Potential of Speech-Pathological Features for Deepfake Speech Detection

Anuwat Chaiwongyen

Suradej Duangpummet

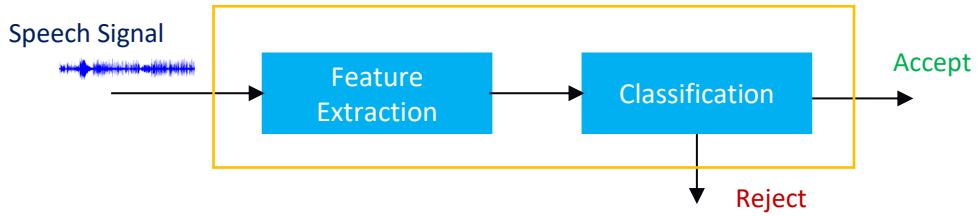
Jessada Karnjana

Waree Kongprawechnon

Masashi Unoki

October 3, 2024

Conventional methods

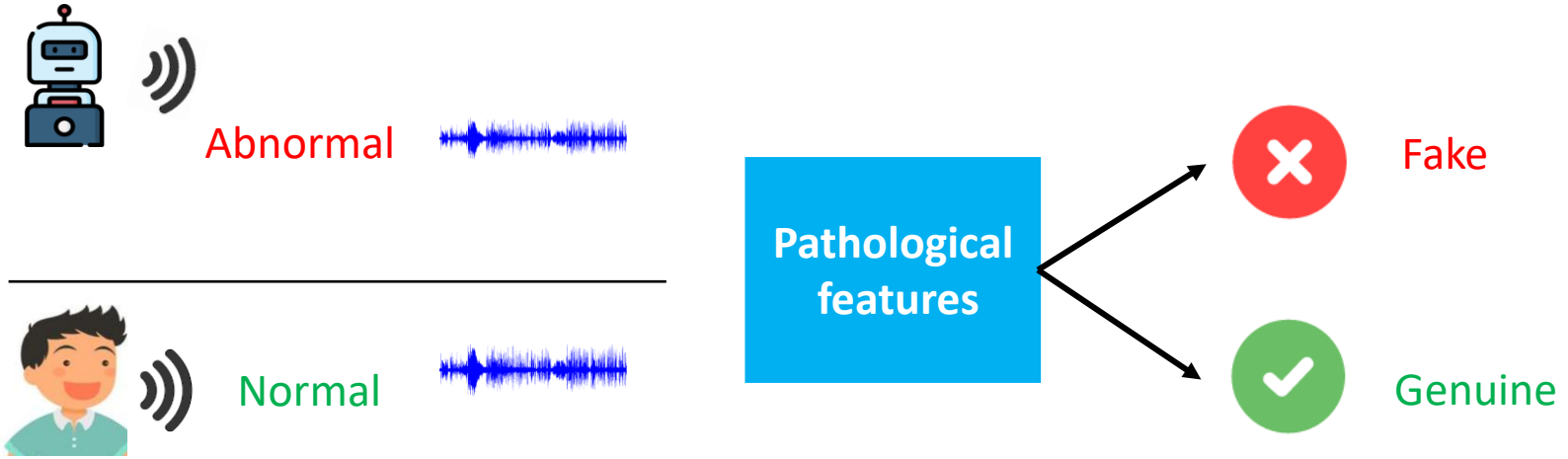


Method	
Feature extraction	**Classification**
1.Linear frequency cepstral coefficients (LFCC)	Gaussian mixture model (GMM) [Baseline]
2.Constant-Q cepstral coefficients (CQCC).	Gaussian mixture model (GMM) [Baseline]
3. LFCC	CNN
4. CQCC and Spectrogram	ResNet
5. Constant Q Transform (CQT)	ResNet-18

- Most conventional features were processed in the phase features, power spectrum features, and cepstral coefficients.

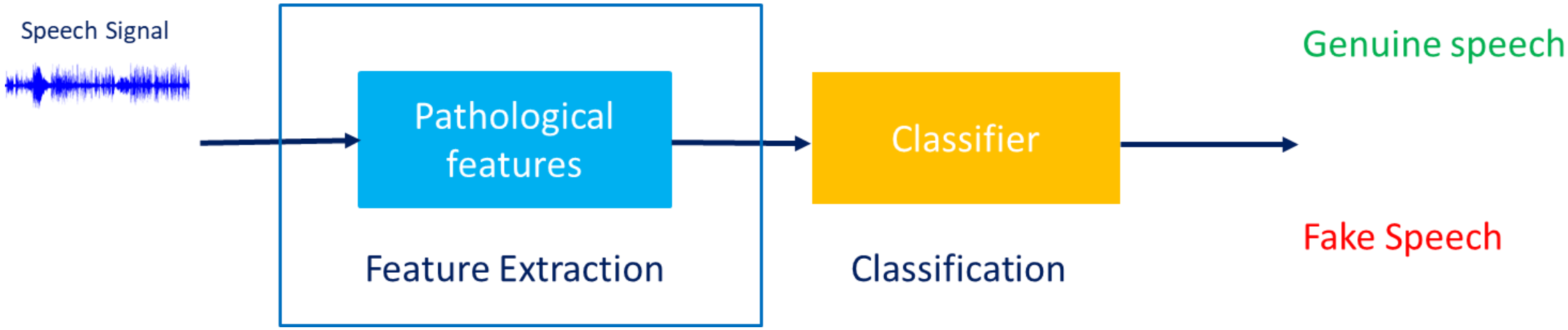
- How about using pathological features to detect deepfake speech?

