Unsupervised Voice Activity Detection by Modeling Source and System Information using Zero Frequency Filtering

Eklavya SARKAR
Research Assistant, Idiap Research Institute and EPFL

Outline

Outline

- Voice Activity Detection

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering
- Proposed Method

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering
- Proposed Method
- Experimental Setup

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering
- Proposed Method
- Experimental Setup
- Baseline Methods

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering
- Proposed Method
- Experimental Setup
- Baseline Methods
- Results and Discussion

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering
- Proposed Method
- Experimental Setup
- Baseline Methods
- Results and Discussion
- Summary

Outline

- Voice Activity Detection
- Background on Zero-Frequency Filtering
- Proposed Method
- Experimental Setup
- Baseline Methods
- Results and Discussion
- Summary
- Future Work

Voice Activity Detection Problem

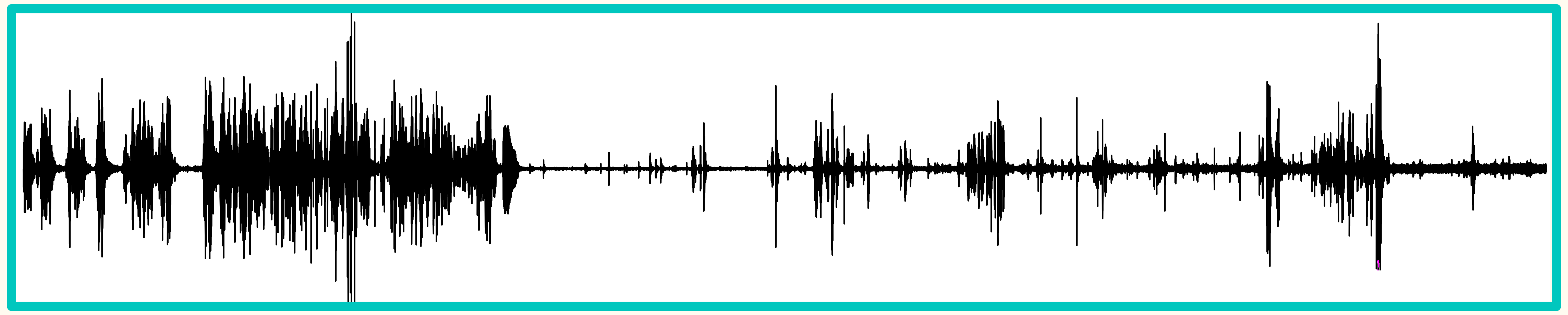
Voice Activity Detection Problem

Task: identify segment boundaries in signals which contain voicing information.

Voice Activity Detection Problem

Task: identify segment boundaries in signals which contain voicing information.

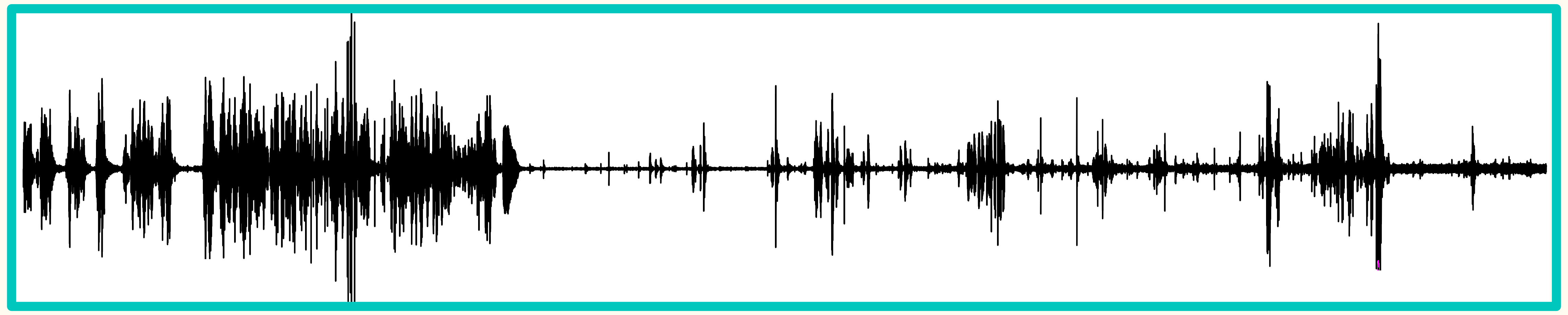
Input: recording containing speech and non-speech.



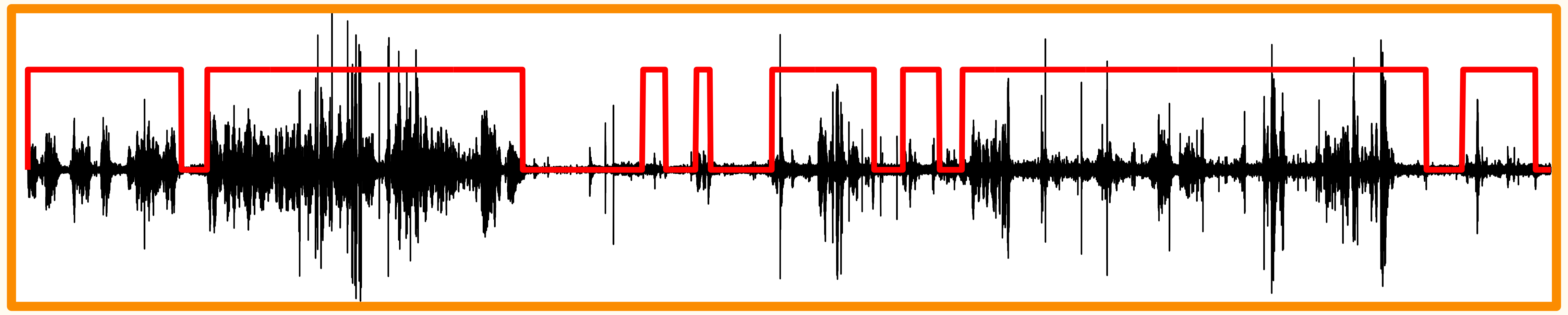
Voice Activity Detection Problem

Task: identify segment boundaries in signals which contain voicing information.

Input: recording containing speech and non-speech.



Output: speech segment boundaries.

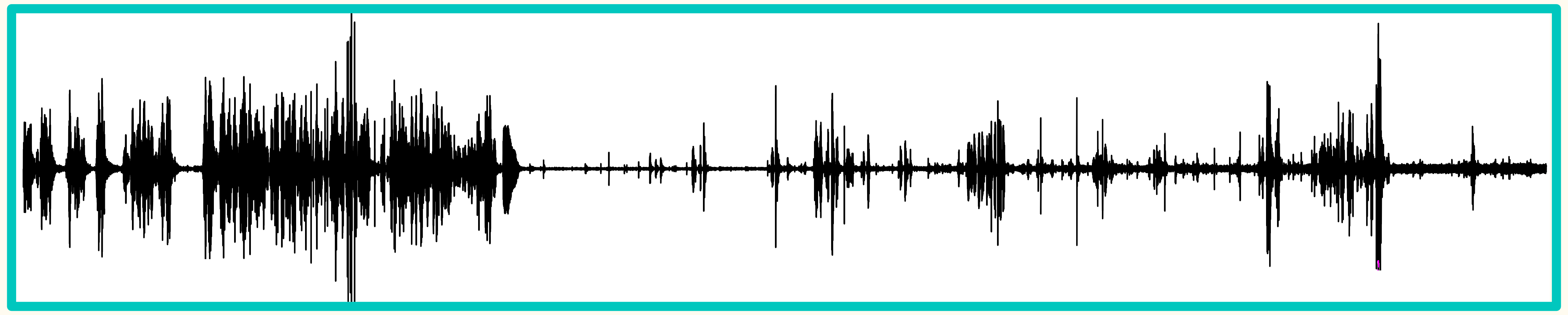


Voice Activity Detection Problem

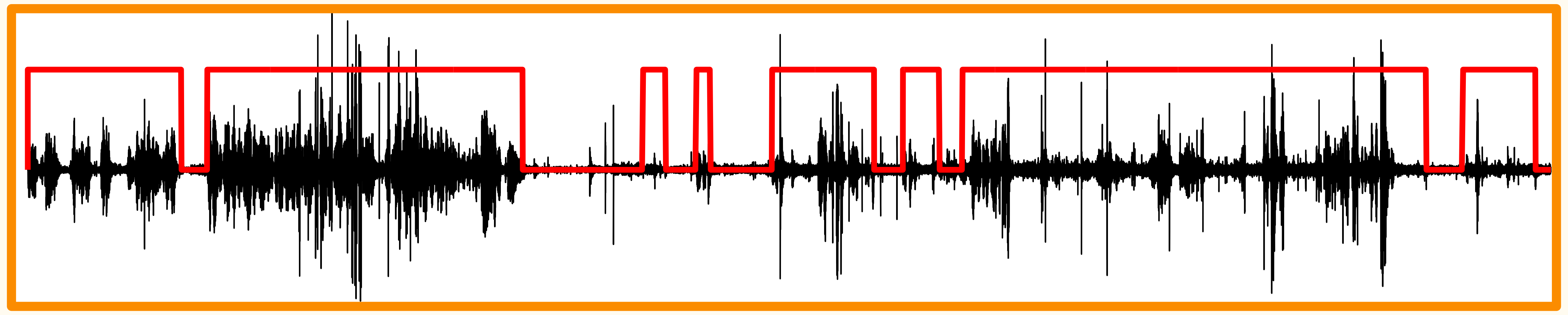
Task: identify segment boundaries in signals which contain voicing information.

- One of the first steps to be carried out in any speech technology.

Input: recording containing speech and non-speech.



Output: speech segment boundaries.

