Heterogeneous Target Speech Separation

Tzinis, Efthymios; Wichern, Gordon; Subramanian, Aswin Shanmugam; Smaragdis, Paris; Le Roux, Jonathan

Abstract

Heterogeneous Target Speech Separation Efthymios Tzinis1,2, Gordon Wichern1, Aswin Subramanian1, Paris Smaragdis2, Jonathan Le Roux1 1Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA 2University of Illinois at Urbana-Champaign, Urbana, IL, USA etzinis2,paris@illinois.edu,wichern,subramanian,leroux @merl.com Abstract We introduce a new paradigm for single-channel target source separation where the sources of interest can be distinguished using non-mutually exclusive concepts (e.g., loudness, gen- der, language, spatial location, etc). Our proposed heteroge- neous separation framework can seamlessly leverage datasets with large distribution shifts and learn cross-domain representations under a variety of concepts used as conditioning. Our ex- periments show that training separation models with heteroge- neous conditions facilitates the generalization to new concepts with unseen out-of-domain data while also performing substan- tially higher than single-domain specialist models. Notably, such training leads to more robust learning of new harder source separation discriminative concepts and can yield improvements over permutation invariant training with oracle source selection. We analyze the intrinsic behavior of source separation train- ing with heterogeneous metadata and propose ways to alleviate emerging problems with challenging separation conditions. We release the collection of preparation recipes for all datasets used to further promote research towards this challenging task.

Interspeech 2022

Heterogeneous Target Speech Separation

Efthymios Tzinis[1,2], *Gordon Wichern*[1], *Aswin Subramanian*[1], *Paris Smaragdis*[2], *Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]University of Illinois at Urbana-Champaign, Urbana, IL, USA

{etzinis2,paris}@illinois.edu,{wichern,subramanian,leroux}@merl.com

Abstract

We introduce a new paradigm for single-channel target source separation where the sources of interest can be distinguished using non-mutually exclusive concepts (e.g., loudness, gender, language, spatial location, etc). Our proposed heterogeneous separation framework can seamlessly leverage datasets with large distribution shifts and learn cross-domain representations under a variety of concepts used as conditioning. Our experiments show that training separation models with heterogeneous conditions facilitates the generalization to new concepts with unseen out-of-domain data while also performing substantially higher than single-domain specialist models. Notably, such training leads to more robust learning of new harder source separation discriminative concepts and can yield improvements over permutation invariant training with oracle source selection. We analyze the intrinsic behavior of source separation training with heterogeneous metadata and propose ways to alleviate emerging problems with challenging separation conditions. We release the collection of preparation recipes for all datasets used to further promote research towards this challenging task.

Index Terms: target speech separation, semi-supervised learning, heterogeneous conditions, conditional inference

1. Introduction

As exemplified in the cocktail party problem [1], humans have the uncanny ability to focus on a source of interest within a complex acoustic scene, and may change the target of their focus depending on the situation, relying on attention mechanisms that modulate the cortical responses to auditory stimuli [2,3]. While the field of source separation has made great strides towards reproducing such abilities in machines, particularly with the advent of deep learning approaches, there is still a gap in terms of the flexibility with which the target source can be determined.

Early works developed "specialist" models intended to isolate only a particular type of sound, such as for speech enhancement [4–7] or instrument demixing [8], where the target was determined by the training scheme and could not be changed at test time. Later works such as deep clustering and permutation invariant training (PIT) [9–11] focused on separating all sources in a mixture without an a priori differentiating factor. However, this lack of explicit bias can lead to instability [12,13], and reintroducing some bias, for example via fixed assignments based on energy or speaker ID [14], may help stabilize training.

Conditioned models, in which the target source of a system is determined based on some semantic information given as input, either via a sound class (e.g., speaker ID, instrument type) or an exemplar (e.g., reference utterance by a target speaker),
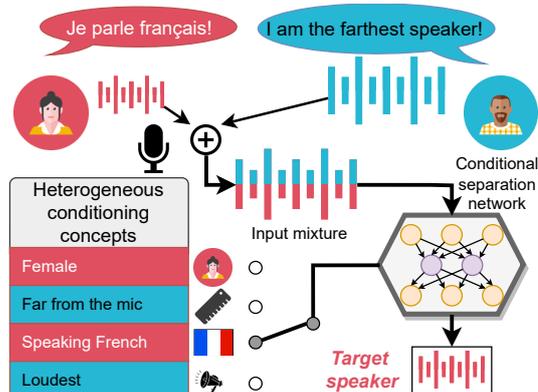
Figure 1: *Illustration of the heterogeneous target speech separation task. Notice that speakers can be separated using any of the semantic concepts and speaker attributes on the left.*

