

OPTIMAL CONDITION TRAINING FOR TARGET SOURCE SEPARATION

Efthymios Tzinis^{1,2}, Gordon Wichern¹, Paris Smaragdis², Jonathan Le Roux¹

¹Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA

²University of Illinois at Urbana-Champaign, Urbana, IL, USA

ABSTRACT

Recent research has shown remarkable performance in leveraging multiple extraneous conditional and non-mutually exclusive semantic concepts for sound source separation, allowing the flexibility to extract a given target source based on multiple different queries. In this work, we propose a new optimal condition training (OCT) method for single-channel target source separation, based on greedy parameter updates using the highest performing condition among equivalent conditions associated with a given target source. Our experiments show that the complementary information carried by the diverse semantic concepts significantly helps to disentangle and isolate sources of interest much more efficiently compared to single-conditioned models. Moreover, we propose a variation of OCT with condition refinement, in which an initial conditional vector is adapted to the given mixture and transformed to a more amenable representation for target source extraction. We showcase the effectiveness of OCT on diverse source separation experiments where it improves upon permutation invariant models with oracle assignment and obtains state-of-the-art performance in the more challenging task of text-based source separation, outperforming even dedicated text-only conditioned models.

Index Terms— conditional sound separation, optimal condition, conditional embedding refinement, text-based separation

1. INTRODUCTION

Humans possess the remarkable ability to isolate sounds from a noisy auditory input stimuli and associate them with objects and actions seamlessly. Auditory machine perception aims to mimic and even enhance this ability in a digitized manner, wherein the main challenge is to find an effective way to train models which are apt for the task of audio source separation.

Early works in deep-learning based audio source separation leveraged fundamental differences between the statistics of the sources of interest and those of other interfering sources in a mixture, making implicit assumptions on their semantic attributes. Thus, one could develop specialist models dedicating an output slot to recover only a given sound of interest, such as for speech enhancement [1–4] or instrument demixing [5]. Eventually, more general training procedures such as deep clustering [6] and permutation invariant training (PIT) [7,8] took over the field, mainly because of their minimal a-priori assumptions on the types of sources. However, PIT’s flexibility in training source separation networks does not come without a price, since PIT can neither solve the source alignment problem nor be used to explicitly specify the source of

interest, and it suffers from instability problems [9]. In contrast to semantically agnostic approaches, conditionally informed systems do not need to fix the order of the output sources and sometimes outperform PIT models [10–12]. Such works include models where an extra input conditional vector might carry information about speaker characteristics, musical instrument type, or general sound-class semantics, as proposed for speech [13–17], music [18–20], and universal sound separation [21–23].

Lately, there has been a resurgence of interest towards conditional separation models [24,25], not only for boosting their performance but also to give the user more flexibility to query the model. In particular, heterogeneous speech separation [24] was recently proposed as a conditional source separation training procedure where non-mutually exclusive concepts are used to discriminate between a mixture’s constituent sources. The resulting model not only can be queried using a diverse set of discriminative concepts (e.g., distance from the microphone, signal-level energy, spoken language, etc.), but also leverages the extra semantic information at training time to outperform PIT. Other follow-up works include single-conditioned models using a natural language description of the sources of interest [26] and/or encoded audio-snippet queries [27].

As the same target source may be queried using multiple equivalent conditions, in this work, we investigate whether for a given input mixture, an initial conditioning may be reformulated into a new conditioning that leads to better separation. As an intermediate step towards that goal, we first consider a system that focuses on reaching the best target extraction performance among all equivalent conditions for a given target, proposing a new training method, OCT, which performs a gradient step using the best performing conditional vector. We then propose OCT++, which combines OCT with an on-the-fly conditional vector refinement module to reformulate, based on the input mixture, an initial query into a representation which can lead to better extraction of the target sources. We also extend the original heterogeneous training framework [24] to conduct experiments on the conditional separation of arbitrary sounds using more diverse and easy-to-use discriminatory semantic concepts such as *text*, *harmonicity*, *energy*, and *source order*. Our experiments show that OCT yields a much higher upper bound for conditional separation based on the complementary semantic information of the diverse associated discriminative concepts surpassing all single-conditioned models and PIT. Moreover, OCT++ yields state-of-the-art performance on text-based sound separation and surprisingly outperforms all dedicated text-based methods by a large margin.

