SIPL
Signal and Image Processing Lab

THE ANDREW & ERNA VITERBI FACULTY OF ELECTRICAL ENGINEERING

TECHNION Israel Institute of Technology

Work in Progress

Acoustic Fence Using Multi-Microphone Speaker Separation

Tomer Fait and Orel Ben-Reuven, Supervised by Amir Ivry

In collaboration with PHOENIX Audio Technologies

Motivation

- Acoustic fencing algorithms separate speakers by their physical location in space
- In conference calls, for instance, they pass speech coming from a zone of interest, and attenuate disturbances from other zones

Simulated conference room with 2 speakers, 2 microphones, and the acoustic fence created to isolate the speakers

Goals

- Develop acoustic fencing system that separates speakers by pre-defined reception zones, and:
  - attenuates speakers outside reception zone
  - passes speakers inside reception zone without distortion
- Create performance measures for desired-speech distortion and disturbance attenuation that correlate well with human subjective evaluation
- Engineer lean system implementation that can be integrated on-device

Challenges

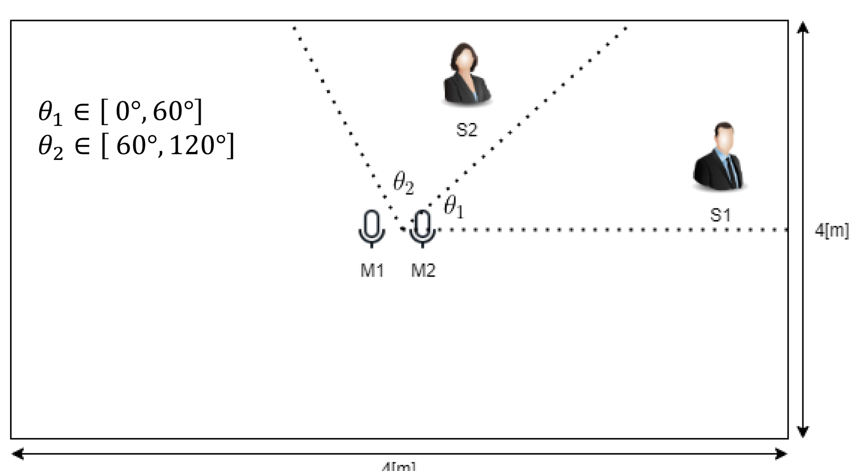
- Speakers outside reception zones may dominate the speakers inside reception zones
- Real conference room recordings are noisy, transient, and reverberant

Acoustic Fencing Setup

- M microphones and N speech sources in a conference room
- $z_m(n)$ is the i_{th} speech signal $s^i(n)$, as captured by the m_{th} microphone:

$$z_m(n) = \sum_{i=1}^{N} s^i(n) * h_m^i(n)$$

- $h_m^i(n)$ is the room impulse response relating the i_{th} speaker and the m_{th} microphone

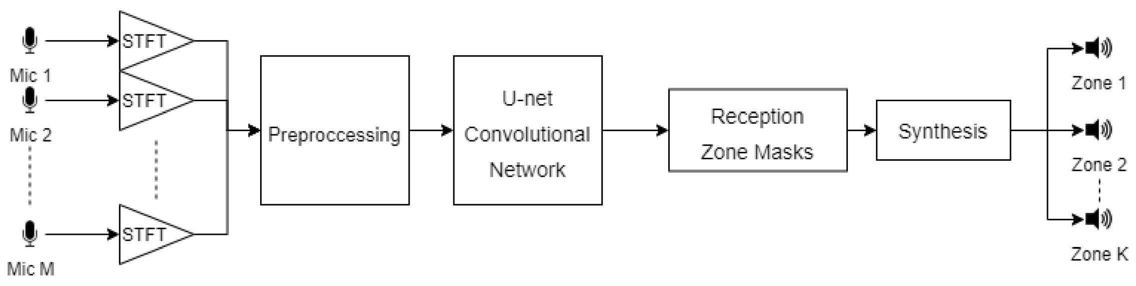


Example setup of a conference room with 2 microphones, 2 speakers, and 2 reception zones.

