Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition

National Institute Of Technology Sikkim

ADVISOR

Dr. Avinash Kumar

STUDENTS

Rishabh Tenguria (B190062EC)

Saurav Kumar (B190065EC)

Manish Kumar (B190056EC)

Outline

INTRODUCTON

Children vs Adult Speech

In speech processing, there are several differences between child and adult speech that can impact how it is processed by automatic speech recognition (ASR) systems. Some differences include:

Pronunciation: Children may have difficulty pronouncing certain sounds and words correctly, which can make it more challenging for ASR systems to recognize the spoken words.

Acoustic variability: Children's voices are often higher-pitched and have more variability in pitch and volume, which can make it harder for ASR systems to distinguish between different words or phrases.

Vocabulary: Children have a more limited vocabulary than adults, which can make it more difficult for ASR systems to recognize less common words or phrases.

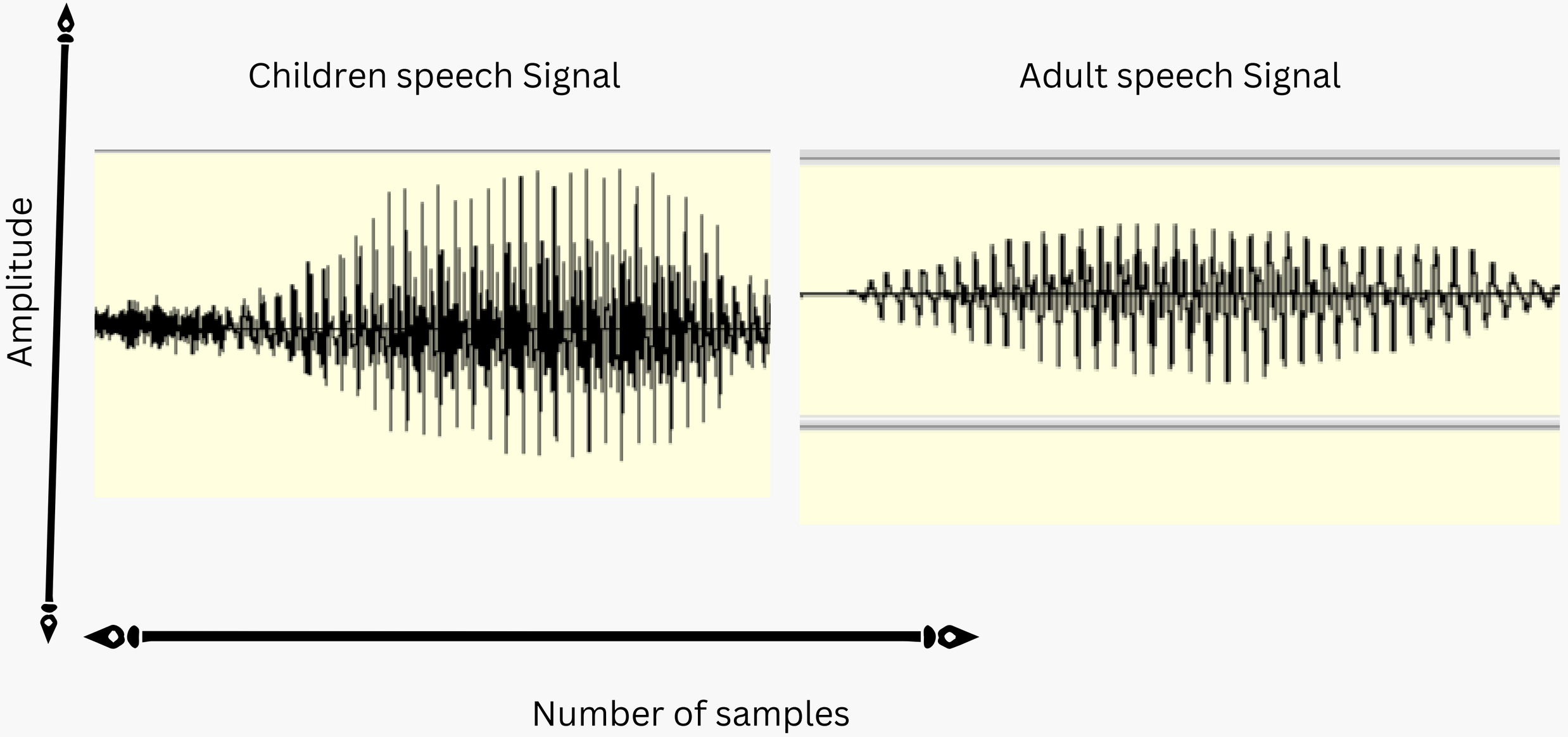


Fig : children speech vs adult speech [8]

PREVIOUS WORK

- In our previous work, we have separated out vowel and non vowel regions form a speech signal.
- Both the region contains different type of information.
- Therefore, it becomes important for us to analyze both the regions saperately.

Literature Review

S.No	Title	Author	Year of Publication	Result
1	Pitch-Adaptive Front-End Features for Robust Children's ASR	S. Shahnawazuddin, Abhishek Dey, Rohit Sinha	2016	a simple technique based on adaptive-liftering for deriving the pitch-robust features. This reduces the sensitivity of the acoustic features to the gross variations in pitch across the speakers.