can also benefit from the good training properties of an explicit bias while allowing some flexibility in choosing the target source at test time. Such methods can provide significant gains over PIT for speech [15–19], music [20–22], and universal sound source separation [23–25], but conditioned models have so far each only considered a single type of condition looking at extracting distinct sources based on mutually-exclusive criteria.

Arguably, the problem of sound source separation is not as well-posed as the way it has been considered in the past. There are many ways to "slice" an acoustic scene depending on the application or the user's intention, where the grouping of sounds considered as target may vary. A first attempt in this direction was to extract target sound sources using multiple abstraction levels of some predefined hierarchical concepts [26]. Here, we explore another direction, by proposing to make the conditioned models mimic humans' flexibility when selecting which source to attend to, by focusing on extracting sounds based on semantic concepts and criteria of different nature, i.e., heterogeneous, such as whether a speaker is near or far from the microphone, being soft or loud, or speaks in a certain language (illustrated in Figure 1). The main contributions of this paper are:

- We introduce a novel heterogeneous target source separation task and publicly release the associated datasets.

- We propose a simple neural network architecture which can effectively separate target speech sources based on several non-mutually exclusive signal characteristic conditions, often outperforming PIT-based models with oracle assignment.

- We make several experimental discoveries: 1) heterogeneous conditioning can help cross-domain generalization, 2) robustness to non-discriminative concepts can be achieved with a small amount of such examples without impacting the overall conditional performance, and 3) adding extra "easy" conditions can lead to better learning on more difficult conditions.
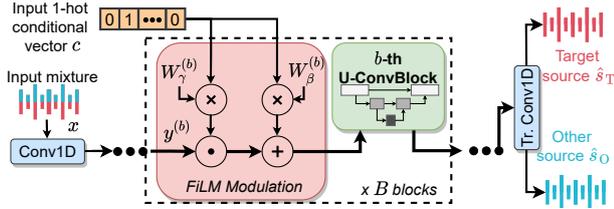
Figure 2: *Illustration of the conditional Sudo rm -rf audio separation architecture. We show the FiLM modulation which is applied to the input $\mathbf{y}^{(b)}$ of the b-th U-ConvBlock in the stack.*

2. Heterogeneous Target Source Separation

2.1. Task formulation

We consider a mixture $\mathbf{x} = \sum_{j=1}^{N} \mathbf{s}_j \in \mathbb{R}^T$ of N source waveforms $\mathbf{s}_1, \ldots, \mathbf{s}_N$, with T time-domain samples. In general, we assume that there exists a signal characteristic condition $\mathcal{C}$ (e.g., the spatial location of a source) in a set $\mathscr{C}$ of conditions, and a desired concept value v for that condition (e.g., *far* or *near*) which belongs to the set $\mathscr{V}$ of all discriminative concepts. Now, given the condition $\mathcal{C}$ and its concept value v, we would like to retrieve from the input mixture $\mathbf{x}$ the *target* submix $\mathbf{s}_T$ of all sources whose condition $\mathcal{C}$ matches the concept value v:

$$\mathbf{s}_T = \sum_{j=1}^{N} \delta(\mathcal{C}(\mathbf{s}_j) = v)\mathbf{s}_j, \quad (1)$$

where δ is an indicator function, and we use the same notation $\mathcal{C}$ to denote a signal characteristic and the function $\mathcal{C} : \mathbb{R}^T \to \mathscr{V}$ which returns the value of that characteristic for an input signal.

