Anonymising Speech in Surveillance - using Speech Masking and Background Separation.

Outline

- Scenario
- Two parts:
  1. **Background Separation**
  2. Anonymising speaker identity & speech content
     - Ambiguity of Short-Term Objective Intelligibility
- Data Sets with increasing realism

Scenario

- Audio surveillance in a **waiting room**,
- Audio **mixture**:

  speech + background sounds.
- Eavesdropping listeners in a **control room**.

Can we...

ensure right to speech privacy,

while simultaneously

ensuring surveillance capability?

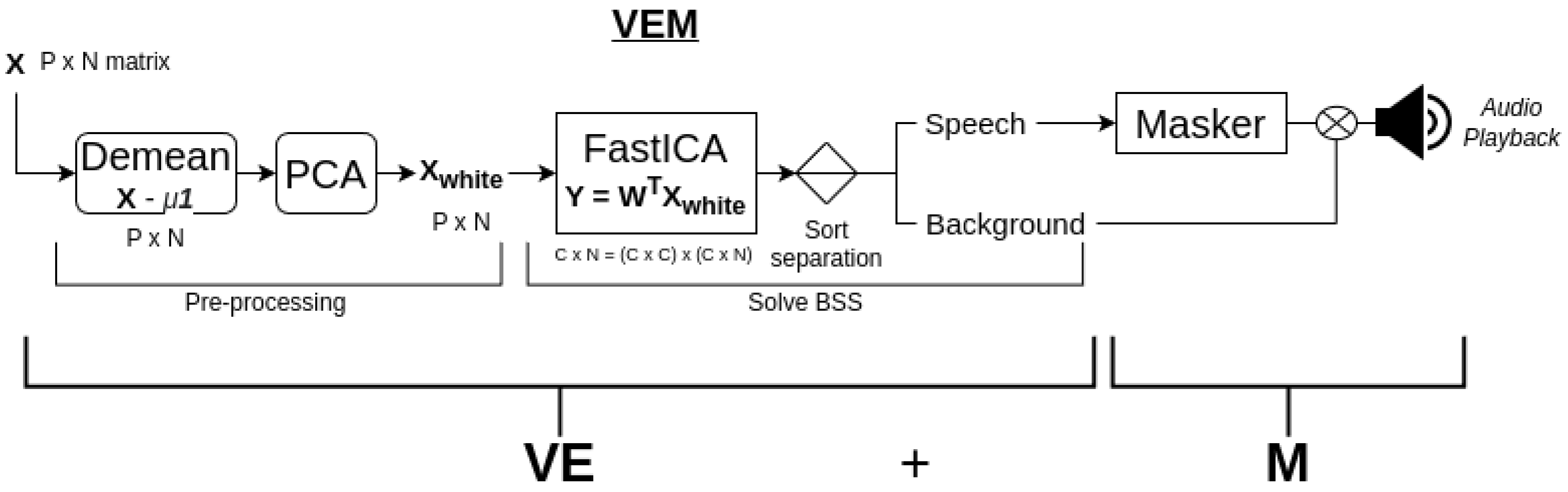
Insights:

- Speech-from-background separation leads to full control of speech signal
- Audio classification happens earlier in the processing pipeline.

So..

- Output is for human ears only
- Separation can be triggered and shut off using **Voice Activity Detection** (VAD)

Model: Voice Extraction and Masking



Part 1: Separation

Semi-Blind Audio Source separation.

Can we with limited a priori audio mixture information perform separation?

BSS Problem Statement

- **Sound sources**: *voiced* and *background*

$$\bar{\mathbf{s}}(t) = \left(s_k\right)_{k=1}^{N} = \left\{v_1, \ldots, v_{N_v}, b_1, \ldots, b_{N_b}\right\} \in \mathbb{R}^N$$

- **Observations** from *P* microphones in an unknown **mixture** $\mathscr{A} : \mathbb{R}^N \mapsto \mathbb{R}^P$

$$\bar{x}(t) = \mathscr{A}\left(\bar{\mathbf{s}}(t)\right) \in \mathbb{R}^P$$

BSS Problem Statement

- We seek the estimated sources, finding unmixing transformation $\mathscr{B} := \mathscr{A}^{-1}$

$$y_i = k_i\big(s_{\sigma(i)}(t)\big),\ i = 1, 2, \ldots, N,$$

- Turns out we cannot know the separation order $\sigma(i)$,
- The resulting source estimate y_i is distorted with k_i

But what about the sources themselves?

Independent Component Analysis

Sound sources viewed random variables

- They are statistically mutually **independent**
- They are all **non-Gaussian**, or all but one.