2	Improvements in the Detection of Vowel Onset and Offset Points in a Speech Sequence.	Avinash Kumar, Syed Shahnawazuddin, Gayadhar Pradhan	2017	This enables us to reduce the sensitivity of the acoustic features to the gross variations in pitch across the speakers.
3	Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition	Ishwar Chandra Yadav, Avinash Kumar, S. Shahnawazuddin and Gayadhar Pradhan	2018	The smoothed spectra thus obtained is used for computing the front-end acoustic features that are more robust towards pitch variations than the existing ones
4	Study of formant modification for children ASR	Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, Mikko Kurimo	2020	The explored technique gives a relative 27% improvement in system performance compared to a hybrid DNN-HMM baseline.
5	Using data augmentation and time-scale modification to improve ASR of children's speech in noisy environments	Hemant Kumar Kathania, Sudarsana Reddy Kadiri, Paavo Alku, Mikko Kurimo	2021	for the most severe noise type (factory noise), the WER values obtained using the baseline system were very poor in all SNR categories (the average WER = 41.29%), but the WER values obtained by the combined system in the "all" scenario dropped to clearly lower levels (the average WER = 14.88%). For the least severe noise type (volvo noise), the corresponding improvement in the averaged WER given by the combined system was from 19.93% to 9.14%

What is Automatic Speech Recognition?

