Signal-Guided Source Separation

Wolfgang Mack, Mohamed Elminshawi, Emanuël A. P. Habets
International Audio Laboratories Erlangen, Germany
{wolfgang, mohamed, emanuel}.{mack, elminshawi, habets}@audiolabs-erlangen.de

***Abstract*—State-of-the-art separation of desired signal components (DSCs) from a mixture is achieved using time-frequency masks or filters estimated by a deep neural network (DNN). The DSCs are typically defined at the time of training, or alternatively during inference via a reference signal (RS). In the latter case, typically, an auxiliary DNN extracts signal characteristics (SCs) from the RS and estimates a set of adaptive weights (AWs) of the first DNN. In both cases, the information of DSCs is stored in the DNN weights. Current methods using audio RSs estimate time-invariant AWs. Applications where the RS and DSCs exhibit time-variant SCs, i.e., they cannot be assigned to a specific class like speech, require time-variant AWs. An example is acoustic echo cancellation with the loudspeaker signal as RS. We propose a method to extract time-variant AWs from a RS and additionally show that current time-invariant AWs methods can be employed for universal source separation. To avoid strong scaling between the estimate and the mixture, we propose to train with the dual scale-invariant signal-to-distortion ratio in a TASNET inspired DNN. We evaluate the proposed AWs systems under various acoustic conditions and show the scenario-dependent advantages of time-variant over time-invariant AWs.**

***Index Terms*—Time-Variant/Invariant Adaptive Weights, Signal-Guided Source Separation, Cancellation, Suppression**

Sound fields are often a composition of different directional sources, reverberation, and background noise. We refer to the recording of such a sound field by a single microphone as a mixture. Extracting one or more sound sources of interest from such a mixture is a highly investigated field with applications ranging from sound enhancement (e.g., [1], [2]) to provide high-quality recordings without noise and interference, to preprocessors for speech recognition systems to reduce the word error rate (e.g., [3], [4]). Historically, signal-processing methods have been used for this purpose. Examples are speech enhancement using a Wiener filter [5], dereverberation [6], [7], or acoustic echo reduction [8]–[10].

Beside signal-processing techniques, deep neural networks (DNNs) have been used to extract desired signals from a mixture [11]–[22]. Typically, a DNN is trained such that one or more gain matrices are obtained. These gain matrices, i.e., the masks or filters [21], [22], are subsequently applied to a representation of the mixture in a transform domain to extract the desired signals. Until recently, these methods required knowledge of the desired signals, such as a signal class, at the time of training the DNN. This constraint has been relaxed with recent developments in universal sound separation [23], [24]. The authors used a permutation invariant loss and estimated masks via a DNN to separate arbitrary sound events. These sound events exhibit different time-invariant (TI) signal characteristics (SCs), like speech, music, animal sounds. If sources exhibit only time-variant (TV) SCs, e.g., radios, the TI SCs based sound separation of [23], [24] is not applicable as TV SCs cannot be assigned to a specific sound event. Hence, the DNN requires TV guidance to perform separation.

Signal-guided DNNs have been proposed for extracting a target speaker from a mixture of several speakers and noise [16], [25], [26]. These methods allow defining the desired speaker based on a reference snippet of the respective speaker during inference. An auxiliary DNN processes the reference snippet to estimate a set of adaptive weights (AWs) for a source extraction DNN such that the desired speaker is extracted. In [27], the authors proposed to use in addition to a speech reference [16], [25], [26] a video of the speaker. However, current methods have only been applied for speech and require audio SCs to be TI (the video stream in [27] allows for TV video SCs). Such characteristics can be a specific speaker, or a particular sound class, like a guitar, or speech. Another application of signal-guided extraction is acoustic echo cancellation (AEC) [28], [29]. In AEC, where the loudspeaker's signal is given as reference, the task is to estimate the near-end speech signal and cancel the loudspeaker's echo. In [28], [29], the reference signal and the mixture are provided to the DNN to estimate a mask to obtain the near-end signal. Incorporating auxiliary information at the input, however, was shown to be inferior to methods using TI AWs [16] for speaker extraction.