Motivation



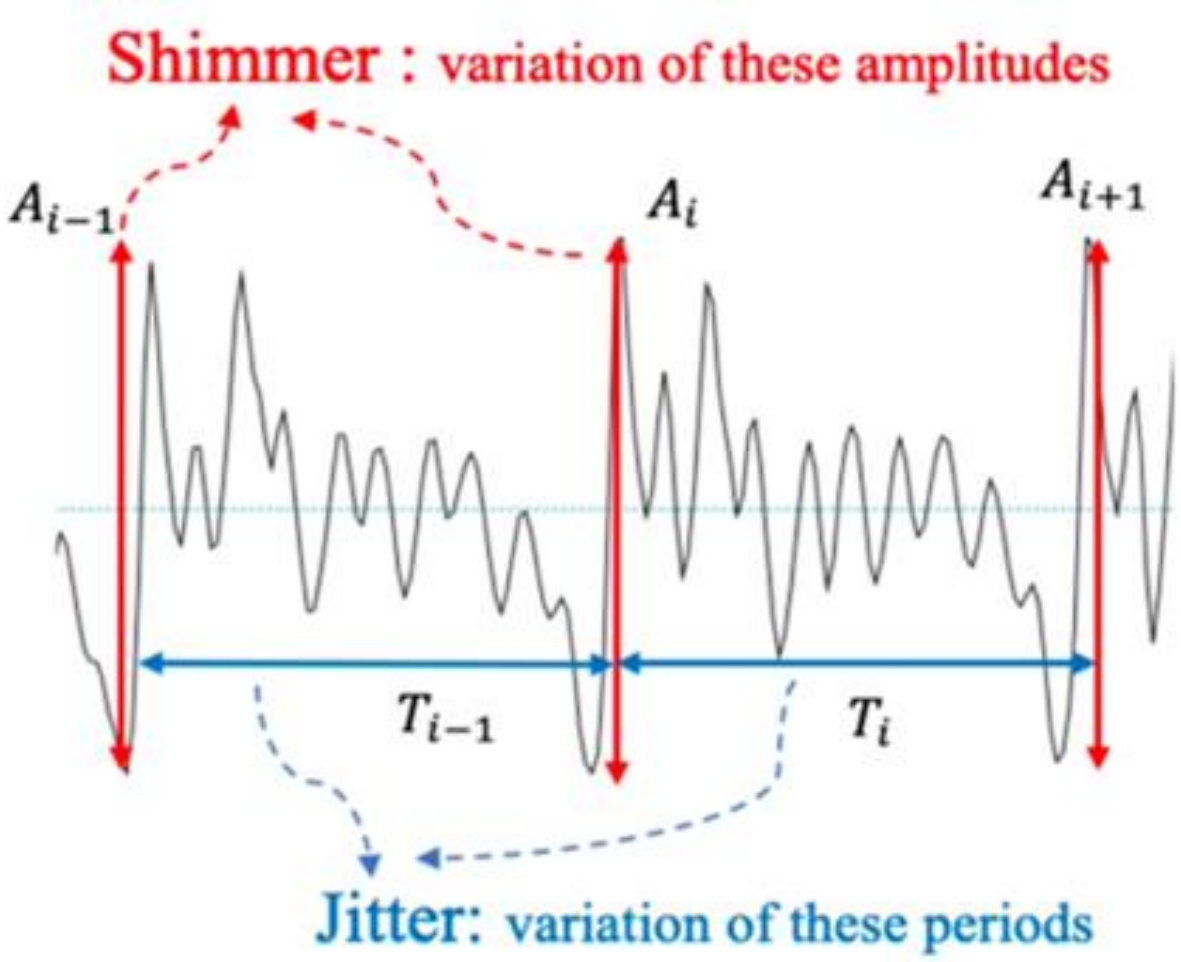
- Speech pathological features are used mainly to discriminate healthy voices from pathological voices (disordered voices).
- The hypothesis is that deepfake speech could possibly be the perceived acoustic quality of the disordered voice.
- Deepfake speech and disordered voice represent the unnaturalness.
- Pathological features can be crucial clues for deepfake speech detection.

Proposed method



1. Jitter
2. Shimmer
3. Harmonics-to-noise ratio (HNR)
4. Cepstral-harmonics-to-noise ratio (CHNR)
5. Normalized noise energy (NNE)
6. Glottal-to-noise excitation ratio (GNE)

Jitter/Shimmer



- ***Jitter*** is the measure of the cycle-to-cycle variations of the fundamental frequency (F0) waveform.
- ***Shimmer*** measures the amplitude variation of a F0 waveform.