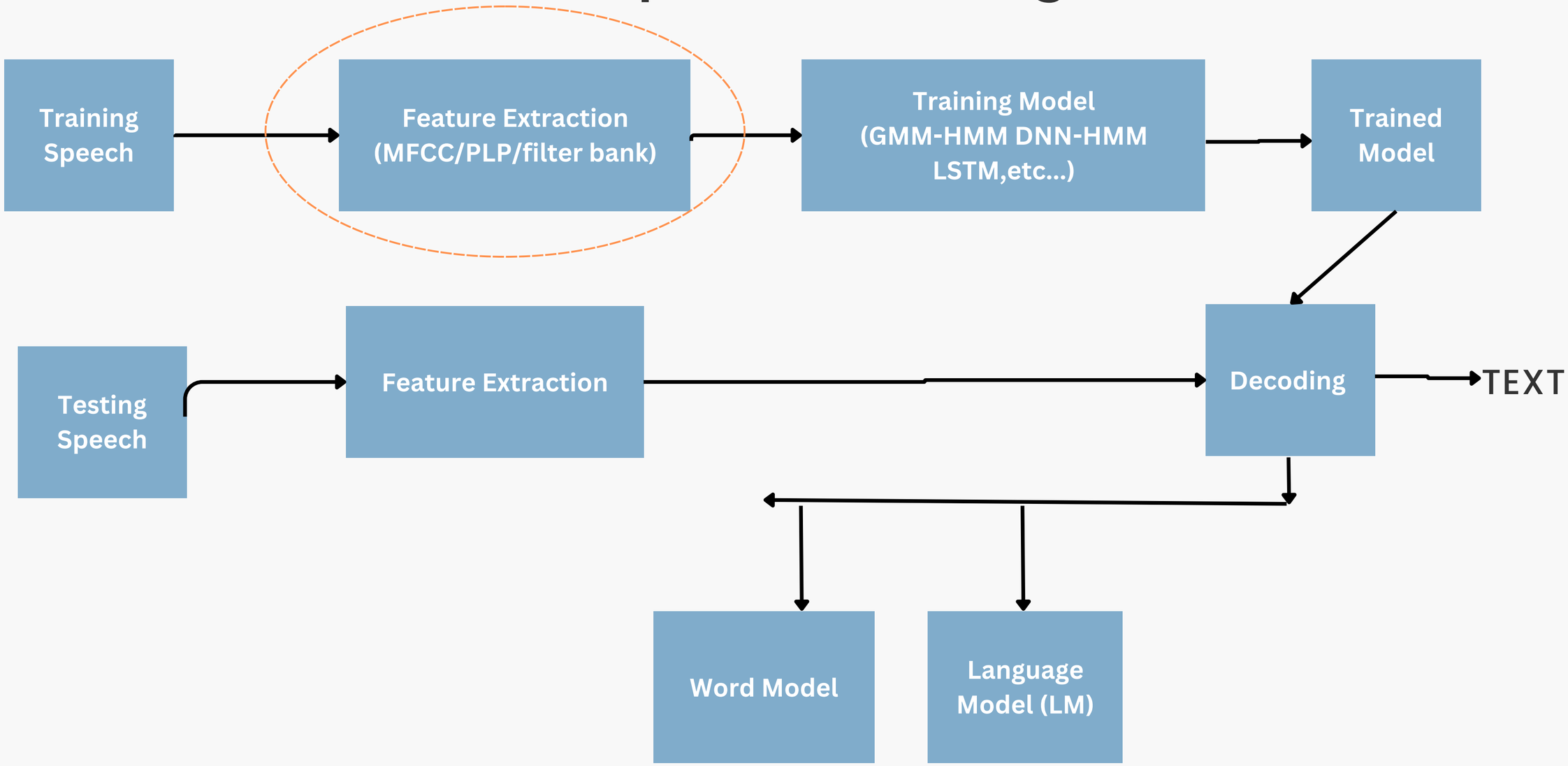


fig- Block diagram of an ASR system [1]

Database

Adults' speech corpus:

- WSJCAM0 is a British English speech corpus for large vocabulary continuous speech recognition.
- The train set derived consists of 15.5 hours of speech data from 92 adult male and female speakers was used for training the acoustic models employed in this study.

Childrens' speech corpus:

- PF-STAR corpus is a British English children's speech database.
- The total duration of speech data was 1.1 hours with a total of 5067 words.
- The developmental set consisted of 150 utterances from 24 child speakers with a total 1.32 hours of speech data.

