Real-Time Speech Enhancement Using An Efficient Convolutional Recurrent Network for Dual-Microphone Mobile Phones in Close-Talk Scenarios

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

[1]The Ohio State University, USA

[2]Inner Mongolia University, China

OUTLINE

1. Background and Motivations
2. Algorithm Description
3. Experiments
4. Conclusion

OUTLINE

1. Background and Motivations

2. Algorithm Description

3. Experiments

4. Conclusion

Dual-Microphone Mobile Phones

- Mobile speech communication has become an increasingly important application for speech enhancement. In an adverse acoustic environment, speech quality and intelligibility can be severely degraded by background noise.

- We focus on speech enhancement for a typical dual-microphone configuration in close-talk scenarios, where a speech signal is picked up with small distance between the primary microphone and the human mouth.

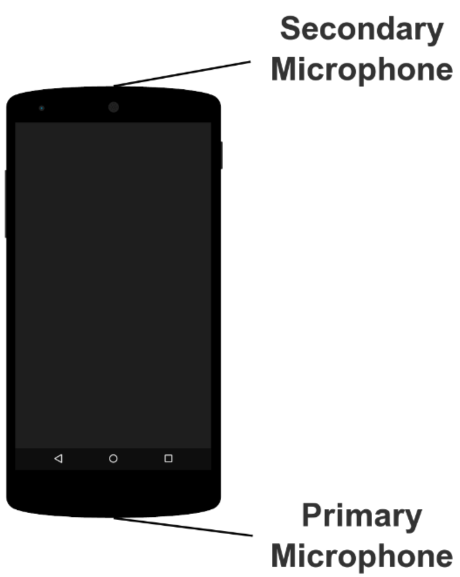

Figure 1: Illustration of a dual-microphone mobile phone.

- In recent studies [1] [2], deep neural networks (DNNs) have been used to perform speech enhancement for dual-microphone mobile phones.

- The experimental results show that the DNN-based approaches significantly outperform several representative traditional algorithms.

[1] I. López-Espejo, et al., "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.
[2] I. López-Espejo, et al., "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.

- Motivated by our recent study [3] on convolutional recurrent networks (CRNs), we propose a novel framework for dual-microphone speech enhancement on mobile phones.

- The proposed CRN model is a causal system. Moreover, the CRN is computationally efficient, and thus is amenable to mobile phone applications.

- The proposed approach substantially outperforms a DNN-based method similar to [1], as well as two traditional methods for speech enhancement.

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

OUTLINE

Problem Formulation

- Let $y_m(k)$, $s_m(k)$ and $n_m(k)$ denote noisy speech, clean speech and background noise, respectively, where m is the channel index.

- The dual-channel signals can be modeled as

$$y_1(k) = s_1(k) + n_1(k) = s(k) + n_1(k)$$
$$y_2(k) = s_2(k) + n_2(k) = s(k) * h_{12}(k) + n_2(k)$$

where $h_{12}(k)$ represents the acoustic impulse response from the primary channel to the secondary channel.

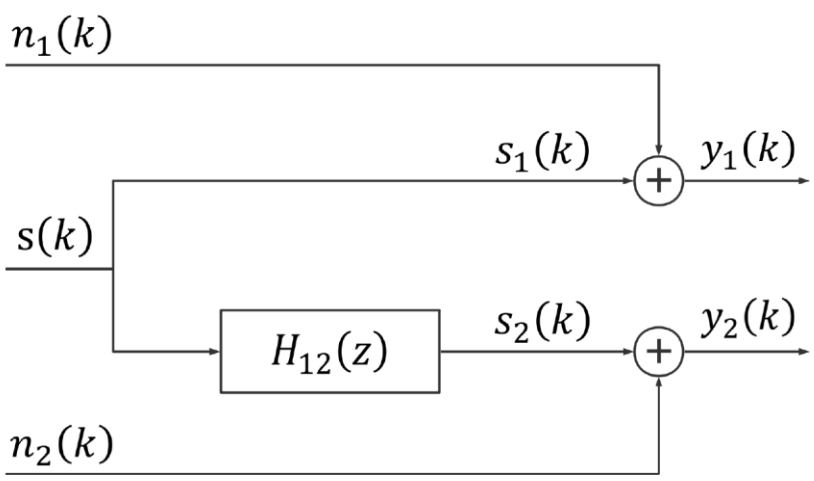

Figure 2: Illustration of the dual-channel signal model.

