Speech Understanding | Major Project

Ayush Abrol B20AI052

Aryan Tiwari B20AI056

Enhancing Voice Activity Detection in Noisy Environments

[Link to Github Repo]

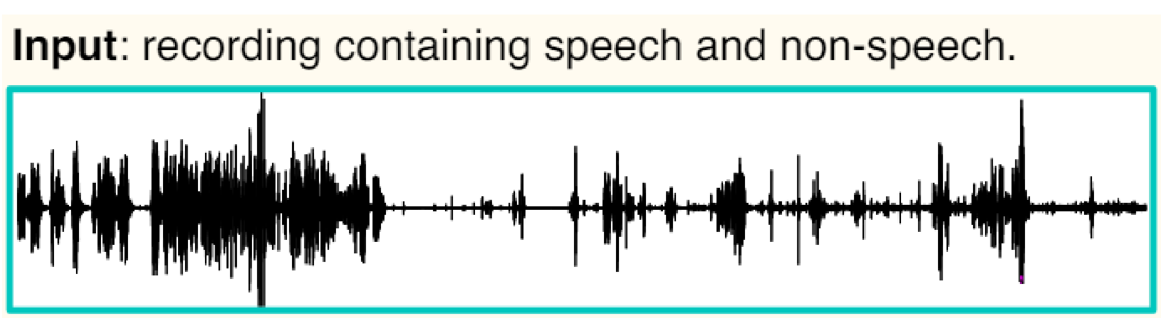
Problem Statement

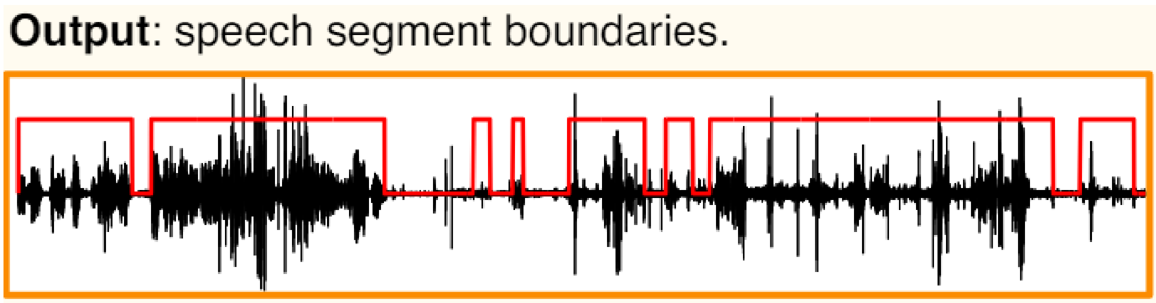
The project, ”Enhancing Voice Activity Detection in Noisy Environments,” addresses the critical need for robust voice-controlled applications in adverse acoustic conditions. In the context of increasing demand for voice assistants and applications, accurately distinguishing speech from noise becomes paramount. Conventional Voice Activity Detection (VAD) systems often falter in noisy environments, impacting user experience. The project aims to explore traditional signal processing technique known as Zero-Frequency Filtering which is an unsupervised method for Voice Activiy Detection.

The project investigates the potential of zero-frequency filtering for jointly modeling voice source and vocal tract system information, and proposes two approaches for Voice Activity Detection (VAD):

1. Demarcating voiced regions using a composite signal composed of different zero-frequency filtered signals.
2. Feeding the composite signal as input to the **rVAD algorithm.**

Our task is to identify segment boundaries in signals which contain voicing information





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 (fo), first formant (F1) and second formant (F2).

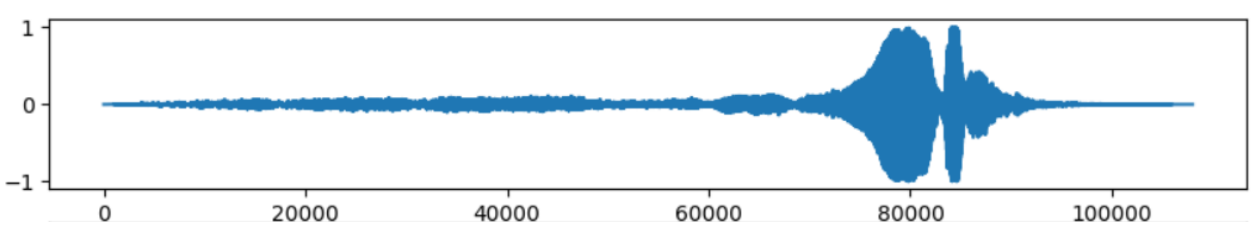
Dataset

The dataset used is Musan Dataset and its noise corpus.

[Dataset Link]

We are using the noise/sound-bible portion and the speech/librivox portion of the corpus.
The noise/sound-bible portion contains 88 recordings. It captures a wide variety of technical and non-technical noises that cannot be considered speech or music. Some recordings feature an ambient environment, e.g., walking through a city. The ambient sounds usually do not feature any intelligible speech. Technical sounds include DTMF tones, various cellphone noises (such as button presses or vibration), dialtones and more. Non-technical sounds include thunder, lighting, clapping, car horns, animal sounds, and more.

The speech/librivox portion contains 175 recordings which contains human speech sentences.



Sample audio from the dataset visualized (Ambulance siren from the sound-bible portion)

Methodology and Algorithm

Zero Frequency Filtering

- ZFF transforms the signal into filtered ones which contain f0, F1, and F2 evidences.