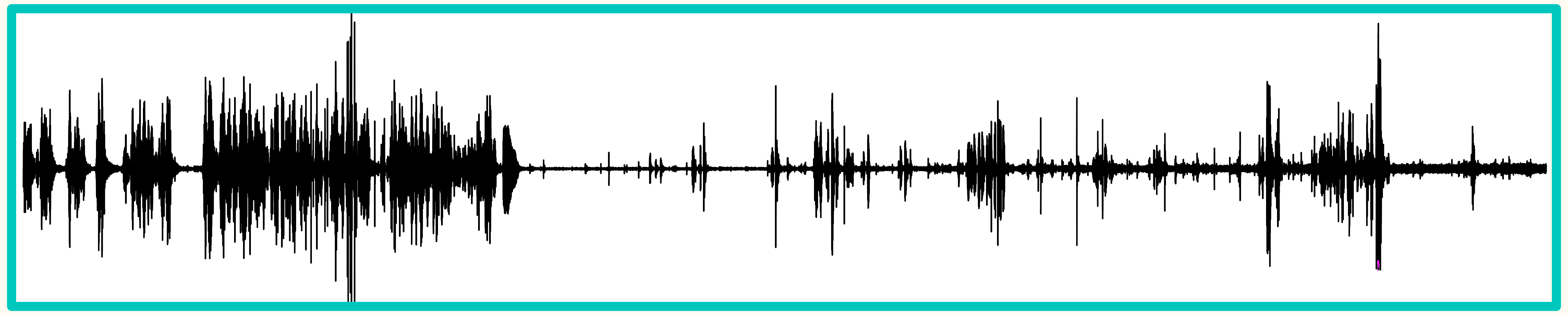


Voice Activity Detection Problem

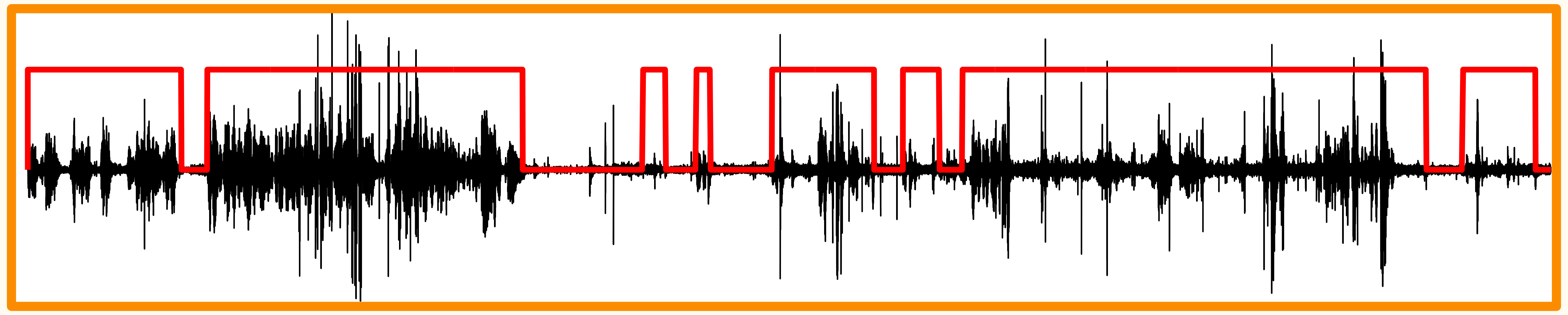
Task: identify segment boundaries in signals which contain voicing information.

- One of the first steps to be carried out in any speech technology.
- Computational efficiency and robustness to noisy data are thus essential pre-requisites for any SOTA VAD.

Input: recording containing speech and non-speech.



Output: speech segment boundaries.



Voice Activity Detection Landscape

Voice Activity Detection Landscape

- Linear Prediction Residual
- Teager Energy Operator (TEO)
- Log Spectral Energy (LSE)
- Perceptual Spectral Flux
- Zero-Frequency Filtering (ZFF)

Unsupervised Methods

Voice Activity Detection Landscape

- Linear Prediction Residual
- Teager Energy Operator (TEO)
- Log Spectral Energy (LSE)
- Perceptual Spectral Flux
- Zero-Frequency Filtering (ZFF)

Unsupervised Methods

- Gaussian Mixture Models
- Neural Networks

Supervised Approaches

Voice Activity Detection Landscape

- Linear Prediction Residual
- Teager Energy Operator (TEO)
- Log Spectral Energy (LSE)
- Perceptual Spectral Flux
- **Zero-Frequency Filtering (ZFF)**

Unsupervised Methods

- Gaussian Mixture Models
- Neural Networks

Supervised Approaches

Background

Background

- In recent years, it has been shown that voice source and vocal tract system information can be extracted using zero-frequency filtering without making any explicit model assumptions about the speech signal, as source-system decomposition does.

Background

- In recent years, it has been shown that voice source and vocal tract system information can be extracted using zero-frequency filtering without making any explicit model assumptions about the speech signal, as source-system decomposition does.
- This paper investigates the potential of zero-frequency filtering for jointly modeling voice source and vocal tract system system information for VAD.

Background

- In recent years, it has been shown that voice source and vocal tract system information can be extracted using zero-frequency filtering without making any explicit model assumptions about the speech signal, as source-system decomposition does.
- This paper investigates the potential of zero-frequency filtering for jointly modeling voice source and vocal tract system system information for VAD.
- Towards that, we demonstrate that voice activity detection can be effectively achieved by combining the outputs of a bank of zero-frequency filters that carry information related to fundamental frequency (f_0), first formant (F_1) and second formant (F_2).

Background

Background

- Zero-frequency filtering (ZFF) was originally proposed in the context of extracting information related to voice source.

Background

- Zero-frequency filtering (ZFF) was originally proposed in the context of extracting information related to voice source.
- In this method, a speech signal is first passed through a cascade of digital resonators centered at 0 Hz, i.e. a zero-frequency filter.

Background

- Zero-frequency filtering (ZFF) was originally proposed in the context of extracting information related to voice source.
- In this method, a speech signal is first passed through a cascade of digital resonators centered at 0 Hz, i.e. a zero-frequency filter.
- The resulting impulse response of these cascaded resonators, implemented as an integrator, is given by eq. (1) and the equivalent transfer function by eq. (2).

Background

- Zero-frequency filtering (ZFF) was originally proposed in the context of extracting information related to voice source.
- In this method, a speech signal is first passed through a cascade of digital resonators centered at 0 Hz, i.e. a zero-frequency filter.
- The resulting impulse response of these cascaded resonators, implemented as an integrator, is given by eq. (1) and the equivalent transfer function by eq. (2).

Background

- Zero-frequency filtering (ZFF) was originally proposed in the context of extracting information related to voice source.
- In this method, a speech signal is first passed through a cascade of digital resonators centered at 0 Hz, i.e. a zero-frequency filter.
- The resulting impulse response of these cascaded resonators, implemented as an integrator, is given by eq. (1) and the equivalent transfer function by eq. (2).

$$x[n] = s[n] - 2x[n-1] + x[n-2] \qquad\qquad H[z] = \frac{1}{1 - 2z^{-1} + z^{-2}}$$

Background

- Zero-frequency filtering (ZFF) was originally proposed in the context of extracting information related to voice source.
- In this method, a speech signal is first passed through a cascade of digital resonators centered at 0 Hz, i.e. a zero-frequency filter.
- The resulting impulse response of these cascaded resonators, implemented as an integrator, is given by eq. (1) and the equivalent transfer function by eq. (2).

$$x[n] = s[n] - 2x[n-1] + x[n-2] \qquad H[z] = \frac{1}{1 - 2z^{-1} + z^{-2}}$$

- A trend removal (i.e. local mean subtraction) step is applied to the previous output to obtain GCI locations and strength of excitation information.

$$y[n] = x[n] - \frac{1}{2N+1} \sum_{k=n-N}^{n+N} x[k]; \qquad N + 1 \le n \le L - N$$

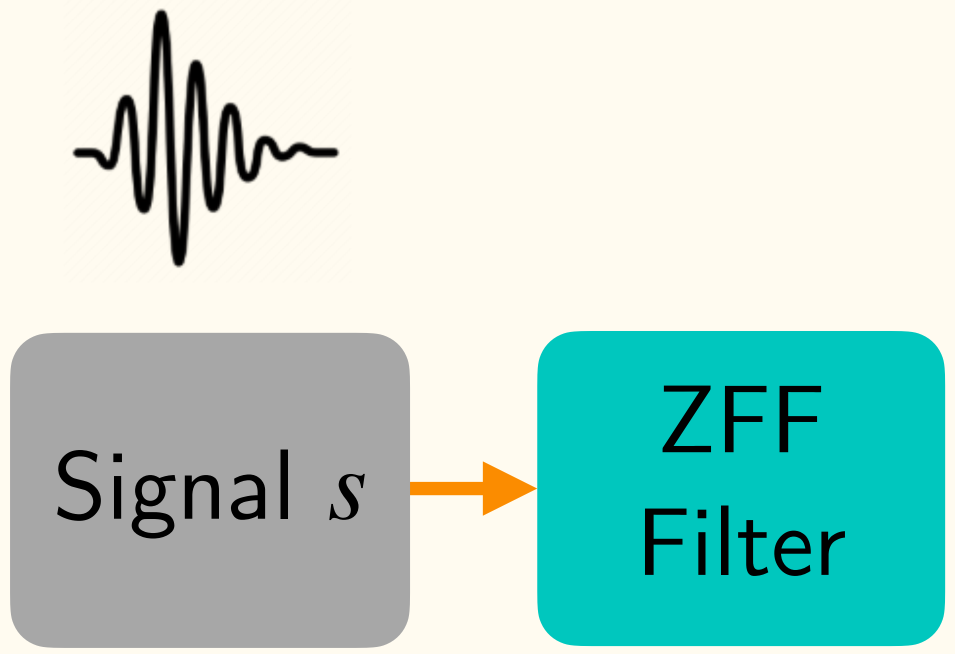

(a)
(b)
Eq. (1)
ZFF Filter
Eq. (2)
Trend Removal
$2N+1$
$\sim$
$[T_0, \frac{T_0}{5}, \frac{T_0}{10}]$
Speech signal $s(n)$
Filtered output $x(n)$
$y_0(n)$
$y_1(n)$
$y_2(n)$
Time [s]
F0
F1
F2
$S(\omega)$
Envelope
Spectrum
$Y_0(\omega)$
F0 (Source)
$Y_1(\omega)$
F1 (System)
$Y_2(\omega)$
F2 (System)
Freq [kHz]