2. METHOD

We formulate the problem of conditional source separation as follows. Given an input mixture x consisting of the sum $x = \sum_{i=1}^M s_i$ of M sources $\mathbf{s} = (s_1, \dots, s_M)$, we consider a target waveform $\mathbf{s}_T = \sum_{j \in \mathcal{A}} s_j$ corresponding to a (potentially empty) subset $\mathcal{A} \subseteq$

Part of this work was performed as E. Tzinis was an intern at MERL. E. Tzinis was partially funded by the Google Ph.D. fellowship. E. Tzinis and P. Smaragdis were partially funded by NIFA grant #2020-67021-32799. Code: https://github.com/etzinis/optimal_condition_training

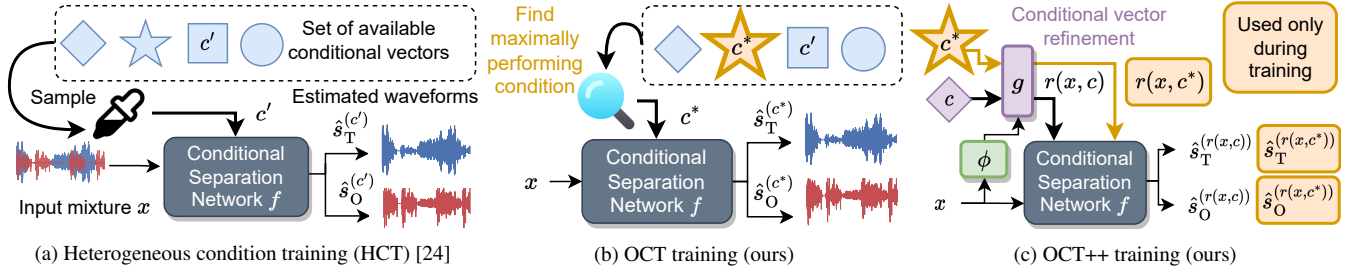


Fig. 1: Different conditional separation training procedures for a given input mixture x . 1a: a heterogeneous condition vector c' (associated with the target waveform s_T) is sampled at random and a gradient step is performed. 1b: all available conditions are first evaluated and the error corresponding to the maximally performing conditional vector c^* is backpropagated. 1c: an initial condition vector of interest c is first converted to a more amenable representation $r(x, c)$ using the trainable mappings ϕ and g , and the parameters are updated based on a regular OCT gradient update as well as the backpropagated errors from the regular path (black).

$\{1, \dots, M\}$ of target sources which can be described as associated with a condition v . Expressing v as a conditional vector $c = c(v)$, we aim to train a model f with parameters θ_f which outputs estimates for both the target submix s_T and the non-target (“other”) submix $s_O = \sum_{j \notin \mathcal{A}} s_j$ of the mixed input sources \mathbf{s} :

$$\hat{s}_T^{(c)}, \hat{s}_O^{(c)} = f(x, c; \theta_f). \quad (1)$$

The condition v could be any discriminative concept which is associated with semantic characteristics of the target waveform s_T . In this work, we consider the set of signal characteristics $\mathcal{C} = \{\mathcal{E}, \mathcal{H}, \mathcal{O}, \mathcal{T}\}$, where \mathcal{E} denotes the signal energy (with values low/high), \mathcal{H} is the harmonicity of the target source (harmonic/percussive), \mathcal{O} is the order of appearance of the source in time (first/second), and \mathcal{T} the text description of the target sound class(es) (e.g., a text embedding representing the words “a dog barking” given a mixture of sounds from an audio recording at a park). Importantly, several conditions v (and the corresponding conditional vector $c(v)$) may be associated with the same target waveform s_T . A schematic representation for all different conditional separation training methods discussed in this work is displayed in Fig. 1.

