ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414346

REAL-TIME SPEECH ENHANCEMENT FOR MOBILE COMMUNICATION BASED ON DUAL-CHANNEL COMPLEX SPECTRAL MAPPING

Ke Tan[1]*, Xueliang Zhang*[2]*, and DeLiang Wang*[1,3]

[1]Department of Computer Science and Engineering, The Ohio State University, USA
[2]Department of Computer Science, Inner Mongolia University, China
[3]Center for Cognitive and Brain Sciences, The Ohio State University, USA
tan.650@osu.edu, cszxl@imu.edu.cn, wang.77@osu.edu

ABSTRACT

Speech quality and intelligibility can be severely degraded by background noise in mobile communication. In order to attenuate background noise, speech enhancement systems have been integrated into mobile phones, and a microphone array is typically deployed to improve the enhancement performance. This paper proposes a novel approach to real-time speech enhancement for dual-microphone mobile phones. Our approach employs a causal densely-connected convolutional recurrent network to perform dual-channel complex spectral mapping. We apply a structured pruning technique for compressing the model without significantly affecting the enhancement performance. This leads to a real-time enhancement system for on-device processing. Evaluation results show that the proposed approach substantially advances the performance of an earlier approach to dual-channel speech enhancement for mobile communication.

Index Terms— real-time speech enhancement, complex spectral mapping, densely-connected convolutional recurrent network, dual-microphone mobile phones, on-device processing

1. INTRODUCTION

In mobile speech communication, speech signals can be severely corrupted by background noise when the far-end talker is in a noisy acoustic environment. Therefore, speech enhancement algorithms have been integrated into most mobile phones for noise reduction. To produce better enhancement results, a two-channel microphone array is typically deployed, where a primary microphone is placed on the bottom of a mobile phone and a secondary microphone on the top. In this study, we focus on noise reduction for such commonly-used dual-microphone mobile phones, and assume that reverberation energy is relatively weak, which is reasonable with relatively small speaker-phone distances in mobile communication.

In the past decade, dual-channel speech enhancement has been extensively studied in the speech processing community. In [1], a Wiener filter was formulated by leveraging the power level difference (PLD) between the signals received by two microphones, which was shown to improve speech quality. Subsequently, a PLD-based noise estimator was designed in [2], which uses the normalized inter-channel PLD as speech presence probability. Using the estimated noise spectrum, a spectral gain is computed and then applied to the noisy spectrum to obtain the enhanced spectrum. The results show that this approach yields better objective intelligibility than the approach in [1]. More recently, Fu *et al.* [3] used a minimum variance distortionless response spatial filter for noise reduction, which was shown to be more robust than the PLD method in [2] against different sensitivities of two microphones. Other related studies include [4] and [5].

Deep learning based speech enhancement for dual-microphone mobile phones has attracted increasing interests in recent years. To our knowledge, the first method was designed by López-Espejo *et al.* [6], in which a deep neural network (DNN) is trained to estimate a binary mask from the log-mel features of the noisy array signals. The enhanced spectrum is produced from the estimated mask through a truncated-Gaussian based imputation algorithm. In a subsequent study [7], a DNN is trained to produce an estimate of the noise spectrum, which is used to compute the primary-channel enhanced spectrum via a vector Taylor series feature compensation method. The enhanced spectrum is then passed into a speech recognizer for evaluation. Experimental results show that the DNN-based approach significantly outperforms several conventional approaches in terms of word accuracy. In a more recent study [8], we proposed a convolutional recurrent network (CRN) for real-time dual-channel speech enhancement, motivated by an earlier study on CRN [9]. The CRN is trained to estimate the phase-sensitive mask (PSM) [10] from magnitude-domain intra- and inter-channel features. The results show that this approach dramatically outperforms several conventional approaches, as well as a simple DNN that estimates the PSM.

Inspired by recent advances in complex-domain speech enhancement [11, 12, 13], we develop a new densely-connected CRN (DC-CRN) to perform dual-channel complex spectral mapping, which directly estimates the real and imaginary spectrograms of the primary-channel clean speech signal from those of the dual-channel noisy mixture. In addition, we propose a structured pruning technique to compress the DC-CRN, which substantially reduces the model size without significantly degrading the enhancement performance. This leads to a low-latency and memory-efficient enhancement system, which is necessary for real-time processing on mobile phones. Our experimental results suggest that the proposed approach consistently improves the enhancement performance over the approach in [8], in terms of short-time objective intelligibility (STOI) [14] and perceptual evaluation of speech quality (PESQ) [15].