Proposed Method

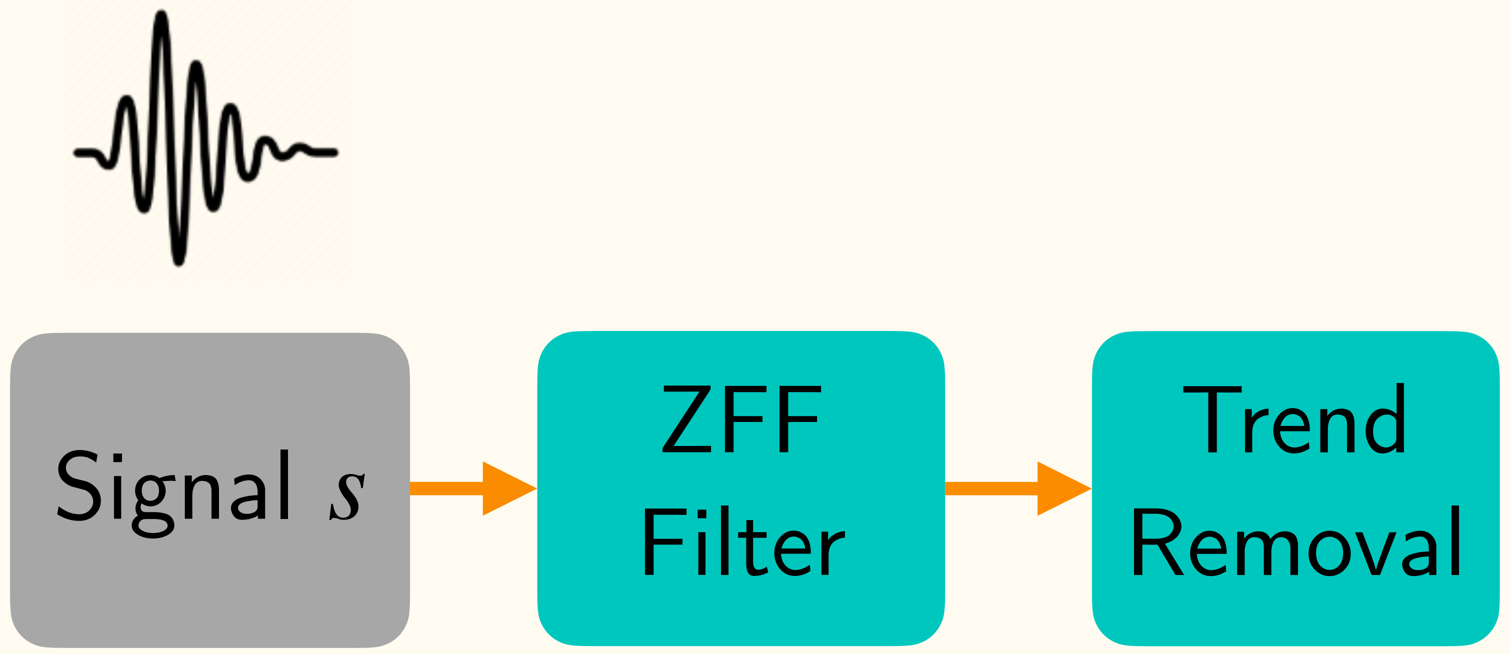
Proposed Method

Signal s

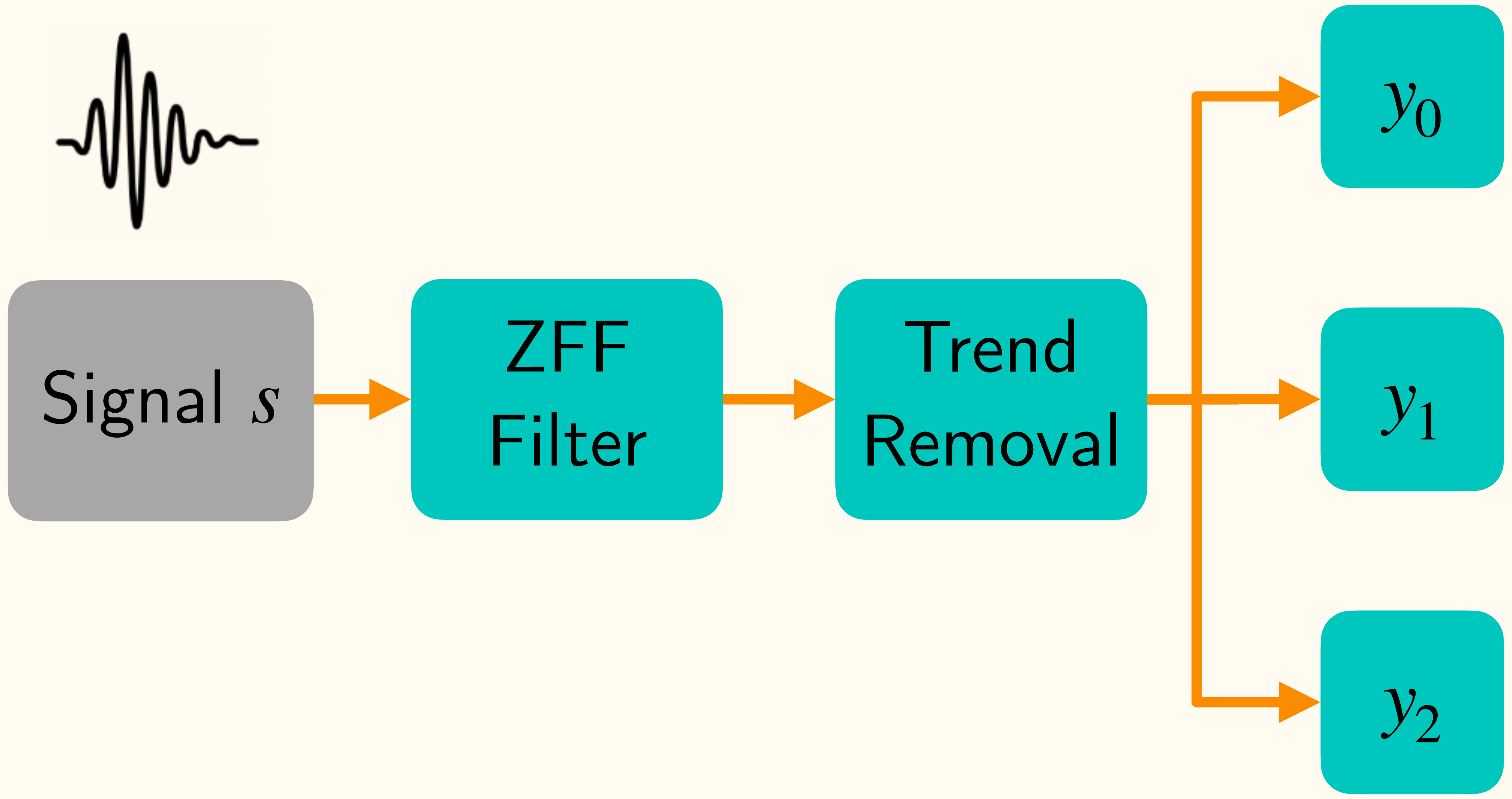
Proposed Method



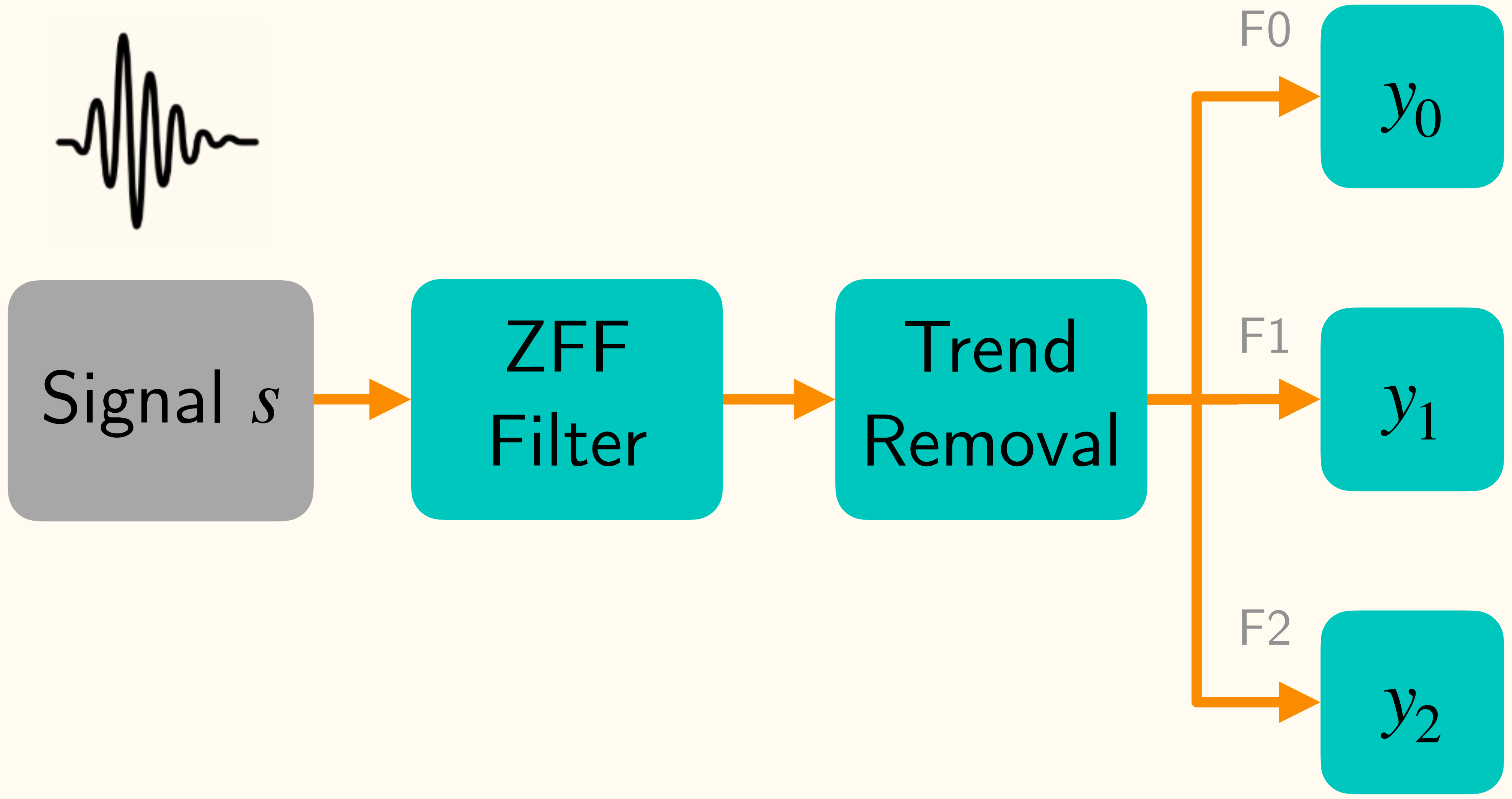
Proposed Method



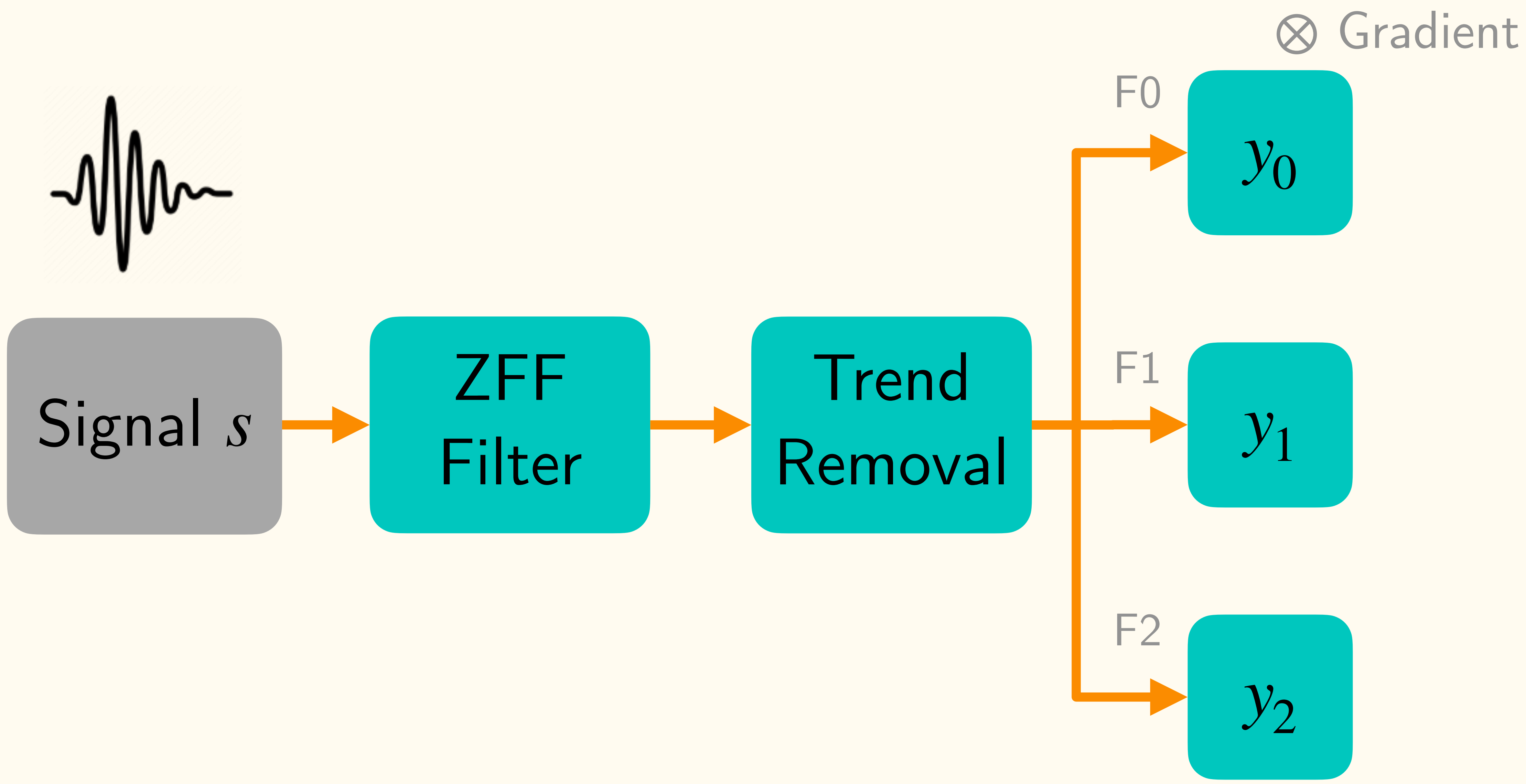
Proposed Method



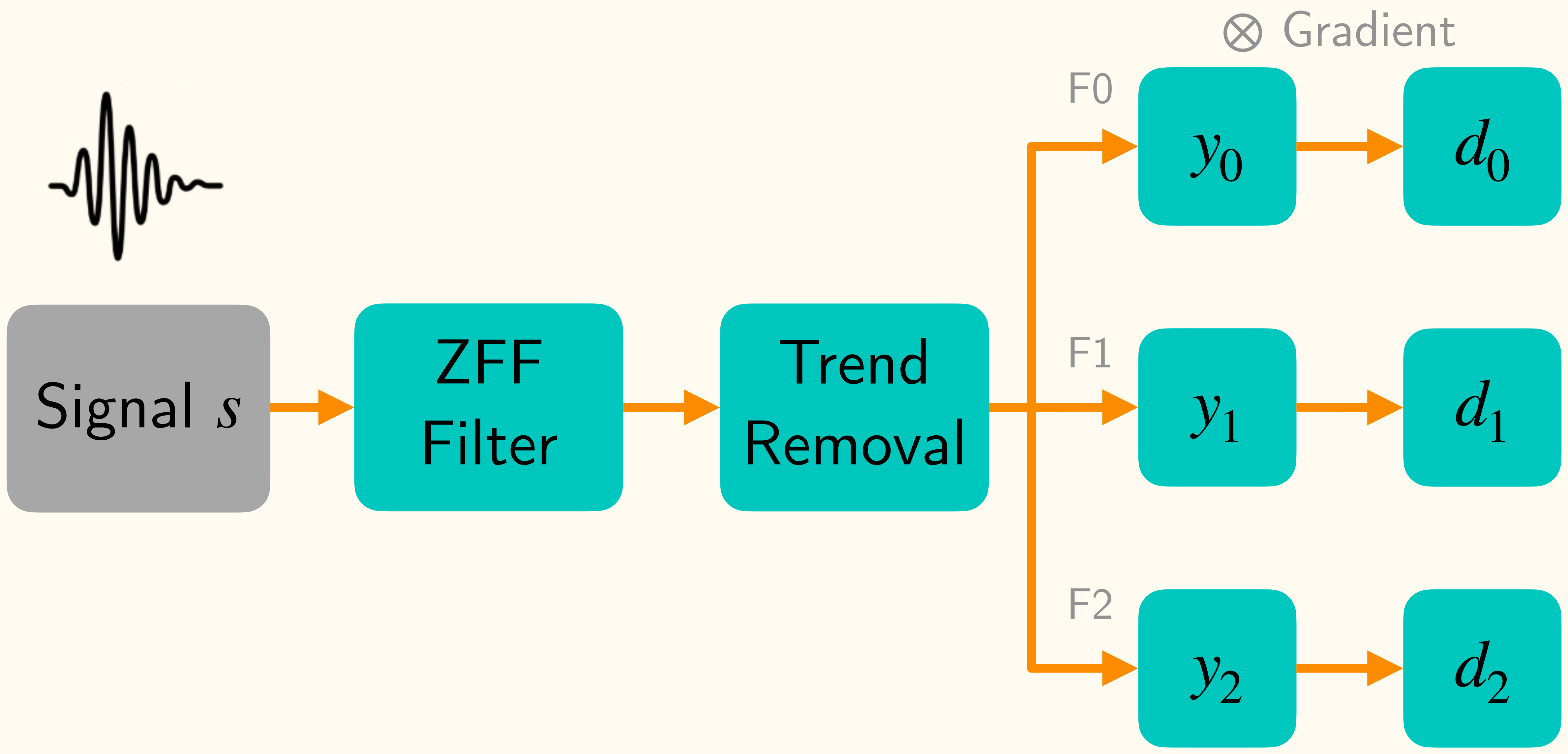
Proposed Method



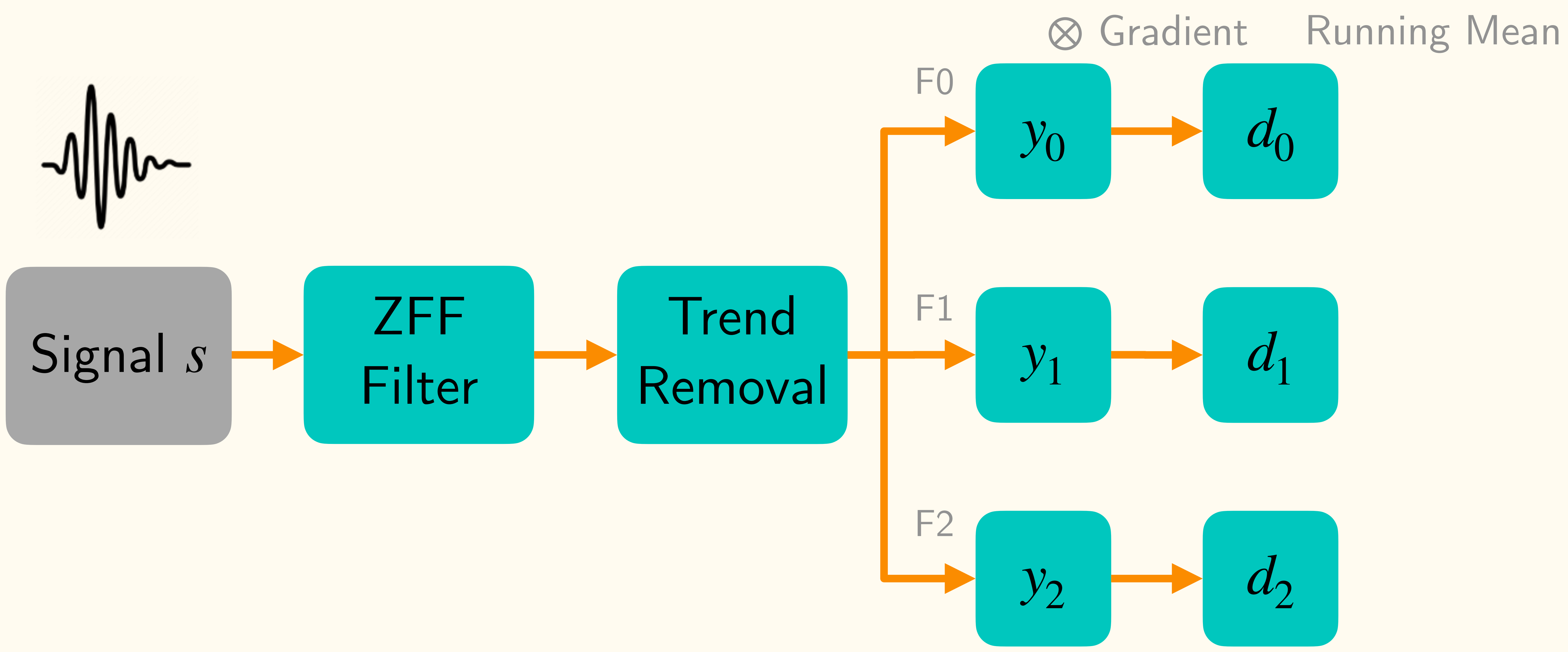