In this work, as an illustration, we focus on the case of speech sources, and consider signal characteristics $\mathcal{C}$ in the set $\mathscr{C} = \{\mathcal{E}, \mathcal{G}, \mathcal{S}, \mathcal{L}\}$, where $\mathcal{E}$ denotes the signal-energy (with values low/high), $\mathcal{G}$ the gender (female/male as self-identified by the dataset's speakers), $\mathcal{S}$ the spatial location (near/far), and $\mathcal{L}$ the language (English/French/German/Spanish). Thus, we can specify a target based on a total of $|\mathscr{V}| = 2 + 2 + 2 + 4 = 10$ concepts. We can encode the semantic discriminative information for the desired concept v in a one-hot vector $\mathbf{c} = \mathbb{1}[v] \in \{0, 1\}^{|\mathscr{V}|}$ which has one only at the corresponding index of the concept v, given some arbitrary ordering of $\mathscr{V}$. The goal of the task is then to train a separation model f, parameterized by $\boldsymbol{\theta}$, which takes as input a mixture of sources $\mathbf{x}$ alongside a conditioning vector $\mathbf{c}$ and estimates the target submix $\widehat{\mathbf{s}}_T$ as follows:

$$\widehat{\mathbf{s}}_T = f(\mathbf{x}, \mathbf{c}; \boldsymbol{\theta}). \quad (2)$$

2.2. Proposed framework

We construct a new conditional separation model based on the efficient Sudo rm -rf default model configuration [27], which obtains high fidelity audio reconstruction for source separation problems with minimal memory and computational requirements. Specifically, we add a FiLM [28] modulation layer at the input of each of the B U-ConvBlocks, as shown in Fig. 2. The extra parameters for the scaling and the bias are B pairs of matrices $(\mathbf{W}_\beta, \mathbf{W}_\gamma)$ with size $|\mathscr{V}| \times C_{\text{in}}$, where $C_{\text{in}} = 512$ is the number of intermediate channels in each processing block. Notably our proposed method, could theoretically scale up to an infinite amount of concepts since there is a dedicated 512-dimensional parameterization for each one-hot representation. We set the network f to produce estimates $\widehat{\mathbf{s}}_T$ and $\widehat{\mathbf{s}}_O$ for $\mathbf{s}_T$ and the submix $\mathbf{s}_O$ of *other* (non-target) sources, enforcing $\widehat{\mathbf{s}}_T + \widehat{\mathbf{s}}_O = \mathbf{x} = \mathbf{s}_T + \mathbf{s}_O$ via a mixture consistency layer [29].

We assume that we have access to a set of domains/datasets $\mathscr{D} = \bigcup_j \mathcal{D}_j$, from where we sample clean sources and their

semantic concepts to train our model by minimizing the l_1 loss:

$$L_{\boldsymbol{\theta}} = |\widehat{\mathbf{s}}_T - \mathbf{s}_T| + |\widehat{\mathbf{s}}_O - \mathbf{s}_O|, \quad \widehat{\mathbf{s}}_T, \widehat{\mathbf{s}}_O = f(\mathbf{x}, \mathbf{c}; \boldsymbol{\theta}),$$
$$\mathbf{x} = \mathbf{s}_1 + \mathbf{s}_2, \quad \mathbf{s}_i \sim \mathcal{D}, \quad \mathcal{D} \in \mathscr{D}, \quad \mathbf{c} \sim P(v), \quad v \in \mathscr{V}, \quad (3)$$

where the target submix $\mathbf{s}_T$ is synthetically constructed as shown in Eq. 1 given the sampled conditioning vector $\mathbf{c}$ and the sources $\mathbf{s}_i$. Notice that the sampling procedures for $\mathbf{c}$ and $\mathbf{s}_i$ are independent and thus, the same mixture $\mathbf{x}$ can heterogeneously disentangled in multiple ways. We argue that controlling the sampling prior of a certain concept $P(v)$ is key to robust learning as well as out-of-domain generalization.

3. Experimental Framework

3.1. Data collections

To show how heterogeneous conditional training behaves in various cross-domain training schemes with large distribution shifts (anechoic vs reverberant) as well as incorporate multiple conditions, we synthesize new conditioning datasets using Pyroomacoustics [30]. Generated mixtures are based on single-speaker utterances from the following three data collections.
Wall Street Journal (WSJ): Anechoic English utterances from the WSJ0 corpus [31] following the same clean audio extraction and partitioning recipe as in the widely used WSJ0-2mix [9].
Spatial LibriSpeech (SLIB): Synthetically reverberant versions of {*train-360*|*dev*|*test*}-*clean* LibriSpeech [32] partitions.
Spatial LibriSpeech (SVOX): We use clean speech recordings available from the multi-lingual Voxforge [33] data collection. We use all available files under "English" (53%), "French" (15%), "German" (16%), and "Spanish" (16%). Since there is no specified partitioning, we randomly split the speakers at a ratio of $8 : 1 : 1$, for train, validation, and test, respectively.

