving music separation by bridging networks," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[22] Y. Mitsufuji, G. Fabbro, S. Uhlich, and F.-R. Stöter, "Music demixing challenge 2021," *arXiv preprint arXiv:2108.13559*, 2021.

[23] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-Unmix - a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019.

[24] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proc. of the AAAI National Conference on Artificial Intelligence (AAAI)*, 2018.

[25] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[26] X. Huang and S. J. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[27] E. Manilow, G. Wichern, P. Seetharaman, and J. Le Roux, "Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity," in *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019.

[28] 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

[29] R. Bittner *et al.*, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research," in *Proc. of the International Society for Music Information Retrieval Conference (ISMIR)*, 2014.

[30] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, "First stereo audio source separation evaluation campaign: Data, algorithms and results," in *Proc. of the International Conference on Independent Component Analysis and Signal Separation (ICA)*, 2007.