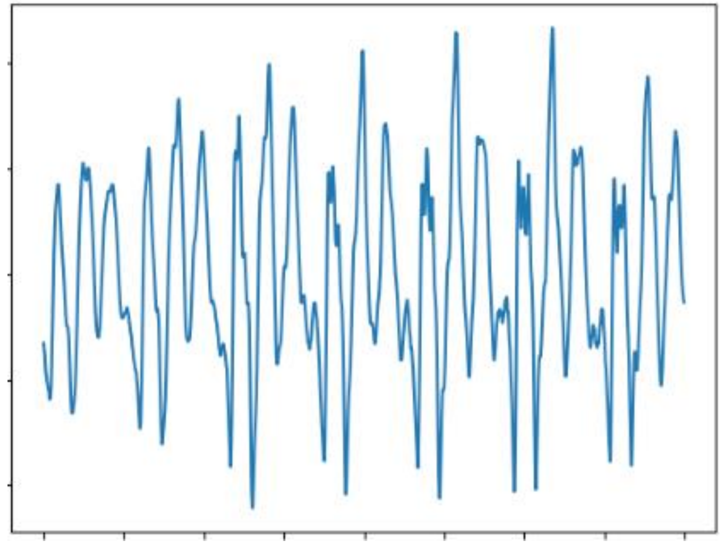
Idea of ICA:

Find components that are furthest away from normality!

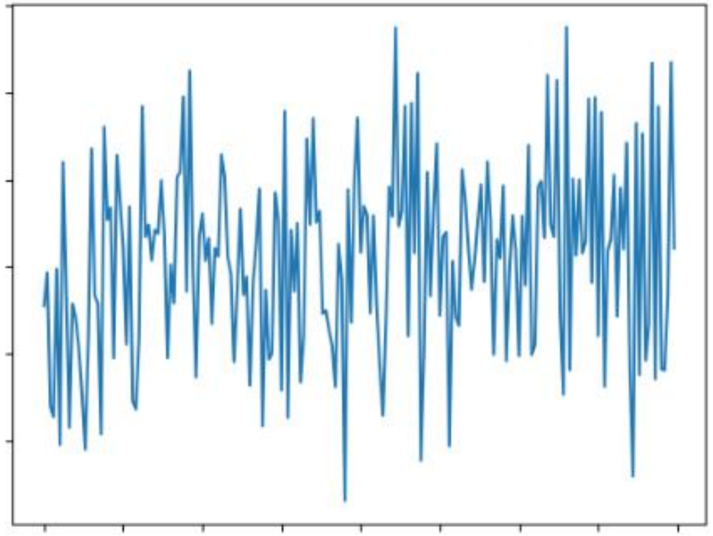
We measure this *distance* with ***negentropy*** (Information Theory)

Speech and frame of reference

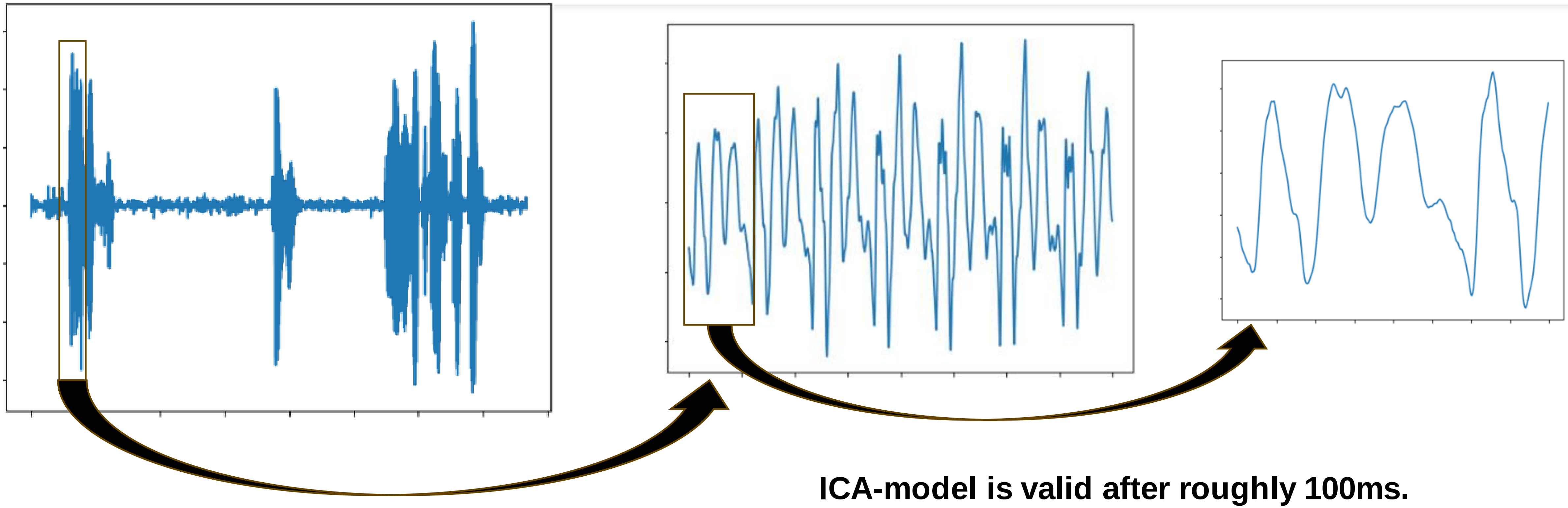
Speech



Gaussian noise



Speech and frame of reference



ICA-model is valid after roughly 100ms.

Brief Mathematical Overview:

- Discrete signal values
- No insight on distributions due to blindness

Distance must be approximated!