$$x[n] = s[n] - 2x[n-1] + x[n-2]$$

$$H(z) = \frac{1}{1 - 2z^{-1} + z^{-2}}$$

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

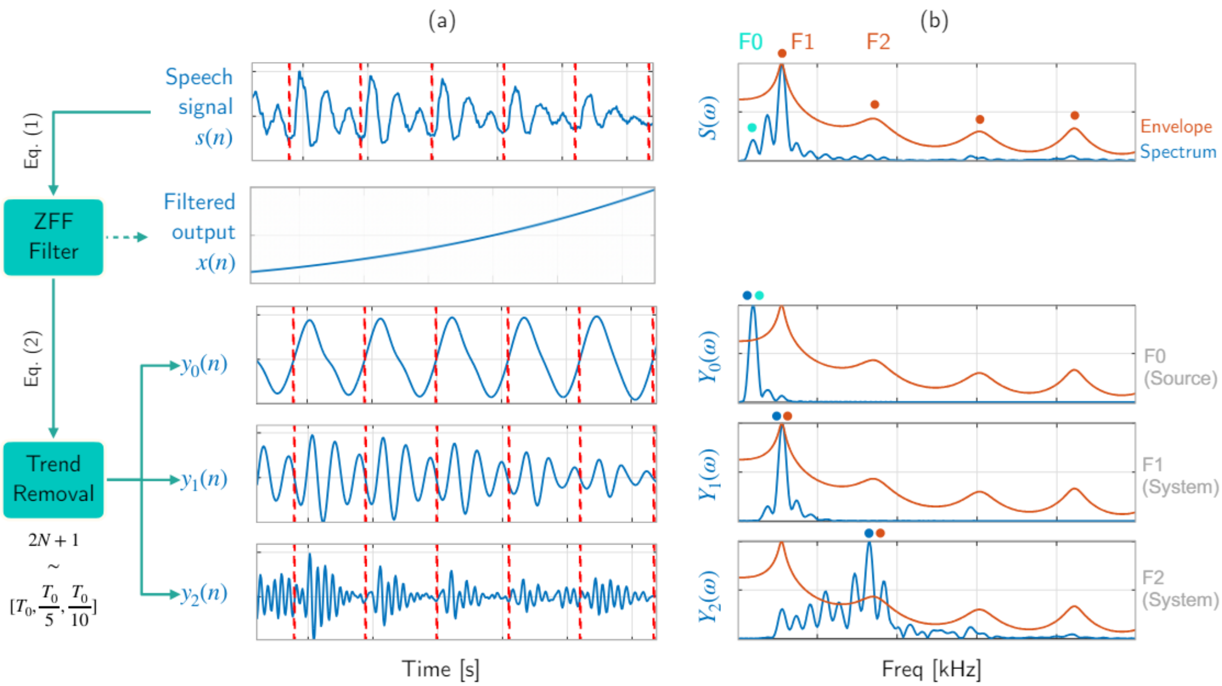


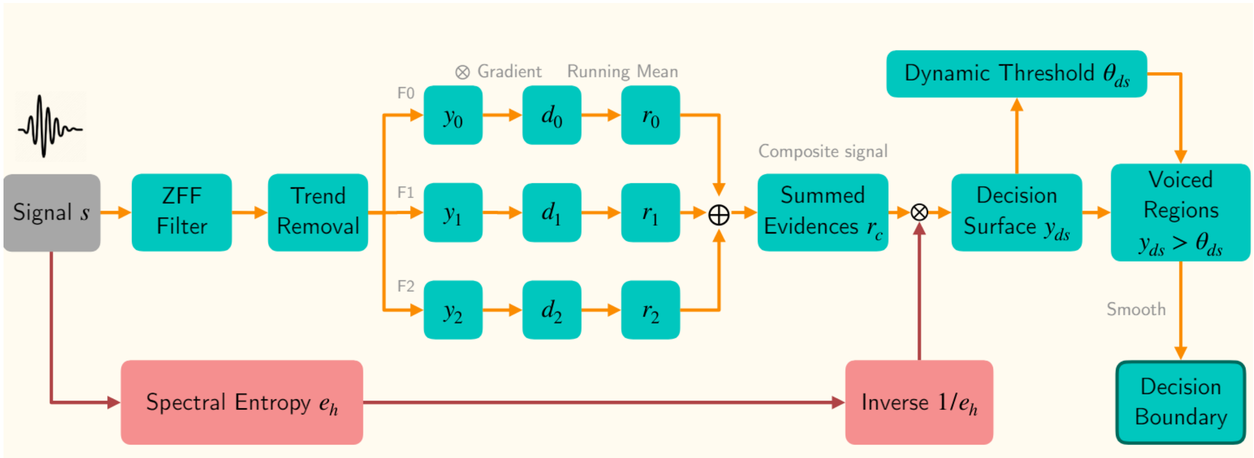
Figure 1. (a1) Speech signal. (a2) Filtered output. (a3–a4) ZFF signals y0(n), y1(n), y2(n).

Figure 2. (b1) S(ω) (—) and its envelope (—). Formant peaks (•). Fundamental frequency peak (•). (b3–b4) Y0(ω), Y1(ω), Y2(ω), and respective peaks (•).

Note: *GCI stands for Glottal Closure Instants. Glottal Closure Instants represent the points in time when the vocal folds come together during speech production. These instants are significant in speech analysis because they provide information about the fundamental frequency (pitch) and the timing of speech sounds.*