2.1. Permutation invariant training (PIT)

Usually, PIT [7, 8] is employed for supervised training of unconditional source separation models by backpropagating the error using the best permutation $\pi \in \mathcal{P}_M$ of the set $\{1, \dots, M\}$ aligning the estimated sources $\hat{\mathbf{s}}$ with the ground-truth sources \mathbf{s} as shown next:

$$\mathcal{L}_{\text{PIT}}(\mathbf{s}, \theta_f) = \min_{\pi \in \mathcal{P}_M} [\sum_{i=1}^M \mathcal{D}(\hat{s}_{\pi(i)}, s_i)], \quad \hat{\mathbf{s}} = f(x, -; \theta_f) \quad (2)$$

where \mathcal{D} is any desired signal-level distance or loss used to penalize the reconstruction error between the estimates and their corresponding targets, and $-$ indicates the absence of conditioning. Notice that for the problem of target source separation, unconditional PIT models need to be considered in combination with a speaker selection scheme, since they do not solve the alignment problem of the estimated sources. Thus, we use the oracle permutation of sources, to study the upper bound of their separation performance.

2.2. Heterogeneous condition training (HCT)

The concept of heterogeneous condition separation, introduced in [24] for conditional speech separation, can be readily extended to general target sound source separation tasks. In essence, the model is fed with an input mixture x as well as a one-hot encoded conditional vector $c_H(v) = \mathbb{1}[v \in \{0, 1\}^{\mathcal{V}}]$ for the desired semantic concept v , where in [24] \mathcal{V} was a set of speaker discriminative concept values such as “highest/lowest energy speaker” or “far/near field speaker.”

During training, a mixture x is drawn or synthetically generated and an associated discriminative concept v (corresponding to an encoded conditional vector $c = c(v)$) is drawn from a sampling prior $P(v)$ to form the desired target submix s_T containing all the sources s_i associated with v . The model tries to faithfully recover the target and non-target waveforms for v by minimizing the following loss:

$$\mathcal{L}_{\text{HCT}}(s_T, s_O, c; \theta_f) = \mathcal{D}(\hat{s}_T^{(c)}, s_T) + \mathcal{D}(\hat{s}_O^{(c)}, s_O), \quad (3)$$

where we explicitly stated s_T, s_O, c as parameters of \mathcal{L}_{HCT} to indicate that multiple combinations of conditions and targets may be considered. In [24], it was shown that when the model is trained with multiple heterogeneous semantic conditional targets, an overall separation performance improvement can be achieved.

2.3. Optimal condition training (OCT)

As the same target waveform may be associated with multiple conditions, the question remains whether some conditions lead to better separation accuracy than others, and whether the system may benefit from modifying the conditioning vector based on the input, in other words to “rephrase the query” in light of the actual input. One reasonable goal to reach when modifying the conditioning would be the conditioning that obtains maximum performance for the given input mixture and target waveforms. A heterogeneous model may however need to balance its performance under multiple conditions, leading to suboptimal separation accuracy for the best conditioning, and thus ultimately for a system relying on modifying an original conditioning by replacing it or making it closer to the best one. Thus, we first consider training a model that solely focuses on optimizing performance for the maximally performing condition.