The rest of this paper is organized as follows. In Section 2, we provide a detailed description of our approach . The experimental setup and evaluation results are presented in Section 3. Section 4 concludes this paper.

This research was supported in part by an NIDCD grant (R01 DC012048), and the Ohio Supercomputer Center. It started when the first author was interning with Elevoc Technology.

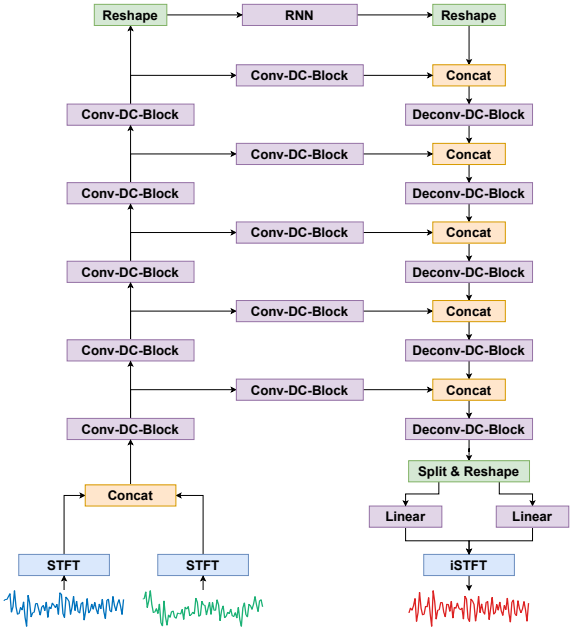


Fig. 1. Diagram of the DC-CRN for dual-channel complex spectral mapping.

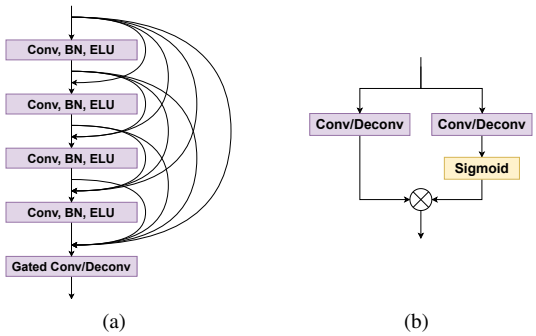
2. SYSTEM DESCRIPTION

2.1. Dual-channel complex spectral mapping

The inter-channel intensity difference (IID) is a spatial cue dominantly used by most studies on dual-channel speech enhancement. Another useful spatial cue is the inter-channel phase difference (IPD) or inter-channel time difference (ITD), which is highly correlated with the direction of arrival with respect to the microphone array. Both IID and IPD (or ITD) can be implicitly exploited by performing multi-channel complex spectral mapping [16], where the IID and the IPD are encoded in the dual-channel complex spectrogram of the noisy mixture. In contrast to conventional beamforming that typically exploits second-order statistics of multiple channels, such an approach has the potential to extract all effective cues in dual-channel complex-domain inputs through deep learning. Furthermore, complex spectral mapping simultaneously enhances magnitude and phase responses of target speech [13], which is advantageous over magnitude-domain approaches that ignore phase. Note that we aim to estimate the clean speech signal captured by the primary microphone in this study.

2.2. Densely-connected convolutional recurrent network

By extending a gated convolutional recurrent network (GCRN) [13] for monaural speech enhancement, we develop a DC-CRN to perform dual-channel complex spectral mapping, which additionally incorporates dense connectivity. As illustrated in Fig. 1, the DC-CRN has an encoder-decoder architecture with skip connections between the encoder and the decoder. In order to compute the input complex spectrograms, we apply short-time Fourier transform (STFT) to the time-domain waveforms of the dual-channel mixtures. The real and imaginary components of the dual-channel spectrograms [11] are concatenated into a 3-dimensional (3-D) representation with four channels. We feed the 3-D representation into a convolutional encoder, which comprises a stack of five convolutional densely-connected (DC) blocks. Subsequently, we reshape the 3-D representation learned by the encoder into a sequence of 1-D features, which is then modeled by a recurrent neural network (RNN). The output of the RNN is reshaped back to a 3-D representation and subsequently passed into a decoder, which is a stack of five deconvolutional DC blocks. The output of the last block is split into two equal-sized 3-D representations along the channel dimension, one for the real spectrum estimation and the other for the imaginary spectrum estimation. We reshape these two 3-D representations individually to a sequence of 1-D features, and then process them with two linear projection layers to estimate the real and imaginary spectrograms of clean speech, respectively. We apply inverse STFT (iSTFT) to the estimated real and imaginary spectrograms to resynthesize the time-domain enhanced signal for the primary channel.