This paper contains three main contributions. First, we show that AWs [16], [25], [26] can be used for universal source separation [23], [24]. Secondly, we propose a TV aggregation method to extract auxiliary information from a reference signal such that the SCs may change over time. TV reference signals are given in, e.g., forensics, reference-noise based speech enhancement [27], [30], acoustic echo cancellation/suppression [8]–[10], [31], auditory attention decoding via EEG signals [32], or audio-visual source extraction [27], [33]. Thirdly, we propose the dual scale-invariant signal-to-distortion ratio (dsi-SDR) training objective. The dsi-SDR penalizes scalings of the estimates w.r.t. the ground-truth such that the residual signal can be obtained by subtracting the estimate from the mixture. Having both signals separated with the original levels allows, for example, to mix them again with adjusted loudness, e.g., when speech in a movie is to be amplified w.r.t. the background sounds.

I. Problem Formulation

We assume a single sensor providing a signal $y \in \mathbb{R}^T$ consisting of 2 components, x_1 and x_2, as

$$y = x_1 + x_2. \tag{1}$$

Additionally, we define reference signals for x_i, denoted as r_i $\forall i \in \{1, 2\}$. To extract x_1, r_1 is required to exhibit

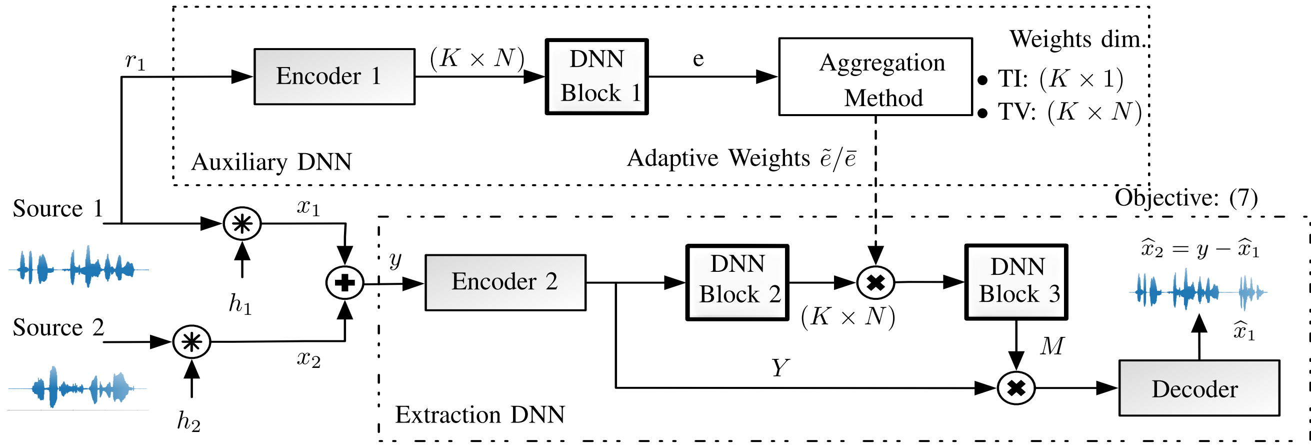

Fig. 1: Proposed DNN for TV/TI AWs learning. Encoder consist of a learnt 1-D conv layer and decoder comprises a 1-D transposed conv layer. For further details see [17], [34]. Each DNN Block consists of a layer-norm, 1×1 conv, a central DNN, PReLU, 1×1 conv and a sigmoid. The 1×1 conv specifies the bottleneck channels B as defined in [17]. In [34], the central DNN consists of TCNNs, whereas we propose to use DPRNNs [19] for performance reasons. The AM specifies the temporal combination of AWs. In [34], the AM is TI averaging [see (4)], whereas we propose a TV AM in Section II.

SCs of x_1 which are different from the SCs of x_2. Possible signals for r_1 could be EEG signals, far-end signals in AEC, noise reference signals, or videos as in [27]. Without loss of generality, we assume all signals to be audio signals and exhibit length T. Based on the extracted SCs from r_1, a mask M is estimated such that when applied to a representation of y in a transform domain, i.e.,

$$\widehat{x}_1 = \mathcal{T}^{-1}(\mathcal{T}(y) \odot M), \tag{2}$$

the estimate $\widehat{x}_1$ is obtained, where $\mathcal{T}$ and $\mathcal{T}^{-1}$ specify the forward transform and the inverse transform, respectively, and $\odot$ is the element-wise product. Subsequently, $\widehat{x}_2$ is obtained by cancellation, i.e.,

$$\widehat{x}_2 = y - \widehat{x}_1. \tag{3}$$

Consequently, the task of TV and TI signal-guided source separation narrows down to the extraction of SCs from r_1, the estimation of M and the selection of $\mathcal{T}$ and $\mathcal{T}^{-1}$.