Intra-Channel and Inter-Channel Features

- Let Y_1 and Y_2 be the short-time Fourier transform (STFT) of the noisy speech signals at the primary channel and the secondary channel, respectively.

- The intra-channel features, i.e. $|Y_1|$ and $|Y_2|$, do not account for inter-channel correlations.

- Hence, the inter-channel features, i.e. $|Y_1 - Y_2|$ and $|Y_1 + Y_2|$ are additionally included, which implicitly incorporate phase correlations between channels.

- The intra-channel and inter-channel features are treated as four different input channels of the CRN.

Training Target

- In this study, we use the phase-sensitive mask (PSM) as the training target, which incorporates the phase information. It is typically defined as [4]

$$PSM(t,f) = Re\left\{\frac{|S_1(t,f)|\exp(j\theta_{s_1})}{|Y_1(t,f)|\exp(j\theta_{y_1})}\right\} = \frac{|S_1(t,f)|}{|Y_1(t,f)|}\cos(\theta_{s_1} - \theta_{y_1})$$

  where $Re\{\cdot\}$ computes the real component.

- Once the PSM is estimated, we apply it to the magnitude spectrogram of noisy speech at the primary channel.

[4] 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 (ICASSP). IEEE, 2015, pp. 708–712.

Training Target

- Based on the analysis of the acoustical environment in [5], we assume that the power level difference (PLD) between the clean speech signals at the two channels is larger than that between the noise signals.

- Hence, the noisy signal difference between channels, i.e. $y_1 - y_2$, may have a higher signal-to-noise ratio (SNR) than y_1, and thus have a cleaner phase.

- We propose to combine the phase of $y_1 - y_2$ with the estimated magnitude to resynthesize waveforms. Thus the PSM should be redefined as

$$PSM(t,f) = Re\left\{\frac{|S_1(t,f)|\exp(j\theta_{s_1})}{|Y_1(t,f)|\exp(j\theta_{y_1-y_2})}\right\} = \frac{|S_1(t,f)|}{|Y_1(t,f)|}\cos(\theta_{s_1} - \theta_{y_1-y_2})$$

[5] 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 (ICASSP). IEEE, 2012, pp. 1693–1696.

- We have recently developed a convolutional recurrent network (CRN) for real-time speech enhancement [6].