Proposed Method



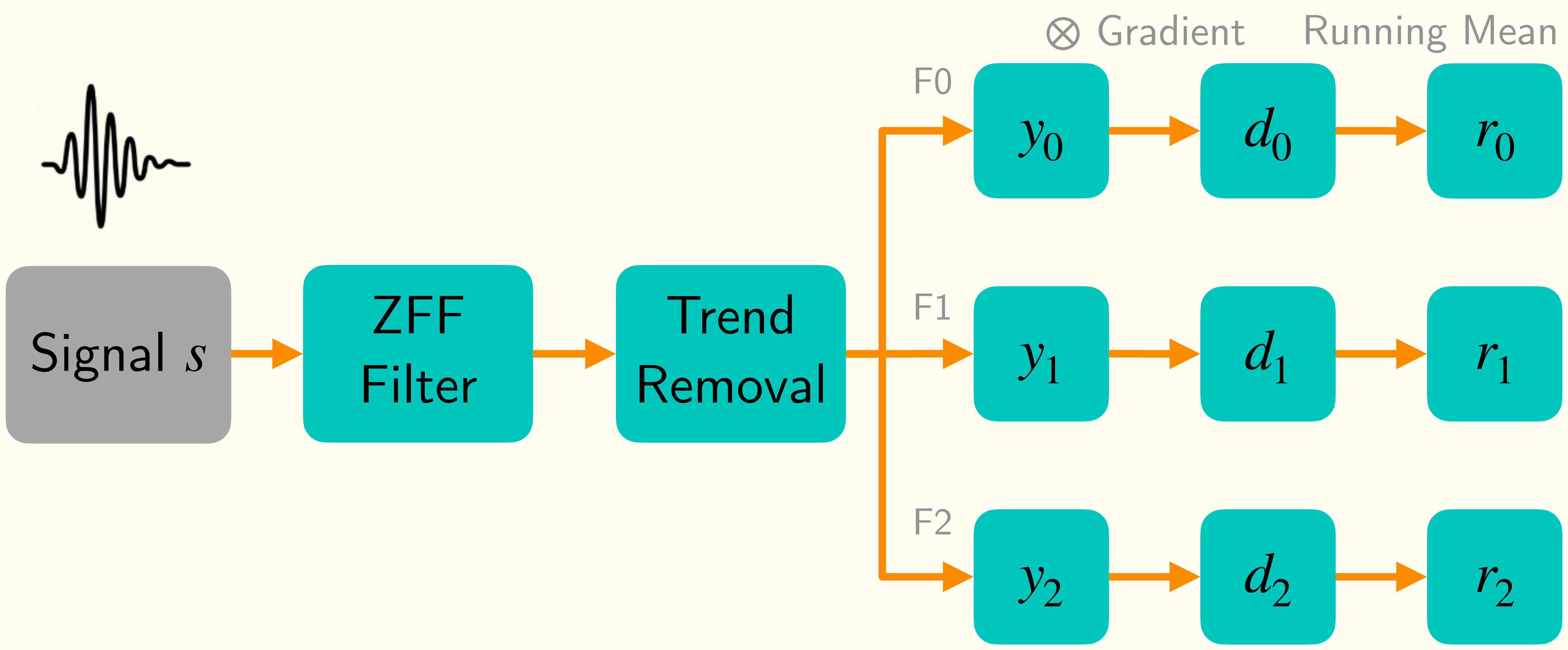
Proposed Method



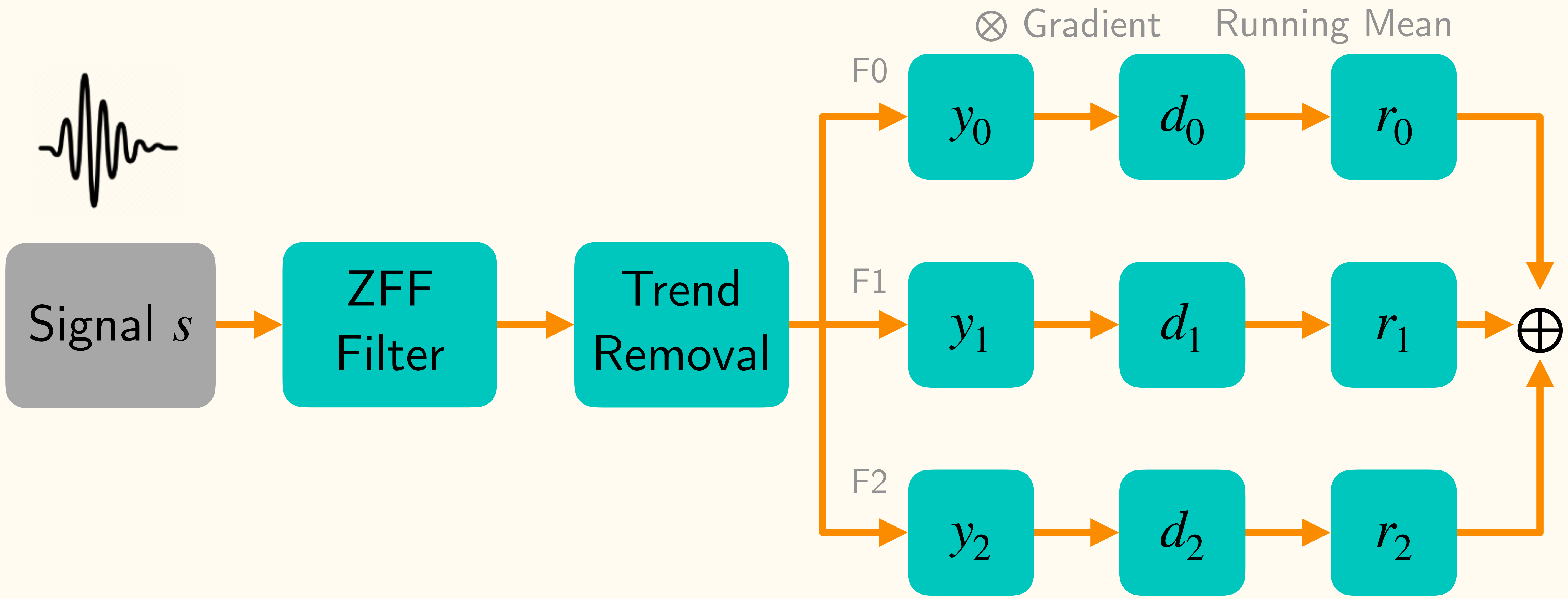
Proposed Method



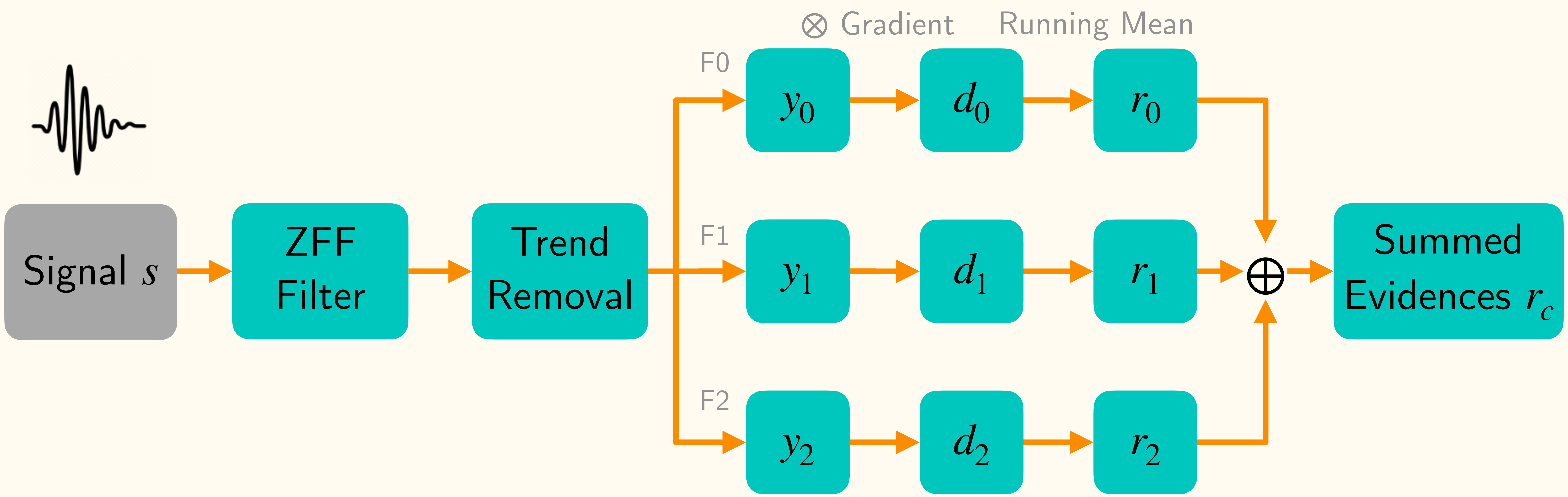
Proposed Method



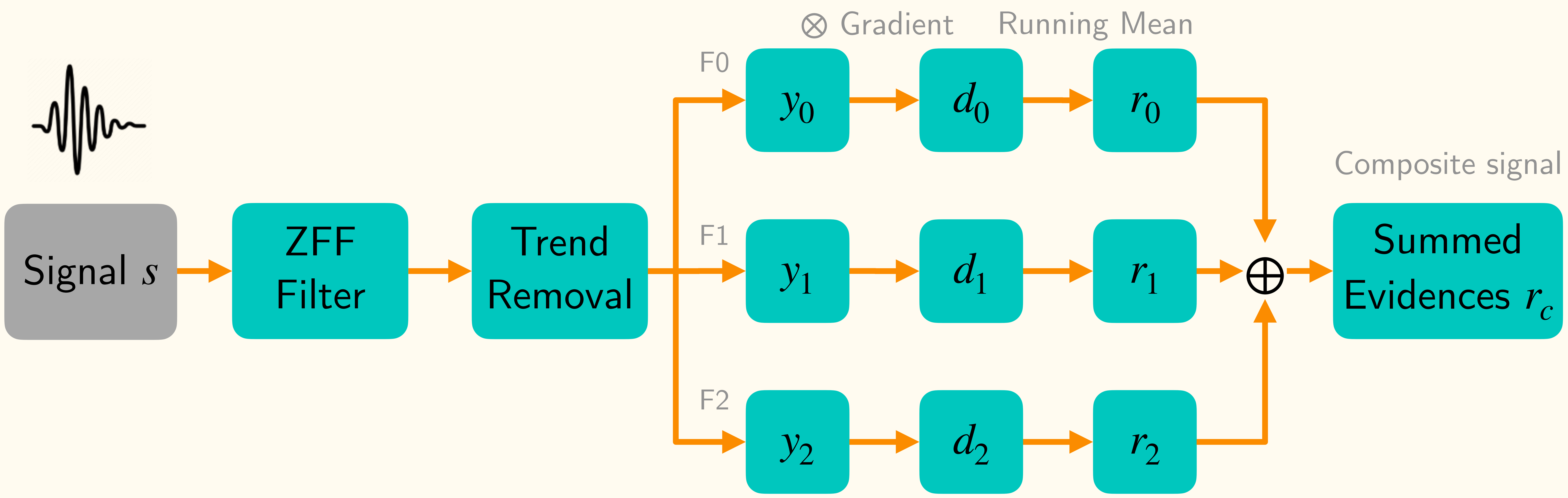
Proposed Method



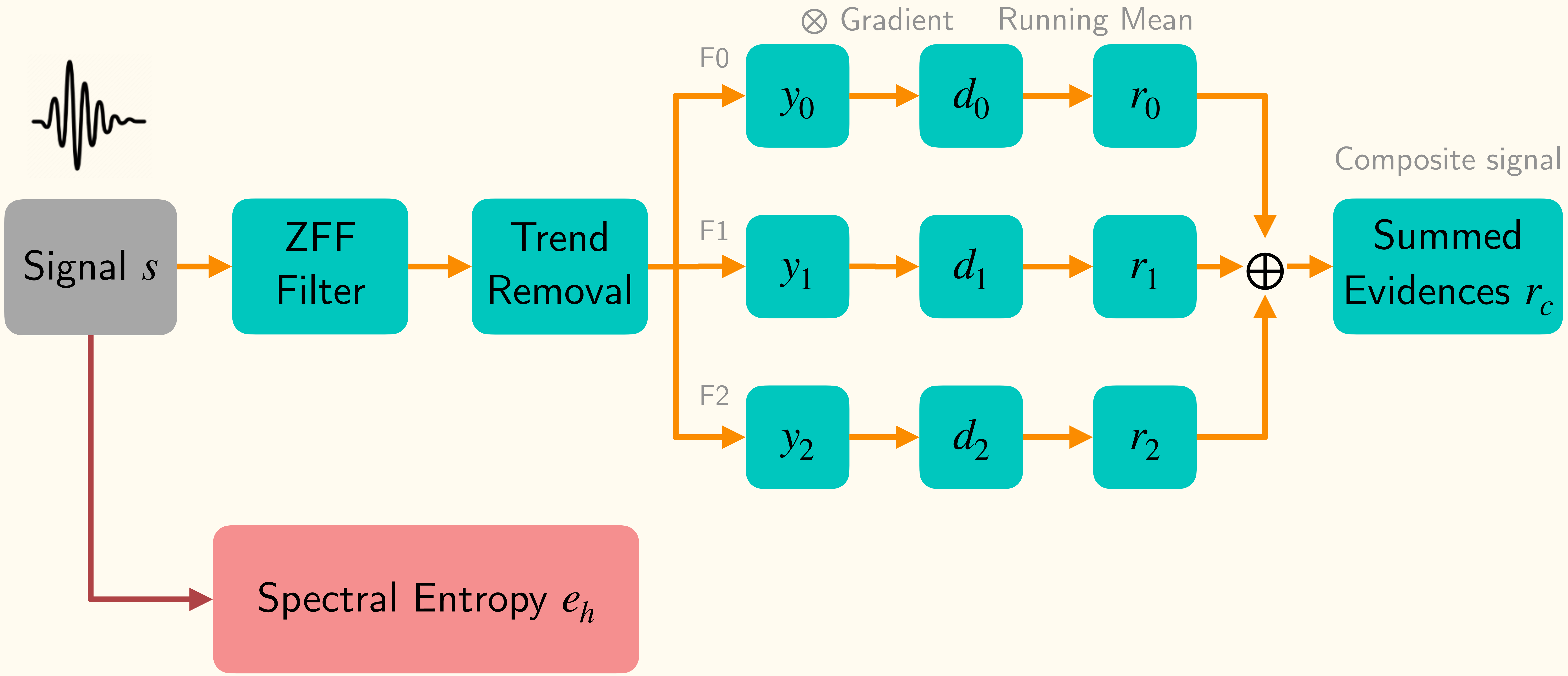
Proposed Method



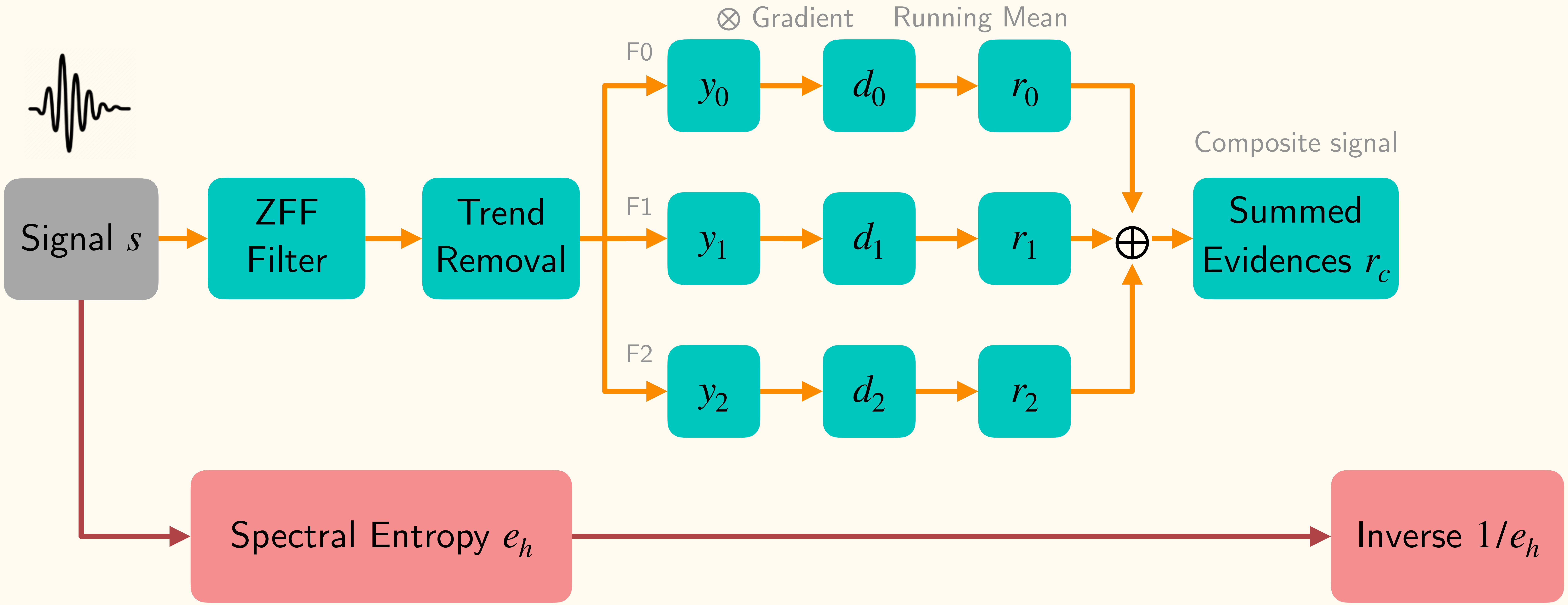