3.2. Mixtures and conditions generation configurations

All mixture datasets are synthetically generated on-the-fly (see the uniform sampling range for each parameter in Table 1), with fixed random seeds for the reproducibility of each partition. The exact data preparation recipes are available online.

For each training epoch, we generate 20,000 new on-the-fly mixtures consisting of two speakers, while for validation and testing we have a fixed-seed generation process of 3,000 mixtures per separation concept $v \in \mathscr{V}$. For each mixture, we sample a domain $\mathcal{D} \in \mathscr{D}$ and a valid separation concept for the chosen domain $\mathcal{D}$, according to a specified prior $v \sim P(v)$. Then, we sample from $\mathcal{D}$ two sources which can be separated using the concept value v, and mix them with a uniformly sampled overlap between $[75, 100]\%$. For cross-domain training we always assume equal prior among the data collections and create only single-domain mixtures. In the case of in- or cross-domain

Table 1: *Data collection metadata.*

Metadata	WSJ	SLIB	SVOX
Conditions $\mathscr{C}$	$\{\mathcal{E}, \mathcal{G}\}$	$\{\mathcal{E}, \mathcal{G}, \mathcal{S}\}$	$\{\mathcal{E}, \mathcal{L}, \mathcal{S}\}$
Room height (m)	-	$\mathcal{U}[2.6, 3.5]$	$\mathcal{U}[2.75, 3.25]$
Room length (m)	-	$\mathcal{U}[9.0, 11.0]$	$\mathcal{U}[8.0, 10.0]$
Room width (m)	-	$\mathcal{U}[9.0, 11.0]$	$\mathcal{U}[8.0, 10.0]$
RT 60 (sec)	-	$\mathcal{U}[0.3, 0.6]$	$\mathcal{U}[0.4, 0.6]$
Microphone location	-	Center	Center
Source height (m)	-	$\mathcal{U}[1.5, 2.0]$	$\mathcal{U}[1.6, 1.9]$
Far field distance (m)	-	$\mathcal{U}[1.7, 3.0]$	$\mathcal{U}[1.5, 2.5]$
Near field distance (m)	-	$\mathcal{U}[0.2, 0.6]$	$\mathcal{U}[0.3, 0.5]$
Number of test recordings	1,770	2,620	11,083
Number of test speakers	18	40	294
Number of train recordings	8,769	132,553	124,937
Number of train speakers	101	1172	2347
Number of val recordings	3,557	2,703	10,244
Number of val speakers	101	40	279

training with the WSJ and SVOX datasets, we create mixtures with an input SNR in $\mathcal{U}[0, 5]$ and $\mathcal{U}[0, 2.5]$, respectively. The SLIB dataset is always used in cross-domain training with either WSJ or SVOX and follows the corresponding input SNR distribution. In order to magnify the domain mismatch for the WSJ-SLIB cross-domain training pair, we assume that SLIB mixtures always contain one near-field and one far-field speaker. On the contrary, when performing cross-domain training with SVOX, we assume that the prior of sampling near/near-field, far/far-field, and near/far-field utterance combinations is equal (except when conditioning on the spatial location $\mathcal{S}$, which needs the sources to be spatially separable, i.e., $\mathcal{S}(s_1) \neq \mathcal{S}(s_2)$). Consequently, we found that SVOX, which is also male-dominated, can become particularly challenging for the language conditioning, where all spatial location combinations are present.

Our method allows us to also create degenerate conditions where the constituent sources map to the same concept value for the sampled condition $\mathcal{C}(s_1) = \mathcal{C}(s_2)$, thus the target submix would become either $\mathbf{s}_T = \mathbf{x}$ or $\mathbf{s}_T = \mathbf{0}$. Except in Section 4.3 where we provide an in-depth analysis of how the prior of these degenerate examples alters the operating point of our model, we always assume that during training the sampled sources can be separated using the concept v. All source audio files have a 4-second length and are downsampled to 8 kHz.

