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)

\{ \text{Unsupervised Methods} \}

- Gaussian Mixture Models
- Neural Networks

\{ \text{Supervised Approaches} \}
Voice Activity Detection Landscape

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

- Gaussian Mixture Models
- Neural Networks

Unsupervised Methods

Supervised Approaches
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.
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.
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.
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).
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).
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).

\[
x[n] = s[n] - 2x[n - 1] + x[n - 2] \quad H[z] = \frac{1}{1 - 2z^{-1} + z^{-2}}
\]
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).

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

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

\[ y[n] = x[n] - \frac{1}{2N + 1} \sum_{k=n-N}^{n+N} x[k]; \quad N + 1 \leq n \leq L - N \]
Removal of Trend

ZFF Filter

\[ y_0(n) \]

\[ y_1(n) \]

\[ y_2(n) \]

\[ s(n) \]

\[ x(n) \]

\[ \frac{T_0}{10} \]

\[ \frac{T_0}{5} \]

\[ 2N + 1 \approx \frac{T_0}{5} \cdot \frac{T_0}{10} \]

\[ \text{Envelope (System)} \]

\[ \text{F1 (Source)} \]

\[ \text{F1 (System)} \]

\[ \text{F2 (System)} \]

\[ \text{S(\omega)} \]

\[ \text{Freq [kHz]} \]

\[ \text{Time [s]} \]
Proposed Method
Proposed Method

Signal $s$
Proposed Method

Signal $s$ $\rightarrow$ ZFF Filter
Proposed Method
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal → $y_0$ → $y_1$ → $y_2$
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

F0 → $y_0$
F1 → $y_1$
F2 → $y_2$
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

$F_0$ → $y_0$
$F_1$ → $y_1$
$F_2$ → $y_2$

$\otimes$ Gradient
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

F0: $y_0 \rightarrow d_0$

F1: $y_1 \rightarrow d_1$

F2: $y_2 \rightarrow d_2$

$\otimes$ Gradient
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

$y_0$ → $d_0$

$y_1$ → $d_1$

$y_2$ → $d_2$

$\otimes$ Gradient → Running Mean

F0, F1, F2
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

- $F_0$: $y_0$ → $d_0$ → $r_0$
- $F_1$: $y_1$ → $d_1$ → $r_1$
- $F_2$: $y_2$ → $d_2$ → $r_2$

$\otimes$ Gradient → Running Mean
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

$y_0$ → $d_0$ → $r_0$

$y_1$ → $d_1$ → $r_1$

$y_2$ → $d_2$ → $r_2$

⊗ Gradient → Running Mean

$F_0$ → $F_1$ → $F_2$
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal → Gradient ⊗ Running Mean → Summed Evidences $r_c$
Proposed Method

Signal $s$ \rightarrow \text{ZFF Filter} \rightarrow \text{Trend Removal} \rightarrow \text{Gradient} \rightarrow \text{Running Mean} \rightarrow \text{Composite signal}

$y_0 \rightarrow d_0 \rightarrow r_0$

$y_1 \rightarrow d_1 \rightarrow r_1$

$y_2 \rightarrow d_2 \rightarrow r_2$

\text{Summed Evidences } r_c
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal

- $F_0$: \( y_0 \rightarrow d_0 \rightarrow r_0 \)
- $F_1$: \( y_1 \rightarrow d_1 \rightarrow r_1 \)
- $F_2$: \( y_2 \rightarrow d_2 \rightarrow r_2 \)

- Gradient: $\otimes$ Gradient $\quad$ Running Mean

Composite signal → Summed Evidences $r_c$

Spectral Entropy $e_h$
Proposed Method

\[ e_h \]  

\[ \text{ZFF Filter} \]  
\[ \text{Trend Removal} \]  
\[ \mathcal{O} \text{ Gradient} \]  
\[ \text{Running Mean} \]  
\[ \oplus \]  
\[ \text{Summed Evidences } r_c \]  

\[ \text{Composite signal} \]  

\[ \text{Spectral Entropy } e_h \]  
\[ \text{Inverse } 1/e_h \]
Proposed Method

\[
\text{Summed Evidences } r_c = \sum_{i=0}^{2} y_i \odot \left( \frac{1}{e} \right) S_{\text{spectral entropy}} \otimes y_i \quad \text{Trend Removal}
\]

\[
\text{Composite signal} = r_0 \oplus r_1 \oplus r_2
\]

Signal \( s \) → ZFF Filter → Trend Removal → Spectral Entropy \( e_h \) → Inverse \( 1/e_h \)

\( F_0 \): Gradient → Running Mean

\( F_1 \): \( y_0 \rightarrow d_0 \rightarrow r_0 \)

\( F_2 \): \( y_1 \rightarrow d_1 \rightarrow r_1 \)

\( y_2 \rightarrow d_2 \rightarrow r_2 \)
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal → $y_0$ → $d_0$ → $r_0$

$y_0$ → $d_0$ → $r_0$

$y_1$ → $d_1$ → $r_1$

$y_2$ → $d_2$ → $r_2$

Gradient ⊗ Running Mean

Composite signal

Summed Evidences $r_c$

Decision Surface $y_{ds}$

Spectral Entropy $e_h$

Inverse $1/e_h$
Proposed Method

- Signal $s$ goes through a ZFF Filter.
- Trend Removal follows the ZFF Filter.
- Spectral Entropy $e_h$ is calculated.
- Dynamic Threshold $\theta_{ds}$ is determined.
- Composite signal is generated.
- Decision Surface $y_{ds}$ is formed.