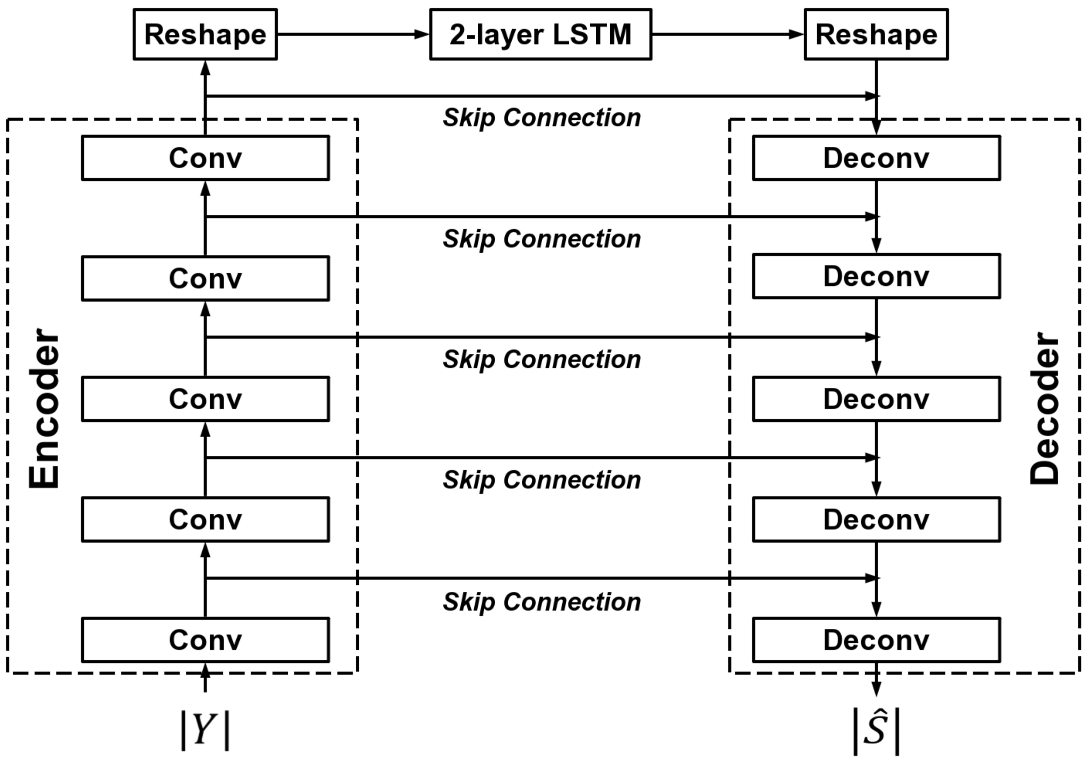

Figure 3: A convolutional recurrent network for real-time speech enhancement

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

OUTLINE

1. Background and Motivations

2. Algorithm Description

3. Experiments

4. Conclusion

- Corpus: WSJ0 SI-84, including 7138 utterances from 83 (= 77 + 6) speakers.

- We consider the target clean speech to be the same as the clean speech signal picked up by the primary microphone ($s_1 = s$). The clean speech signal at the secondary microphone is generated by the acoustic path h_{12} from the primary channel to the secondary channel ($s_2 = s * h_{12}$).

- The acoustic path h_{12} is modeled as a time-invariant finite impulse response (FIR) filter, of which the coefficients are estimated by minimizing the mean squared error (MSE), i.e. $\mathbb{E}[e^2(k)]$, where

$$e(k) = s_2^{(rec)}(k) - \sum_{l=0}^{p} \hat{h}_{12}(l) s_1^{(rec)}(k-l)$$

Here $s_1^{(rec)}$ and $s_2^{(rec)}$ are clean speech signals recorded by a dual-microphone mobile phone that is mounted on a dummy head in an anechoic environment.

- We use 6 different mobile phones: 6 different inter-channel acoustic paths (five for training, one for testing).

- Two different noise fields: diffuse noise and point-source noise.

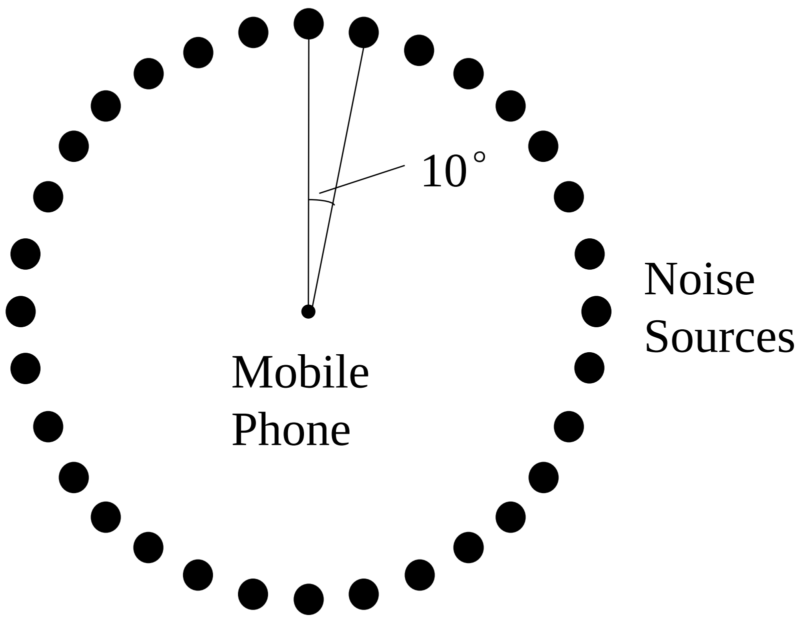

Figure 4: Simulation of diffuse noise.

Experimental Setup

- Training: 10,000 noises from a sound effect library. The SNRs are randomly sampled from {-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5} dB. We create 320,000 mixtures in total.

- Testing: babble and cafeteria noises. SNRs: -5, 0, 5 and 10 dB. We create 150 (= 25 $\times$ 6) mixtures for each SNR.