3.3. Experimental details

Model: We use a conditional Sudo rm -rf model (see Section 2.2) with $B = 16$ processing blocks. For efficient processing, we set the encoder and decoder filter lengths to 41 taps with a hop size of 20 time-samples [27]. The number of learnable encoder/decoder bases and the number of intermediate channels are set to 512. For all other parameters, we use the default values from the unconditional Sudo rm -rf model [27]. This setup leads to only a slight increase in trainable parameters compared to the unconditional version, from $9.66 \rightarrow 9.84$ millions.

Training: The conditional models minimize the l_1 loss defined in Eq. (3) while the unconditional models are optimizing the conventional PIT [9–11] l_1 loss after computing the optimal assignment of the estimates to the target sources. We train all of the models using a batch size of 6 using the Adam [34] optimizer with an initial learning rate of 10^{-3}, halving it every 20 epochs. We empirically found that 120 epochs is enough training time for both PIT and conditional models to converge.

Evaluation: We evaluate the source reconstruction fidelity of all the models at 120 epochs using the median scale-invariant signal to distortion ratio (SI-SDR) [35] between the estimate $\widehat{\mathbf{s}}_T$ and the ground-truth target $\mathbf{s}_T$, as shown below:

$$\text{SI-SDR}(\widehat{\mathbf{s}}_T, \mathbf{s}_T) = -20 \log_{10}\left(\|\alpha \mathbf{s}_T\| / \|\alpha \mathbf{s}_T - \widehat{\mathbf{s}}_T\|\right), \quad (4)$$

where $\alpha = \mathbf{s}_T^\top \widehat{\mathbf{s}}_T / \|\widehat{\mathbf{s}}\|^2$ is a scalar which makes the metric invariant to scaling. The oracle PIT models are evaluated with SI-SDR but using the best possible assignment of the estimated sources to the target waveform. For the cases where the target submix is zero, $\mathbf{s}_T = \mathbf{0}$, thus, $\mathbf{s}_O = \mathbf{x}$, we report the non-trivial SI-SDR$(\widehat{\mathbf{s}}_O, \mathbf{s}_O)$ which equivalently measures the reconstruction of the mixture at the non-target slot and is a meaningful metric since we enforce mixture consistency at the output of our models, namely $\mathbf{x} = \widehat{\mathbf{s}}_T + \widehat{\mathbf{s}}_O$. We want to underline that although all of our experiments display a similar behavior under the mean aggregation, we choose to report the median because of the susceptibility of SI-SDR to outliers. This metric instability becomes even more evident for conditional models that need to perform combined source ordering and separation.

Table 2: *Median SI-SDR (dB) results of our proposed heterogeneous conditioning method on gender ($\mathcal{G}$), spatial location ($\mathcal{S}$), and language ($\mathcal{L}$) on the SVOX and SLIB dataset test partitions, for various combinations of training domains ($\mathcal{D}$) and conditions ($\mathcal{C}$). We also show PIT unconditional models with oracle target selection. * indicates specialist systems which are each trained and tested on a given pair of domain and condition.*

Training method	$|\mathcal{D}|$	$|\mathcal{C}|$	Train condition priors (%)				Test conditions			
			SLIB		SVOX		SLIB		SVOX	
			$\mathcal{G}$	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{S}$	$\mathcal{G}$	$\mathcal{S}$	$\mathcal{L}$	$\mathcal{S}$
Conditioned*	1	1	100	100	100	100	**11.4**	**11.2**	2.5	**9.1**
PIT (Oracle)*	1	1	100	100	100	100	11.0	10.7	4.6	7.5
In-domain heterogeneous	1	2	50	50			10.9	10.7	−0.5	8.6
					50	50	−0.6	6.2	3.2	6.8
PIT (Oracle)	1	2	50	50			9.5	8.9	**5.6**	6.8
					50	50	5.2	4.5	4.6	5.6
Cross-domain heterogeneous	2	2		50	25	25	−1.4	9.2	4.3	8.2
			25	25		50	9.9	9.9	−0.7	9.0
			50			50	10.1	8.9	−0.9	9.0
				50	50		−0.5	8.4	4.3	6.8
	2	3	25	25	25	25	8.9	8.7	4.4	7.8
PIT (Oracle)	2	3	25	25	25	25	8.0	7.3	5.5	6.5