In [34], the authors proposed a solution to address TI AWs learning, to extract a single desired speaker from a mixture of speakers. An overview of the architecture is given in Figure 1. In [34], r_1 is a snippet of the desired speaker speaking a different utterance than in the mixture. They proposed to pass r_1 through an auxiliary DNN, i.e., $e = \mathcal{A}(r_1)$, where r_1 is mapped to $e \in \mathbb{R}^{K \times N}$, K specifies the number of filter-channels introduced by $\mathcal{A}$, and N is the total number of time-frames in the transform domain. Subsequently, global weights, denoted as $\bar{e} \in \mathbb{R}^{K \times 1}$ are aggregated, i.e.,

$$\bar{e} = \frac{1}{N} \cdot \sum_{1}^{N} e[:, n], \tag{4}$$

where n is the time-frame index. In parallel, the mixture y is transformed via the learnt encoder $\mathcal{E}_2$ to $Y = \mathcal{E}_2(y) \in \mathbb{R}^{K \times N}$, where $\mathcal{E}_2$ may differ from $\mathcal{E}_1$. Then, Y is fed into a DNN to estimate the mask $M \in \mathbb{R}^{K \times N}$ which is subsequently applied to Y, i.e., $\widehat{X}_1 = Y \odot M \in \mathbb{R}^{K \times N}$, where $\widehat{X}_1$ is the representation of $\widehat{x}_1$ in the transform domain. The guidance to x_1 is performed by incorporating the AWs in the masking DNN, i.e.,

$$M = \mathcal{B}_3(\mathcal{B}_2(Y) \odot \bar{e}), \tag{5}$$

where $\mathcal{B}$ specifies the respective DNN Block in Figure 1. The estimate $\widehat{x}_1$ is obtained by passing $\widehat{X}_1$ through a learnt decoder, $\widehat{x}_1 = \mathcal{D}(\widehat{X}_1)$. In [34], the authors trained their approach with the scale-invariant signal-to-distortion ratio (si-SDR) [35], [36], loss $= -\text{si-SDR}(x_1, \widehat{x}_1)$. Due to large unpredictable scaling mismatches (> 100) between x_1 and $\widehat{x}_1$, when using the si-SDR, (3) yields $\widehat{x}_2 \approx -\widehat{x}_1$. However, in [34], only x_1 was of interest.

Note that, to separate a specific class, e may be independent of n as in [34] whereas for TV SCs, it must be dependent on n. As an example, consider a mixture of 2 speakers where the reference signal r_1 is a video of the lips of speaker 1 while speaking. The lips movement contains information on which speaker to focus on. As both speakers have access to the same vocabulary and facial characteristics are not necessarily related to speech, local SCs are required.

II. Proposed Method

To estimate TV and TI AWs for signal-guided source separation, we propose several modifications for the DNN framework in [34] and a novel cancellation-enabling loss function to be able to use (3).

A. Algorithm

Firstly, as dual-path recurrent neural networks (DPRNNs) [19] were shown to outperform temporal convolution neural networks (TCNNs) [37], we propose to replace the TCNNs used in [34] for all DNN Blocks in Figure 1 by DPRNNs. Secondly, the averaging in (4) requires SCs to be TI. For the TV aggregation method (AM), we propose to replace the averaging with a long short-term memory (LSTM) [38] or bidirectional LSTM (BLSTM) DNN,

$$\tilde{e} = \text{(B)LSTM(e)}, \tag{6}$$

such that $M = \mathcal{B}_3(\mathcal{B}_2(Y) \odot \tilde{e})$. Note that $\tilde{e}$ is of size $(K \times N)$ unlike the TI $\bar{e}$. The memory of the (B)LSTM allows to exploit global SCs as well as local SCs and to address SC changes.