1. Jitter

A) **Jitter (local)** is the percentage of the average absolute difference between consecutive periods divided by the average period.

$$\text{Jitter } (local) = \frac{\frac{100}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}|}{\frac{1}{N}\sum_{i=1}^{N} T_i},$$

B) Jitter x-point (PPQx) is the difference between the frequency of each index (Ti) and an average of the x-point closest neighbors around Ti.

$$\text{Jitter } (PPQx) = \frac{\frac{100}{N-m+1}\sum_{i=m}^{N-m}\left|T_i - \left(\frac{1}{x}\sum_{n=i-m}^{i+m} T_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} T_i}, \quad m = \frac{x-1}{2}$$

Jitter PPQ3, and PPQ5 are used.

2. Shimmer

A) **Shimmer (local)** refers to the average of absolute differences between the source-signal amplitude in each index (Ai) and its next neighbor (Ai+1) divided by the average of the signal amplitudes.

$$\text{Shimmer } (local) = \frac{\frac{100}{N-1}\sum_{i=1}^{N-1} |Ai - Ai_{+1}|}{\frac{1}{N}\sum_{i=1}^{N} A_i},$$

B) Shimmer x-point (APQx):. represents the average absolute difference between a period of its average and its x−point closest neighbors, divided by the average period.

$$\text{Shimmer } (APQx) = \frac{\frac{100}{N-m+1}\sum_{i=m}^{N-m} \left|A_i - \left(\frac{1}{x}\sum_{n=i-m}^{i+m} A_n\right)\right|}{\frac{1}{N}\sum_{i=1}^{N} A_i},$$

Shimmer APQ3, APQ5, and APQ11 are used.

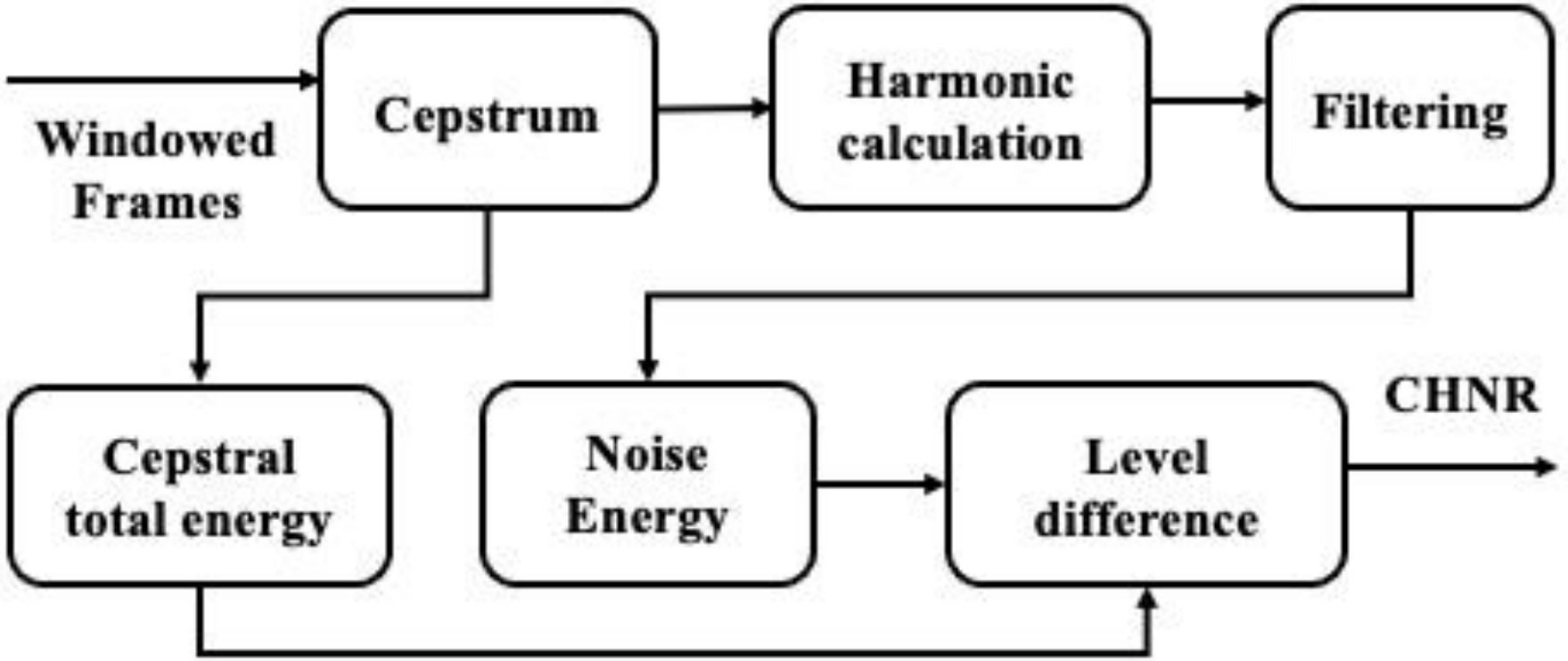
3. Harmonics-to-noise ratio

- The HNR is a measure of the proportion of the harmonic and noise components of speech.
- The noise (ιEn) is computed as the energy of the residual produced after subtracting the average waveform from each individual cycle.
- The harmonic energy (γEn) is determined as the energy of an average waveform of a
  frame.

