Towards Real-Time Single-Channel Speech Separation in Noisy and Reverberant Environments

Julian Neri[1] Sebastian Braun[2]

[1]McGill University, CIRMMT, Montréal, QC, Canada
[2]Microsoft Research, Redmond, WA, USA

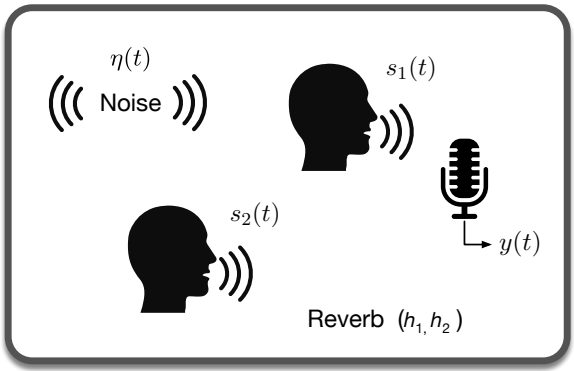
June 8, 2023

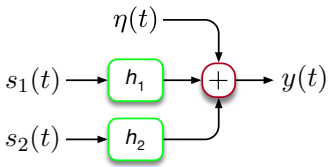

McGill

CIRMMT Centre for Interdisciplinary Research in Music Media and Technology

Introduction - problem

Given a single-channel recording of people talking in a reverberant and noisy space.





Retrieve separate, de-reverberated and noise-free signals of each talker, s_1, s_2.
Applications: re-mixing, hearing aids, dictation, and augmented reality.

Introduction - review

Offline

No constraint on causality, model size, or computational complexity.

- Seen immense progress with deep learning, and is mature enough for separating speech with high quality.

Examples: [Luo and Mesgarani, 2019] [Luo et al., 2020] [Subakan et al., 2021] [Cord-Landwehr et al., 2022]

Real-time

Must be causal with minimal "look-ahead", compact, and efficient.

- Harder than non-real-time because models must be compact, and only use current and past signal information.

Introduction - overview of methodology

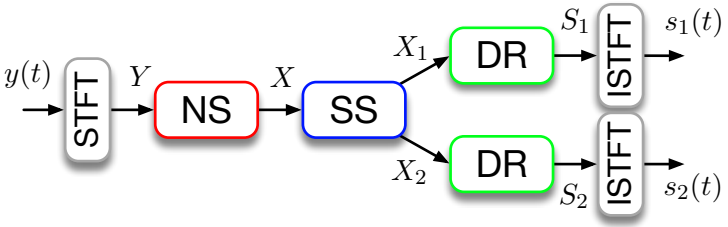
We propose real-time speech separation deep neural network architectures that are

- ✓ causal,
- ✓ resource-efficient,
- ✓ order of magnitude cheaper than current state-of-the-art (SOTA),
- ✓ and require only a small look-ahead.

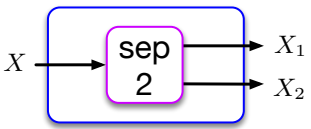
For training and testing, we

- explore different loss functions and soft thresholds,
- test models on real recordings,
- and propose a new non-intrusive channel separation estimate.

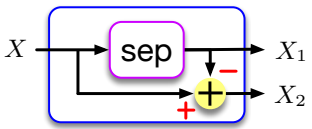
Methodology - modular system architecture



- $x(t)$: reverberant speech mixture; noise suppressed version of $y(t)$
- $x_r(t)$: reverberant speech source r



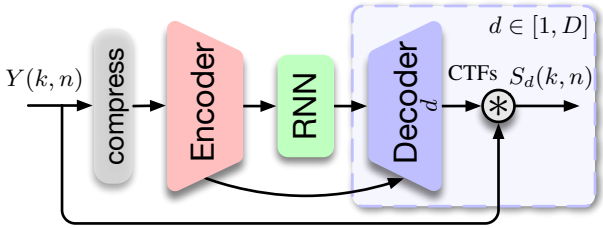
(a) Multi-decoder



(b) Subtractive (SUB)

Methodology - system architecture

Modules are multi-decoder generalizations of the convolutional recurrent U-net (CRUSE) [Braun et al., 2021].