To avoid errors introduced by incorrect scaling of x_1 in $\widehat{x}_2$ after the cancellation in (3), a scale-preserved $\widehat{x}_1$ w.r.t. x_1 is required. Consequently, we propose to train by maximizing the dual scale-invariant SDR (dsi-SDR),

$$\text{dsi-SDR} = \text{si-SDR}(x_1, \widehat{x}_1) + \text{si-SDR}(x_2, \widehat{x}_2) \tag{7}$$

which optimizes simultaneously for $\widehat{x}_1$ and $\widehat{x}_2$. We noticed that the proposed loss, by chance, can lead to a strongly scaled $\widehat{x}_1$ w.r.t. x_1 such that (3) yields $\widehat{x}_2 \approx -\widehat{x}_1$. To avoid this local optimum, we propose to train in the first epoch using loss $= -\text{SDR}(x_1, \widehat{x}_1)$ [35], [36] instead of (7) to fix the scaling of $\widehat{x}_1$ w.r.t. x_1 close to 1. Assuming small errors, such that $x_1 \neq \widehat{x}_1$, the SDR (also the MSE) is optimized when x_1 in $\widehat{x}_1$ is slightly down-scaled [36]. In combination with (3), such a downscaling harms the performance w.r.t. $\widehat{x}_2$ as it contains residuals of x_1. Consequently, from the second epoch on, we replace the SDR with the proposed dsi-SDR loss to train simultaneously for $\widehat{x}_1$ and $\widehat{x}_2$. As shown in Section IV, optimizing for $\widehat{x}_1$ and $\widehat{x}_2$ simultaneously is non-contradictory and the optimum of both can be reached with the proposed loss.

B. Implementation

We consider a causal and an acausal implementation of the proposed model. The causal implementation uses a single-layer LSTM as AM (hidden dim = 256) and the causal version of the DPRNN [19]. Following the notation in [19], for the DPRNN, we set RNN hidden dim = 128, $K = 16$, $B = 2$. Additionally, we set the number of bottleneck channels of the 1×1 conv described in Figure 1 to 64. The channels are normalized similar to global layer normalization (gLN) [17], however, independently per time-frame. In the acausal version, the AM is BLSTM-based (hidden dim = 128), $K = 90$, gLN, acausal DPRNN. In the encoder/decoder, we used $K = 256$ channels and a window length of 16 samples as in [19], following our notation, and a hop-size of 8 samples. We used a sigmoid activation after DNN Block 3 in Figure 1 such that $M[n, k] \in [0, 1]$. We trained the causal models with the proposed TV aggregation and the acausal models with the TI and TV AMs. For training, we used ADAM [39] with a learning rate of 10^{-3}, a weight-decay of 10^{-5}, a gradient clipping of 5. We trained 300 epochs with a batch-size of 8, a learning rate decay of 50% with a patience of 10 epochs and early stopping with a patience of 20 epochs w.r.t. the validation loss.

III. Data Sets

Here, we give an overview of the datasets. The SC-connection of r_1 to x_1 is ensured by $x_1 = r_1 * h_1$, where $*$ is the convolution operator and $h_1 \in \mathbb{R}^{T_h}$ is a room impulse response (RIR) of length T_h. All signals (x_1, x_2, r_1) were cut to four seconds and resampled to 8 kHz.

A. Room Impulse Response Parameters

We simulated different RIRs for training, validation, and test sets using the source-image method [40], [41]. The room sizes were $\{[2, 4, 2.7], [6, 6, 2.7], [10, 4, 2.7], [7, 3, 2.7], [8, 10, 2.7]\}$, $\{[5, 6, 2.7], [4, 3, 2.7], [8, 9, 2.7]\}$, $\{[5, 6, 3], [4, 3, 3], [8, 9, 3]\}$ [m]. The reverberation times were $[0.2, ..., 0.5]$, $[0.23, ..., 0.53]$, $[0.25, ..., 0.45]$ [s] with an increment of 0.1 s each. The source-microphone distances were $[0.5, 0.7, ..., 1.9]$, $[0.55, 1.05, ..., 2.05]$, $[0.85, 1.35, 1.85]$ [m]. Each of the 3 sets corresponds to training, validation, and test, respectively. To generate a single RIR the intra-set parameters are randomly combined and source and microphone are positioned randomly in the room. The minimum object distance to walls was set to 1 m. The total number of different training, validation and test RIRs is 17760, 1776, 999, respectively.