Corollary 2.2.1 (Approximation of Negentropy). *[27] Assuming X is a r.v with zero-mean and unit variance and similarly $X^* \sim \mathcal{N}(0,1)$. Then the negentropy of X can be approximated with*

$$J(X) = [\mathbb{E}[G(X)] - \mathbb{E}[G(X^*)]]^2,$$

where G is a non-quadratic and preferably slowly increasing function, as suggestions for $G(x)$:

$$G_1(x) = \log\cosh \alpha x, \ \alpha \in [1,2] \tag{11}$$

$$G_2(x) = -\exp -\frac{x^2}{2} \tag{12}$$

$$G_3(x) = x^3 \tag{13}$$

FastICA Method

Assume linear model

$$\bar{x} = A\bar{s}$$
$$\bar{y} = W\bar{x}$$
$$W = A^{-1}$$

Several components extraction

means decorrelation weights **W**

Algorithm 1 FastICA for Several Components Extraction

Input: $N \times P$ pre-whitened data matrix $\tilde{X}$
Input: Desired number of independent components $M \leq P$
Output: $M \times M$ matrix W of unmixing matrix estimate
Initialisation: Random initialisation of the unmixing matrix W
Repeat until convergence:
Compute projections $W^T\tilde{X}$
Compute $g(W^T\tilde{X}) = \tanh(W^T\tilde{X})$ and $g'(W^T\tilde{X}) = 1 - \tanh^2(W^T\tilde{X})$
Compute new estimate W^+

$$W^+ = \mathbb{E}[\tilde{X}g(W^T\tilde{X})] - \mathbb{E}[g'(W^T\tilde{X})W]$$
$$W^+ \leftarrow \frac{W^+}{||W^+||}$$

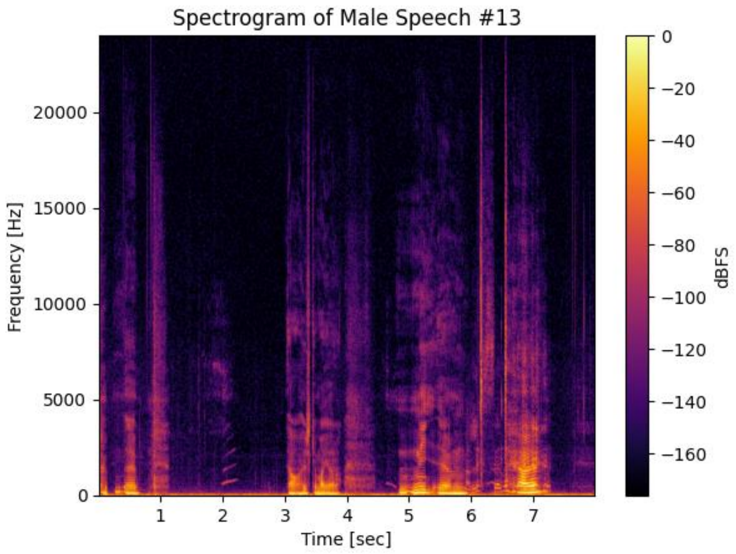
Decorrelate W^+ with respect to previously unmixing matrix estimates W

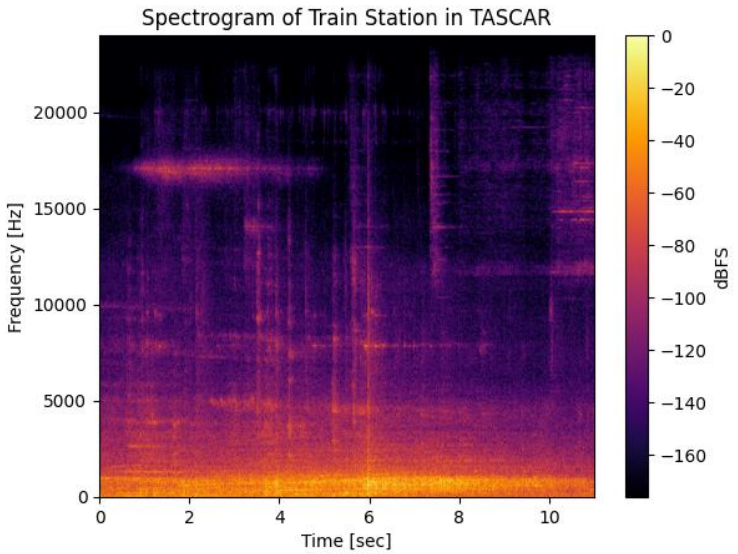
$$E, D \leftarrow PCA(W)$$
$$W^+ \leftarrow ED^{-1/2}E^T$$