- Y: input short-time Fourier transform (STFT)
- D: number of decoders
- S_d : output STFT of decoder d
- CTFs: convolutive transfer functions

Loss functions - overview

Utterance-level permutation invariant training (uPIT) [Kolbæk et al., 2017]:

$$\mathcal{L}_{\text{uPIT}} = \min \left(\mathcal{L}\left(\boldsymbol{S}_1, \widehat{\boldsymbol{S}}_1\right) + \mathcal{L}\left(\boldsymbol{S}_2, \widehat{\boldsymbol{S}}_2\right), \mathcal{L}\left(\boldsymbol{S}_1, \widehat{\boldsymbol{S}}_2\right) + \mathcal{L}\left(\boldsymbol{S}_2, \widehat{\boldsymbol{S}}_1\right) \right) . \tag{1}$$

Baseline $\mathcal{L}$: scale-invariant signal-to-distortion ratio (SISDR) [Roux et al., 2019], the SOTA distance metric.

Loss functions - distance metric and soft threshold

The **complex compressed MSE (CCMSE)** $\mathcal{L}$ between target source r and estimate d, with compression c and weight λ, is

$$\mathcal{L}(\boldsymbol{S}_r, \widehat{\boldsymbol{S}}_d) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{n=1}^{N} (1-\lambda) \Big| |S_r(k,n)|^c - |\widehat{S}_d(k,n)|^c \Big|^2 + \lambda \Big| |S_r(k,n)|^c e^{j\Phi_{S_r(k,n)}} - |\widehat{S}_d(k,n)|^c e^{j\Phi_{\widehat{S}_d(k,n)}} \Big|^2 . \quad (2)$$

Soft threshold CCMSE with $\tau \in \mathbb{R}_{\geq 0}$,

$$\mathcal{L}_\tau(\boldsymbol{S}_r, \widehat{\boldsymbol{S}}_d) = 10 \log_{10} \left(\mathcal{L}(\boldsymbol{S}_r, \widehat{\boldsymbol{S}}_d) + \tau \right) . \quad (3)$$

Datasets and settings

Train and validate

- Generated mixtures of 10 s duration at 16 kHz.
- 700 h speech, 246 h noise, 128k RIRs simulated in 2000 rooms.
- Training data is made "on-the-fly".
- Validate speech corpora different from train.

Test

- REAL-M [Subakan et al., 2022] - 1436 real-world single-microphone recordings of two talkers in different acoustic environments from laptops, smartphones, etc.

STFT 20 ms window, 50% overlap, $N_{\mathsf{FFT}} = 320$.
Network input: 161 power-compressed complex frequency components.

Metrics and baseline

DNSMOS [Reddy et al., 2022]: non-intrusive estimator of mean opinion score (MOS) for signal quality (SIG), background noise (BAK) and overall (OVR).

We propose a novel, non-intrusive metric, the **channel separation estimate (CSE)**,

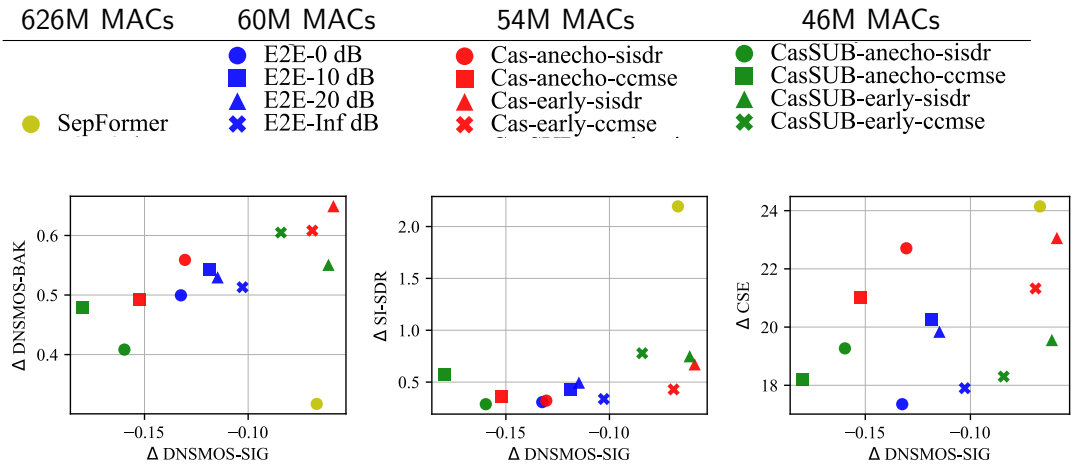
$$\text{CSE} = -20 \log_{10} \frac{|\widehat{\boldsymbol{s}}_r^{\mathsf{T}} \widehat{\boldsymbol{s}}_{r'}|}{\|\widehat{\boldsymbol{s}}_r\|^2 + \|\widehat{\boldsymbol{s}}_{r'}\|^2} \,. \tag{4}$$