Proposed Method



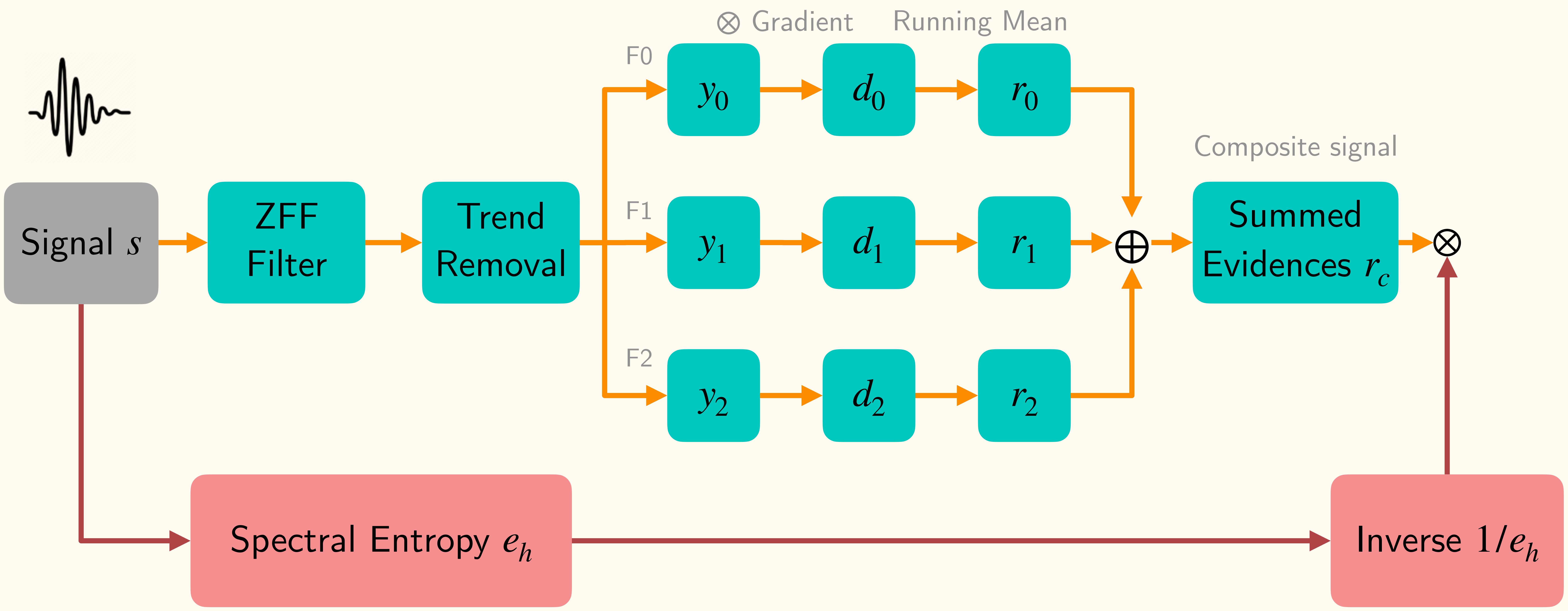
Proposed Method



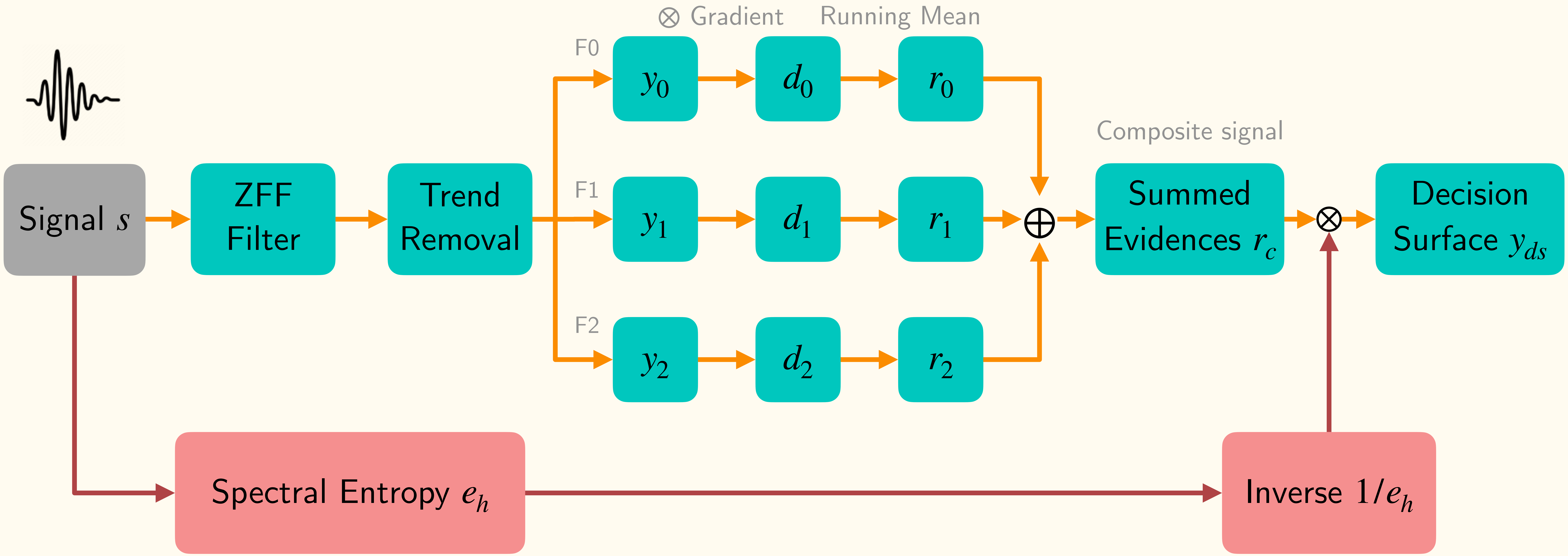
Proposed Method



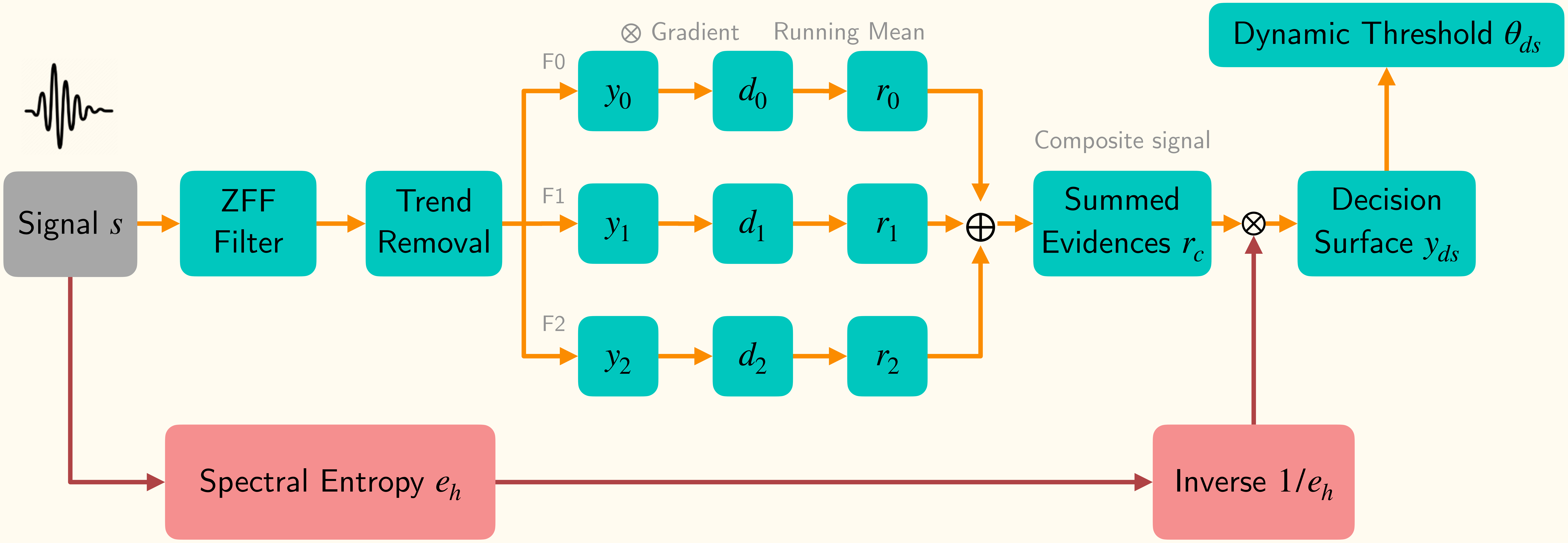
Proposed Method



Proposed Method



Proposed Method



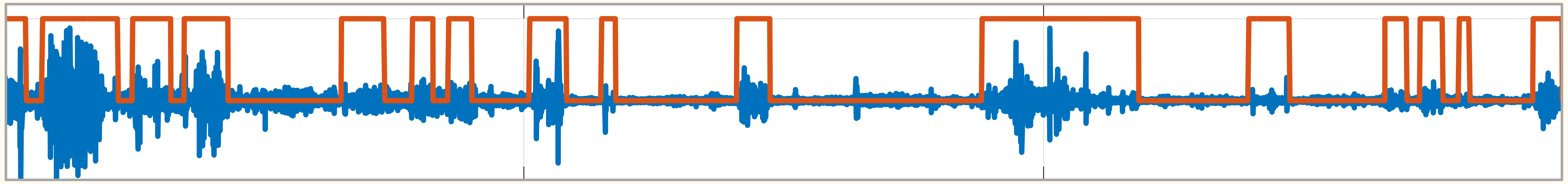
Proposed Method

Proposed Method

Proposed Algorithm

Proposed Algorithm