Update unmixing matrix: $W \leftarrow W^+$
End

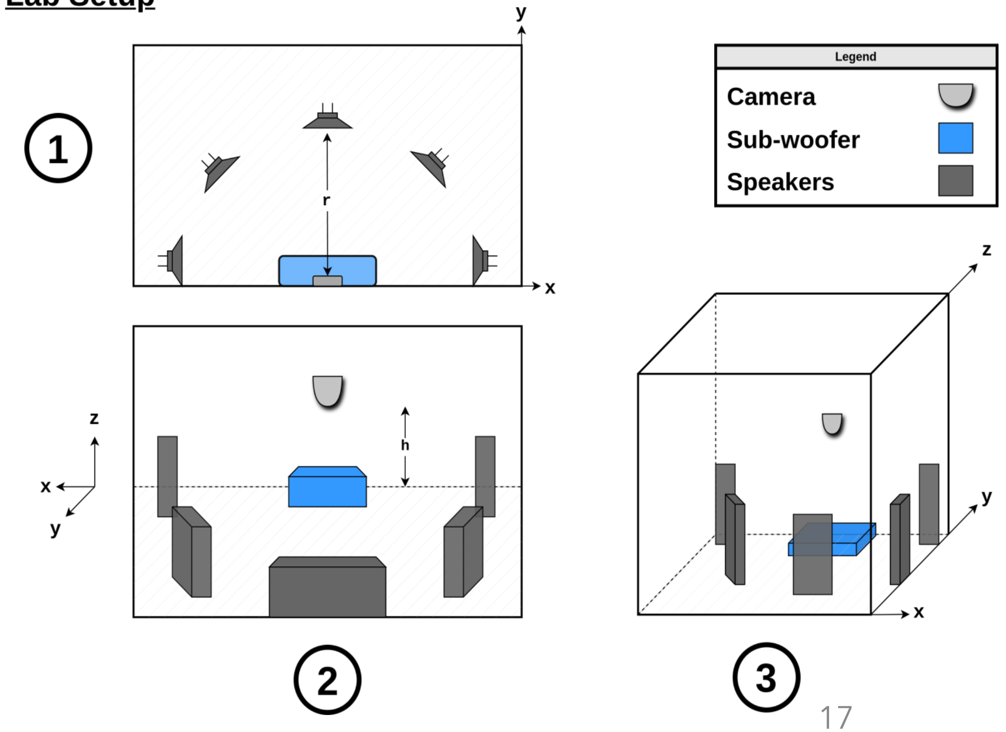
Data sets

1. Synthetic mixtures
2. TASCAR Simulations
3. Camera Lab recordings





Lab Setup



Data set components

- Clean speech of male and female (5 min. each)
- Background sounds
    - Urban city sounds
    - Train station
    - Park
    - Forest
    - Traffic
    - Crowd murmur
    - Office

Data set 1 – Static Speech & SNR:s

- Speech standing still from angle, DOA –59 degrees
- Backgrounds are diffuse sources.
- SNR-levels:
  - 0dB
  - –6 dB
  - -12dB

Can we separate a basis case and at somewhat real conditions?

Data set 2 – Moving Source & Reverb.

- Speech source traces a rectangle path in front of the camera,
- Dynamic parameters:
  - DOA, also reflective sound paths / Room Impulse Response (RIR)
  - SNR

Can we handle more than one dynamic parameter?

Data set 2 increases realism of scenario.