MFCC

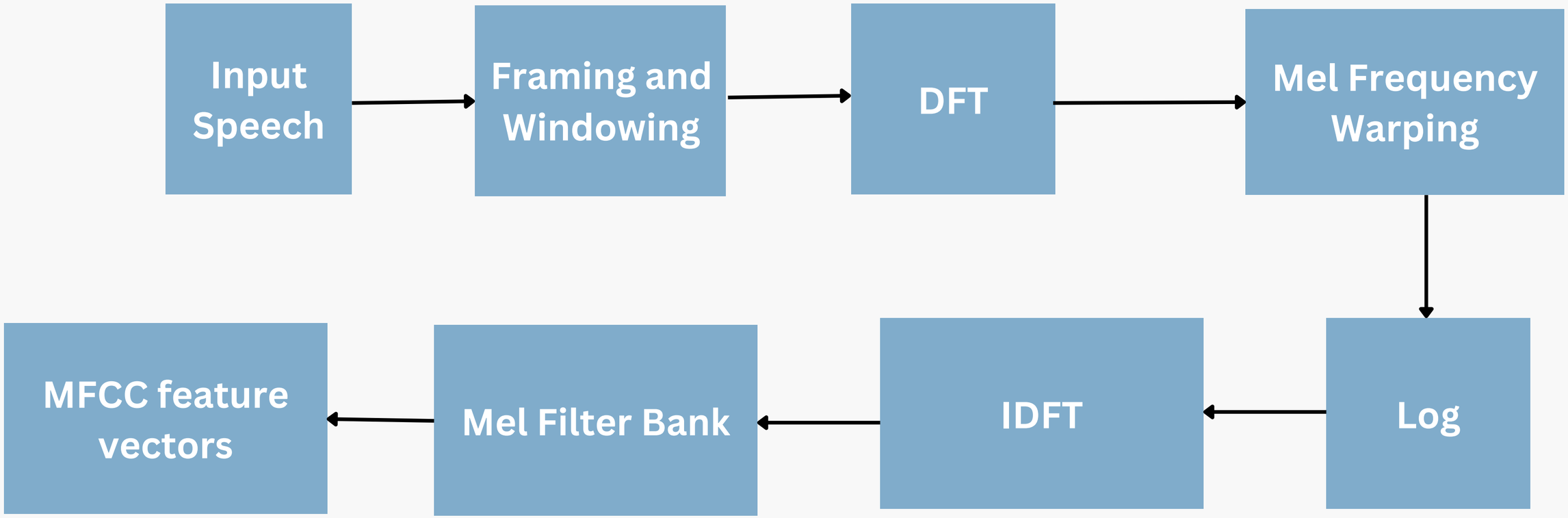
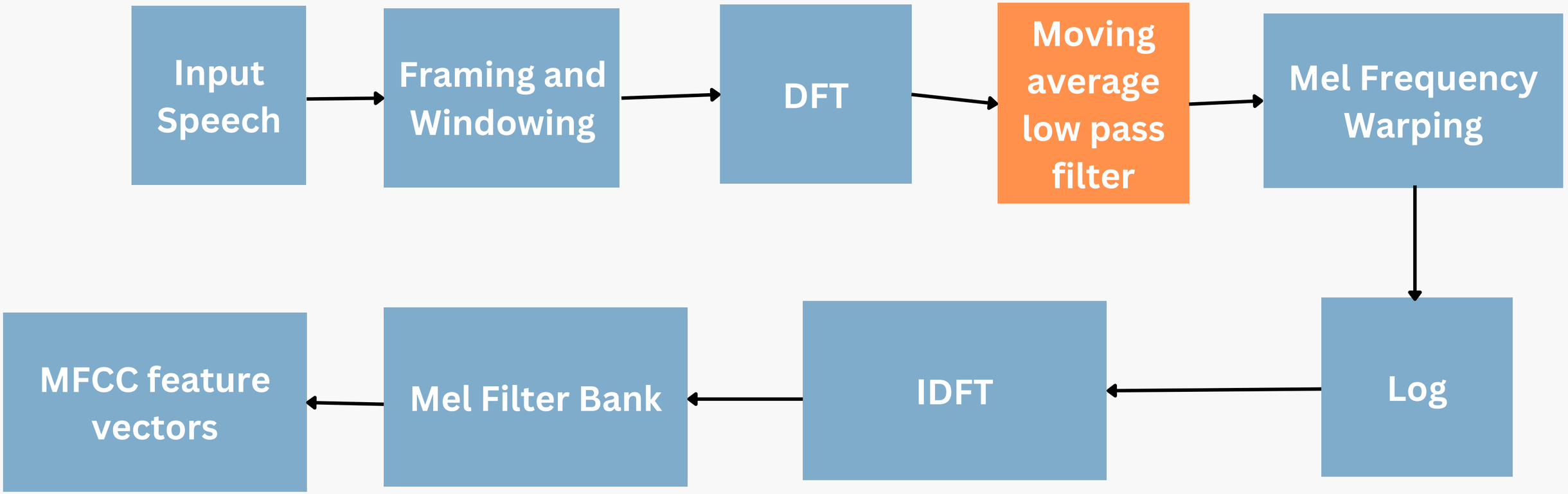