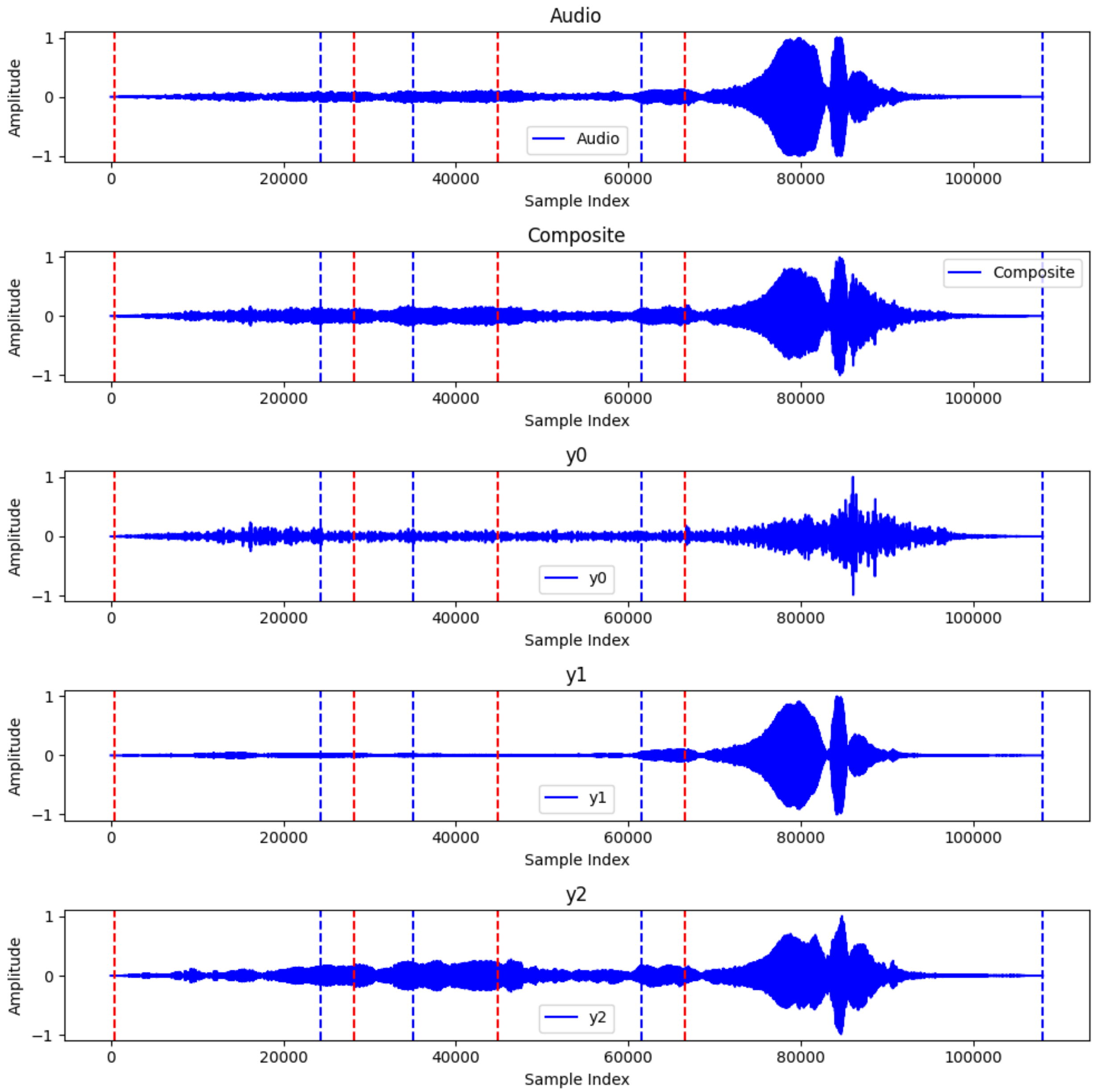
- In this method, a speech signal is first passed through a cascade of digital resonators centered at 0 Hz, i.e. a zero-frequency filter.
- The resulting impulse response of these cascaded resonators, implemented as an integrator, is given by eq. (1) and the equivalent transfer function by eq. (2) above.
- A trend removal (i.e. local mean subtraction) step is applied to the previous output to obtain GCI locations and strength of excitation information.