B. Training, Validation, Test Data

We simulated training, validation, and test sets using the respective subsets of LibriSpeech (LS) [42], the YouTube-based FSDnoisy18k (FN) [43] and the RIRs described in Section III-A. The training set of FN was divided 70/30 between a training and a validation set due to the lack of a validation set. LS contains several 100 hours of clean speech. From FN, we used the 'Acoustic Guitar', 'Bass Guitar', 'Piano', 'Rain', and 'Engine' classes to cover noisy and harmonic signals. For training, we used a single data set with mixtures of four kinds in form of x_1-x_2: 1. Speaker-Speaker (SS), 2. Speaker-Noise (SN), 3. Noise-Speaker (NS) 4. Noise-Noise (NN), where Noise is a (non-speech) signal from FN and Speaker from LS. Each mixing was selected with a probability of 25%. In SS, the speakers are different and in NN the noise classes are different. In each scenario, we selected different RIRs from the same room to simulate x_1 and x_2. The signal r_1 is the anechoic version of x_1. Additionally, we repeated all experiments in non-reverberant environments, such that $x_1 = r_1$, using the same but non-reverberant training, validation, and test set to show an upper bound of the proposed method. In the following, the respective anechoic (A) sets are marked via $\bullet_{\text{A}}$ and the reverberant (R) sets are marked via $\bullet_{\text{R}}$. The training samples are continuously generated at runtime with a signal-to-interference ratio (SIR) $\in [-5, 5]$ dB. An epoch is defined as 10 k training samples. The validation set contains 2 k and the test set 4 k samples equally distributed over the four mixing scenarios with an SIR of 0 dB.

IV. Performance Evaluation

We trained three baseline models [28], [34] and six proposed models (TI/TV $\times$ causal/acausal $\times$ $\bullet_{\text{R}}$/$\bullet_{\text{A}}$ dataset)[1]. The signal model and the proposed architecture are depicted in Figure 1.

The results of the proposed models are summarized in Table I. The model trained for $\bullet_{\text{A}}$ shows an upper bound of the proposed algorithm when $x_1 = r_1$. The si-SDR values obtained for x_1 and x_2 are both very high and similar, which shows that the proposed loss function was able to optimize for both. For comparison, we trained the proposed TV acausal

[1] Audio examples are available at https://www.audiolabs-erlangen.de/resources/2020-EUSIPCO-Signal-Guided-Source-Separation

TABLE I: Results si-SDR in dB on the test sets for the causal (C) and acausal (AC) versions of the proposed method trained with the dsi-SDR (7). For the evaluation with the si-SDR, $\widehat{x}_2$ is obtained via (3).

		$\mathrm{SS_R}$		$\mathrm{SN_R}$		$\mathrm{NS_R}$		$\mathrm{NN_R}$		$\mathrm{SS_A}$		$\mathrm{SN_A}$		$\mathrm{NS_A}$		$\mathrm{NN_A}$	
Agg.	Evaluation	C	AC	C	AC	C	AC	C	AC	C	AC	C	AC	C	AC	C	AC
TI (4)	si-SDR(x_2, $\widehat{x}_2$)		8.7		12.9		12.3		11.4		14.5		16.0		15.5		12.6
	si-SDR(x_1, $\widehat{x}_1$)		8.7		12.9		12.3		11.4		14.5		16.0		15.5		12.6
TV (6)	si-SDR(x_2, $\widehat{x}_2$)	**11.5**	**12.4**	**14.0**	**15.0**	**14.4**	**15.5**	**13.4**	**14.7**	41.0	49.2	39.9	47.7	42.7	50.3	39.6	46.2
	si-SDR(x_1, $\widehat{x}_1$)	**11.5**	**12.4**	**14.0**	**15.0**	**14.4**	**15.5**	**13.5**	**14.7**	**41.3**	**50.3**	**40.2**	**51.2**	**42.9**	**51.5**	**40.0**	**50.3**

TABLE II: Baseline si-SDR results in dB on the test sets. The evaluation in the baseline papers only covers the speaker-speaker scenario. All models are acausal, also SpeakerBeam for the specific scenario in Figure 1 due to the TI AM. Using the notation in [34], we set $N = 256$, $L = 16$, $B = 64$, $H = 96$, $R = 2$, $X = 8$, $P = 3$, norm=gLN for SpeakerBeam. SpeakerBeam [34] is trained with the si-SDR for x_1 and Zhang [28] is optimized for a mask for x_2, as proposed in the respective papers. For evaluation, the respective residual signal is obtained by subtraction of the respective estimated signal from the mixture similar to (3).