OCT follows a greedy approach in which instead of sampling a heterogeneous conditional vector c and training the separation system, several (potentially all) possible conditional vectors $c' \in \mathcal{C}$ associated with the target waveform s_T are first evaluated, and we update the parameters of the network based on the condition that minimizes the overall error. Formally, we write the following loss function for updating the parameters of the conditional network as:

$$c^* = \operatorname{argmin}_{c' \in \mathcal{C}} [\mathcal{D}(\hat{s}_T^{(c')}, s_T) + \mathcal{D}(\hat{s}_O^{(c')}, s_O)], \quad (4)$$

$$\mathcal{L}_{\text{OCT}}(s_T, s_O; \theta_f) = \mathcal{L}_{\text{HCT}}(s_T, s_O, c^*; \theta_f),$$

where c^* is the optimal condition (i.e., the one obtaining the smallest loss) for the input mixture x and the target s_T . We consider updating the model’s parameters using conditional target vectors describing the ground-truth target submix s_T under various contexts sampled from the available signal characteristics $\mathcal{C} = \{\mathcal{E}, \mathcal{H}, \mathcal{O}, \mathcal{T}\}$. For example, if one wants to train a conditional separation system based

on text queries \mathcal{T} , there might be more effective ways to disentangle and isolate the same sources of interest based on complementary semantic information like the energy, the harmonicity, or the order of appearance of the sources. The evaluation of the ideal conditional target c^* is straightforward since we have access to the model f and the ground-truth waveforms s_T and s_O during training. Of course, at inference time, one does not have access to the set of equivalent conditions to a given condition v , so focusing on improving only the optimal condition is not guaranteed to be a viable solution. This procedure was intended to serve as the basis for a method in which an auxiliary network refines an original condition by mapping it to the optimal equivalent one in light of the input mixture. One may in fact expect that focusing solely on maximally performing conditions, or in other words the easiest queries, may harm performance for other conditions. Surprisingly, the final conditional model learns how to associate the sources of interest with the corresponding semantic concepts and the overfitting problem can be easily avoided using an extra gradient update based on the condition of interest. OCT models can also perform better compared to dedicated systems trained and tested on the same input conditional information.

2.4. OCT++: OCT with embedding refinement

Going a step further, there are cases where the input conditional information might not be informative enough by itself to lead to a conditioning vector that appropriately specifies the sources of interest, and one may hope to obtain an improved conditioning vector by letting the system look at both the input mixture and the original conditioning vector to output an improved conditioning vector. We thus consider introducing a learnable transformation $g(\cdot)$ of the conditional vector c to refine the conditional information so that it may be better utilized by the framework. For example, if the input mixture contains a guitar and a bass with different starting times, a query that corresponds to which instrument was played first (c_O : source order query) could be more informative than the textual description of the target musical instrument (c_T : text query). In that case, even if the user gives as an input *retrieve the bass*, the learnable transformation g could be used to map the less informative textual conditional input c_T to something that resembles the ideal (oracle) conditional target $c^* = c_O$. That transformation would in effect relieve the extraction network from making a difficult source selection and let it focus on the extraction. We let the learnable mapping g take into account information about both the input mixture, via a time-invariant encoded representation $\phi(x; \theta_\phi)$, and the initial conditional target c , computing the refined (or reassigned) conditional vector $r(x, c)$ as:

$$r(x, c) = g(\text{concat}(\phi(x; \theta_\phi), c); \theta_g). \quad (5)$$

The final loss to be minimized combines the heterogeneous loss of Eq. 3 on the refined condition $r(x, c)$ and the OCT loss of Eq. 4, where c^* is the condition which leads to maximal performance after refinement. The loss is computed based on the refined counterpart $r(x, c^*)$, as well as an extra regularizer term which aims to promote consistency at the conditional refinement mapping g (e.g. steer the refined conditional target $r(x, c)$ towards the ideal one $r(x, c^*)$):

$$c^* = \underset{c \in \mathcal{C}}{\text{argmin}} \left[\mathcal{D}(\hat{s}_T^{(r(x, c^*))}, s_T) + \mathcal{D}(\hat{s}_O^{(r(x, c^*))}, s_O) \right], \quad (6)$$

$$\mathcal{L}_{\text{OCT++}}(s_T, s_O, c; \theta) = \mathcal{L}_{\text{HCT}}(s_T, s_O, r(x, c); \theta_f) + \mathcal{L}_{\text{HCT}}(s_T, s_O, r(x, c^*); \theta_f) + \|r(x, c) - r(x, c^*)\|^2,$$

where the set of trainable parameters $\theta = \{\theta_f, \theta_g, \theta_\phi\}$ contains all the main network's f parameters, the parameters of the conditional refinement mapping g , and the parameters of the mixture encoder ϕ .