(a) (b)

Fig. 2. Diagrams of the densely-connected block (a) and the gated convolution/deconvolution (b). The symbol $\otimes$ represents the element-wise multiplication, and "BN" and "ELU" denote batch normalization and the exponential linear unit, respectively.

Inspired by U-Net++ [17] for image segmentation, we use a convolutional DC block to process the features learned by each DC block in the encoder, prior to concatenating them with the output of the corresponding DC block in the decoder. The introduction of such DC block based skip pathways can enrich the feature maps from the encoder, which would help to increase the similarity between the feature maps from the encoder and the decoder and thus improves their fusion.

As shown in Fig. 2(a), we propose to use a dense connectivity pattern in each DC block, i.e. introducing direct connections from any layer to all subsequent layers. Such dense connections improve the information flow between layers. Specifically, each DC block has five layers, where each of the first four layers comprises a 2-D convolutional layer successively followed by batch normalization and exponential linear activation function. The last layer is a gated convolutional or deconvolutional layer as depicted in Fig. 2(b), which incorporates gated linear units [18]. Note that "Conv-DC-Block" in Fig. 1 performs gated convolution in the last layer, and "Deconv-DC-Block" gated deconvolution in the last layer.

2.3. Network configurations

To systematically investigate the proposed architecture, we first configure the DC-CRN into a noncausal system with a reasonably large model size. In each convolutional or deconvolutional DC block, each of the first four layers has 8 output channels with a kernel size of 1×3 (time$\times$frequency). For the DC blocks in the encoder and the decoder, the last layer in each of them has a kernel size of 1×4, where a stride of 2 and a zero-padding of 1 (for each side) is applied along the frequency dimension. Note that the kernel size is set to 1×4 rather than 1×3 in order to mitigate the checkerboard artifacts [19], which arise when the kernel size of a strided deconvolution is not divisible by the stride. Moreover, the DC blocks in the encoder have 16, 32, 64, 128 and 256 output channels successively, and those in the

decoder have 256, 128, 64, 32 and 16 output channels successively. The convolutional DC blocks in the skip pathways have the same hyperparameters as those in the encoder, except that the last layer uses a stride of 1 and a kernel size of 1×3. Analogously, these DC blocks have 16, 32, 64, 128 and 256 output channels successively. In addition, the RNN used for sequential modeling is a two-layer bidirectional long short-term memory (BLSTM), of which each layer contains 640 units in each direction. Akin to [13], we adopt a grouping strategy [20] to reduce the number of trainable parameters in the BLSTM without significantly affecting the performance. The number of groups is empirically set to 2.

A causal and lightweight DC-CRN can be easily obtained by simply changing the network configurations. First, we change the number of output channels of all DC blocks to 16, except that the last DC block in the decoder only has 2 output channels. Second, we replace the BLSTM by a two-layer unidirectional LSTM, which has 80 units in each layer. All other settings are the same as in the noncausal DC-CRN.

2.4. Iterative structured pruning

To further reduce the number of trainable parameters, we propose a structured pruning method to compress the causal DC-CRN. We first define the pruning granularity as follow. For each of the convolutional and deconvolutional layers, we treat each kernel (i.e. a 2-D matrix) as a weight group for pruning. In the implementation of LSTM, the weight matrices for four gates (i.e. input, forget, cell and output gates) are typically concatenated, which amounts to two larger matrices, one for the layer input and the other for the hidden state from the last time step. Each column of these matrices is treated as a weight group for pruning. Similarly, we treat each column of the weight matrix of each linear layer as a weight group for pruning. Note that we only prune weights, as the number of biases is trivial relative to that of weights.