Pipeline

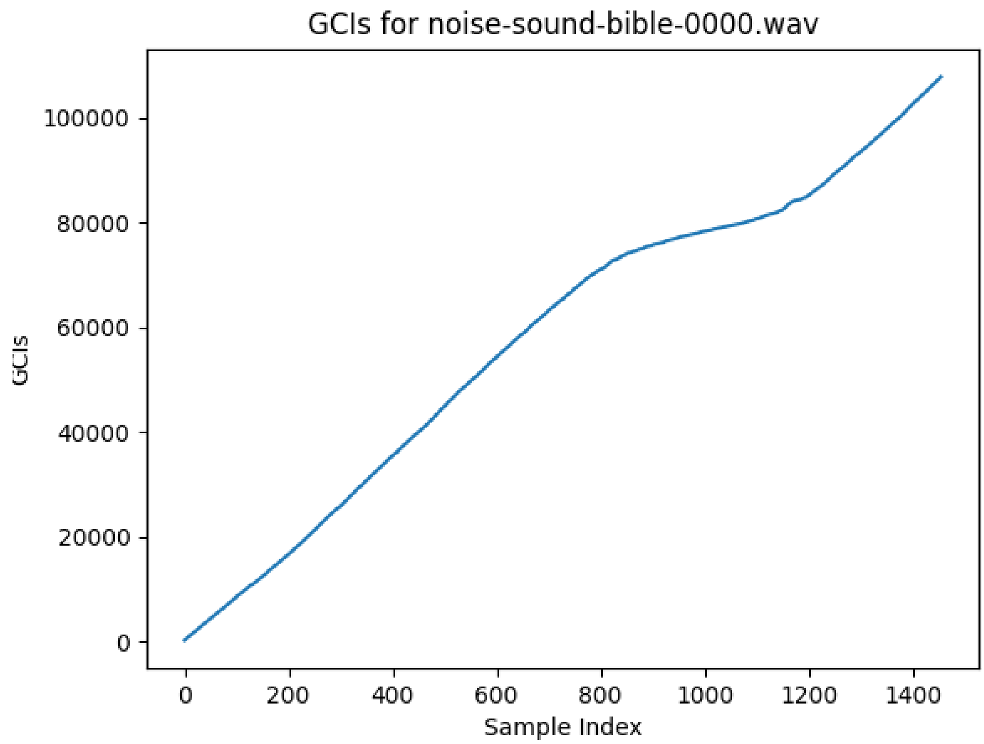


Results and Analysis

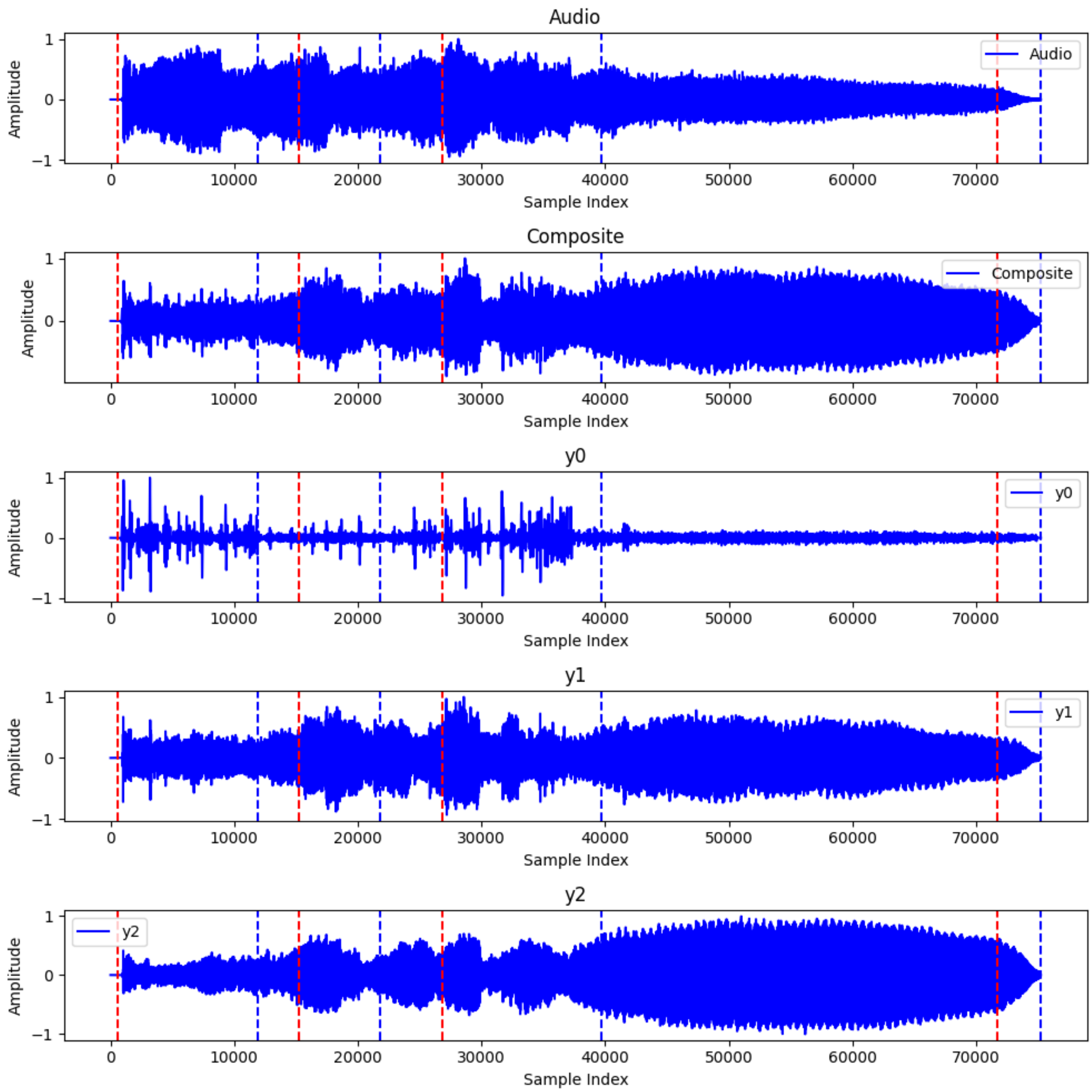
Sound-Bible Corpus for Musan Dataset

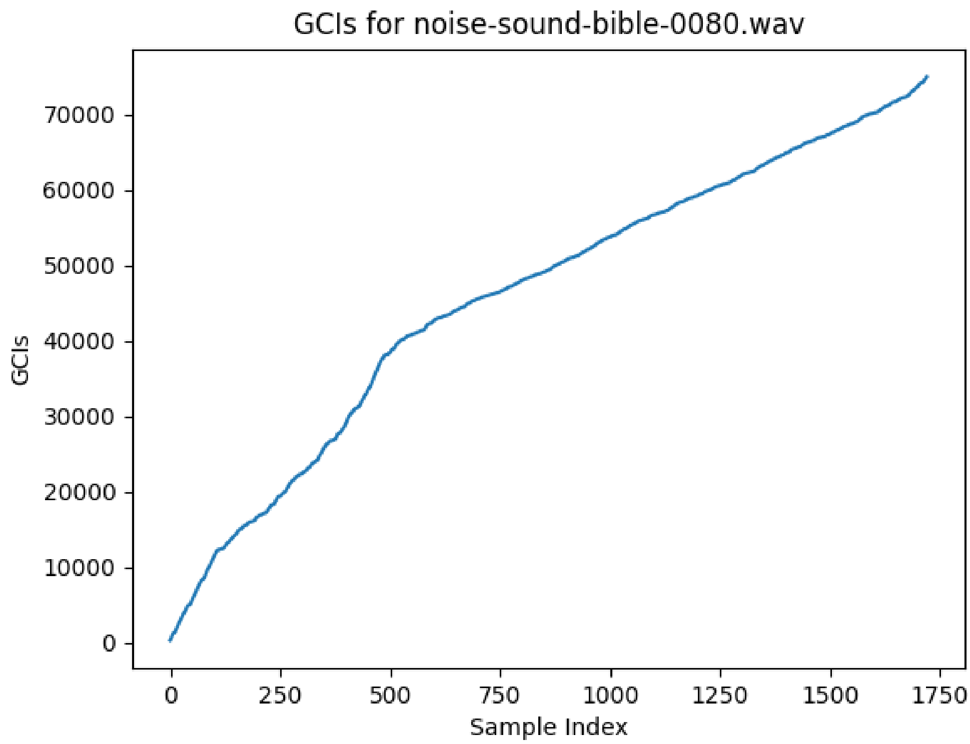