In this case, the model tries to both optimize the separation performance of its estimate \hat{s}_T as well as the reassignment mapping g as it tries to make the conditional input vector look mostly like the highest performing conditional query after the transformation $r(x, c^*)$.

3. EXPERIMENTAL FRAMEWORK

3.1. Datasets

We extract the following three mixing datasets based on different portions of the FSD50K [28] audio data collection, which consists of 200 sound classes. Each training epoch consists of the on-the-fly generation of 20,000 mixtures of 5 s length, sampled at 8 kHz and mixed at random input SNRs $\mathcal{U}[0, 2.5]$ dB with at least 80% overlap (harder set) or $\mathcal{U}[0, 5]$ dB with at least 60% overlap (easier set). The validation and test sets for each one of the following datasets are similarly generated, with 3,000 and 5,000 mixtures, respectively.

Random super-classes: We first randomly sample two distinct sound classes (out of the available 200), then sample a representative source waveform for each class and mix them together.

Different super-classes: We select a subset of classes from the FSD50K ontology corresponding to six diverse and more challenging to separate super-classes of sounds, namely: *Animal* (21 subclasses), *Musical Instrument* (35 subclasses), *Vehicle* (15 subclasses), *Domestic & Home Sounds* (26 subclasses), *Speech* (5 subclasses) and *Water Sounds* (6 subclasses). Each mixture contains two sound waveforms that belong to distinct super-classes.

Same super-class: Following the super-class definition from above, we force each mixture to consist of sources that belong to the same abstract category of sounds to test the ability of text-conditioned models in extremely challenging scenarios.

3.2. Separation Model

We follow [24] and use the same conditional Sudo rm -rf model [29] with a trainable FiLM [30] layer before each U-ConvBlock, with a mixture consistency layer at the output sources [31], except that we here use only $B = 8$ U-ConvBlocks since they were empirically found to be adequate for our universal conditional separation experiments. For the OCT++ embedding refinement part, we use as ϕ the downsampling encoder part of one U-ConvBlock block with a similar configuration of 512 intermediate channels and four 4-strided depth-wise convolutional layers, and we reduce the time axis using a two-head attention pooling similar to [32]. The resulting vector is concatenated with the conditional vector c and passed through g , which is a two-layer MLP with ReLU intermediate activations to form the refined conditional vector $r(x, c)$.

3.3. Baseline systems

Text-based separation [26]: We follow the previous state-of-the-art text-based source separation system proposed in [26] and use a pre-trained BERT [33] encoding for the class of each sound. The final class encoding is computed after passing the first output token of the sequence model through a linear layer with a ReLU activation.

Proposed text-based separation: We also propose a stronger baseline for the text-based separation, wherein we replace the language model with a sentence-BERT model [34] and the first token with a mean average pooling operation and a trainable linear layer on top which better describes the linguistic information for shorter sentences like in audio-class based information (see results in Table 1).

HCT [24]: We train the system with equal sampling probability over all the available signal characteristics $\mathcal{C} = \{\mathcal{E}, \mathcal{H}, \mathcal{O}, \mathcal{T}\}$.

3.4. Training and evaluation details

We train all models using the losses described in Sec. 2. For the OCT text-based separation experiments we always perform a gradient update with both the text-query c_T and the best performing condition c^* to avoid overfitting to the rest of the heterogeneous conditions. We use a batch size of 6 and the Adam [35] optimizer with an initial learning rate of 10^{-3} , halving it every 15 epochs.