- Let $\{\theta_k\}_{k=1}^{K}, \{y_k(n)\}_{k=1}^{K}$ be K reception zones and their respective generated signals
- The goal of the system – separate $\{z_m(n)\}_{m=1}^{M}$ to K signals $\{\hat{y}_k(n)\}_{k=1}^{K}$ such that only speakers located inside θ_k are in $\hat{y}_k(n)$ without distortion

DDESS

- DDESS – Deep Direction Estimation for Speech Separation, is an acoustic-fencing algorithm that operates in the STFT domain
- Assumes W-disjoint orthogonality. That is, each time-frequency (TF) bin in the STFT transform, is dominated by a single speaker
- The algorithm classifies each TF bin into a reception zone



Block diagram of DDESS algorithm:
The microphones signals are transformed into STFT domain, features are extracted, and masks are generated by the U-Net. Finally, the signal is reconstructed by the masks and reference microphone.

Evaluation Criteria

- Unlike existing measures, we separately evaluate the systems' distortion of the desired signal, and suppression of interference
- We applied the output mask to 3 STFT signals:
  - Original input signal (double-talk), $r(t)$
  - Desired signal only, $p(t)$
  - Interference signal only, $b(t)$

$$r(t) = b(t) + p(t)$$

- The resulting 3 signals are:
  - Double-talk fencing, $r'(n)$
  - Pass signal fencing, $p'(t)$
  - Block signal fencing, $b'(t)$

$$\begin{array}{ccc} & Mask & \\ r(t) & \rightarrow & r'(t) \\ p(t) & \rightarrow & p'(t) \\ b(t) & \rightarrow & b'(t) \end{array}$$

- We define the following evaluation criteria:
  - $\boldsymbol{Pass\ Signal}\ SegSNR$
$$= \frac{10}{M}\sum_{m=0}^{M-1} \log_{10} \frac{\sum_{l=0}^{L-1} p_{m\cdot L+l}^2}{\sum_{l=0}^{L-1}\left(p_{m\cdot L+l} - p'_{m\cdot L+l}\right)^2 + \sigma^2}$$
where L , M are the window's length and number, and σ prevents division by zero.
  - $\boldsymbol{Block\ Signal}\ Attenuation = 10\log_{10}\frac{\|b'\|^2}{\|b\|^2}$
  - $\boldsymbol{Double-Talk}$ perceptual evaluation of speech quality (PESQ):

PESQ Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

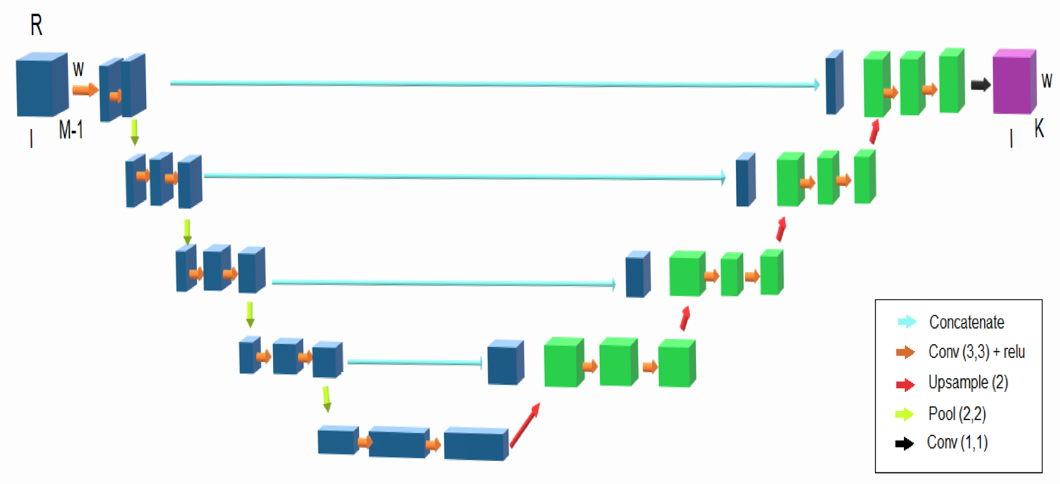
PESQ score practically ranges from -0.5 (worst) up to 4.5 (best)