Multiply-accumulate (MAC) operations per 10 ms of audio.

Baseline models:

- SepFormer [Subakan et al., 2021] large offline speech separator
- E2E - end-to-end version, similar complexity to cascade (CAS) version.

Results


626M MACs
60M MACs
54M MACs
46M MACs
SepFormer
E2E-0 dB
E2E-10 dB
E2E-20 dB
E2E-Inf dB
Cas-anecho-sisdr
Cas-anecho-ccmse
Cas-early-sisdr
Cas-early-ccmse
CasSUB-anecho-sisdr
CasSUB-anecho-ccmse
CasSUB-early-sisdr
CasSUB-early-ccmse
Δ DNSMOS-BAK
Δ DNSMOS-SIG
Δ SI-SDR
Δ DNSMOS-SIG
Δ CSE
Δ DNSMOS-SIG

Results - ablation for cascaded modules

Stage	NS	NS + SS	NS + SS + DR	
			early	anechoic
Δ DNSMOS-SIG	0.17	-0.08	-0.07	-0.15
Δ DNSMOS-BAK	0.50	0.52	0.61	0.49
Δ DNSMOS-OVR	0.32	0.11	0.17	-0.04
Δ CSE	0.00	20.31	21.33	21.02

Table: Output improvement from each stage of a cascade model trained with CCMSE.

Results - balancing weaker signals with soft threshold loss

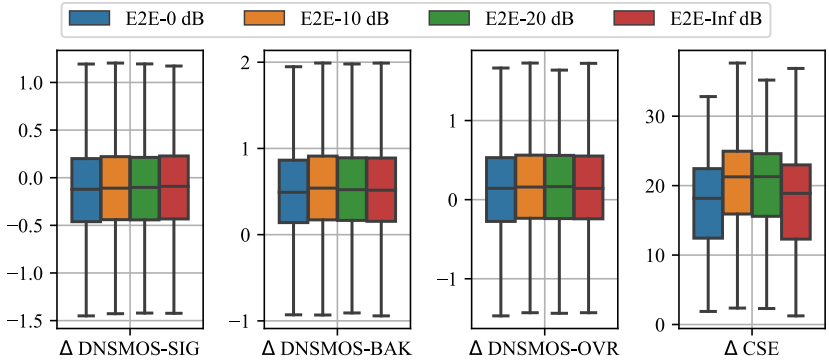
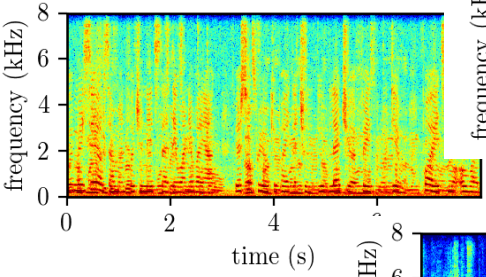


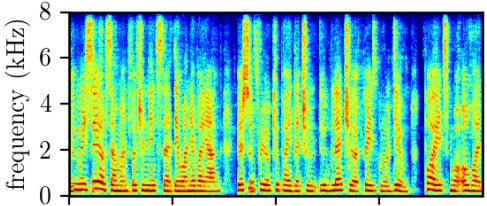
Figure: Result of using different soft thresholds τ in dB when training the E2E model using CCMSE loss on early reflection targets.