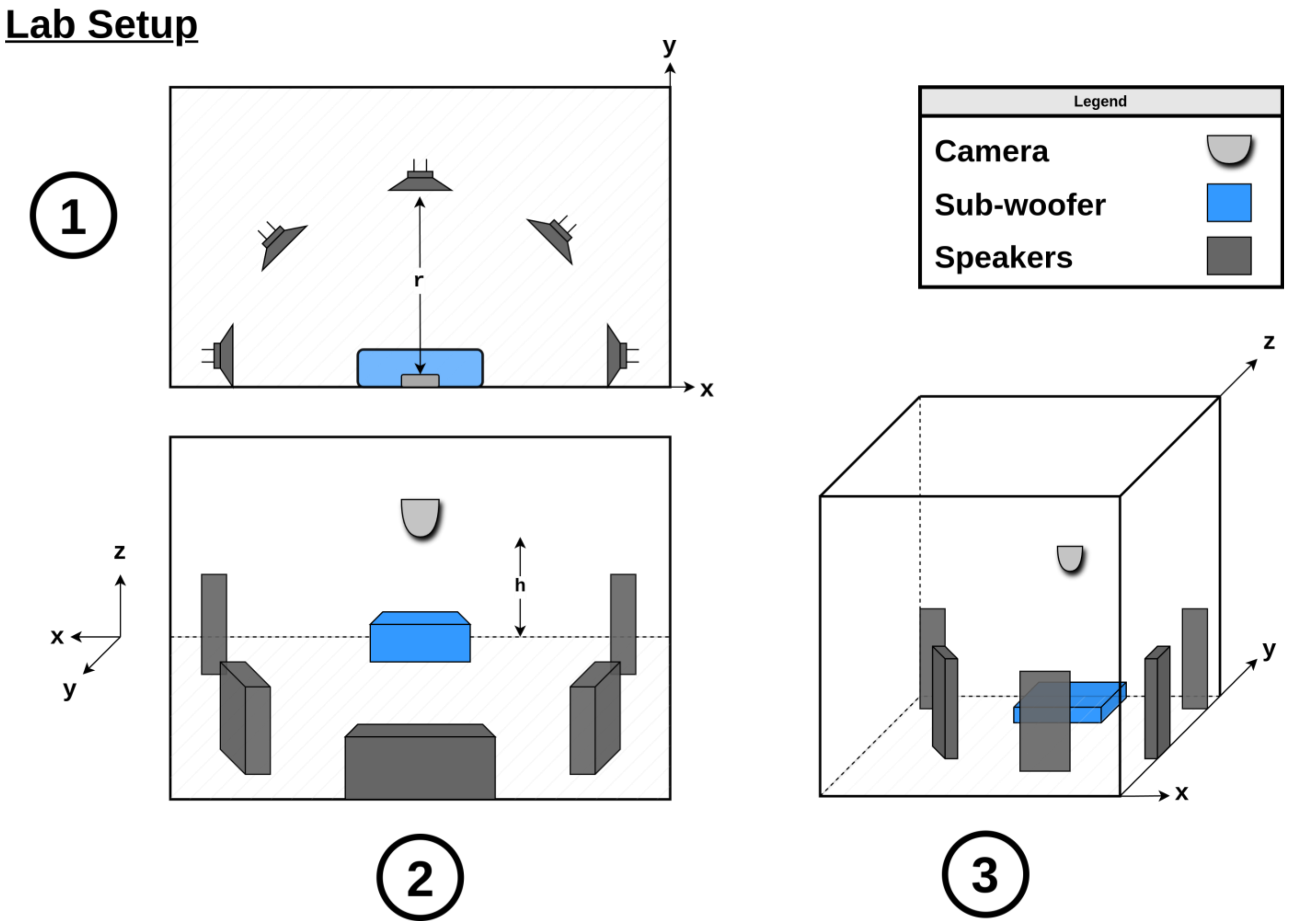
Data set 3 – Camera Lab recordings

- TASCAR scenes are played back in lab, are 10x longer
- Dynamic parameters:
  - DOA, also reflective sound paths / Room Impulse Response (RIR)
  - SNR
  - Additive static noise from analogue-to-digital conversion
  - Scene duration increased tenfold, then locations in scene of former time duration are examined.

Real data!

Data set 3

Setup of Lab



Performance metrics

How intelligible is the separated speech to the original speech input?

Separation quality of speech:

(Extended) Short-Term Objective Intelligibility

$$ESTOI(x, y) \mapsto [-1, 1]$$

How similar is the separated background to the original background sound input?

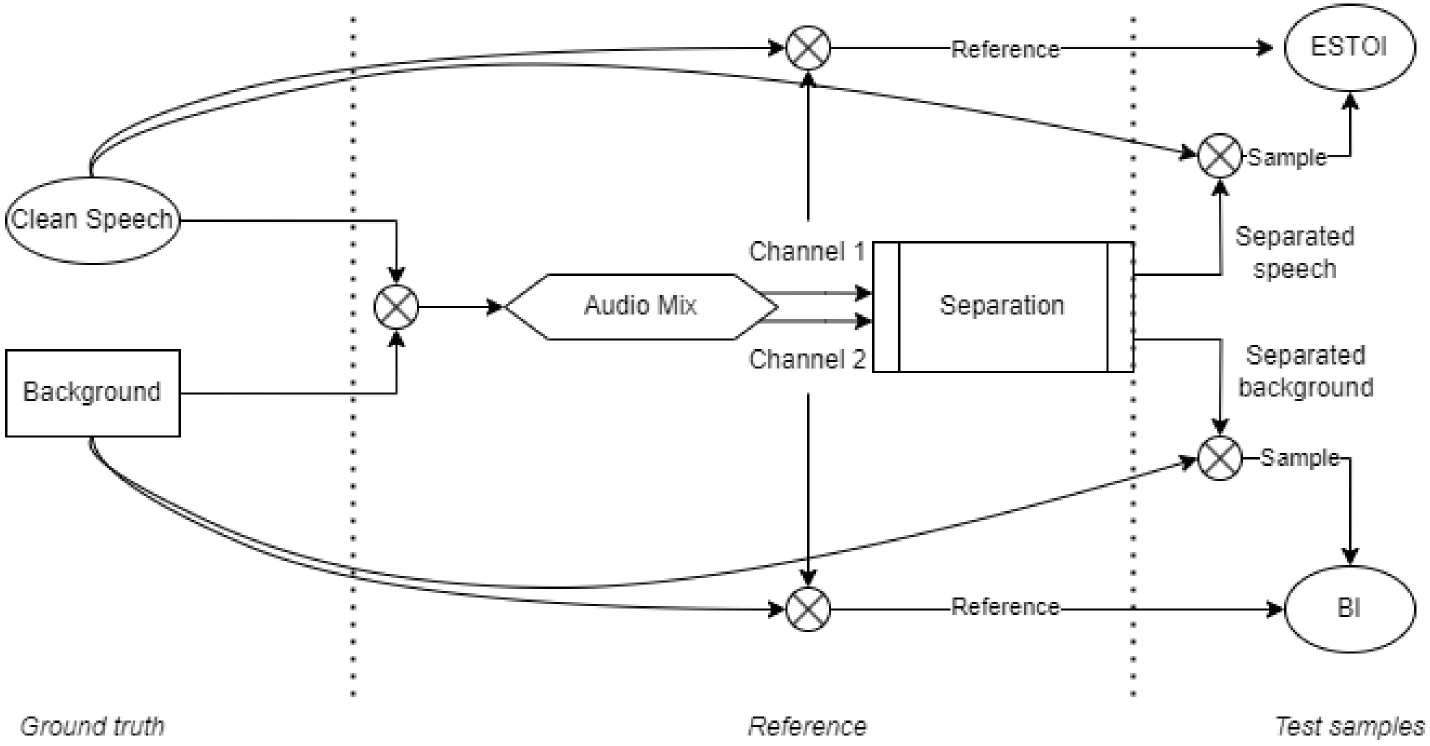
Separation quality of background:

Magnitude of norm. Cross-corr.

$$BI(\boldsymbol{b}, \boldsymbol{y}) := \left| \frac{\sum_i^N b_i y_i}{\sigma_b \sigma_y} \right| \mapsto [0, 1],$$

Performance metrics and their reference

- How intelligible is the separated speech to the original speech input?
- How similar is the separated background to the original background sound input?
- **… Compared to the input mix?**

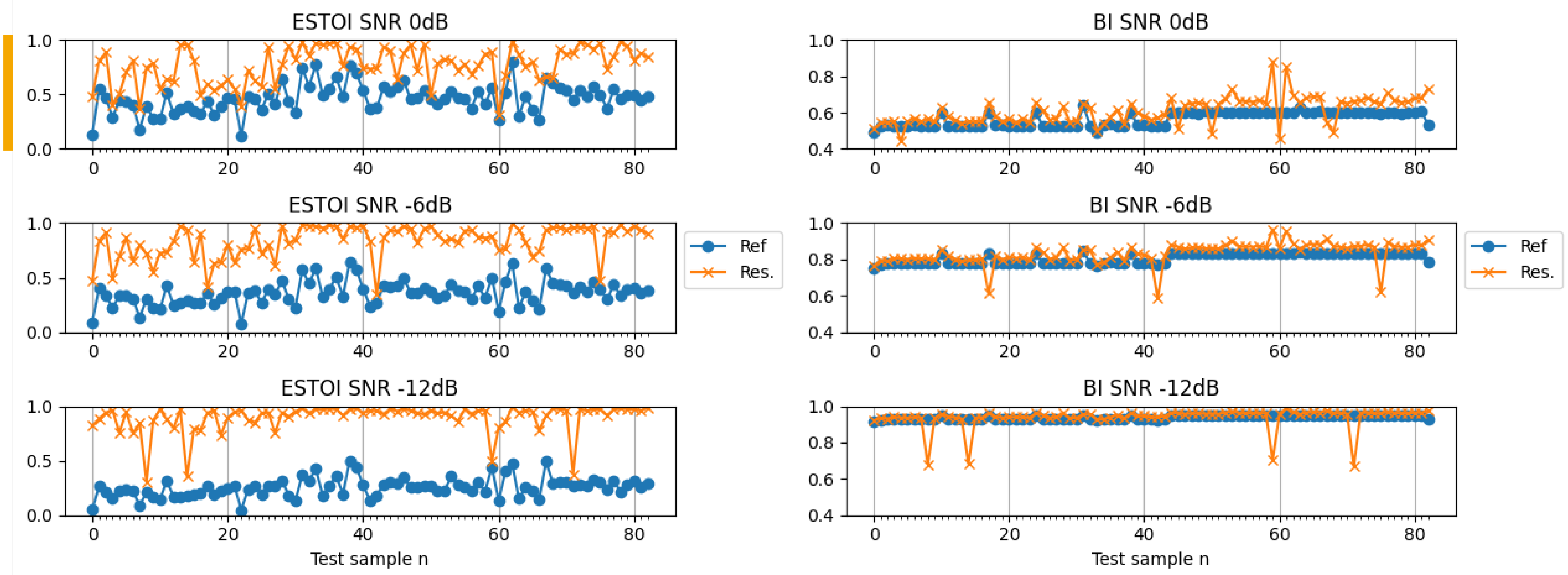


Data Set 1:

Input →Speech & background separation

Results: Speech Separation

Data Set 1:



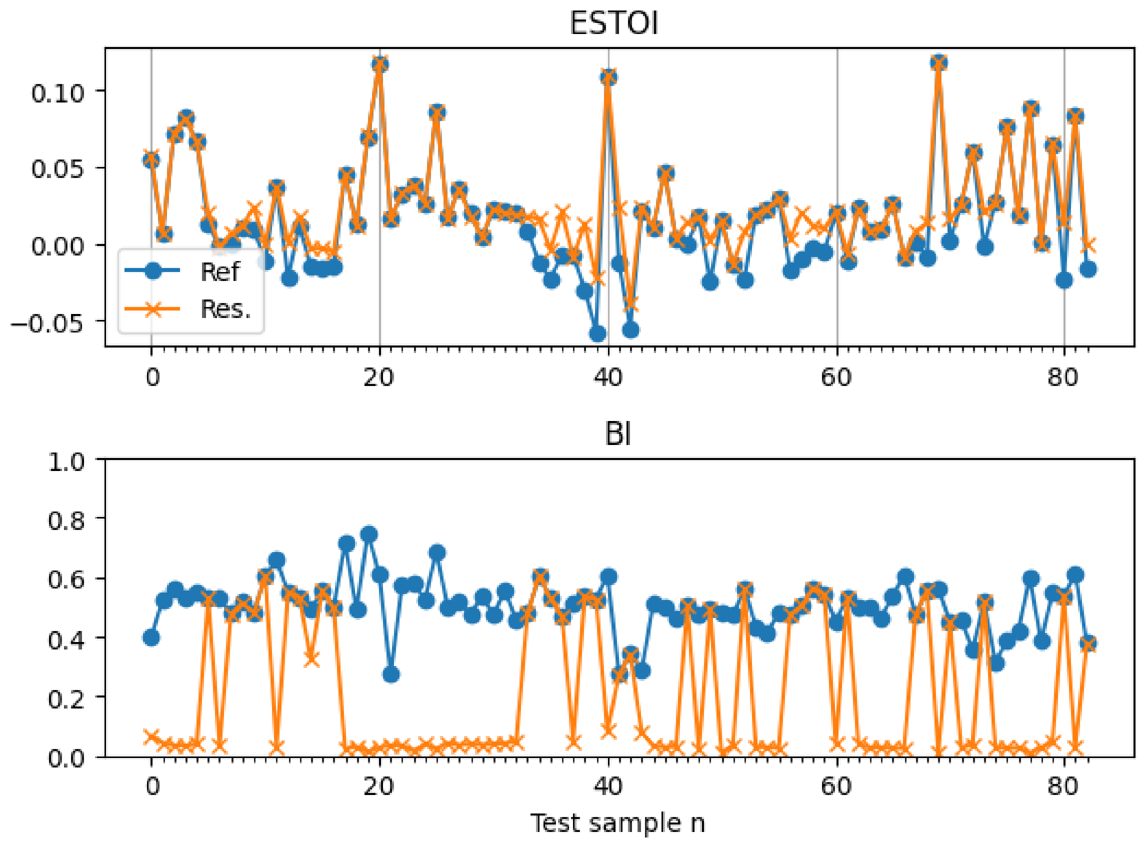
Data Set 2:

Speech is removed in one channel...

Results: Speech Separation

Data Set 2:

- **One channel without speech**
- **One channel unaltered**



Results: Speech Separation

Data Set 1: Ideal separation of speech and background

Data Set 2: Sufficient background separation

Data set 3: Separation **unsuccessful**

Discussion: Speech Separation

- Separation quality deteriorates when data aligns to reality
  - Linear model breaks down
  - Determined system $\longrightarrow$ Underdetermined system
  - Insufficient pre-processing
- Intelligibility measure can indicate successful separation
- Background Intactness deteriorates with increasing spectral subtraction

Part 2:
Speech & Identity Masking

Speech Content Masker.

Removing cues in speech.

Auditory and Speech Masking

- Decrease intelligibility of speech with presence of a **masker sound**
    - Tonal and temporal masking
- Informational masking
    - Masking sound is **speech-like**
    - Possesses similar characteristics in temporal and spectral structure.
    - Hinders access to speech cues and information in speech

Proposed Speech Masker

Aim:

Minimise *speech intelligibility* and *induced irritation* for listener

Idea:

Use target speech as seed for masker sound

Limitation:

Masker sound production must be computationally lightweight for real-time use.

Proposed Speech Masker Structure

Create a cheap speech-like masker:

- **Time-reversion**: Locally flip speech frames
- **Phase-less:** Nullify phase of STFT and go back

Local speech frames are time-reversed and rendered phase-less.

Intelligibility as an Objective Measure

ESTOI is proposed to objectively measure speech intelligibility of speech affected by modulating masking sound, **e.g., another speech-like sound**.

We evaluate this claim by computing STOI and ESTOI on the proposed masker sound.

Masker sound:

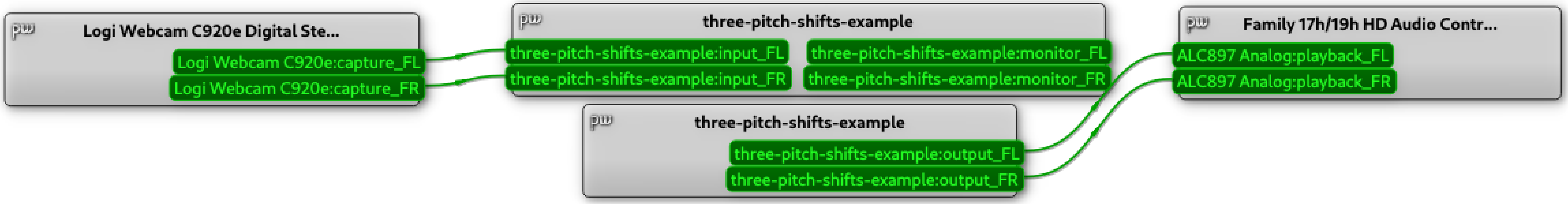
Time-reversed and Phase-less Masker: **Insufficient.**

Speaker Identity Anonymizer.