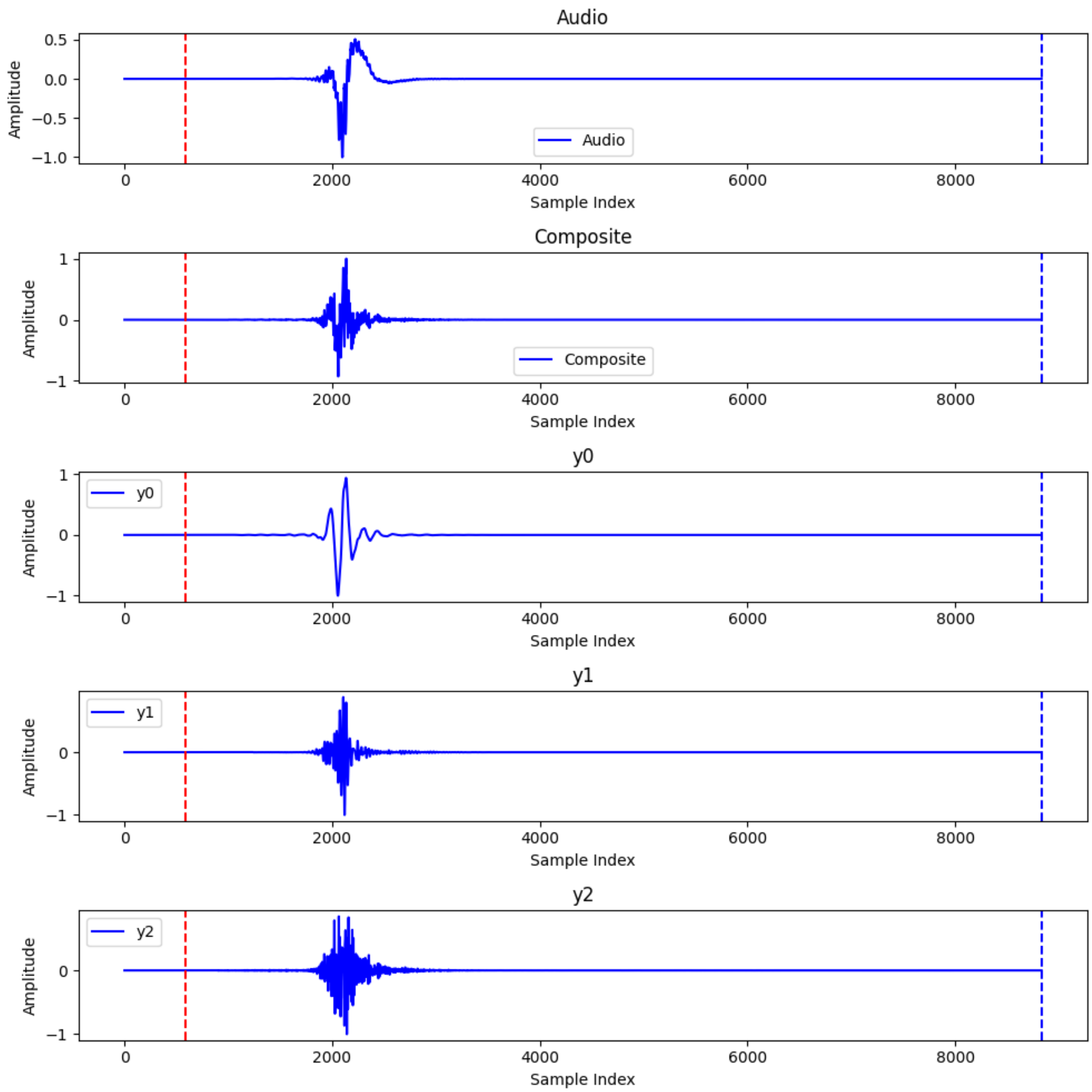
Here, the area between a red line and a blue line represents the decision boundary of the part of speech.

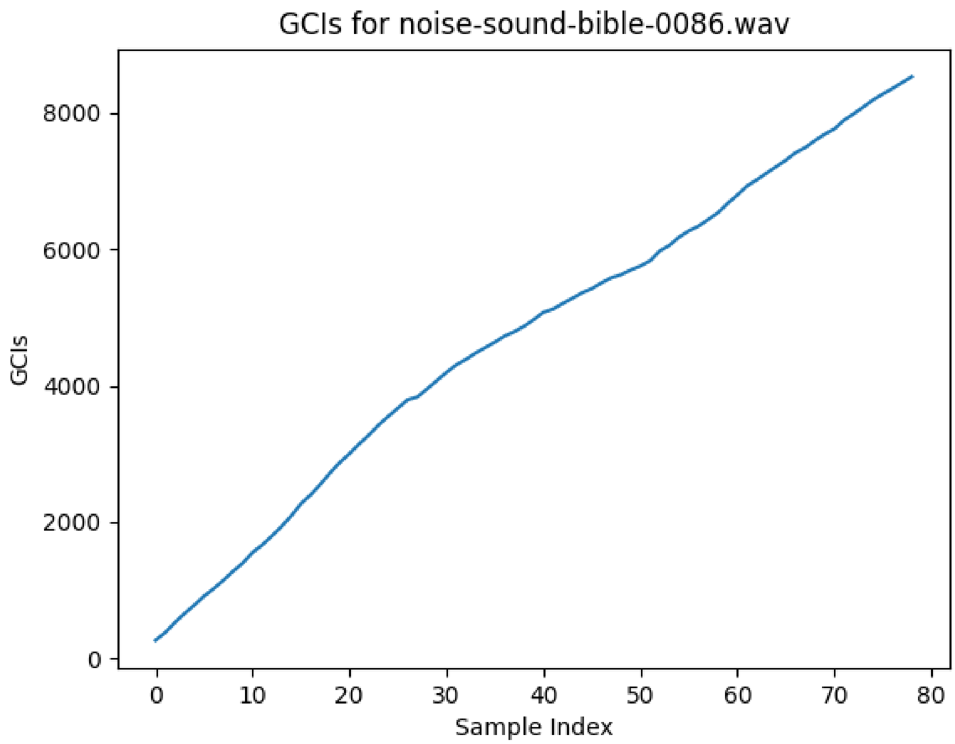


Some more examples:

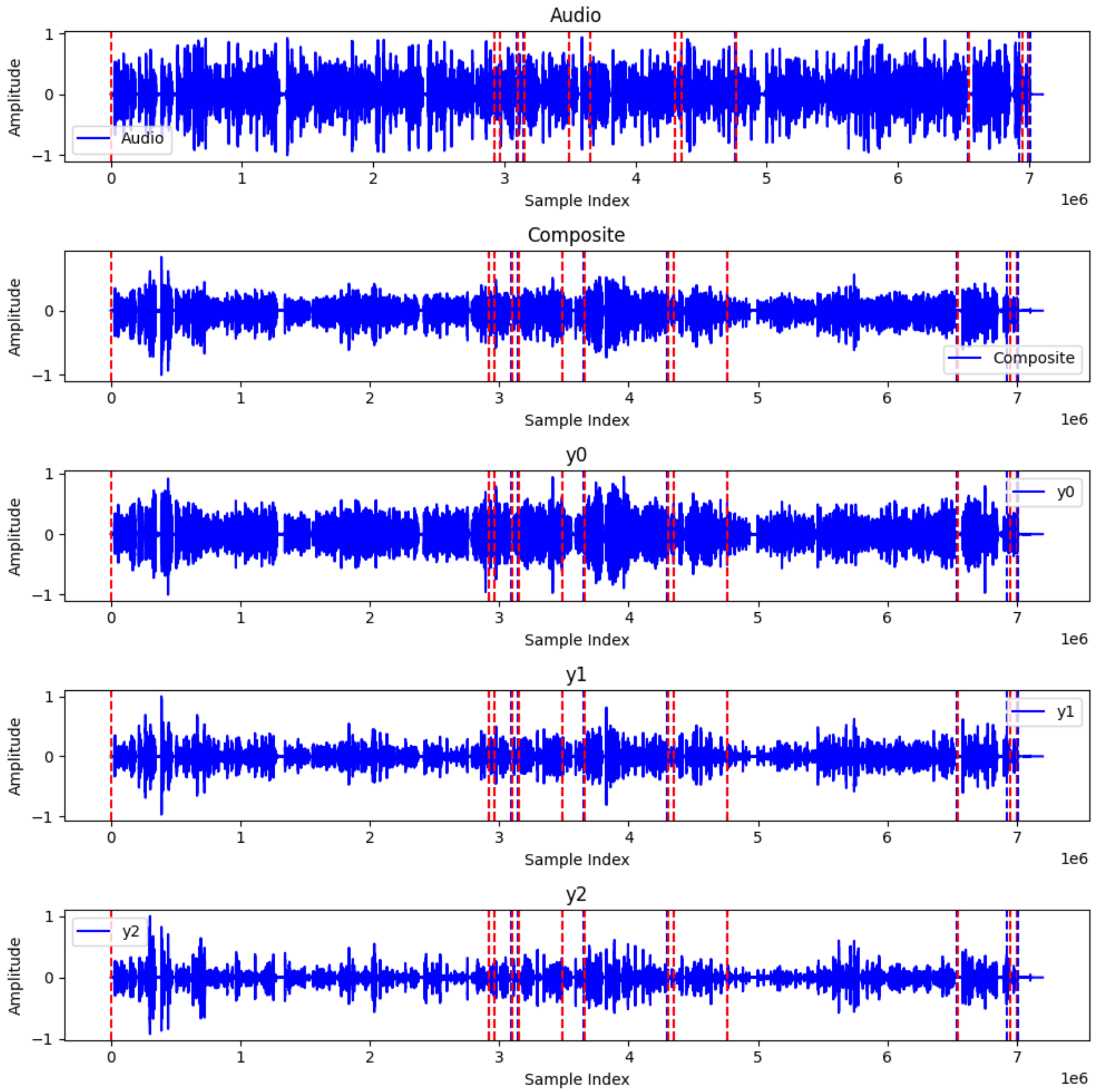

Audio
Amplitude
Sample Index
Composite
y0
y1
y2
1
0
−1
0
10000
20000
30000
40000
50000
60000
70000


GCIs for noise-sound-bible-0080.wav
GCIs
0
10000
20000
30000
40000
50000
60000
70000
0
250
500
750
1000
1250
1500
1750
Sample Index

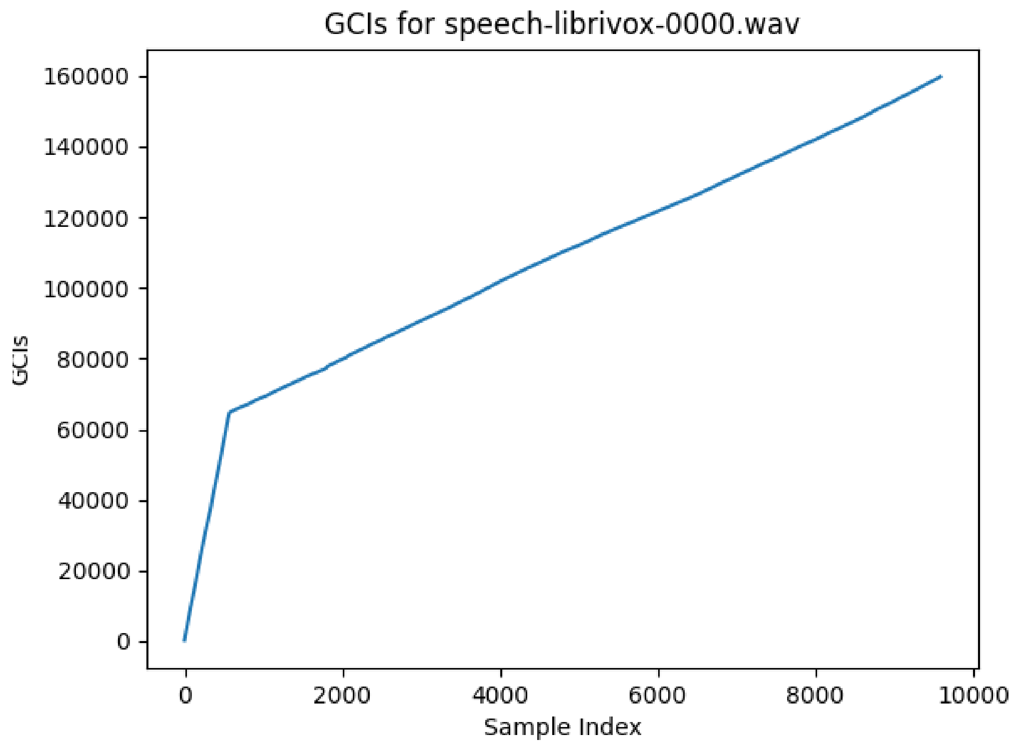

Audio
Amplitude
0.5
0.0
−0.5
−1.0
Audio
0
2000
4000
6000
8000
Sample Index
Composite
Amplitude
1
0
−1
Composite
0
2000
4000
6000
8000
Sample Index
y0
Amplitude
1
0
−1
y0
0
2000
4000
6000
8000
Sample Index
y1
Amplitude
0
−1
y1
0
2000
4000
6000
8000
Sample Index
y2
Amplitude
0
−1
y2
0
2000
4000
6000
8000
Sample Index