fig- MFCC Feature Extraction [8]

NUSS-MFCC



NUSS-MFCC Feature Extraction [8]

Moving average low pass Filter :

A moving average low-pass filter is a recursive filter that is used as you should expect for a low-pass filter, the output is a smooth rise to the steady state level, it is often used to seek trends in noisy signals.

Approach:

```
windowSize = 1;
b=(1/windowSize)*ones(1,windowSize);
a = 1;
smooth = filter(b,a,m);
```

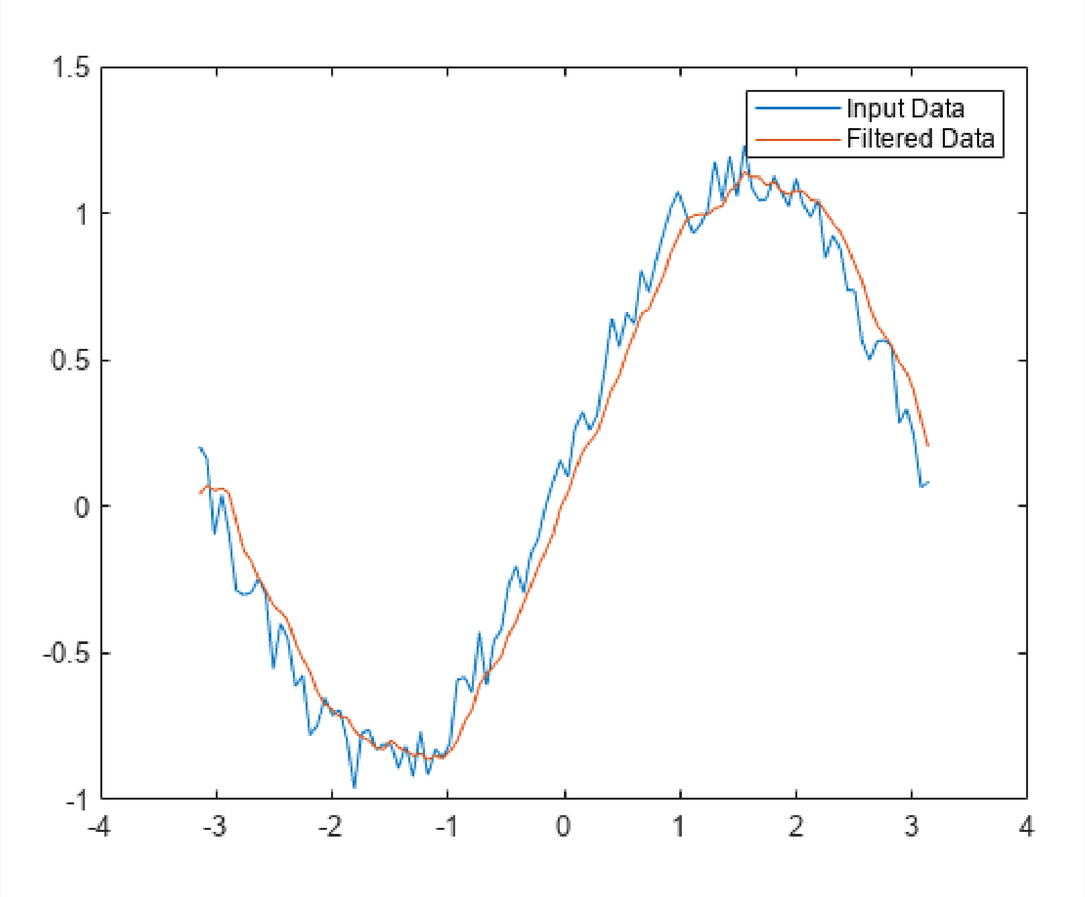


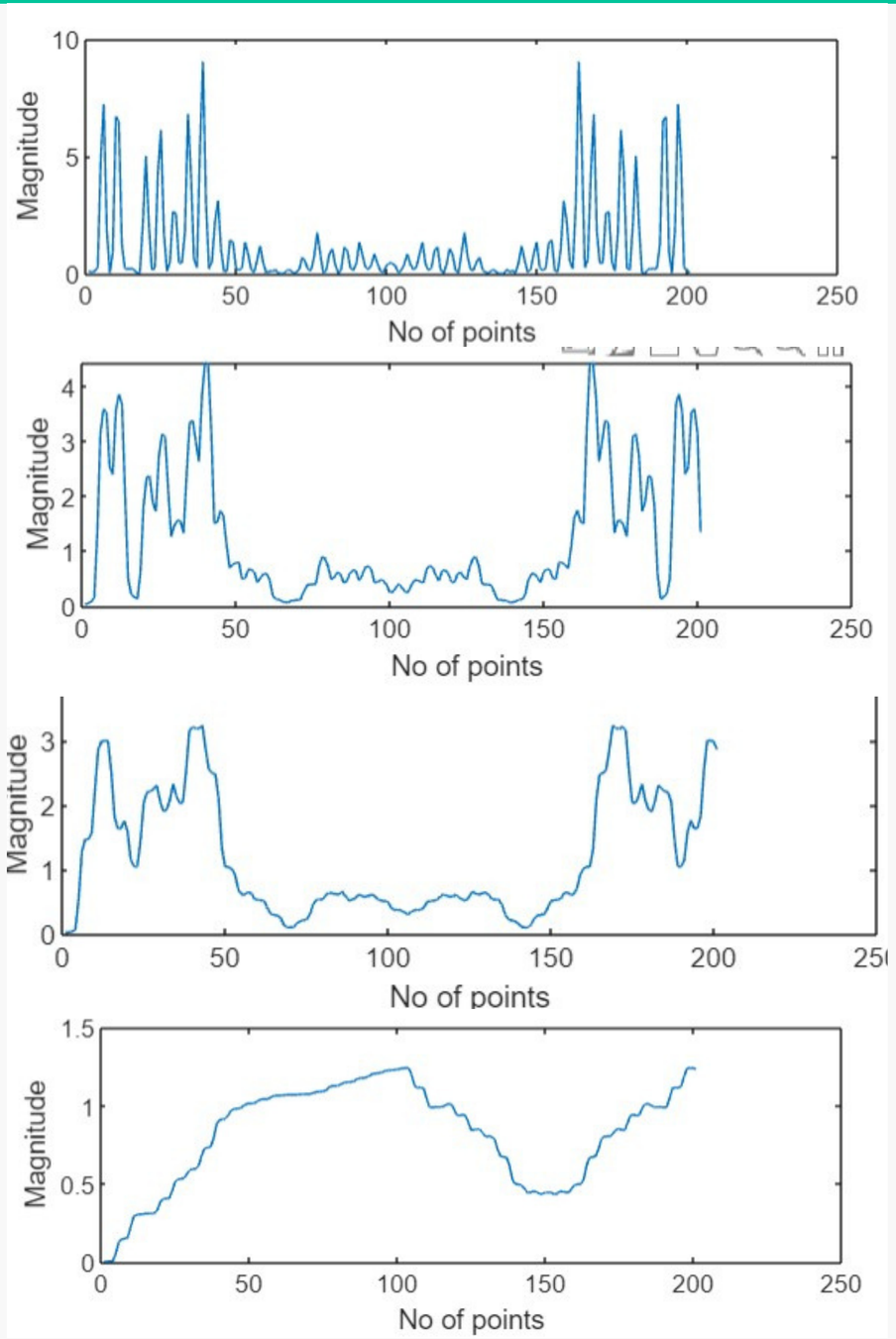
Figure shows input data vs filtered data [1]