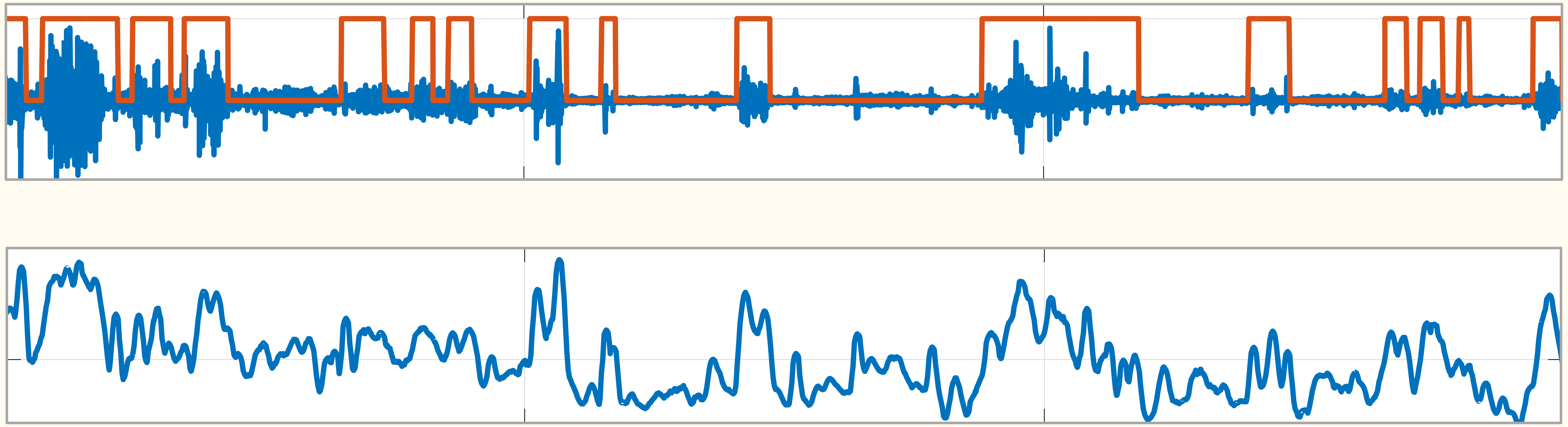
Speech signal s
Decision Boundary



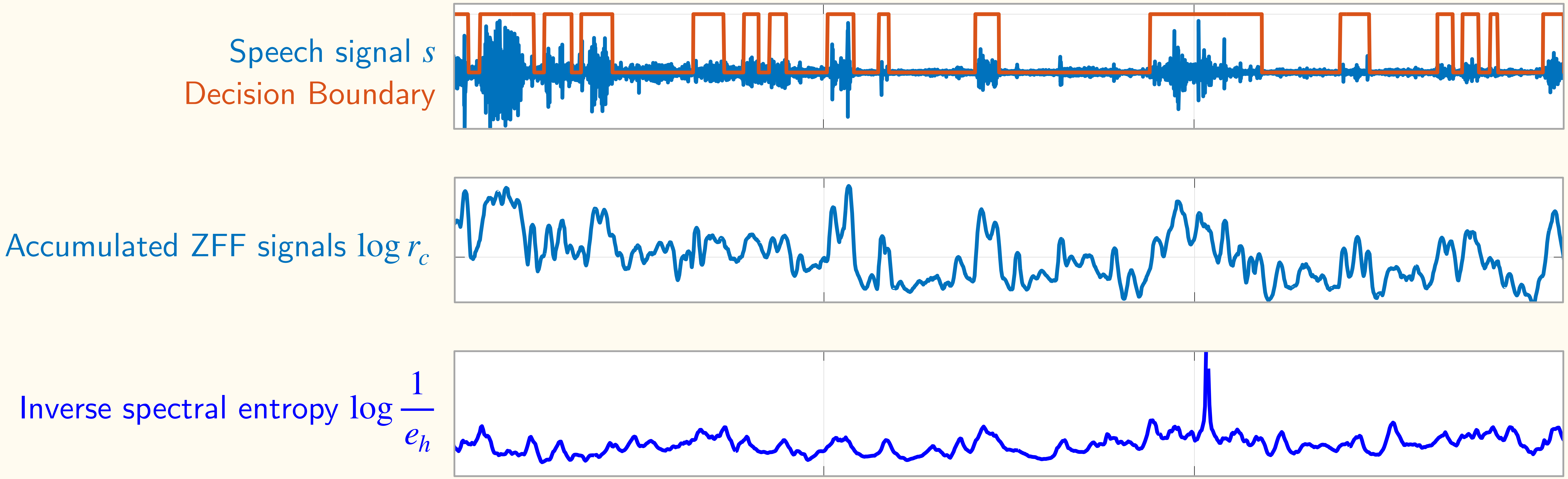
Proposed Algorithm

Speech signal s
Decision Boundary

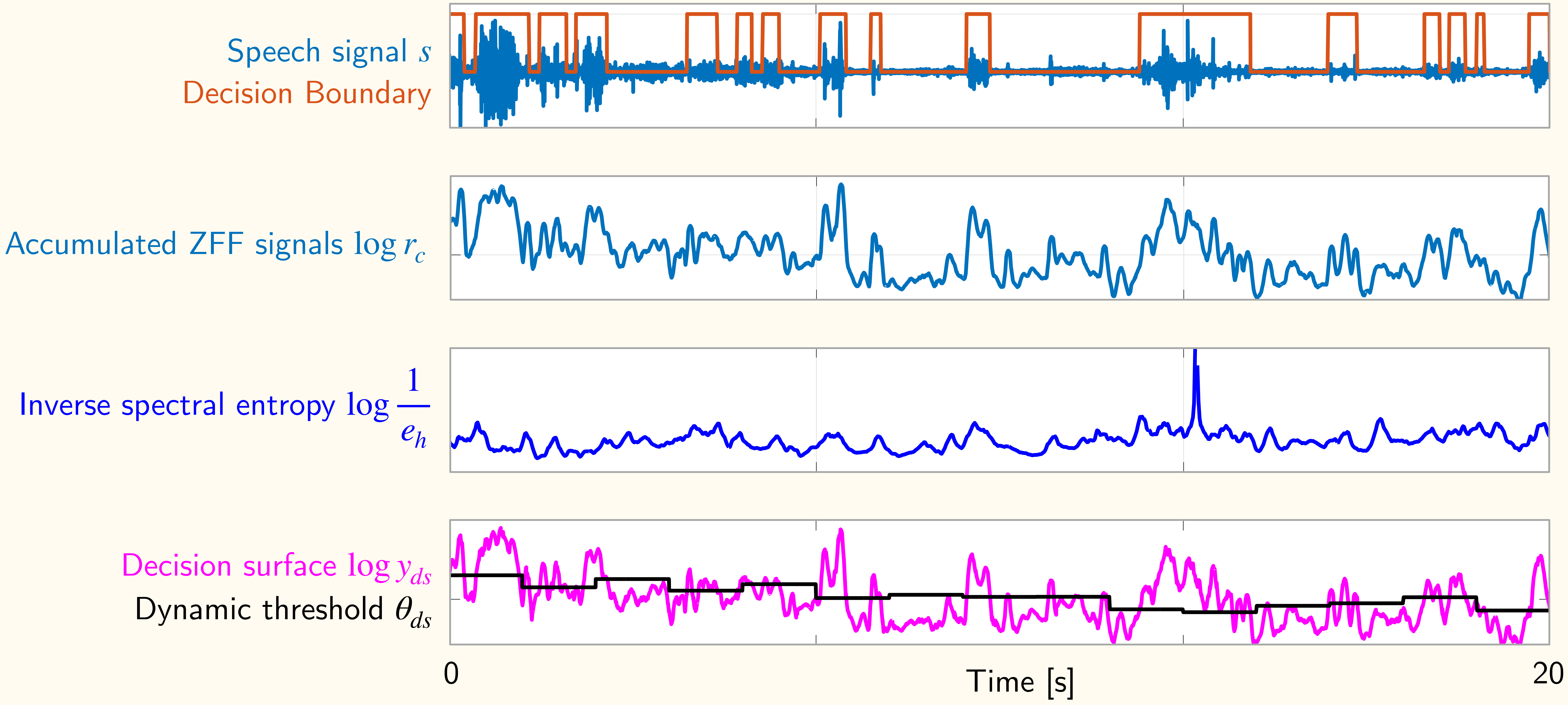
Accumulated ZFF signals $\log r_c$



Proposed Algorithm



Proposed Algorithm



Experimental Setup

Experimental Setup

Database:

Experimental Setup

Database:

- Aurora-2

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Metrics:

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Metrics:

- F1-Score $$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}; \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}; \quad \mathrm{F1} = 2 \cdot \frac{\mathrm{P}\cdot\mathrm{R}}{\mathrm{P}+\mathrm{R}}$$