In order to increase the compression rate, we use a group sparse regularization technique [21] to impose the group-level sparsity of weight tensors. Specifically, we introduce the following sparse group lasso (SGL) regularizer:

$$\mathcal{R}_{\text{SGL}} = \frac{\lambda_1}{n(\mathcal{W})} \sum_{w\in\mathcal{W}} |w| + \frac{\lambda_2}{n(\mathcal{G})} \sum_{\mathbf{g}\in\mathcal{G}} \sqrt{p_{\mathbf{g}}} \left\|\mathbf{g}\right\|_2, \tag{1}$$

where $\mathcal{W}$ and $\mathcal{G}$ denote the set of all weights and that of all weight groups, respectively. The function $n(\cdot)$ calculates the cardinality of a set, and $\left\|\cdot\right\|_2$ the ℓ_2 norm. The symbol $p_{\mathbf{g}}$ denotes the number of weights in each weight group $\mathbf{g}$. Here λ_1 and λ_2 are predefined weighting factors. Hence, the loss function can be written as $\mathcal{L} = \mathcal{L}_{\text{RI+Mag}} + \mathcal{R}_{\text{SGL}}$, where $\mathcal{L}_{\text{RI+Mag}}$ is the loss function developed in [16].

On a validation set $\mathcal{V}$, the importance of a specific set $\mathcal{U}$ of weight groups can be quantified by the increase in the loss induced by removing it:

$$\mathcal{I}_{\mathcal{U}} = \mathcal{L}_{\text{RI+Mag}}(\mathcal{V}, \Theta|\mathbf{g}=\mathbf{0}, \forall \mathbf{g}\in\mathcal{U}) - \mathcal{L}_{\text{RI+Mag}}(\mathcal{V}, \Theta), \tag{2}$$

where Θ is the set of all trainable parameters in the model, and $\mathcal{U}$ can be any subset of $\mathcal{G}$. We perform a per-tensor sensitivity analysis to determine the pruning ratio for each layer, following Algorithm 1. Subsequently, group-level pruning is applied to each weight tensor as per the tensor-wise pruning ratios. We then fine-tune the pruned model to maintain the enhancement performance. The fine-tuned model is evaluated on the validation set by two metrics, i.e. STOI and PESQ. We repeat this procedure until the number of pruned weights becomes trivial in an iteration or a significant drop in STOI or PESQ is observed on the validation set. During this procedure, the parameter set Θ becomes smaller after each iteration.

Algorithm 1 Per-tensor sensitivity analysis

Input: (1) Validation set $\mathcal{V}$; (2) set $\mathcal{G}_l$ of all nonzero weight groups in the l-th weight tensor $\mathbf{W}_l$, $\forall l$; (3) loss function $\mathcal{L}_{\text{RI+Mag}}(\mathcal{V}, \Theta)$, where Θ is the set of all nonzero trainable parameters in the model; (4) predefined tolerance value α.

Output: Pruning ratio β_l for weight tensor $\mathbf{W}_l$, $\forall l$.

for each tensor $\mathbf{W}_l$ **do**
  for β in $\{0\%, 5\%, 10\%, \ldots, 90\%, 95\%, 100\%\}$ **do**
    Let $\mathcal{U} \subseteq \mathcal{G}_l$ be the set of the $\beta(\%)$ of nonzero weight groups with the smallest ℓ_1 norms in tensor $\mathbf{W}_l$;
    $\mathcal{I}_{\mathcal{U}} \leftarrow \mathcal{L}_{\text{RI+Mag}}(\mathcal{V}, \Theta|\mathbf{g}=\mathbf{0}, \forall \mathbf{g}\in\mathcal{U}) - \mathcal{L}_{\text{RI+Mag}}(\mathcal{V}, \Theta)$;
    if $\mathcal{I}_{\mathcal{U}} > \alpha$ **then**
      $\beta_l \leftarrow \beta - 5\%$;
      break
    end if
  end for
  if β_l is not assigned any value **then**
    $\beta_l \leftarrow 100\%$;
  end if
end for
return β_l for weight tensor $\mathbf{W}_l$, $\forall l$