We evaluate the source reconstruction fidelity at 110 epochs, after empirically finding that all models had converged, using the mean scale-invariant signal-to-distortion ratio (SI-SDR) [36] between the estimate \hat{s}_T and the ground-truth target s_T . For the unconditional PIT oracle models, we measure the permutation invariant SI-SDR.

4. RESULTS

4.1. Importance of the appropriate conditional vector

In Fig. 2, we show the performance of several single-condition models (trained to only handle a single type of query) and their oracle ensemble (where, for a given target, we select the query type leading to the best separation among all queries associated with the target) versus our proposed oracle OCT approach for target sound extraction. It is evident that several of the conditions fail dramatically on challenging data, while the best performing condition remains more robust, which indicates the importance of providing the right context for the task of target sound separation. For instance, the energy condition cannot be used when there is an ambiguity regarding the loudest source, as in cases where the input SNR is close to 0 dB (see Fig. 2a). Notably, the text-based condition, which is the most convenient to be used, performs poorly in the more challenging setups where the super-classes of sounds being mixed are similar or restricted, which enhances our belief that one needs to steer the conditional embedding vector towards the highest performing condition based on the given input mixture. Surprisingly, the OCT oracle model manages to perform better than the oracle best single-conditioned model which hints that integrating sound sources’ semantic information through gradient-based updates can be an effective way for more robust source separation.

4.2. OCT against state-of-the-art methods

We choose text-based separation as our main benchmark since it is the most challenging condition and simultaneously the one that a user would likely use to describe the sources of interest. We measure the separation performance for the three universal separation datasets, as summarized in Table 1. It is evident that the oracle OCT method gives the best results even compared to the PIT oracle, which does not solve the estimated source alignment problem. We can thus assume that the complementary conditional information might be used to better disentangle the sources. Although our proposed single-conditioned text-based model surpasses the previous state-of-the-art text-based condition method [26] under all dataset configurations, it still performs poorly, especially for the harder to disentangle mixtures with input SNR in the $[0, 2.5]$ dB range. Surprisingly, OCT, which was trained using the error signal from the best condition (which could be different from the text query), outperforms the dedicated text-based models, leveraging the complimentary information from the rest of the discriminative semantic concepts. OCT yields a significant improvement over heterogeneous training, which indicates that it is potentially a more efficient way of performing cross-semantic information training for source separation. Finally, our proposed embedding refinement method OCT++ outperforms

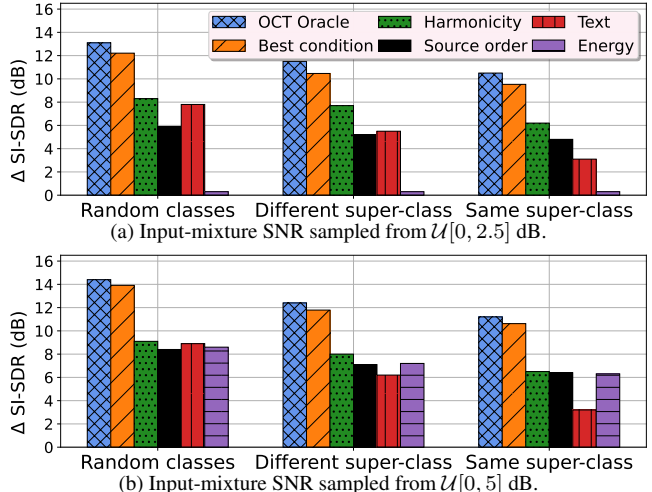


Fig. 2: Mean test SI-SDR (dB) for target source separation with single-conditioned models using either *harmonicity* (\mathcal{H}), *source order* (\mathcal{O}), *text* (\mathcal{T}), or *signal-level energy* (\mathcal{E}) versus their oracle ensemble, which uses the best conditional vector, and an oracle OCT model trained and tested with the highest performing query.