Spectral Smoothing

Spectral smoothing means to reduce high frequency noise (e.g., removing "spikes" in the spectrum). When smoothing a spectrum, one must be careful not to remove high frequency components that contain useful information. A suitable smoothing window size must be selected to avoid the same.

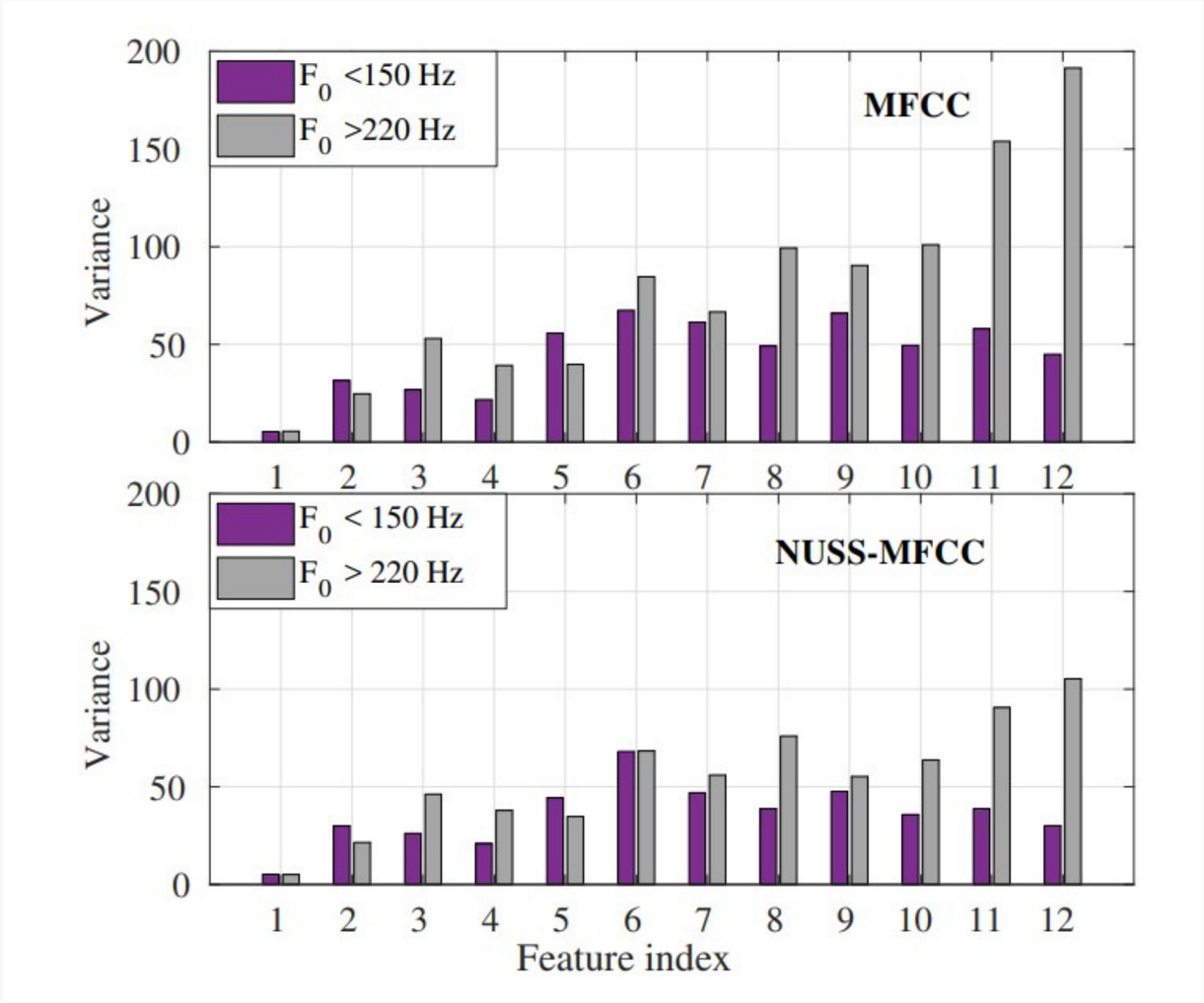
How spectral Smoothing is affected by our approach ?