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Metrics:

- F1-Score $$\mathsf{P} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FP}}; \quad \mathsf{R} = \frac{\mathsf{TP}}{\mathsf{TP} + \mathsf{FN}}; \quad \mathsf{F1} = 2 \cdot \frac{\mathsf{P} \cdot \mathsf{R}}{\mathsf{P} + \mathsf{R}}$$

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Metrics:

- F1-Score $$\mathrm{P} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}; \quad \mathrm{R} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}; \quad \mathrm{F1} = 2 \cdot \frac{\mathrm{P}\cdot\mathrm{R}}{\mathrm{P}+\mathrm{R}}$$

Task:

Experimental Setup

Database:

- Aurora-2
- Sets: *Train*, *Test A*, *Test B*, *Test C*
- SNRs: clean, 20, 15, 10, 5, 0, -5
- Labels: obtained using a HTK recognizer (trained on 12 MFCC coefficients, $\Delta + \Delta\Delta$s + log-energy, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Metrics:

- F1-Score $$\mathsf{P} = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FP}}; \quad \mathsf{R} = \frac{\mathsf{TP}}{\mathsf{TP}+\mathsf{FN}}; \quad \mathsf{F1} = 2 \cdot \frac{\mathsf{P}\cdot\mathsf{R}}{\mathsf{P}+\mathsf{R}}$$

Task:

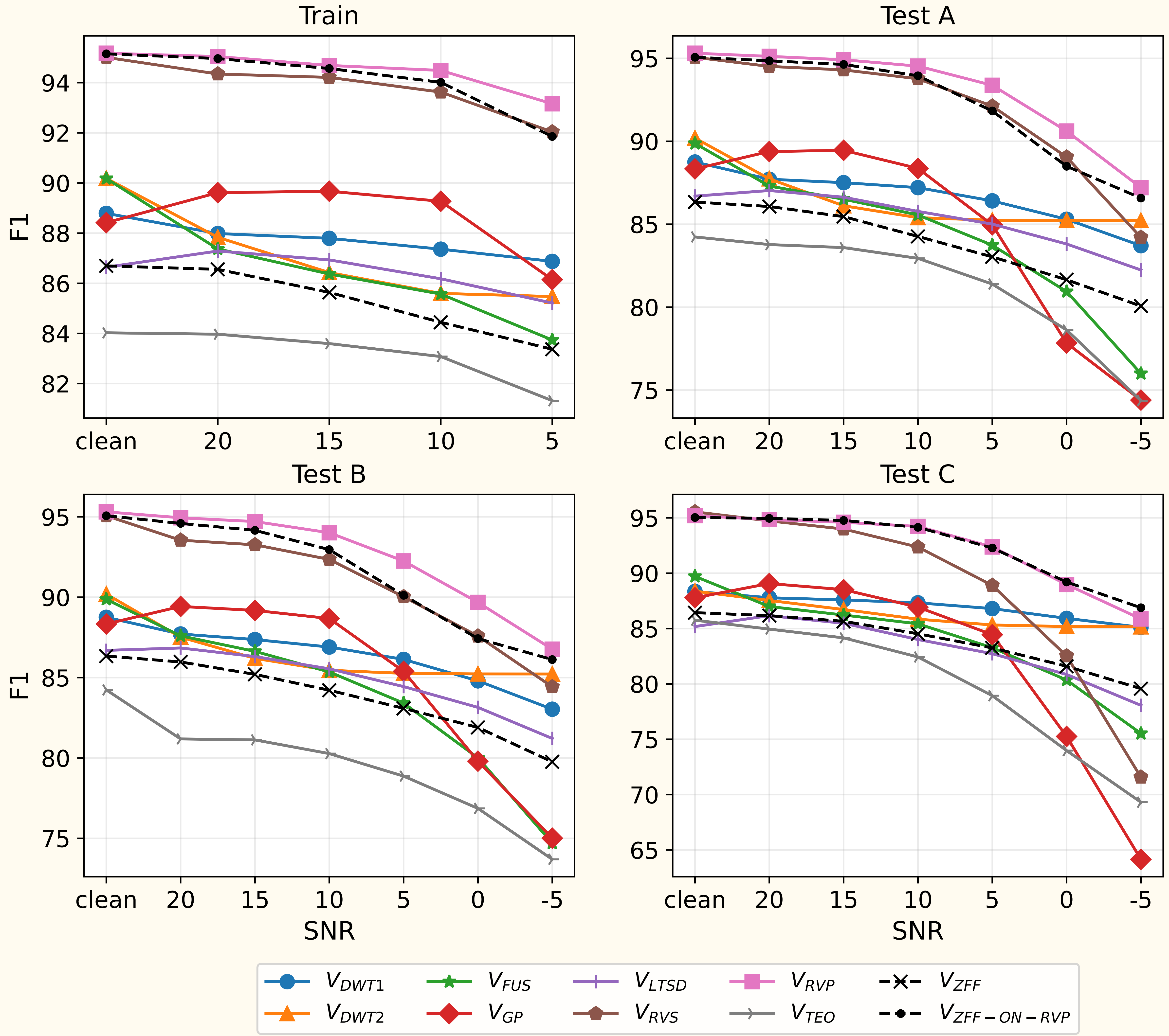
- Binary classification task (speech vs. non-speech) at sample-level.

VAD Baseline Methods

- rVAD (V_{RVP})
- rVAD-Fast (V_{RVS})
- GP-VAD (V_{GP})
- LTSD (V_{LTSD})
- Fusion (V_{FUS})
- Wavlet ($V_{\text{DWT1,2}}$)
- LSD (V_{LSD})
- TEO (V_{TEO})
- LSE (V_{LSE})

Results and Discussion

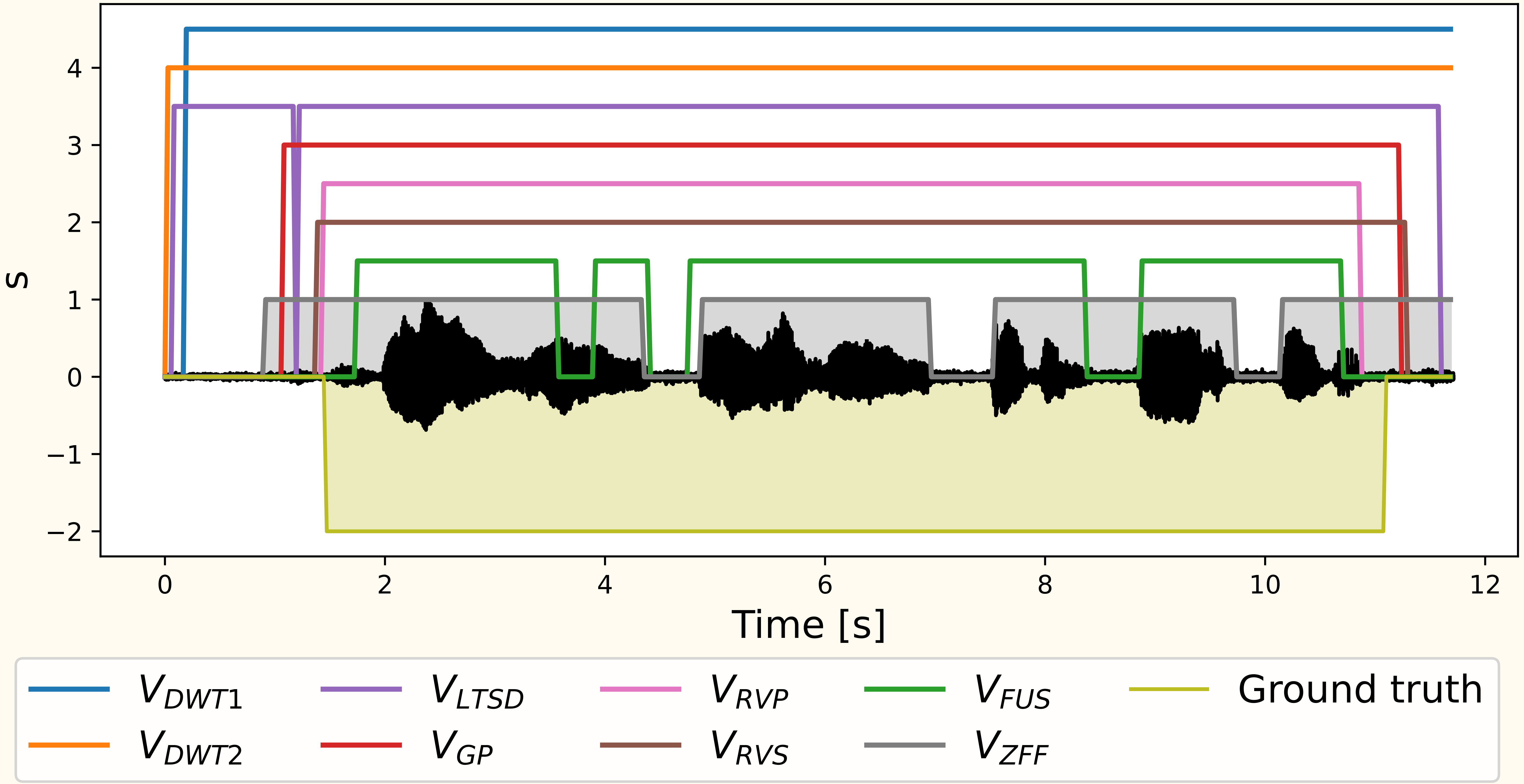
Performance of methods on Aurora-2 across all SNRs and sets.



Method	σ_{F1}
V_{DWT}	1.6
V_{LSD}	1.7
V_{LTSD}	2.0
V_{ZFF}	2.2
V_{LSE}	2.8
V_{RVP}	3.0
$V_{ZFF-ON-RVP}$	3.2
V_{TEO}	3.7
V_{RVS}	4.3
V_{FUS}	4.5
V_{GP}	5.7

Across all test sets

Results and Discussion



- V_{ZFF} remains invariant to added interferences across a range of SNRs.
- V_{ZFF} segments the signal into significantly tighter intervals than other baselines as well the ground truth.

Summary

- Investigated jointly modelling source and system information using ZFF for VAD.
- Proposed and validated two approaches for VAD on the Aurora-2 dataset.
- Investigations demonstrated that VAD can effectively by performed by:
  - Combining filter outputs together to compose a composite signal carrying f_0, F_1, F_2 information, and then applying a dynamic threshold after spectral entropy-based weighting.
  - Passing the composite signal to another VAD.

Summary

- Proposed method produces more refined boundaries compared to other supervised and unsupervised baselines methods in the literature and is robust against degradation as well as channel characteristics.
- First approach operates in time-domain and is relatively less complex to implement.
- Second approach illustrates that the composite signal is an effective representation of speech characteristics, and hence can be used in conjunction with other VADs.

Future Work

- Advantage of proposed method: it does not explicitly assume any mathematical model for the produced speech signal in order to acquire source and system information.
- It can thus also be extended to other types of audio signals, such as animal and bird vocalizations.
- We can also model the composite signal using the raw waveform neural network based modeling approach for supervised voice activity detection.

Thank you !

Idiap Research Institute

www.idiap.ch/~esarkar/

+ 41 27 72 06 322

eklavya.sarkar@idiap.ch