Table 1: Mean test SI-SDR (dB) results for text-based sound source separation using mixing strategies with two levels of difficulty: input-SNRs $\mathcal{U}[0, 2.5]$ dB and at least 80% overlap (left) and $\mathcal{U}[0, 5]$ dB with at least 60% overlap (right). **Bolded** and *Italic* numbers denote the best non-oracle and oracle models trained to perform text-based source separation, respectively.

Training method * Denotes our implementation.	Input-SNR $\mathcal{U}[0, 2.5]$ dB			Input-SNR $\mathcal{U}[0, 5]$ dB		
	Super-classes in-mixture Random	Diff.	Same	Super-classes in-mixture Random	Diff.	Same
Text only [26]*	6.1	3.9	2.2	8.6	6.0	2.9
Text only (ours)	7.9	5.6	3.1	9.0	6.3	3.3
HCT [24]	6.8	4.8	2.3	7.0	4.3	2.4
(Proposed) OCT (No ϕ and g)	8.4	6.0	3.3	9.3	6.5	3.6
(Proposed) OCT++	8.7	6.2	3.6	9.3	6.7	3.7
(Oracle) OCT (No ϕ and g)	13.1	<i>11.5</i>	10.5	14.4	12.4	11.2
(Oracle) OCT++	<i>13.2</i>	<i>11.5</i>	<i>10.6</i>	<i>14.7</i>	<i>12.6</i>	<i>11.5</i>
(Oracle) PIT [8]	12.4	10.7	9.8	12.4	10.7	9.8

the previous state-of-the-art text-based separation method by 0.7 to 2.6 dB SI-SDR and yields a consistent improvement on top of the OCT by converting the conditional vector to a more amenable representation for text-based separation. We hypothesize that future work could provide much larger improvements by employing more sophisticated mixture encoders ϕ and refinement embedding maps g .

5. CONCLUSION

We have introduced a new training method for source separation which leverages the backpropagation of the optimal conditional vector signal. OCT outperforms all previous state-of-the-art single- and multi-condition (aka heterogeneous) training methods for the more challenging and easier-to-use text-based conditioning. Oracle OCT also outperforms unconditional models trained and evaluated with permutation invariance. OCT++ enables further refinement by transformation of the conditional information vectors to a more amenable to separation form adapted to the input mixture. In the future, we aim to pair the proposed training methods with self-supervised approaches and explore in more detail the effectiveness of OCT.