Baseline	Evaluation	$\mathrm{SS_R}$	$\mathrm{SN_R}$	$\mathrm{NS_R}$	$\mathrm{NN_R}$
BL	si-SDR(x_2, $\widehat{x}_2$)	12.0	14.7	15.2	14.4
Speakerb. [34]	si-SDR(x_1,$\widehat{x}_1$)	7.2	11.9	11.4	10.9
Speakerb. [34]	si-SDR(x_2,$\widehat{x}_2$)	-12.7	-17.2	-16.9	-16.6
Zhang [28]	si-SDR(x_1, $\widehat{x}_1$)	6.7	9.5	9.0	9.1
Zhang [28]	si-SDR(x_2, $\widehat{x}_2$)	6.4	9.2	8.7	8.8

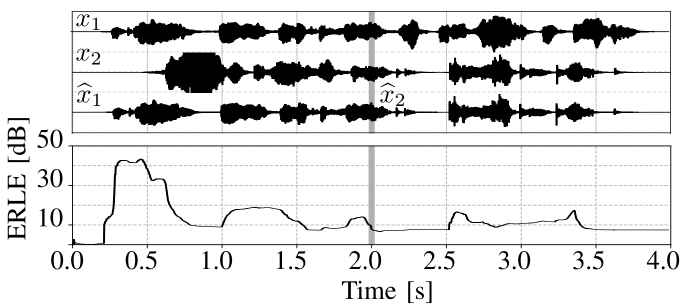

Fig. 2: ERLE over time for the TV acausal model. After 2 s, the used reference switches from r_1 to r_2.

model to directly yield $\widehat{x}_2$ without using (3) with a baseline loss to maximize si-SDR$(x_2, \widehat{x}_2)$. We refer to this model as BL and report the results in Table II. Training with the proposed dsi-SDR performs slightly better compared to the BL. Consequently, the cancellation in (3) and training with the dsi-SDR do not harm performance but allow simultaneously estimating x_1 and x_2.

The TV AM clearly outperforms the TI AM for all cases, as expected. With TV AM, it is possible to focus on temporal and global SCs, whereas with TI AM, only on global SCs. Evidently, temporal SCs from the reference can only be exploited if they are the same in x_1. If the temporal SCs were different, as in [34], where the reference is a different snippet from the same speaker, these temporal SCs could not be exploited. According to the performance differences in TI and TV, using the temporal SCs improves the si-SDR by at least $\approx$ 2 dB and more than 30 dB for anechoic signals $\bullet_\mathrm{A}$.

The acausal models perform slightly better than the causal models. The representation of r_1 in the mixture, x_1, is an attenuated, delayed, and reverberated version. The delay implies that to process temporal SCs, only the reference of the past is required. Nevertheless, the acausal models are slightly better, possibly because the processing of the complete reference allows for better processing of global SCs, which do not change in our scenario.

The proposed TI acausal model also outperforms SpeakerBeam [34] as shown in Table II. This shows that using DPRNNs instead of TCNNs improves performance. Moreover, the $\mathrm{NN_R}$ separation results using the TI models show that global SCs can be used to separate sound classes. Hence, the proposed method and SpeakerBeam can be used for universal source extraction. In particular, training with the proposed loss also enables universal source separation where x_1 and x_2 are of interest. As expected, the si-SDR of SpeakerBeam for x_2 is very low as SpeakerBeam was trained with the si-SDR on x_1, only. Consequently, $\widehat{x}_1$ is strongly scaled compared to x_1 and $\widehat{x}_2 \approx -\widehat{x}_1$ after (3).

A comparison to [28] shows that the proposed TV AWs method outperforms the input-concatenation of the reference method by approximately 6 dB for the acausal system. Similar results have been found in [16]. In contrast to [34], [28] trained to optimize a time-frequency mask, which was applied to a mixture representation in the short-time Fourier transform domain. As this objective is not scale-invariant as the si-SDR, the si-SDR difference for x_1 and x_2 is smaller compared to [34].