Masking Speech Identity..

Speaker Identity Masking

- Three pitch-shifting threads
- Running LADSPA-plugins into PipeWire
- Scales pitches up and down, always ensuring tonal masking



DEMO:

Can you count the number of speakers and their gender?

DEMO:

Correct answer:
4 people: *Male 1→ Female 1→ Male 2→ Male 3*

Conclusion: Speaker Identity & Speech Masking

Speech Masking:

Low latency vs. efficient masking.

Right to speech privacy attainable through **speaker anonymisation**.

Ambiguity of ESTOI:

Right figure: Effect of adding short reverb.

Subjective Intelligibility:

No considerable difference.

Objective Intelligibility:

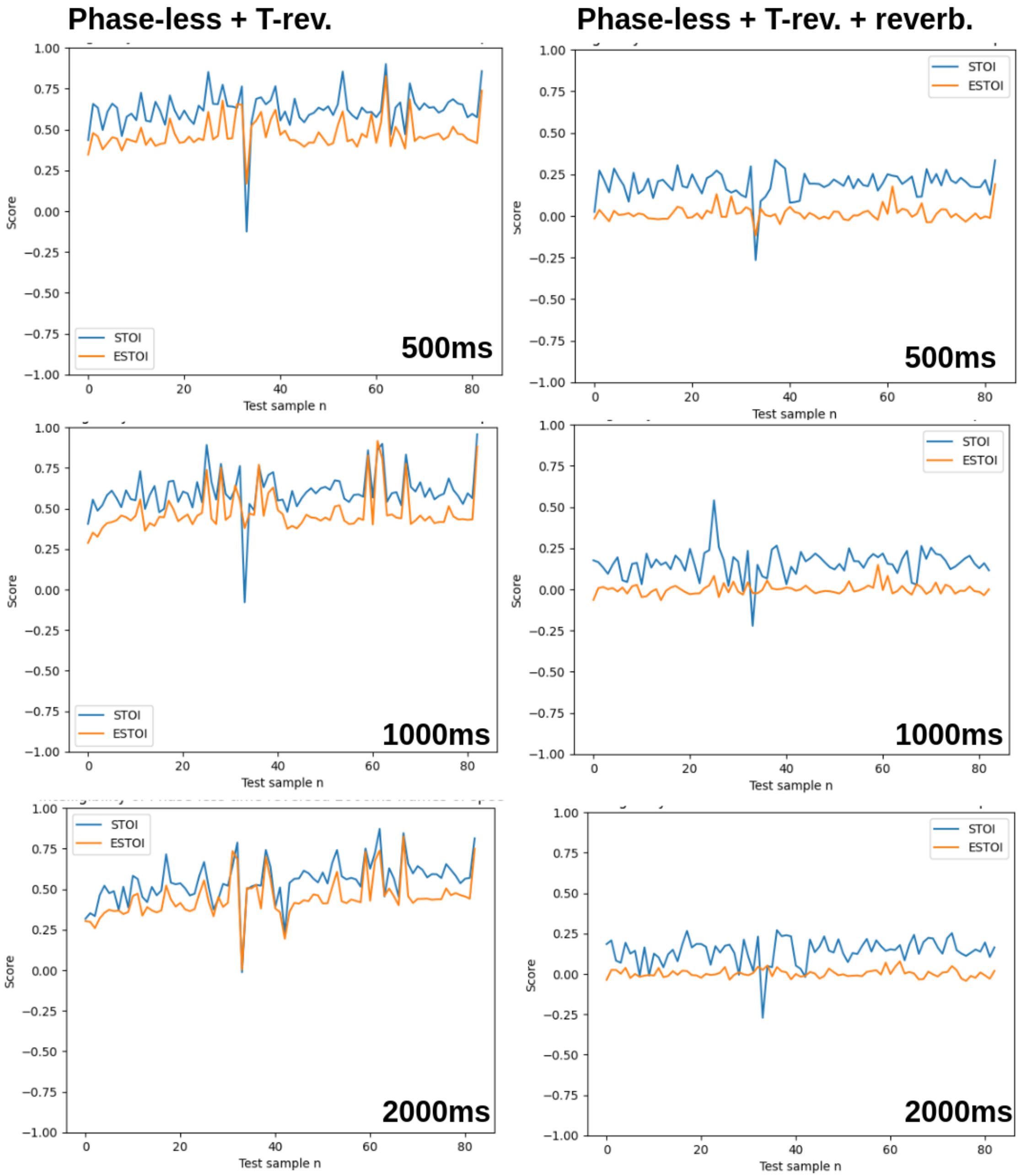
As unintelligible as

- Pitch shifting
- Input mix reference

Almost as unintelligible as

ESTOI(speech, non-speech sound) ??

Intelligibility of Second Speech Masker



Future work:

- **ICA**
  - Dynamic MIMO model
  - Better suitable contrasts?
  - Online implementation and permutation problem

- **Speech & Speaker Anonymisation**
  - Reversibility?
    - Saving nullified phases as keys for each window
    - Reversing several pitch shifts

- **Objective Measures for Intelligibility**
  - Disambiguate

Thank you!