Results - audio example - DNS2022

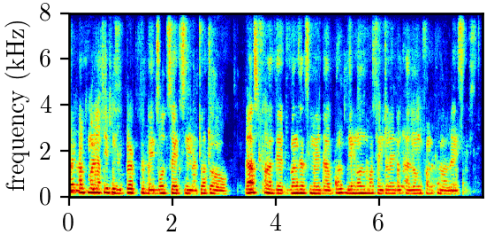
Mixture



Cascade SUB 1

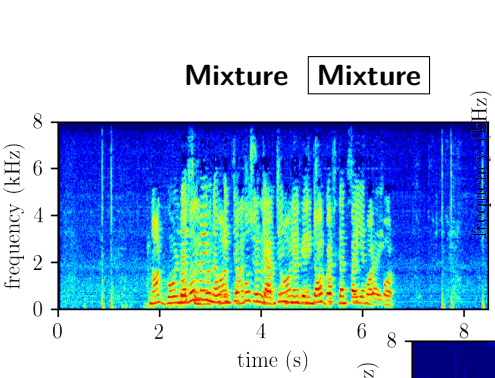


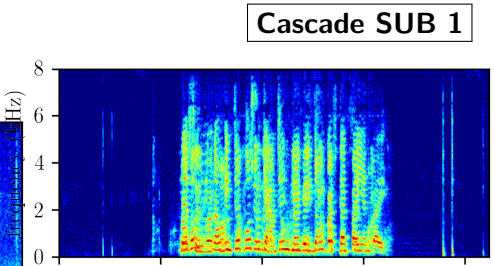
Cascade SUB 2

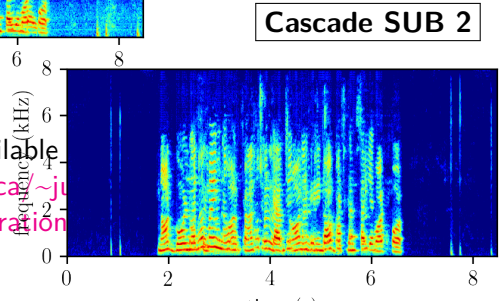


Audio examples are ava
https://www.music.mcgill.
real-time-speech-sepa

Results - audio example - REAL-M







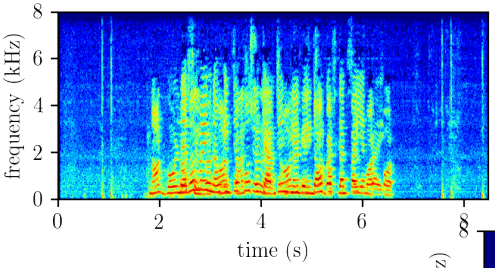
Audio examples are available

https://www.music.mcgill.ca/~ju

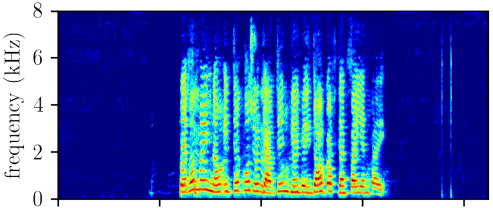
real-time-speech-separation

Results - audio example - REAL-M

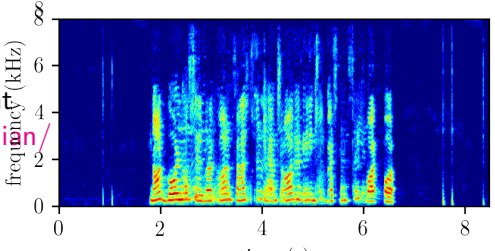
Mixture **Mixture**



Cascade 1



Cascade 2



Audio examples are available at
https://www.music.mcgill.ca/~julian/real-time-speech-separation

Conclusion & Future Work

Conclusions

- task-splitting noise suppression (NS), reverberant speech separation (SS), and de-reverberation (DR) is more efficient than the end-to-end model in terms of separation and speech quality
- including early reflections in target outputs leads to better signal quality
- Subtractive separation is most efficient
- -10 dB threshold $\rightarrow$ better separation

Future work

- listening study to measure subjective quality on real recordings
- evaluate CSE metric with a listening study
- apply subtractive separation to recursively separate several talkers