Finally, we investigate the effect of a change in the reference from r_1 to r_2 after 2 s. This change of reference includes a change in the RIR of the source to extract from h_1 to h_2. We evaluate the scenario for a speaker-speaker mixture in Figure 2 using the TV causal model. We use the echo return loss enhancement (ERLE) [9] for evaluation, as the tested scenario is similar to AEC. As expected, a change in the RIR/reference does not affect the proposed method severely due to the use of TV SCs. This is especially interesting as most conventional methods for AEC exhibit strong performance degradations when the echo path changes abruptly, especially during double talk.

V. Conclusion

We proposed a time-variant adaptive weights learning system, which can be applied for various suppression and cancellation tasks, such as found in AEC. To train the DNN, we proposed the dual scale-invariant signal-to-distortion ratio as a learning objective to penalize estimate scaling as required to enable cancellation. We also showed that adaptive weights learning can be used for universal source separation. An experimental evaluation of the proposed systems showed the advantages of time-variant over time-invariant adaptive weights learning, given a temporal connection between the reference and the desired signal in the mixture.

References

[1] T. Gerkmann and E. Vincent, "Spectral masking and filtering," in *Audio source separation and speech enhancement*, E. Vincent, T. Virtanen, and S. Gannot, Eds. New York, NY, USA: John Wiley & Sons, 2018, pp. 65–86.

[2] J. Benesty, J. Chen, and E. A. P. Habets, *Speech Enhancement in the STFT Domain*, ser. SpringerBriefs in Electrical and Computer Engineering. Springer-Verlag, 2011.

[3] R. Haeb-Umbach, J. Heymann, L. Drude, S. Watanabe, M. Delcroix, and T. Nakatani, "Far-field automatic speech recognition," *Proc. of the IEEE*, pp. 1–25, 2020, to appear.

[4] A. Moore, P. Peso Parada, and P. Naylor, "Speech enhancement for robust automatic speech recognition: Evaluation using a baseline system and instrumental measures," *Comp. Sp. and Lang.*, vol. 46, pp. 574 – 584, 2017.

[5] N. Wiener, *The Extrapolation, Interpolation and Smoothing of Stationary Time Series*. New York, NY, USA: John Wiley & Sons, Inc., 1949.

[6] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*. London, UK: Springer, 2010.

[7] K. Kinoshita, M. Delcroix, S. Gannot, E. A. P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka, "A summary of the REVERB challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. on Adv. in Sig. Proc.*, vol. 2016, no. 1, pp. 1–19, 2016.

[8] E. Hänsler and G. Schmidt, *Acoustic echo and noise control: A practical approach*. New York, NY, USA: John Wiley & Sons, 2004.

[9] G. Enzner, H. Buchner, A. Favrot, and F. Kuech, "Acoustic echo control," in *Academic Press Library in Signal Processing*, J. Trussell, A. Srivastava, A. K. Roy-Chowdhury, A. Srivastava, P. A. Naylor, R. Chellappa, and S. Theodoridis, Eds. Amsterdam, The Netherlands: Elsevier Science Publishers B. V., 2014, vol. 4, pp. 807–877.

[10] J. Benesty, T. Gänsler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Springer-Verlag Berlin Heidelberg, 2001.

[11] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, 2014.

[12] J. R. Hershey, Z. Chen, J. L. Roux, and S. Watanabe, "Deep clustering: Discriminative embeddings for segmentation and separation," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2016, pp. 31–35.

[13] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE Trans. Aud., Sp., Lang. Proc.*, vol. 24, no. 3, pp. 483–492, 2016.

[14] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE Trans. Aud., Sp., Lang. Proc.*, vol. 25, no. 10, pp. 1901–1913, 2017.

[15] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE Trans. Aud., Sp., Lang. Proc.*, vol. 26, no. 10, pp. 1702–1726, 2018.

[16] K. Zmolikova, M. Delcroix, K. Kinoshita, T. Ochiai, T. Nakatani, L. Burget, and J. Cernocky, "SpeakerBeam: Speaker aware neural network for target speaker extraction in speech mixtures," *IEEE J. sel. Top. in Sig. Proc.*, vol. 13, no. 4, pp. 800–814, 2019.

[17] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation," *IEEE Trans. Aud., Sp., Lang. Proc.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[18] Z.-Q. Wang, K. Tan, and D. Wang, "Deep learning based phase reconstruction for speaker separation: A trigonometric perspective," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2019, pp. 71–75.

[19] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: Efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2020, pp. 46–50.