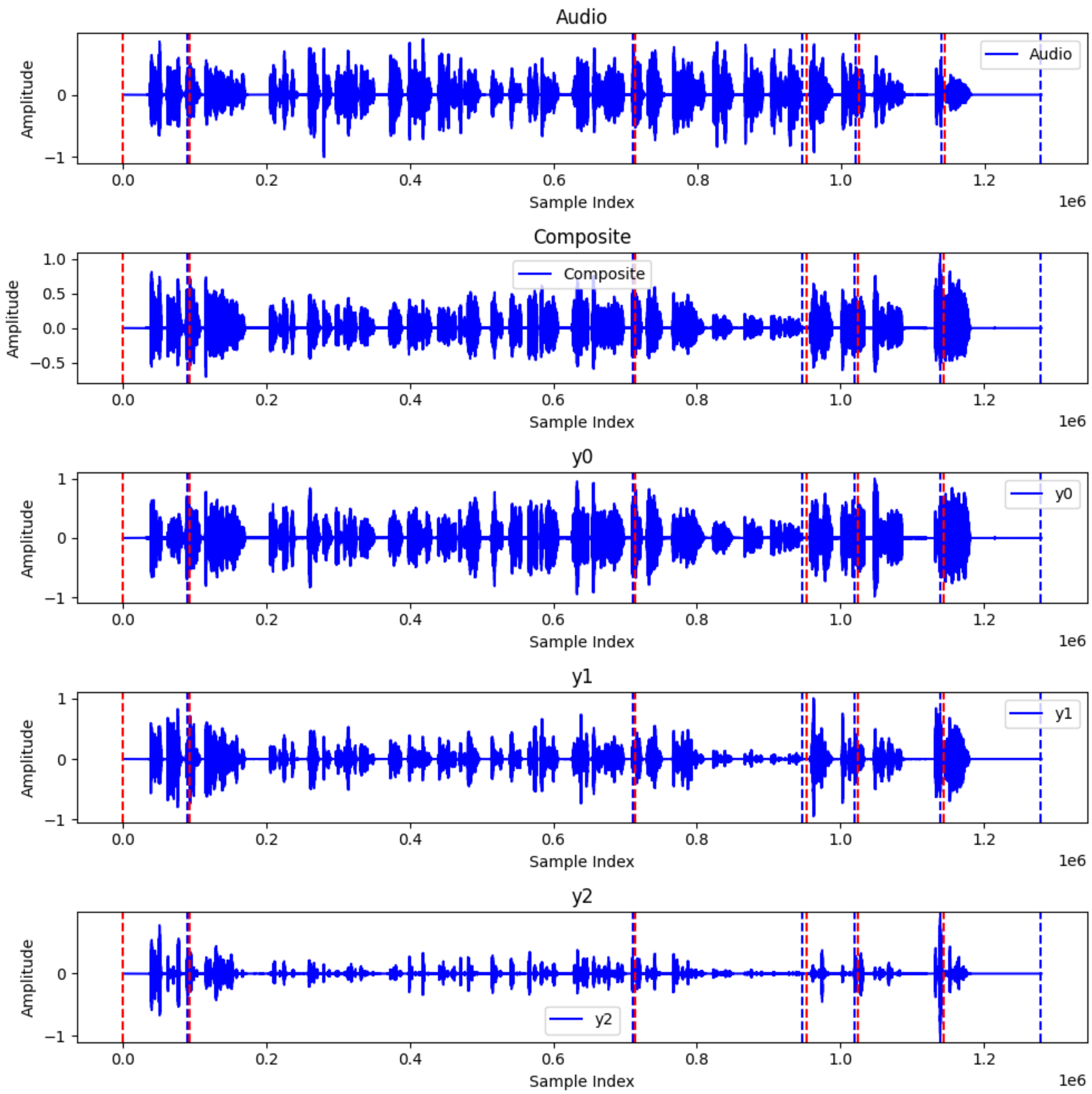
Librivox dataset of Human Speech Sentences