3. EVALUATION AND ANALYSIS

3.1. Experimental setup

In the experiments, we simulate a rectangular room with a size of $10\times7\times3$ m^3 using the image method [22]. The target source (mouth) is located at the center of the room, and the primary microphone is placed on a sphere centered at the target source, of which the radius is randomly sampled between 0.01 m and 0.15 m. Such a distance range covers both of two mobile phone use scenarios, i.e. hand-held and hands-free scenarios. The geometry of the dual-channel microphone array is fixed, and the distance between microphones is set to 0.1 m. Thus the location of the secondary microphone is randomly selected on a sphere with a radius of 0.1 m, which is centered at the primary microphone. We randomly sample the reverberation time (T_{60}) between 0.2 s and 0.5 s. Following this procedure, we simulate a set of 5000 dual-channel room impulse responses (RIRs) for training and validation data, and another set of 846 dual-channel RIRs for test data. We use the training set of the WSJ0 dataset [23] as the speech corpus, which consists of 12776 utterances from 101 speakers. These speakers are split into three groups (i.e. 89, 6 and 6) for generating training, validation and test data, respectively.

Following [24], we simulate a diffuse babble noise field. Specifically, we first concatenate the utterances spoken by each of the 630 speakers in the TIMIT corpus [25], and then split them into 480 and 150 speakers for training and testing. We randomly choose 72 speech clips from 72 randomly chosen speakers, and place them on a horizontal circle centered at the primary microphone, where the azimuths range from 0° to 355° with a step of 5°. The distance between the primary microphone and each of the interfering sources is 2 m.

We simulate a training set including 40000 mixtures, each of which is created by mixing a diffuse babble noise and a randomly sampled WSJ0 utterance convolved with a randomly selected RIR. The signal-to-noise ratio (SNR) is randomly sampled between -5 and 0 dB. Similarly, we create a validation set consisting of 846 mixtures. For each of four SNRs, i.e. -5, 0, 5 and 10 dB, we create a test set including 846 mixtures. In order to mimic the head shadow effect in hand-held scenarios, we downscale the amplitude of the speech

Table 1. Comparisons of alternative models in STOI and PESQ. Here ✓ indicates causal model, and ✗ indicates noncausal model.

Test SNR	-5 dB		0 dB		5 dB		10 dB		# Param.	Causal
Metric	STOI (%)	PESQ	STOI (%)	PESQ	STOI (%)	PESQ	STOI (%)	PESQ		
Unprocessed	58.71	1.49	72.08	1.73	83.53	2.04	91.41	2.38	-	-
NC-CRN-PSM	85.48	2.20	90.79	2.60	93.82	2.93	95.47	3.17	12.99 M	✗
NC-DC-CRN-RI	**92.77**	**3.07**	**96.09**	**3.41**	**97.66**	**3.63**	**98.45**	**3.78**	8.36 M	✗
IRM	92.02	2.83	94.21	3.10	96.24	3.39	97.74	3.68	-	-
PSM	94.08	3.16	96.26	3.40	97.87	3.66	98.87	3.88	-	-
C-CRN-PSM	78.77	1.76	86.80	2.18	91.53	2.56	94.05	2.88	73.15 K	✓
C-DC-CRN-RI	**87.57**	**2.56**	**93.36**	**2.99**	**96.35**	**3.30**	**97.74**	**3.53**	290.44 K	✓
C-DC-CRN-RI-P1	86.88	2.54	93.08	2.97	96.16	3.26	97.63	3.46	124.96 K	✓
C-DC-CRN-RI-P2	87.13	2.56	93.10	2.98	96.14	3.27	97.62	3.47	113.68 K	✓
C-DC-CRN-RI-P3	86.64	2.52	92.89	2.95	96.07	3.26	97.61	3.47	108.77 K	✓
C-DC-CRN-RI-P4	86.63	2.49	92.85	2.91	96.03	3.22	97.59	3.44	106.21 K	✓
C-DC-CRN-RI-P5	86.63	2.48	92.86	2.90	96.07	3.20	97.65	3.43	104.76 K	✓
C-DC-CRN-RI-P6	86.45	2.51	92.64	2.94	95.88	3.27	97.47	3.51	103.07 K	✓

signal at the secondary channel prior to mixing, where the downscaling ratio is randomly sampled between -10 and 0 dB. Such a data simulation method accounts for various ways of holding a mobile phone, which is more robust than using close-talk inter-channel relative transfer functions [8].