[20] J. Le Roux, G. Wichern, S. Watanabe, A. Sarroff, and J. R. Hershey, "Phasebook and friends: Leveraging discrete representations for source separation," *IEEE J. sel. Top. in Sig. Proc.*, vol. 13, no. 2, pp. 370–382, 2019.

[21] W. Mack and E. A. P. Habets, "Deep filtering: Signal extraction and reconstruction using complex time-frequency filters," *IEEE Sig. Proc. Lett.*, vol. 27, pp. 61–65, 2020.

[22] H. Schröter, T. Rosenkranz, A. Escalante Banuelos, M. Aubreville, and A. Maier, "CLCNet: Deep learning-based noise reduction for hearing aids using complex linear coding," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2020, pp. 6949–6953.

[23] E. Tzinis, S. Wisdom, J. R. Hershey, A. Jansen, and D. P. W. Ellis, "Improving universal sound separation using sound classification," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2020, pp. 96–100.

[24] I. Kavalerov, S. Wisdom, H. Erdogan, B. Patton, K. Wilson, J. L. Roux, and J. R. Hershey, "Universal sound separation," in *Proc. IEEE W. on Appl. of Sig. Proc. to Aud. and Ac. (WASPAA)*, 2019, pp. 175–179.

[25] M. Delcroix, K. Zmolikova, T. Ochiai, K. Kinoshita, S. Araki, and T. Nakatani, "Compact network for speakerbeam target speaker extraction," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2019, pp. 6965–6969.

[26] M. Delcroix, K. Zmolikova, K. Kinoshita, A. Ogawa, and T. Nakatani, "Single channel target speaker extraction and recognition with speaker beam," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2018, pp. 5554–5558.

[27] T. Ochiai, M. Delcroix, K. Kinoshita, A. Ogawa, and T. Nakatani, "Multimodal SpeakerBeam: Single channel target speech extraction with audio-visual speaker clues," in *Proc. Interspeech Conf.*, 2019, pp. 2718–2722.

[28] H. Zhang and D. L. Wang, "Deep learning for acoustic echo cancellation in noisy and double-talk scenarios," in *Proc. Interspeech Conf.*, 2018, pp. 3239–3243.

[29] H. Zhang, K. Tan, and D. L. Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions," in *Proc. Interspeech Conf.*, 2019, pp. 4255–4259.

[30] W. Harrison, Jae Lim, and E. Singer, "A new application of adaptive noise cancellation," *IEEE Trans. Ac. , Speech, Sig. Proc.*, vol. 34, no. 1, pp. 21–27, 1986.

[31] M. M. Halimeh, C. Huemmer, and W. Kellermann, "A neural network-based nonlinear acoustic echo canceller," *IEEE Sig. Proc. Lett.*, vol. 26, no. 12, pp. 1827–1831, 2019.

[32] N. Das, J. Zegers, T. Francart, A. Bertrand *et al.*, "Linear versus deep learning methods for noisy speech separation for EEG-informed attention decoding," *J. of Neur. Eng.*, vol. 17, no. 4, 2020.

[33] T. Afouras, J. S. Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," in *Proc. Interspeech Conf.*, 2018, pp. 3244–3248.

[34] M. Delcroix, T. Ochiai, K. Zmolikova, K. Kinoshita, N. Tawara, T. Nakatani, and S. Araki, "Improving speaker discrimination of target speech extraction with time-domain speakerbeam," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2020, pp. 691–695.

[35] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Aud., Sp., Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[36] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?" in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2019, pp. 626–630.

[37] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Europ. Conf. on Comp. Vis.*, 2016, pp. 47–54.

[38] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neur. Comp.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[39] J. B. D. Kingma, "Adam: A method for stochastic optimization," in *Proc. IEEE Intl. Conf. on Learn. Repr. (ICLR)*, 2015, pp. 1–15.

[40] E. A. P. Habets. (2016, May) Room impulse response (RIR) generator. [Online]. Available: https://github.com/ehabets/RIR-Generator

[41] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Ac. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[42] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2015, pp. 5206–5210.

[43] E. Fonseca, M. Plakal, D. P. Ellis, F. Font, X. Favory, and X. Serra, "Learning sound event classifiers from web audio with noisy labels," in *Proc. IEEE Intl. Conf. on Ac., Sp. and Sig. Proc. (ICASSP)*, 2019, pp. 21–25.