4. Results and Discussion

4.1. Cross-domain multi-conditioning

Table 2 compares the proposed heterogeneous framework with a system using PIT training and oracle speaker selection on the SLIB and SVOX datasets. For the *specialist* systems (annotated with *), which are trained to focus on a single signal characteristic, the *conditioned* models outperform their PIT counterparts with oracle selection for all conditions except SVOX language, likely because of the difficulty of simultaneously learning a harder concept like language and separating under challenging recording conditions. We also note only a small drop in performance between the specialist and in-domain heterogeneous models, meaning a single model can learn to jointly perform targeted separation based on multiple signal characteristics. In addition, the most general multi-condition and multi-domain pre-trained network almost always outperforms unconditional separation with oracle source assignment (last 2 rows). Finally, heterogeneous training can aptly perform domain adaptation (see *cross-domain heterogeneous* rows), since learning spatial conditioning appears to work well for SLIB when trained using spatial labels only for the SVOX data, and vice-versa.

4.2. Learning to generalize with bridge conditioning

Table 3 further explores domain adaptation, but in more severe conditions, as WSJ is anechoic and SLIB is reverberant. The proposed heterogeneous training method achieves 7.1 dB SI-SDR for the gender conditioned mixtures on SLIB, even though no SLIB gender-conditioned mixtures were used during training, compared to 5.8 and 4.2 dB when seeing no in-domain SLIB training data at all. We ablate the benefit of the proposed approach across two dimensions. First, we investigate the importance of having a "bridge" condition (here, the signal-energy characteristic $\mathcal{E}$), which is a common signal characteristic used to train the model across domains. Removing the bridge condition leads to a performance drop in SLIB gender conditioned cases from 7.1 to 5.5 dB. Second, we evaluate the importance of the bridge condition concept to be as discriminative as possible. If the signal-energy bridge condition includes ambiguous energy mixtures (e.g., sources mixed at ≈ 0 dB which do not contain a clearly louder speaker), performance on SLIB's gender-conditioned test set drops from 7.1 to 6.2 dB.

Training method	Train condition priors (%)				Test conditions			
	WSJ		SLIB		WSJ		SLIB	
	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$	$\mathcal{G}$	$\mathcal{E}$
Proposed	25	25		50	13.3	12.4	7.1	8.8
(-) Bridge condition	50			50	14.5	7.4	5.5	9.2
(-) Exclude amb. $\mathcal{E}$ cases	25	25		50	13.0	11.8	6.2	8.4
(-) In-domain data	100				**17.3**	−2.4	5.8	−2.3
	50	50			15.2	**14.3**	4.2	3.0
PIT (Oracle)*	100	100	100	100	**17.3**	13.6	**10.9**	**10.2**
PIT (Oracle)	25	25	25	25	12.9	11.9	9.3	8.5

4.3. Learning from degenerate conditions

A heterogeneous target speech separation model allows a user to select the signal characteristic most relevant for isolating the desired speaker in a specific mixture. However, to make a truly robust system, our model should still behave as expected when the query is a degenerate, meaning that the target for a given query is a zero waveform $s_T = \mathbf{0}$ (e.g., we condition on "French" and the mixture contains only "English" speakers), or when the query matches all sources in the mixture itself. To do so, we control the prior distribution for a concept $P(v)$ to be non-discriminative, where $C(s_1) = C(s_2)$. To the best of our knowledge, shifting the operating point for controlling the robustness of separation systems has only been studied for unconditional separation models [36] and a few works have considered degenerate conditions in target speech extraction but without thoroughly investigating the influence on performance of the ratio of such conditions in the training data [37, 38].

Because separating degenerate conditions as a task by itself is trivial, i.e., the target is either silence or the mixture, such samples can easily saturate the loss function during training. Figure 3 shows the trade-off between the percentage of degenerate gender condition cases used during training with WSJ and the performance on various conditioning tasks on WSJ test sets. We evaluate the performance on cases where the concept is discriminative (top-panel) and non-discriminative (bottom panel). We see that including no degenerate causes the system to fail on same gender mixtures, but including even 1% degenerate examples during training degrades the performance for cross-gender mixtures. For this dataset, there appears to be a sweet-spot when training with 0.4% degenerate condition examples.