$$\text{HNR} = 20 \log \frac{\iota En}{\gamma En}$$

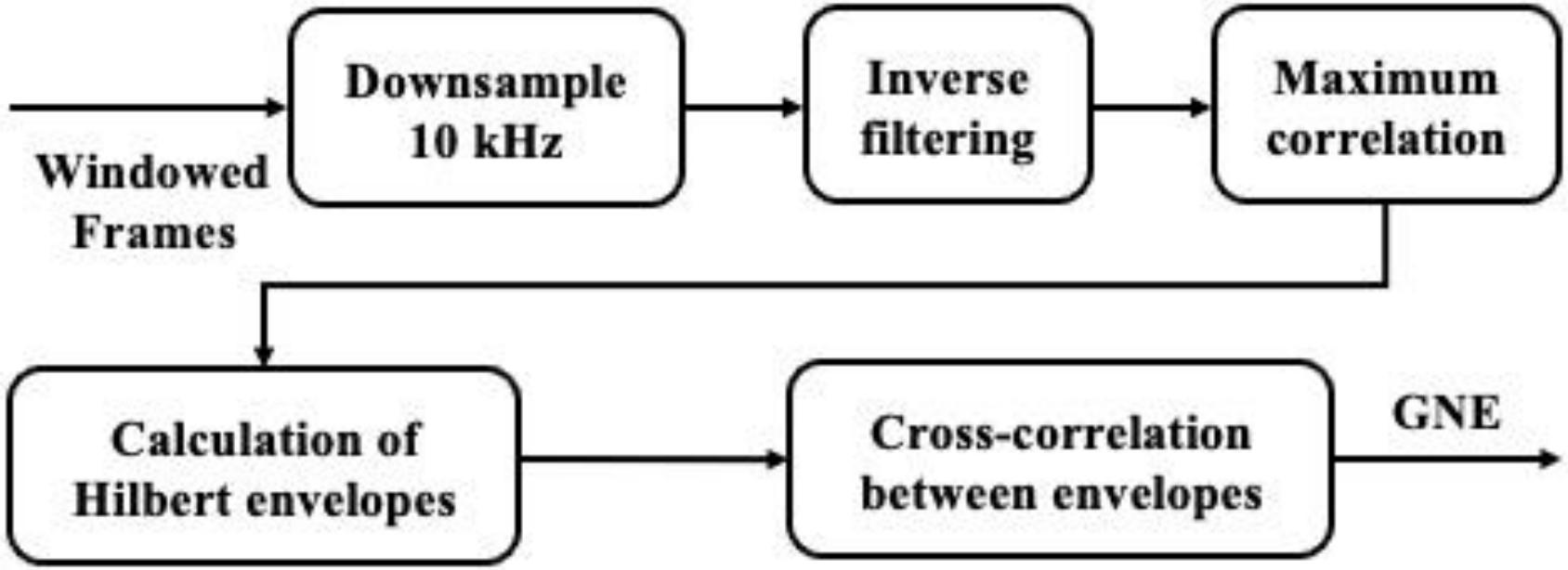
4. Cepstral harmonics-to-noise ratio

- CHNR is to calculate HNR as the difference in level between the cepstral total energy and the noise energy.