- In the case of speech, the ripples in the magnitude spectrum are mostly due to the excitation source information.
- The excitation source information is undesirable for ASR and, therefore, should be effectively removed.
- A large window size over smoothens our speech signal which results in information loss.
- A suitable window size = 4 is chosen.
- Spectral smoothing via the moving average low pass filter helps in removing the source information to a large extent, because of the low pass filtering effect.



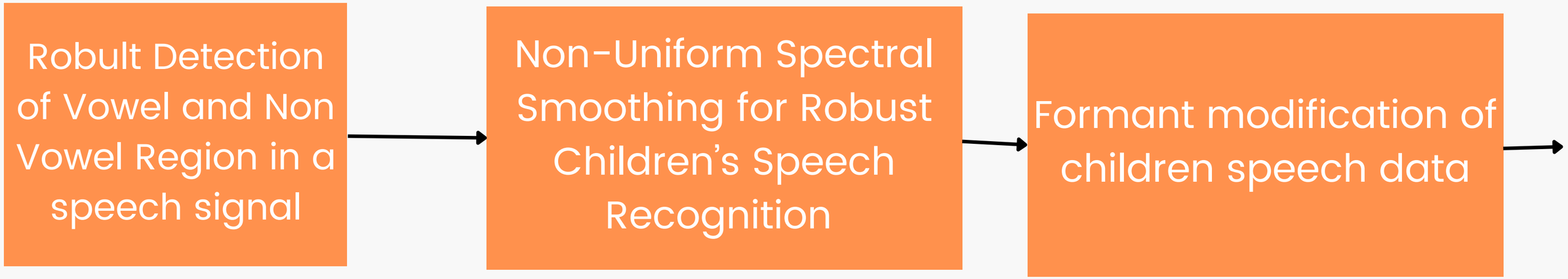
Result and discussions

There is a significant mismatch in the variance of higher-indexed cepstral coefficients across the two pitch groups



[FIG] Variance plots of cestral coefficients for vowel /IY/ [3]

Timeline and Future Work



Refrences

[1] Pitch-Adaptive Front-End Features for Robust Children's ASR.
S Shahnawazuddin, A Dey, R Sinha - Interspeech, (2016).

[2] Kumar, A., Shahnawazuddin, S. & Pradhan, G. Improvements in the Detection of Vowel Onset and Offset Points in a Speech Sequence. Circuits Syst Signal Process 36, 2315–2340 (2017).

[3] Non-Uniform Spectral Smoothing for Robust Children's Speech Recognition.
IC Yadav, A Kumar, S Shahnawazuddin, G Pradhan - Interspeech,(2018)

[4] Study of formant modification for children ASR
HK Kathania, SR Kadiri, P Alku, M Kurimo - ICASSP 2020-2020 IEEE International Conference on, (2020).

[5] Using data augmentation and time-scale modification to improve asr of children's speech in noisy environments
HK Kathania, SR Kadiri, P Alku, M Kurimo - Applied Sciences, (2021)

[6] https://kaldi-asr.org/

[7] https://www.kaggle.com/datasets/nltkdata/timitcorpus.

[8] https://in.mathworks.com/products/matlab.html

[9] https://wavesurfer-js.org/