The models are trained on 4-second segments using the AMSGrad optimizer [26] with a minibatch size of 16. The learning rate is initialized to 0.001, which decays by 0.98 every two epochs. The validation set is used for both selecting the best model among different epochs and performing the pruning sensitivity analysis. Other training configurations are the same as [8]. For structured pruning, the initial values of λ_1 and λ_2 (see Eq. (1)) are empirically set to 1 and 0.1, both of which decay by 10% every pruning iteration. The tolerance value α for sensitivity analysis is set to 0.02.

3.2. Experimental results

Table 1 shows comprehensive comparisons among alternative models in terms of STOI and PESQ, in which the numbers represent the averages over the test set in each condition. The proposed models with noncausal and causal DC-CRNs are represented by "NC-DC-CRN-RI" and "C-DC-CRN-RI", respectively. The pruned DC-CRN model for the k-th iteration is denoted as "C-DC-CRN-RI-Pk". In addition, "C-CRN-PSM" represents the approach in [8], and "NC-CRN-PSM" a noncausal version of it. In the noncausal CRN, the numbers of output channels for the layers in the encoder are changed to 16, 32, 64, 128 and 256 successively, and those for each layer in the decoder to 128, 64, 32, 16 and 1 successively. The two-layer LSTM is replaced by a two-layer BLSTM, of which each layer contains 512 units in each direction.

We can see that that our proposed approach substantially outperforms the approach in [8] in terms of both STOI and PESQ. At -5 dB SNR, for example, "NC-DC-CRN-RI" yields a 7.6% STOI improvement and a 0.89 PESQ improvement over "NC-CRN-PSM". Similar improvements are observed for "C-DC-CRN-RI" over "C-CRN-PSM". Moreover, we compare the proposed approach with two ideal masks, i.e. the PSM and the ideal ratio mask (IRM). As presented in Table 1, our noncausal enhancement system ("NC-DC-CRN-RI") produces higher STOI and PESQ than the IRM, and slightly lower STOI and PESQ than the PSM. In addition, our pruning method substantially compress the model size without significantly sacrificing the performance. As shown in Table 1, the causal DC-CRN originally has 290.44 K parameters. After 6 iterations of pruning, the number of parameters in the DC-CRN is reduced to 103.07 K, which is comparable to that of the CRN in [8], i.e. 73.15 K.

In Table 2, we conduct an ablation study to examine the effects of dense connectivity in the DC-CRN. Three variants of the causal DC-CRN are created: (i) replacing the DC block based skip pathways by skip connections as in [8]; (ii) replacing each DC block in the encoder and the decoder by a corresponding gated convolutional or deconvolutional layer, as in [13]; (iii) doing both (i) and (ii). We can observe that all these variants underperform the proposed causal DC-CRN, which suggests the effectiveness of dense connectivity. For example, STOI decreases by 1.31% and PESQ by 0.14, when dense connectivity in the encoder and the decoder is removed.

Table 2. Effects of dense connectivity at -5 dB SNR.

Test SNR	-5 dB			# Param.
Metric	STOI (%)	PESQ	SNR (dB)	
Unprocessed	58.71	1.49	-5.03	-
C-DC-CRN-RI	**87.57**	**2.56**	**8.61**	290.44 K
− DC_{Skip} (i)	87.23	2.53	8.49	253.32 K
− DC_{ED} (ii)	86.26	2.42	8.02	218.69 K
− DC_{Skip} − DC_{ED} (iii)	82.77	2.10	6.37	181.57 K

Table 3. Investigation of inter-channel features for magnitude- and complex-domain approaches. "ICFs" represent the inter-channel features.

Test SNR	-5 dB			Domain
Metric	STOI (%)	PESQ	SNR (dB)	
Unprocessed	58.71	1.49	-5.03	-
C-CRN-PSM w/ ICFs	78.77	1.76	5.13	Magnitude
C-CRN-PSM w/o ICFs	76.14	1.67	4.56	Magnitude
C-DC-CRN-RI w/ ICFs	87.64	2.56	8.44	Complex
C-DC-CRN-RI w/o ICFs	87.44	2.56	8.61	Complex

We now investigate the inclusion of inter-channel features for both magnitude- and complex-domain approaches. As shown in Table 3, the inclusion of inter-channel features significantly improves STOI and PESQ for the magnitude-domain approaches. For the complex-domain approach (i.e. our approach), we use the STFTs of the inter-channel noisy signal difference and summation as the inter-channel features, similar to the approach in [8]. With multi-channel complex spectral mapping, the explicit use of these inter-channel features does not produce performance gain, as shown in Table 3. This suggests that inter-channel features can be captured implicitly through a DNN that is trained for multi-channel complex spectral mapping, which is consistent with [16] for speech dereverberation.