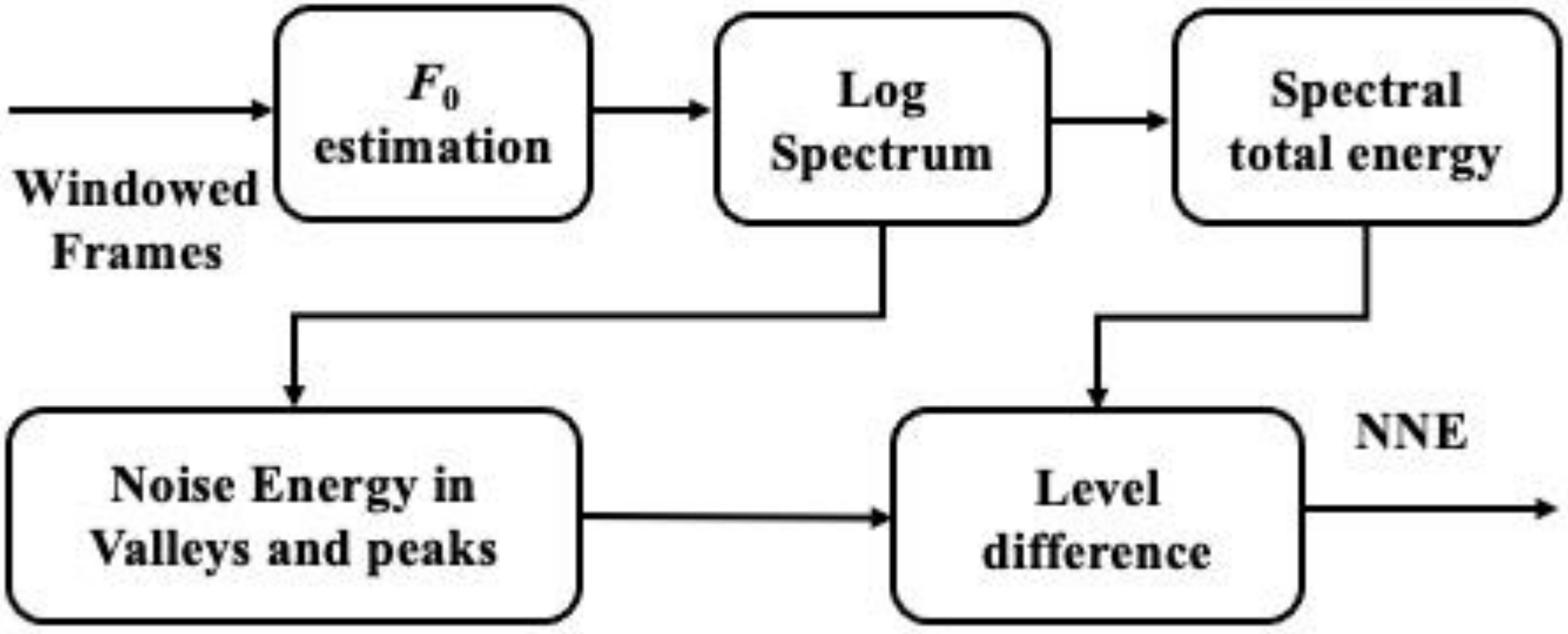
5. Glottal-to-noise excitation ratio

- GNE is used to describe turbulent noise while disregarding modulation effects.
- GNE is assumed that glottal pulses produce a simultaneous and synchronous excitation of multiple frequency channels.



6. Normalized noise energy (NNE)

- Normalized noise energy (NNE) is defined as the ratio of the energy of the noise to the total energy of the signal for each frame of analysis.