Mathematical expressions:

- Decision Surface: $y_{ds}$
- Summed Evidences: $r_c$
- Inverse: $1/e_h$
- Dynamic Threshold: $\theta_{ds}$

Symbols:

- $y_0$, $d_0$, $r_0$
- $y_1$, $d_1$, $r_1$
- $y_2$, $d_2$, $r_2$
- Gradient
- Running Mean

Equations:

- $y_0 \oplus d_0 \oplus r_0$
- $y_1 \oplus d_1 \oplus r_1$
- $y_2 \oplus d_2 \oplus r_2$

Diagram:

- Signal $s$ → ZFF Filter → Trend Removal → Spectral Entropy $e_h$ → Inverse $1/e_h$ → Dynamic Threshold $\theta_{ds}$ → Decision Surface $y_{ds}$
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal → Summed Evidences $r_c$ → Decision Surface $y_{ds}$ → Voiced Regions $y_{ds} > \theta_{ds}$

- Spectral Entropy $e_h$
- Inverse $1/e_h$
- Dynamic Threshold $\theta_{ds}$

- $F_0$: $y_0 \rightarrow d_0 \rightarrow r_0$
- $F_1$: $y_1 \rightarrow d_1 \rightarrow r_1$
- $F_2$: $y_2 \rightarrow d_2 \rightarrow r_2$

- Gradient $\otimes$
- Running Mean $\oplus$
- Composite signal $\otimes$

- Trend Removal $\rightarrow$ ZFF Filter $\rightarrow$ Signal $s$
Proposed Method

Signal $s$ → ZFF Filter → Trend Removal → $y_0$, $d_0$, $r_0$

$y_1$, $d_1$, $r_1$ → Summed Evidences $r_c$ → Decision Surface $y_{ds}$

Dynamic Threshold $\theta_{ds}$ → Voiced Regions $y_{ds} > \theta_{ds}$

$y_2$, $d_2$, $r_2$ → Inverse $1/e_h$

Spectral Entropy $e_h$ → Smooth

Composite signal $\otimes$ Gradient $\otimes$ Running Mean
Proposed Algorithm
Proposed Algorithm

Speech signal $s$
Decision Boundary
Proposed Algorithm

Speech signal $s$
Decision Boundary

Accumulated ZFF signals $\log r_c$
Proposed Algorithm

Speech signal $s$

Decision Boundary

Accumulated ZFF signals $\log r_c$

Inverse spectral entropy $\log \frac{1}{e_h}$
Proposed Algorithm

Speech signal $s$

Decision Boundary

Accumulated ZFF signals $\log r_c$

Inverse spectral entropy $\log \frac{1}{e^h}$

Decision surface $\log y_{ds}$

Dynamic threshold $\theta_{ds}$

Time [s]
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: \textit{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, $\Delta + \Delta \Delta s + \log\text{-energy}$, 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, $\Delta + \Delta\Delta s + \text{log-energy}$, 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
  - $P = \frac{TP}{TP + FP}$; $R = \frac{TP}{TP + FN}$; $F1 = 2 \cdot \frac{P \cdot R}{P + R}$
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, $\Delta + \Delta \Delta s + \log\text{-energy}$, computed over *Train*, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

Metrics:
- F1-Score
  \[ P = \frac{TP}{TP + FP}; \quad R = \frac{TP}{TP + FN}; \quad F1 = 2 \cdot \frac{P \cdot R}{P + R} \]
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
  \[ P = \frac{TP}{TP + FP}; \quad R = \frac{TP}{TP + FN}; \quad F1 = 2 \cdot \frac{P \cdot R}{P + R} \]

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, $\Delta + \Delta \Delta s + \text{log-energy}$, computed over Train, modeled by 16 HMMs states, each represented by 3 Gaussian mixtures).

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

Task:
- Binary classification task (speech vs. non-speech) at sample-level.
VAD Baseline Methods

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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>$\sigma_{F_1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{DWT}$</td>
<td>1.6</td>
</tr>
<tr>
<td>$V_{LSD}$</td>
<td>1.7</td>
</tr>
<tr>
<td>$V_{LTSD}$</td>
<td>2.0</td>
</tr>
<tr>
<td>$V_{ZFF}$</td>
<td>2.2</td>
</tr>
<tr>
<td>$V_{LSE}$</td>
<td>2.8</td>
</tr>
<tr>
<td>$V_{RV_{P}}$</td>
<td>3.0</td>
</tr>
<tr>
<td>$V_{ZFF-RV_{P}}$</td>
<td>3.2</td>
</tr>
<tr>
<td>$V_{TEO}$</td>
<td>3.7</td>
</tr>
<tr>
<td>$V_{RV_{S}}$</td>
<td>4.3</td>
</tr>
<tr>
<td>$V_{FUS}$</td>
<td>4.5</td>
</tr>
<tr>
<td>$V_{GP}$</td>
<td>5.7</td>
</tr>
</tbody>
</table>

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.
● 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 be 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