4.4. Robust learning of new concepts using more conditions

As shown in Table 2, SVOX conditioned on language has the lowest SI-SDR scores, possibly because we have four possible languages, thus, there is a higher prior probability of picking a speaker with the incorrect language. In Fig. 4, we explore how adding an easier signal characteristic, spatial location $\mathcal{S}$, improves performance on the more difficult language conditioning $\mathcal{L}$. Surprisingly, the best performance for language conditioning is achieved when only 20% of the training mixtures are conditioned using $\mathcal{L}$ and 80% are conditioned using $\mathcal{S}$. Unfortunately, the inverse relationship does not hold, i.e., including the more difficult condition $\mathcal{L}$ negatively impacts performance for
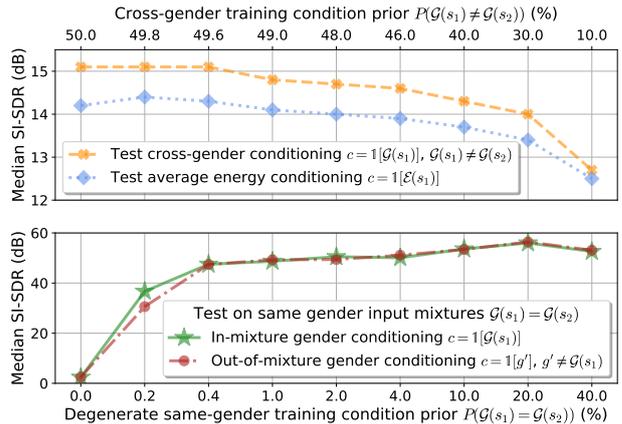

Figure 3: *Median SI-SDR performance on the WSJ test set for various conditions when training with a static prior of sampling an energy condition vector $P(c = \mathbb{1}(\mathcal{E}(s_1))) = 50\%$ and sweeping through the percentage of degenerate gender conditioning. In essence, the leftmost points correspond to training a model with only easy cross-gender mixtures whereas as we move towards the rightmost points, the percentage of the degenerate gender conditioning with same-gender inputs increases.*
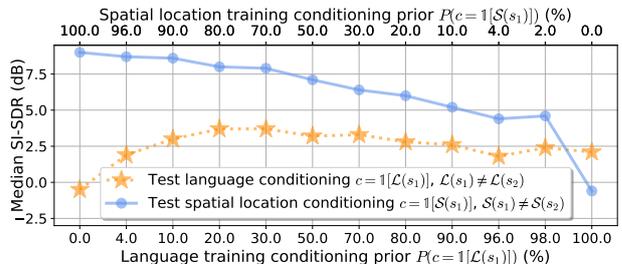

Figure 4: *Evaluation of the median SI-SDR performance on the SVOX test set for language and spatial conditioning when training with different priors of the input conditional vectors. From left to right we increase the amount of the harder language conditioning $P(c = \mathbb{1}(\mathcal{L}(s_1))) : 0 \to 1$ and (inversely $P(c = \mathbb{1}(\mathcal{S}(s_1))) : 1 \to 0$ for the complementary easier spatial conditioning). Note that the best model for the harder language condition is not the one trained solely for this task.*

the easier condition $\mathcal{S}$. We suspect this is due to model parameters that are unrelated to the conditioning, such as the learned encoder and decoder receiving large gradient updates for cases when the incorrect language is chosen, which may cause them to be less effective for speech separation in general. Mitigating this issue is an important topic of future work.

5. Conclusion

In this paper, we have introduced the heterogeneous target source separation task based on several non-mutually exclusive signal characteristic conditions. We have experimentally shown that the proposed heterogeneous condition training framework provides benefits in terms of domain generalization and has the ability to leverage "easy" conditions to facilitate training with more difficult concepts. Heterogeneous conditioning can be made robust to degenerate cases and often performs better than unconditional models with oracle source assignment. In the future, we plan to extend our method to incorporate a variable number of speakers and training with self-supervised schemes.

6. References

[1] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma, "Auditory attention—focusing the searchlight on sound," *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[3] N. Mesgarani and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature*, vol. 485, no. 7397, pp. 233–236, 2012.

[4] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, 2014.

[5] F. J. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, Dec. 2014, pp. 577–581.

[6] H. Erdogan, J. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, Apr. 2015, pp. 708–712.

[7] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[8] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep U-Net convolutional networks," in *Proc. ISMIR*, Oct. 2017, pp. 23–27.

[9] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. ICASSP*, 2016, pp. 31–35.

[10] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proc. Interspeech*, Sep. 2016, pp. 545–549.