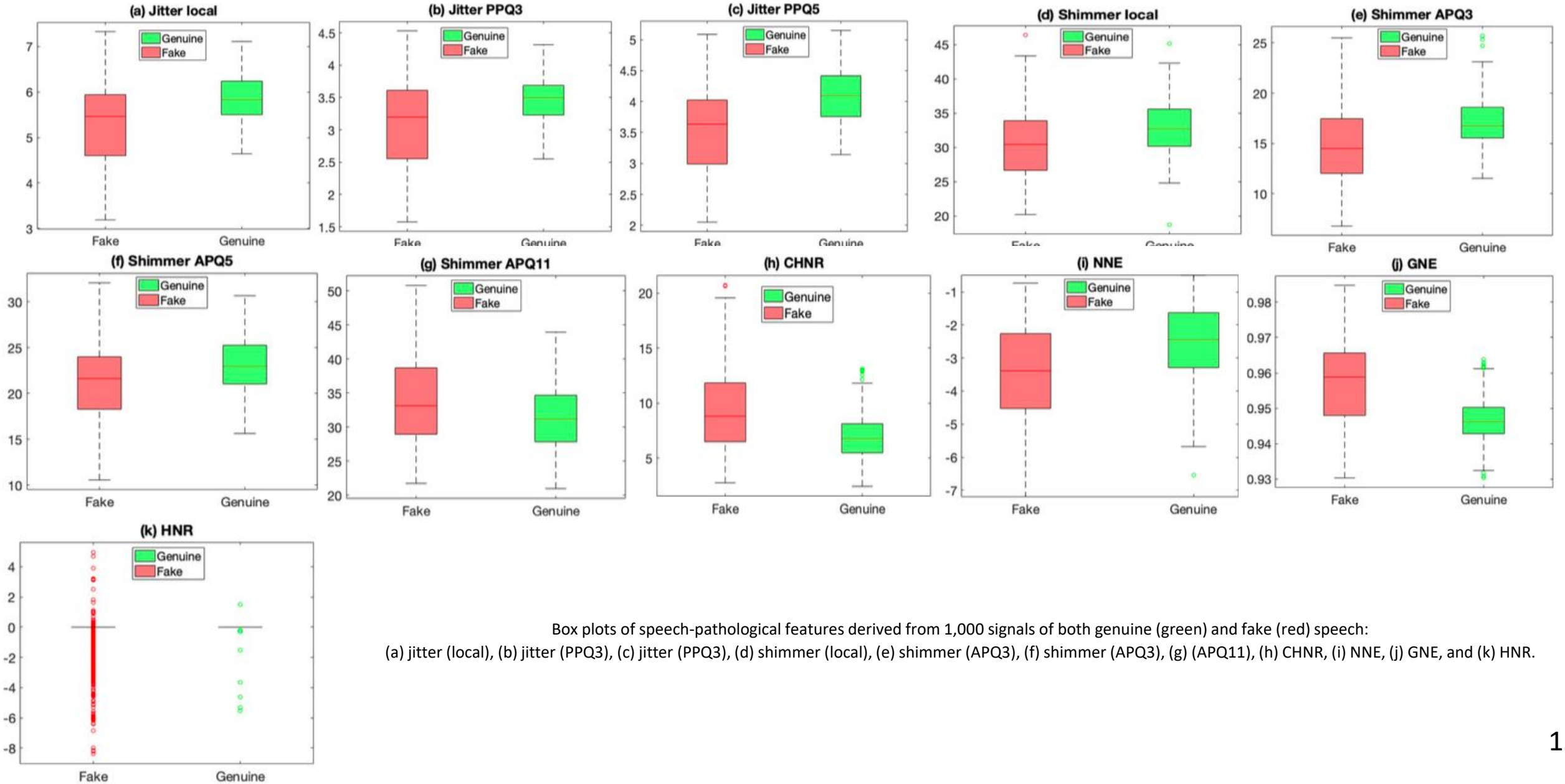


Dataset: ASVspoof2019 and 2021 (LA)

- Two ASVspoof datasets have no background noise.

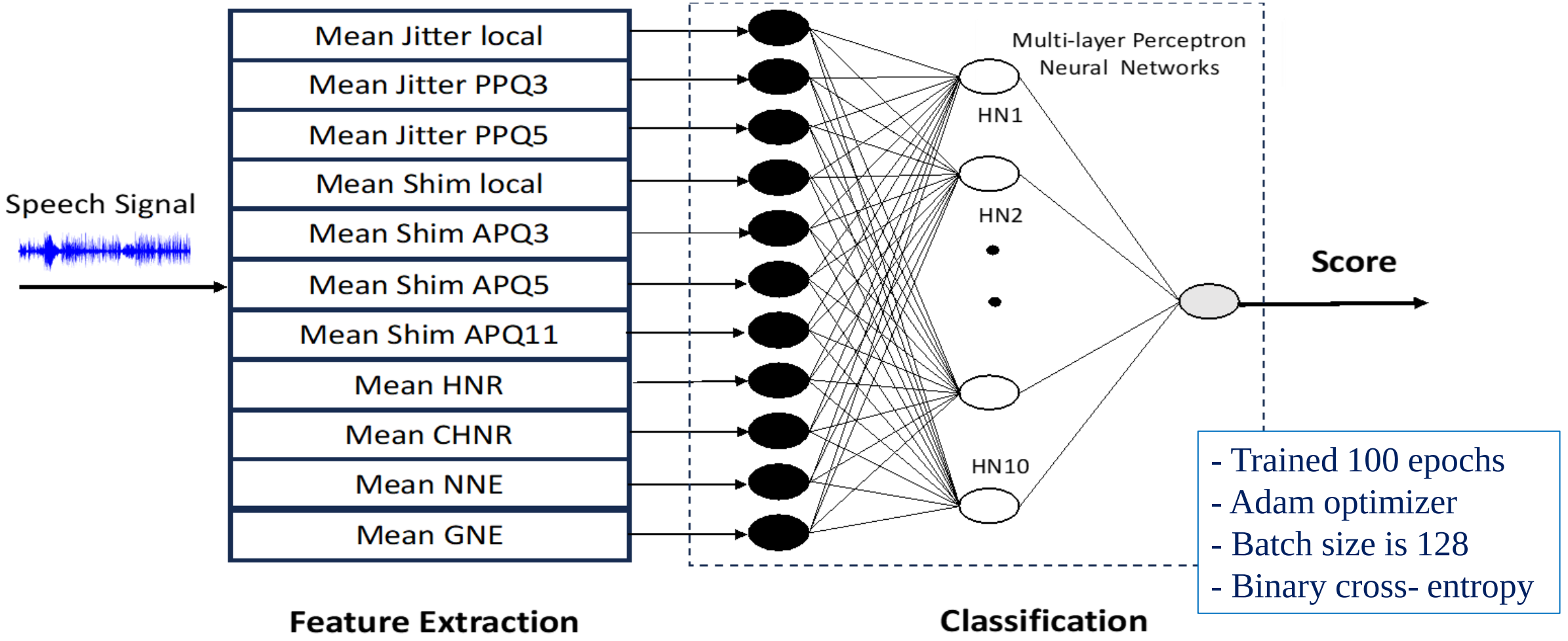
Dataset		Number of utterances		
		Genuine	spoofed	Total
ASVspoof 2019	Training	2,580	22,800	25,380
	Development	2,548	22,296	24,844
	Evaluation	7,355	64,578	71,933
ASVspoof 2021	Evaluation	18,452	163,114	181,566

Feature analysis



Box plots of speech-pathological features derived from 1,000 signals of both genuine (green) and fake (red) speech:
(a) jitter (local), (b) jitter (PPQ3), (c) jitter (PPQ3), (d) shimmer (local), (e) shimmer (APQ3), (f) shimmer (APQ3), (g) (APQ11), (h) CHNR, (i) NNE, (j) GNE, and (k) HNR.

