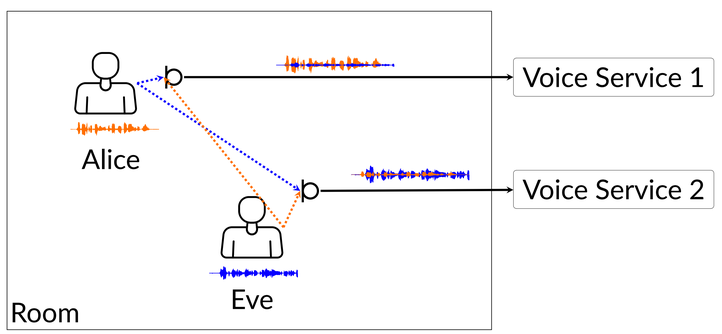
Privacy and Quality Improvements in Open Offices Using Multi-Device Speech Enhancement

Silas Rech Mohammad Hassan Vali Tom Bäckström

Aalto University
School of Electrical
Engineering

Department of Information and Communications Engineering, Aalto University, Finland

1. Problem



- **Problem**: Independent conversations can leak into microphones
- **Challenge:** Remove non-relevant speakers without prior knowledge
- **Assumption:** The microphone of a person is closer than others
- **Idea:** Consider the microphone closest to a speaker as "clean" signal

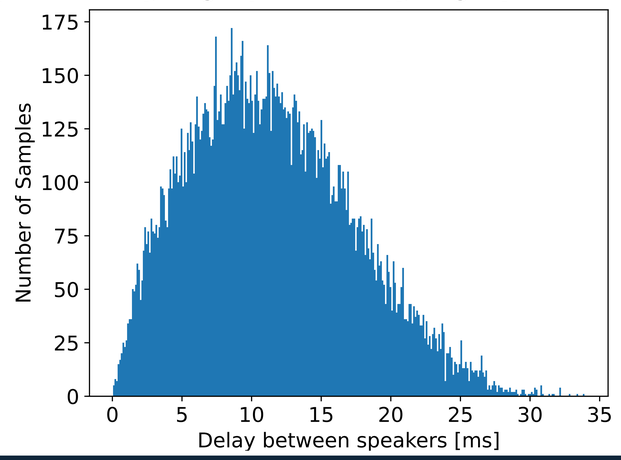
2. Signal Model

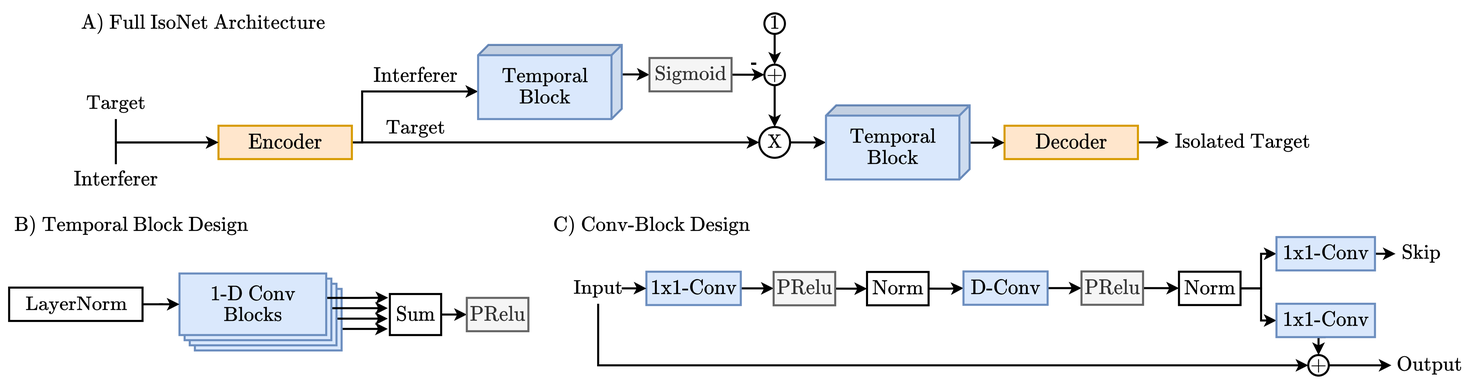
$$x_{\text{mix}}(n) = \sum_{c=1}^{N_s} s_c(n) + s_n(n) \qquad s_c(n) = x_c(n) * h_c(n)$$

- $s_c(n)$ c-th speech source
- $x_c(n)$ $c-th$ clean audio signal
- $s_n(n)$ non-speech background noise
- $h_c(n)$ the impulse response

3. Database

- LibriSpeech Dev-Clean with $fs = \mathbf{16\,kHz}$
- Simulate office scenarios with PyRoom Acoustics package
- Width [5 m to 10 m], length [5 m to 10 m], height [2.5 m to 5 m]





4. IsoNet Architecture

- Encoder, mask-generator, enhancer, and decoder
- Input: Two audio signals
- Output: Single channel enhanced audio
- Loss function:

$$\text{SI-SNR} := 10 \cdot \log_{10} \frac{||s_{target}||^2}{||e_{noise}||^2},$$

$s_{target} := \frac{(\hat{s}^T s)s}{||s||^2}$. s is the reference and $\hat{s}$ is the estimated isolated signal. $e_{noise} := \hat{s} - s_{target}$ and $||\cdot||^2$ is the power of the signal.

5. Results

Method	PESQ	△ SI-SNR	△ SDR	Size
WaveFilter	-	-	10.45	5.9 M*
PercepNet	2.412	-	-	8.5/26.5 M
VoiceFilter	-	12.6	-	18.8 M*
Conv-TasNet	3.22	12.2*	12.7*	5.1 M
SepFormer	-	16.5*	17*	26 M
TDANet	-	17.4*	17.9*	2.3 M
IsoNet(ours)	3.7	18.6	14.1	3.7 M

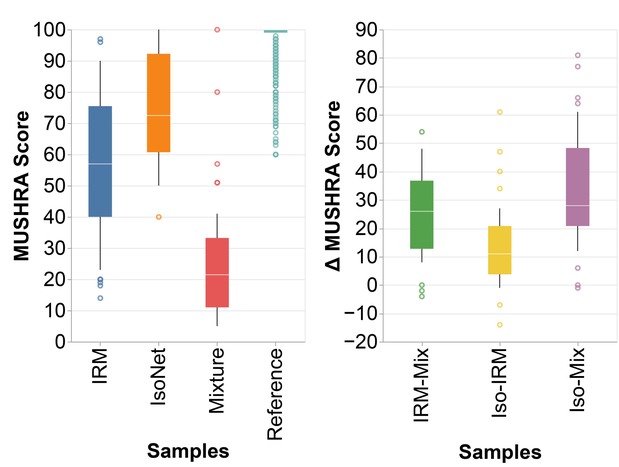
∗ describes baseline models which are trained on the Libri2Mix Dataset. The reported values are the original ones from the paper, as there was no public model available for retraining.

- Mutual information (MI) as indicator for residual of leaked speaker

	MI [bits]	Variance
Before Iso-Net	3.52	0.177
After Iso-Net	1.39	0.151

6. Listening Tests

- Mushra Listening Test with $N = 26$ listeners.
- Improvement from original to IsoNet isolated signal is 47.72 Mushra points



7. Conclusions

- Speech enhancement can contribute to privacy improvements
- Voice isolation can easily be done with simple assumptions without prior information
- Iso-Net is a small, real-time capable network with 3.7M parameters