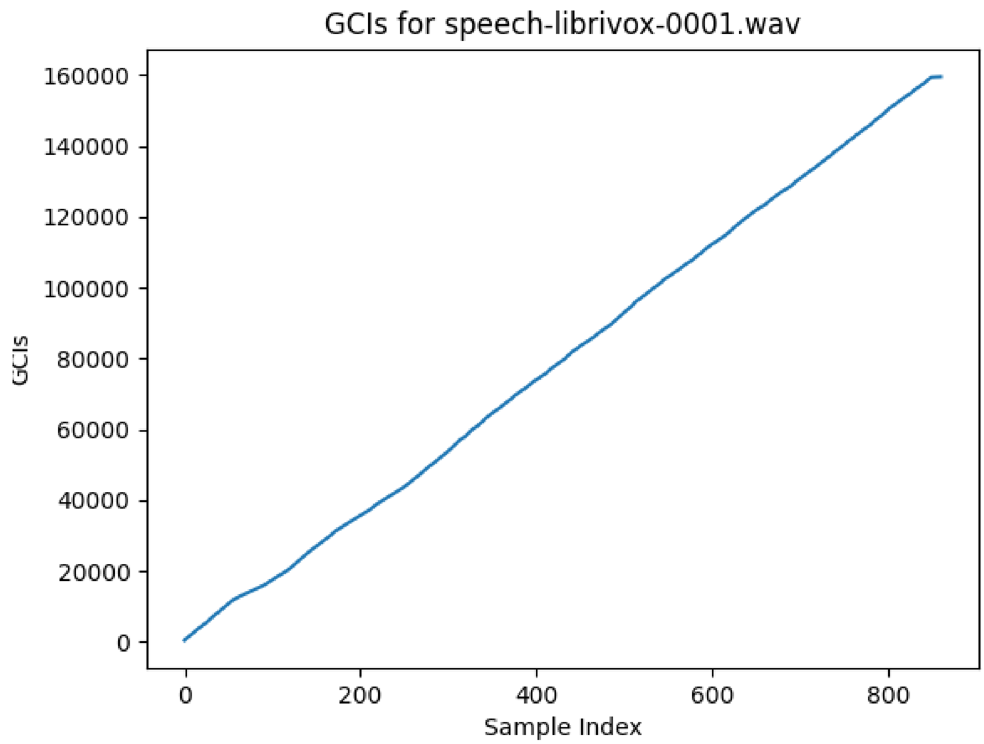


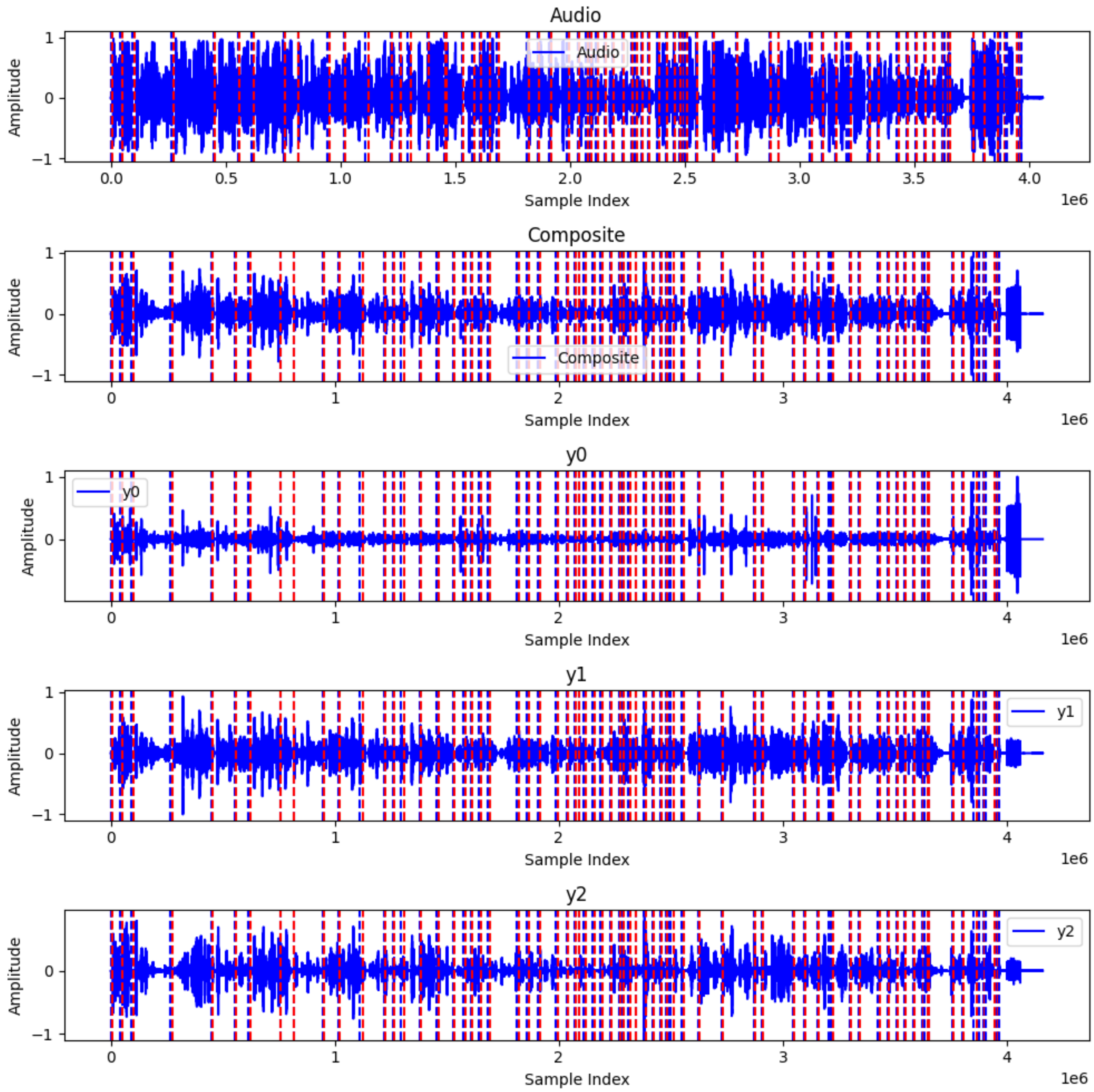
speech-librivox-0000 m english

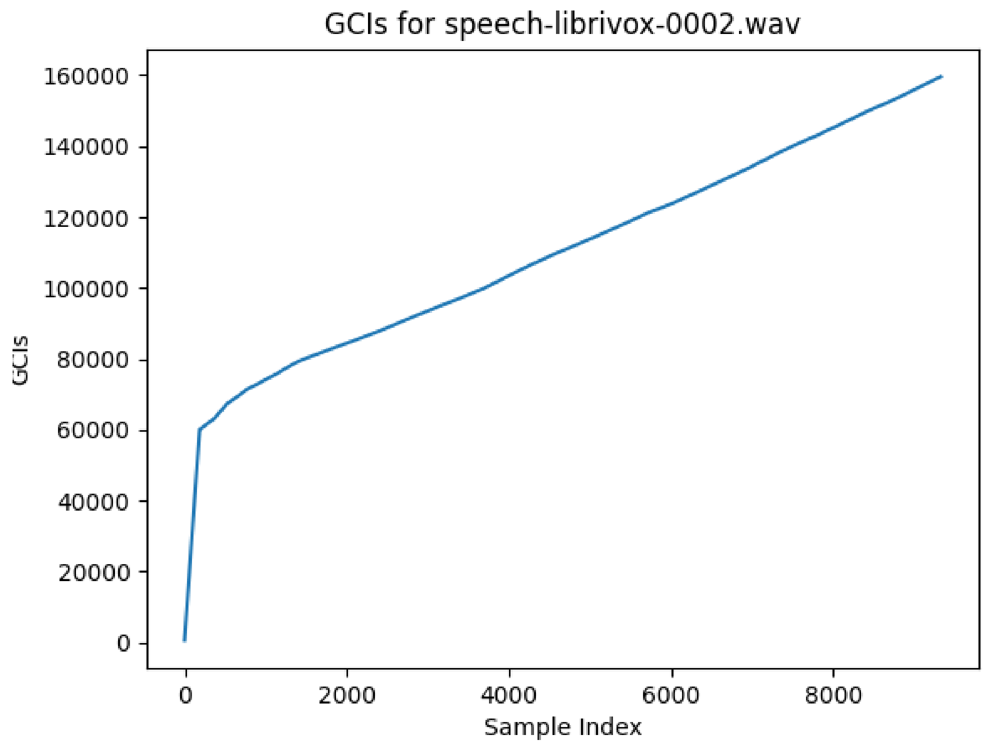


Some more exemplary results


Audio
Amplitude
Sample Index
1e6
Composite
y0
y1
y2
0.0
0.2
0.4
0.6
0.8
1.0
1.2


GCIs for speech-librivox-0001.wav
GCIs
Sample Index
0
20000
40000
60000
80000
100000
120000
140000
160000
0
200
400
600
800


Audio
Amplitude
Sample Index
1e6
Composite
y0
y1
y2



Conclusion

- This method produces refined boundaries 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

References

- https://www.isca-archive.org/interspeech_2022/sarkar22_interspeech.html
- https://speechprocessingbook.aalto.fi/Recognition/Voice_activity_detection.html#:~:text=Voice activity detection (VAD) (,presence probability (SPP) estimation.
- https://arxiv.org/html/2312.05815v1