Block diagram of proposed method



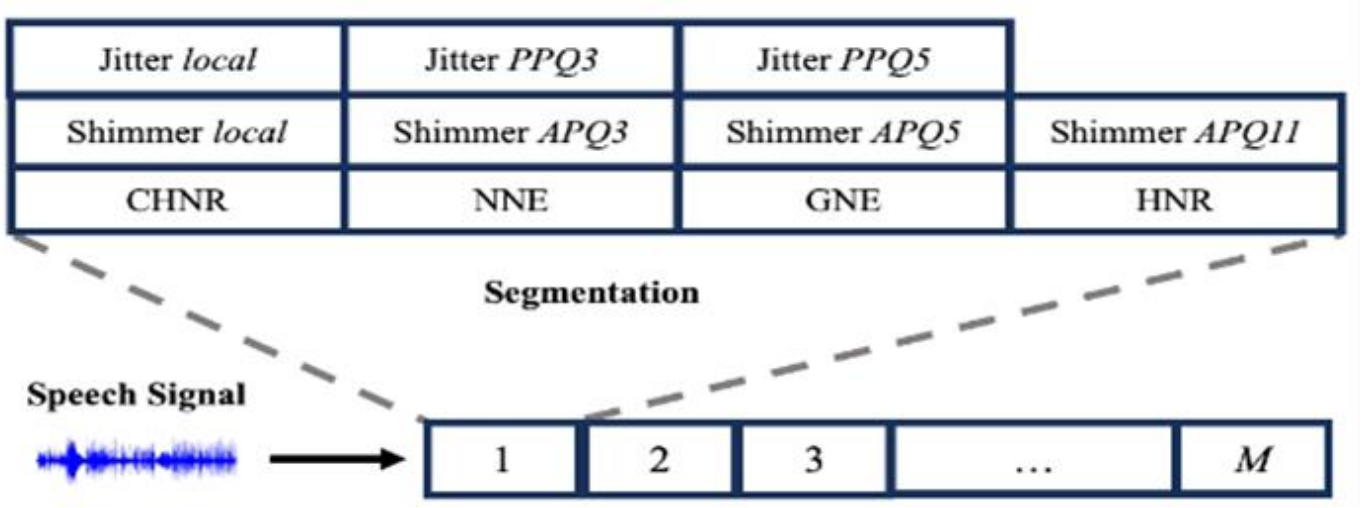
- Trained 100 epochs
- Adam optimizer
- Batch size is 128
- Binary cross- entropy

Evaluation results (averaged)

Development set of ASVspoof2019

Speech-pathological features	Accuracy (%)	Balanced accuracy(%)	Precision (%)	Recall (%)	F1-score	F2-score (%)
1. Jitter (*local*)	68.19	54.56	90.93	71.70	80.18	74.74
2. Jitter ($PPQ3$)	61.15	60.28	92.94	61.31	73.93	65.68
3. Jitter ($PPQ5$)	66.24	55.53	91.24	69.01	78.57	72.54
4. Shimmer (*local*)	75.33	51.90	90.17	81.38	85.56	83.00
5. Shimmer ($APQ3$)	85.64	53.17	90.37	94.03	92.16	**93.27**
6. Shimmer ($APQ5$)	58.16	50.25	89.82	60.20	78.08	65.45
7. Shimmer ($APQ11$)	**86.81**	52.00	90.13	**95.97**	**92.87**	94.6
8. CHNR	84.21	54.61	90.67	91.84	91.25	91.61
9. NNE	73.51	**62.84**	**92.95**	76.26	83.78	79.11
10. GNE	85.56	59.75	91.73	92.21	91.97	92.11
11. HNR	21.11	55.92	99.74	12.13	21.63	14.71
12. Combining 10 features (except HNR)	89.94	61.82	92.04	97.20	94.55	96.12
13. Combining all 11 features	89.17	60.61	91.81	96.54	94.12	95.66