Data Corpus

- TIMIT - Acoustic-Phonetic Continuous Speech Corpus
- Contains (in English): 6300 sentences, 630 speakers, 8 major dialect, 2000+ textually different sentences
- The test set is 27% (about 40 minutes) of the data set

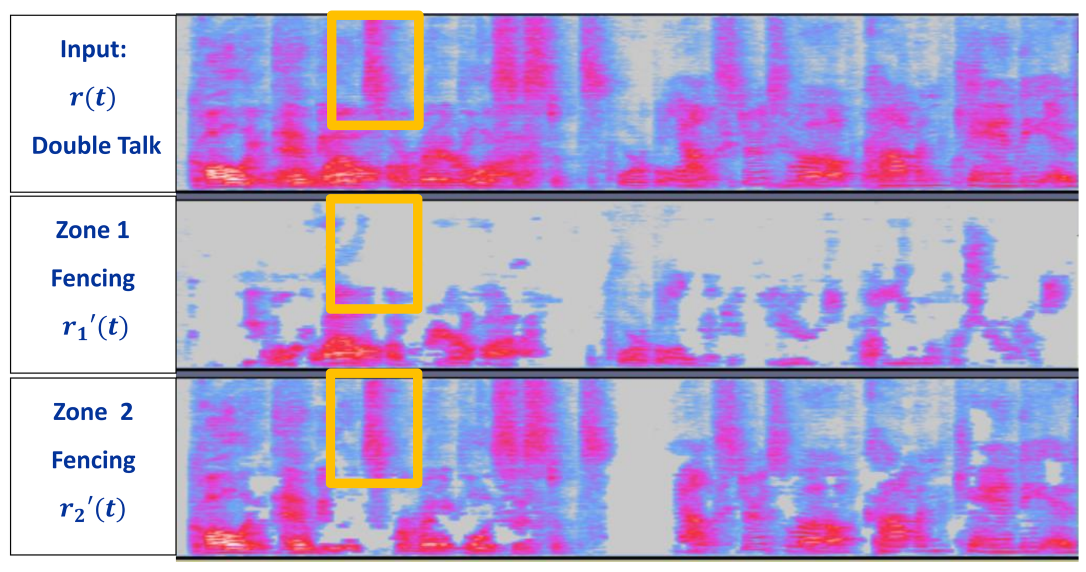
U-net architecture

- The DDESS algorithm utilizes the U-net architecture in order to address the speech separation task
- U-net is trained to classify each TF bin of the STFT image to one reception zone
- The tag for each TF bin is created according to the reception zone with the highest energy



U-net convolutional network

Results



Spectrograms of input and outputs signals. The spectral shape in the input's orange rectangular is only passed to zone 2.

Criterion	DDESS	Conv TasNet
Seg-SNR [dB]	13.69	9.12
Attenuation [dB]	12.38	3.46
PESQ	3	0.8

Evaluation criteria results for DDESS and Conv TasNet on the TIMIT database. Conv TasNet is a state-of-the-art speech separation algorithm (though intended for larger datasets).

On Device Implementation

Our implementation enables on-device integration, e.g., using the NDP120 neural processor by Syntiant

Criterion	Value
Network Parameters	~1M
Parameters Memory	4.14 MB
Inference Time	30ms
System Latency	48ms
Number of FLOPs	< 4M

Conclusion

- We proposed an acoustic fencing algorithm that outperformed leading speech separators and obtained:
  - Higher attenuation of interfering signals
  - Lower distortion of desired signals
- We constructed two performance measures that showed consistency with human objective evaluation
- We engineered a lean implementation that allows practical embedding of our system into on-device platforms

June 2021