4. CONCLUSION

We have proposed a novel framework for real-time speech enhancement on dual-microphone mobile phones. The framework employs a causal DC-CRN to perform dual-channel complex spectral mapping, which leverages both spectral and spatial cues in dual-channel complex-domain inputs. In addition, we apply iterative structured pruning to the DC-CRN, which yields a low-latency and memory-efficient enhancement system that is amenable to real-time processing on mobile phones. Evaluation results show that the proposed approach substantially outperforms an earlier approach to dual-channel speech enhancement for mobile phones, in terms of both STOI and PESQ.

5. REFERENCES

[1] N. Yousefian, A. Akbari, and M. Rahmani, "Using power level difference for near field dual-microphone speech enhancement," *Applied Acoustics*, vol. 70, no. 11-12, pp. 1412–1421, 2009.

[2] M. Jeub, C. Herglotz, C. Nelke, C. Beaugeant, and P. Vary, "Noise reduction for dual-microphone mobile phones exploiting power level differences," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2012, pp. 1693–1696.

[3] Z.-H. Fu, F. Fan, and J.-D. Huang, "Dual-microphone noise reduction for mobile phone application," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7239–7243.

[4] J. Zhang, R. Xia, Z. Fu, J. Li, and Y. Yan, "A fast two-microphone noise reduction algorithm based on power level ratio for mobile phone," in *8th International Symposium on Chinese Spoken Language Processing*. IEEE, 2012, pp. 206–209.

[5] Y.-Y. Chen, "Speech enhancement of mobile devices based on the integration of a dual microphone array and a background noise elimination algorithm," *Sensors*, vol. 18, no. 5, pp. 1467, 2018.

[6] I. López-Espejo, J. A. González, Á. M. Gómez, and A. M. Peinado, "A deep neural network approach for missing-data mask estimation on dual-microphone smartphones: application to noise-robust speech recognition," in *Advances in Speech and Language Technologies for Iberian Languages*, pp. 119–128. Springer, 2014.

[7] I. López-Espejo, A. M. Peinado, A. M. Gomez, and J. M. Martín-Doñas, "Deep neural network-based noise estimation for robust ASR in dual-microphone smartphones," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 117–127.

[8] K. Tan, X. Zhang, and D. L. Wang, "Real-time speech enhancement using an efficient convolutional recurrent network for dual-microphone mobile phones in close-talk scenarios," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 5751–5755.

[9] K. Tan and D. L. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, 2018, pp. 3229–3233.

[10] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2015, pp. 708–712.

[11] D. S. Williamson, Y. Wang, and D. L. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[12] S.-W. Fu, T.-y. Hu, Y. Tsao, and X. Lu, "Complex spectrogram enhancement by convolutional neural network with multi-metrics learning," in *IEEE 27th International Workshop on Machine Learning for Signal Processing*. IEEE, 2017, pp. 1–6.

[13] K. Tan and D. L. Wang, "Learning complex spectral mapping with gated convolutional recurrent networks for monaural speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 380–390, 2020.

[14] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[15] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (Cat. No. 01CH37221)*. IEEE, 2001, vol. 2, pp. 749–752.

[16] Z.-Q. Wang and D. L. Wang, "Multi-microphone complex spectral mapping for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 486–490.

[17] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11. Springer, 2018.

[18] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *34th International Conference on Machine Learning*, 2017, vol. 70, pp. 933–941.

[19] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi, "Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize," *arXiv preprint arXiv:1707.02937*, 2017.

[20] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 799–808.

[21] S. Scardapane, D. Comminiello, A. Hussain, and A. Uncini, "Group sparse regularization for deep neural networks," *Neurocomputing*, vol. 241, pp. 81–89, 2017.

[22] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[23] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) complete LDC93S6A," *Web Download. Philadelphia: Linguistic Data Consortium*, vol. 83, 1993.

[24] X. Zhang and D. L. Wang, "Deep learning based binaural speech separation in reverberant environments," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 1075–1084, 2017.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *STIN*, vol. 93, pp. 27403, 1993.

[26] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of adam and beyond," in *International Conference on Learning Representations*, 2018.