6. REFERENCES

- [1] 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.
- [2] 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*, 2014, pp. 577–581.
- [3] 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*, 2015, pp. 708–712.
- [4] 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.
- [5] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner *et al.*, “Singing voice separation with deep U-Net convolutional networks,” in *Proc. ISMIR*, 2017, pp. 23–27.
- [6] 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.
- [7] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe *et al.*, “Single-channel multi-speaker separation using deep clustering,” in *Proc. Interspeech*, 2016, pp. 545–549.
- [8] 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*, 2017, pp. 241–245.
- [9] G.-P. Yang, S.-L. Wu, Y.-W. Mao, H.-y. Lee *et al.*, “Interrupted and cascaded permutation invariant training for speech separation,” in *Proc. ICASSP*, 2020, pp. 6369–6373.
- [10] L. Le Magoarou, A. Ozerov, and N. Q. Duong, “Text-informed audio source separation. example-based approach using non-negative matrix partial co-factorization,” *Journal of Signal Processing Systems*, vol. 79, no. 2, pp. 117–131, 2015.
- [11] S. Parekh, S. Essid, A. Ozerov, N. Q. Duong *et al.*, “Motion informed audio source separation,” in *Proc. ICASSP*, 2017, pp. 6–10.
- [12] K. Schulze-Forster, C. Doire, G. Richard, and R. Badeau, “Weakly informed audio source separation,” in *Proc. WASPAA*, 2019, pp. 273–277.
- [13] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa *et al.*, “Single channel target speaker extraction and recognition with speaker beam,” in *Proc. ICASSP*, 2018, pp. 5554–5558.
- [14] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa *et al.*, “A unified framework for neural speech separation and extraction,” in *Proc. ICASSP*, 2019, pp. 6975–6979.
- [15] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar *et al.*, “Voice-Filter: Targeted voice separation by speaker-conditioned spectrogram masking,” in *Proc. Interspeech*, 2019, pp. 2728–2732.
- [16] X. Xiao, Z. Chen, T. Yoshioka, H. Erdogan *et al.*, “Single-channel speech extraction using speaker inventory and attention network,” in *Proc. ICASSP*, 2019, pp. 86–90.
- [17] Z. Chen, X. Xiao, T. Yoshioka, H. Erdogan *et al.*, “Multi-channel overlapped speech recognition with location guided speech extraction network,” in *Proc. SLT*, 2018, pp. 558–565.
- [18] P. Seetharaman, G. Wichern, S. Venkataramani, and J. Le Roux, “Class-conditional embeddings for music source separation,” in *Proc. ICASSP*, 2019, pp. 301–305.
- [19] G. Meseguer-Brocal and G. Peeters, “Conditioned-U-Net: Introducing a control mechanism in the U-Net for multiple source separations,” in *Proc. ISMIR*, 2019, pp. 159–165.
- [20] 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.
- [21] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen *et al.*, “Improving universal sound separation using sound classification,” in *Proc. ICASSP*, 2020, pp. 96–100.
- [22] T. Ochiai, M. Delcroix, Y. Koizumi, H. Ito *et al.*, “Listen to what you want: Neural network-based universal sound selector,” in *Proc. Interspeech*, 2020, pp. 1441–1445.
- [23] Y. Okamoto, S. Horiguchi, M. Yamamoto, K. Imoto *et al.*, “Environmental sound extraction using onomatopoeia,” *arXiv preprint arXiv:2112.00209*, 2021.
- [24] E. Tzinis, G. Wichern, A. S. Subramanian, P. Smaragdis *et al.*, “Heterogeneous target speech separation,” in *Proc. Interspeech*, 2022, pp. 1796–1800.
- [25] Y. Ohishi, M. Delcroix, T. Ochiai, S. Araki *et al.*, “Concept-beam: Concept driven target speech extraction,” in *Proc. ACM Multimedia*, 2022, pp. 4252–4260.
- [26] X. Liu, H. Liu, Q. Kong, X. Mei *et al.*, “Separate what you describe: Language-queried audio source separation,” in *Proc. Interspeech*, 2022, pp. 1801–1805.
- [27] K. Kilgour, B. Gfeller, Q. Huang, A. Jansen *et al.*, “Text-driven separation of arbitrary sounds,” in *Proc. Interspeech*, 2022, pp. 5403–5407.
- [28] E. Fonseca, X. Favory, J. Pons, F. Font *et al.*, “Fsd50k: an open dataset of human-labeled sound events,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.
- [29] 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.
- [30] E. Perez, F. Strub, H. De Vries, V. Dumoulin *et al.*, “FiLM: Visual reasoning with a general conditioning layer,” in *Proc. AAAI*, 2018, pp. 3942–3951.
- [31] S. Wisdom, J. R. Hershey, K. Wilson, J. Thorpe *et al.*, “Differentiable consistency constraints for improved deep speech enhancement,” in *Proc. ICASSP*, 2019, pp. 900–904.
- [32] E. Tzinis, S. Wisdom, T. Remez, and J. R. Hershey, “Audioscopev2: Audio-visual attention architectures for calibrated open-domain on-screen sound separation,” in *Proc. ECCV*, 2022.
- [33] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. ACL*, 2019, pp. 4171–4186.
- [34] F. Iandola, A. Shaw, R. Krishna, and K. Keutzer, “SqueezeBERT: What can computer vision teach nlp about efficient neural networks?” in *Proc. of SustainNLP Workshop*, 2020, pp. 124–135.
- [35] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, “SDR—half-baked or well done?” in *Proc. ICASSP*, 2019, pp. 626–630.