- In close-talk scenarios, the direct-to-reverberant ratio (DRR) of the speech signal is high, so that the reverberation from it can be omitted.

Table 1: Comparisons of different approaches for diffuse noise.

metrics	STOI (in %)				PESQ			
SNR	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
noisy	57.58	69.66	80.71	89.19	1.49	1.77	2.09	2.43
MMSE	52.88	65.45	76.67	85.74	1.48	1.81	2.15	2.45
MS	54.30	67.05	79.05	87.84	1.49	1.83	2.17	2.47
DNN	80.80	87.07	91.81	95.00	2.18	2.54	2.87	3.18
Prop.	**92.52**	**94.95**	**96.66**	**97.88**	**2.89**	**3.20**	**3.48**	**3.70**

Table 2: Comparisons of different approaches for point-source noise.

metrics	STOI (in %)				PESQ			
SNR	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
noisy	57.65	69.82	80.87	89.27	1.51	1.77	2.09	2.42
MMSE	53.08	65.47	76.63	85.83	1.50	1.83	2.15	2.45
MS	54.35	67.42	79.29	87.87	1.51	1.83	2.16	2.45
DNN	80.49	87.04	91.82	95.03	2.16	2.53	2.87	3.18
Prop.	**91.81**	**94.68**	**96.54**	**97.83**	**2.85**	**3.17**	**3.45**	**3.68**

MMSE: minimum mean squared error based speech enhancement
MS: minimum statistics
DNN: three hidden layers, (3+1)×161×2, 64, 64, 64, 161
CRN (Prop.): encoder, LSTM, decoder

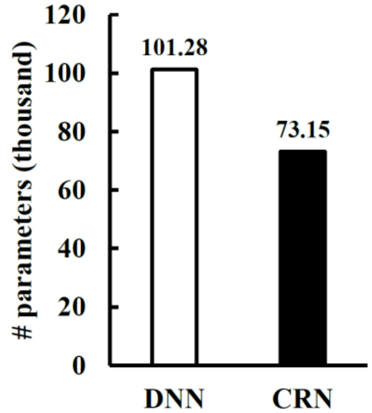

Figure 4: The number of trainable parameters (unit: thousand).

Table 3: Evaluation of the inter-channels features and the phase of noisy signal difference between channels.

metrics	STOI (in %)				PESQ			
SNR	-5 dB	0 dB	5 dB	10 dB	-5 dB	0 dB	5 dB	10 dB
noisy	57.62	69.74	80.79	89.23	1.50	1.77	2.09	2.43
(i)	83.67	89.00	93.04	95.79	2.38	2.71	3.02	3.32
(ii)	86.75	91.36	94.65	96.84	2.56	2.88	3.21	2.50
(iii)	88.96	92.44	95.02	96.85	2.65	2.97	3.25	3.50
(iv)	**92.17**	**94.82**	**96.60**	**97.86**	**2.87**	**3.19**	**3.47**	**3.69**

(i) intra-channel features + the phase of y_1;

(ii) both intra-channel and inter-channel features + the phase of y_1;

(iii) intra-channel features + the phase of $y_1 - y_2$;

(iv) both intra-channel and inter-channel features + the phase of $y_1 - y_2$.

Demos - Diffuse Noise

- Babble diffuse noise, -5 dB
  untrained female speaker:
  - Unprocessed (dual channels):
  - Unprocessed (primary channel):
  - MMSE:
  - MS:
  - DNN:
  - CRN (Prop.):
  - Clean:

Demos – Point-Source Noise

- Cafeteria point-source noise, -5 dB untrained male speaker:
  - Unprocessed (dual channels):
  - Unprocessed (primary channel):
  - MMSE:
  - MS:
  - DNN:
  - CRN (Prop.):
  - Clean:

OUTLINE

1. Background and Motivations

2. Algorithm Description

3. Experiments

4. Conclusion

Conclusion

- We have proposed a new deep learning based framework for real-time speech enhancement on dual-microphone mobile phones in a close-talk scenario.
- The proposed framework incorporates a computationally efficient CRN, which is trained from both intra-channel and inter-channel features.
- In addition, we propose to use the phase of noisy signal difference between channels to resynthesize the waveform.
- The experimental results show that the proposed approach consistently outperforms a DNN-based method, as well as two traditional speech enhancement methods.