[11] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *Proc. ICASSP*, Mar. 2017, pp. 241–245.

[12] M. Yousefi, S. Khorram, and J. H. Hansen, "Probabilistic permutation invariant training for speech separation," in *Proc. Interspeech*, Sep. 2019, pp. 4604–4608.

[13] S.-F. Huang, S.-P. Chuang, D.-R. Liu, Y.-C. Chen, G.-P. Yang, and H.-y. Lee, "Stabilizing label assignment for speech separation by self-supervised pre-training," in *Proc. Interspeech*, Aug. 2021, pp. 3056–3060.

[14] G.-P. Yang, S.-L. Wu, Y.-W. Mao, H.-y. Lee, and L.-s. Lee, "Interrupted and cascaded permutation invariant training for speech separation," in *Proc. ICASSP*, May 2020, pp. 6369–6373.

[15] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. ICASSP*, Apr. 2018, pp. 5554–5558.

[16] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "A unified framework for neural speech separation and extraction," in *Proc. ICASSP*, May 2019, pp. 6975–6979.

[17] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking," in *Proc. Interspeech*, Sep. 2019, pp. 2728–-2732.

[18] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan, C. Liu, D. Dimitriadis, J. Droppo, and Y. Gong, "Single-channel speech extraction using speaker inventory and attention network," in *Proc. ICASSP*, May 2019, pp. 86–90.

[19] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan, J. Li, and Y. Gong, "Multi-channel overlapped speech recognition with location guided speech extraction network," in *Proc. SLT*, Dec. 2018, pp. 558–565.

[20] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, "Class-conditional embeddings for music source separation," in *Proc. ICASSP*, May 2019, pp. 301–305.

[21] G. Meseguer-Brocal and G. Peeters, "Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations," in *Proc. ISMIR*, Nov. 2019, pp. 159–165.

[22] O. Slizovskaia, G. Haro, and E. Gomez Gutierrez, "Conditioned source separation for musical instrument performances," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2083–2095, 2021.

[23] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. Ellis, "Improving universal sound separation using sound classification," in *Proc. ICASSP*, May 2020, pp. 96–100.

[24] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito, K. Kinoshita, and S. Araki, "Listen to what you want: Neural network-based universal sound selector," in *Proc. Interspeech*, Oct. 2020, pp. 1441–1445.

[25] Y. Okamoto, S. Horiguchi, M. Yamamoto, K. Imoto, and Y. Kawaguchi, "Environmental sound extraction using onomatopoeia," *arXiv preprint arXiv:2112.00209*, 2021.

[26] E. Manilow, G. Wichern, and J. Le Roux, "Hierarchical musical source separation," in *Proc. ISMIR*, Oct. 2020, pp. 376–383.

[27] E. Tzinis, Z. Wang, X. Jiang, and P. Smaragdis, "Compute and memory efficient universal sound source separation," *Journal of Signal Processing Systems*, vol. 94, no. 2, pp. 245–259, 2022.

[28] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, Feb. 2018, pp. 3942–3951.

[29] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe, M. Chinen, B. Patton, and R. A. Saurous, "Differentiable consistency constraints for improved deep speech enhancement," in *Proc. ICASSP*, May 2019, pp. 900–904.

[30] R. Scheibler, E. Bezzam, and I. Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. ICASSP*, Apr. 2018, pp. 351–355.

[31] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," 1993, Web Download. Philadelphia: Linguistic Data Consortium.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, Apr. 2015, pp. 5206–5210.

[33] "Voxforge dataset," *http://www.voxforge.org/*.

[34] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[35] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. ICASSP*, May 2019, pp. 626–630.

[36] E. Tzinis, D. Bralios, and P. Smaragdis, "Unified gradient reweighting for model biasing with applications to source separation," in *Proc. ICASSP*, Jun. 2021, pp. 531–535.

[37] Z. Zhang, B. He, and Z. Zhang, "X-TaSNet: Robust and accurate time-domain speaker extraction network," in *Proc. Interspeech*, Oct. 2020, pp. 1421–1425.

[38] M. Borsdorf, C. Xu, H. Li, and T. Schultz, "Universal speaker extraction in the presence and absence of target speakers for speech of one and two talkers," in *Proc. Interspeech*, Aug. 2021, pp. 1469–1473.