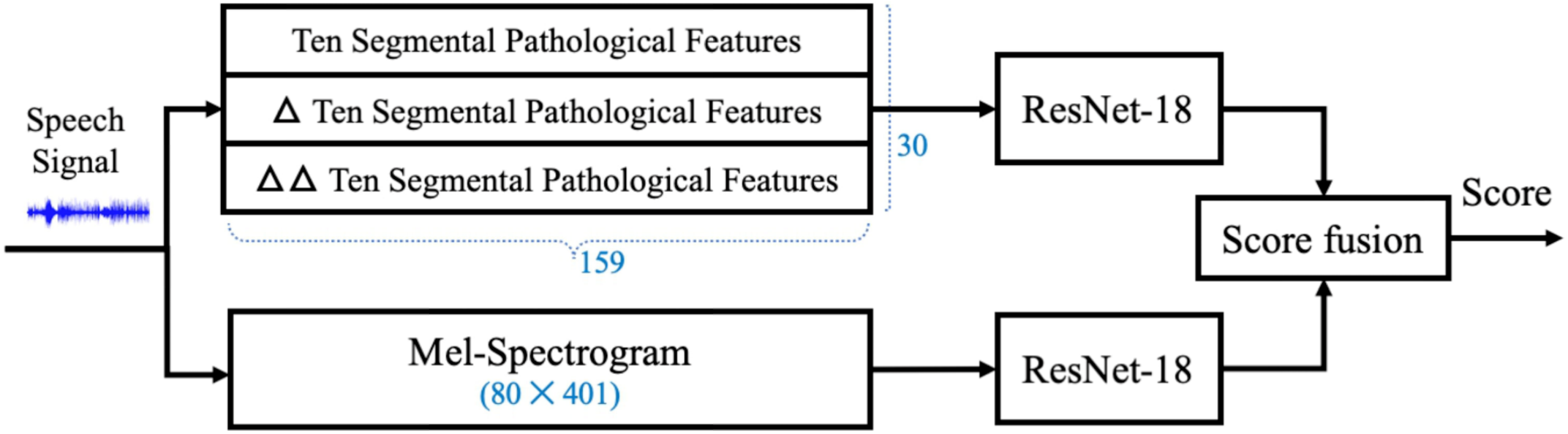
Extension: frame-based analysis



- Length of the speech to 4 seconds
- Sample rate to 16000
- Window frames to 50 milliseconds
- Overlap of 25 milliseconds

Features	Accuracy (%)	Balanced accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	F2-score (%)
1.Average (10 features)	89.94	61.82	92.04	97.20	94.55	96.12
2. Segmental frames of analysis (10 features) with ResNet-18	**96.67**	**85.81**	**96.60**	**99.47**	**98.17**	**99.17**

Extended: proposed method



- PF is ten segmental pathological features
- Δ is the first order derivative of ten segmental pathological features
- ΔΔ is the second order derivative of ten segmental pathological features.

- Trained 100 epochs
- Adam optimizer
- Batch size is 32
- Binary cross- entropy

Evaluation results (frame-based)

Evaluation set of ASVspoof2019

Method	Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60 × 265)	90.10	**89.94**	**98.27**	90.15	94.23	91.73	10.06
2. Mel-spectrogram (80 × 401)	94.36	86.71	97.33	96.36	96.84	98.36	**8.44**
3. *PF* (10 × 159)	91.36	72.99	94.33	96.14	95.23	96.77	11.33
4. Δ (10 × 159)	92.69	72.64	94.17	97.91	96.01	97.14	13.39
5. ΔΔ (10 × 159)	91.65	70.17	93.67	97.26	95.43	96.52	12.69
6. *PF*+Δ (20 × 159)	92.77	72.63	94.16	98.01	96.05	97.21	12.44
7. *PF*+ΔΔ (20 × 159)	93.59	74.65	94.55	98.64	96.65	97.80	10.34
8. Δ+ΔΔ (20 × 159)	92.48	71.91	94.02	97.82	95.89	97.04	12.86
9. *PF*+Δ+ΔΔ (30 × 159)	93.96	73.86	94.36	99.19	96.71	98.51	10.22
10. **Proposed method**	**95.06**	77.70	97.30	**99.46**	**97.30**	**98.59**	10.19

Evaluation results (frame-based)

Evaluation set of ASVspoof2021

- PF is ten segmental pathological features
- Δ is the first order derivative of ten segmental pathological features
- ΔΔ is the second order derivative of ten segmental pathological features.

Method	Evaluation set (%)						
	Accuracy	Balanced accuracy	Precision	Recall	F1-score	F2-score	EER
1. LFCC (60 × 265)	85.22	**83.44**	**97.58**	85.68	91.24	87.82	16.55
2. Mel-spectrogram (80 × 401)	92.50	66.37	92.96	99.16	95.96	97.86	20.92
3. *PF*+Δ+ΔΔ (30 × 159)	**92.60**	67.61	93.12	99.09	96.01	97.84	**15.97**
4. **Proposed method**	91.87	59.97	91.69	**99.96**	**96.65**	**98.18**	**15.97**

Summary

- Speech-pathological features to detect deepfake speech has been proposed.
- Using only 10 numerical features with an MLP neural network could potentially detect deepfake speech.
- This work proposed segmental frames of analysis for speech-pathological features, and the overall performance has significantly improved.
- To compare the speech-pathological features with Mel-spectrogram and LFCC, the speech-pathological features outperform in terms of recall, while the differences in other metrics are insignificant in both datasets.
- The dimension of the proposed feature is only 30×159, while the Mel-spectrogram is 80×401 and the LFCC is 60×265.
- When combining the proposed feature with the Mel-spectrogram, almost all metrics are improved.

Publications

International Journal

1. A. Chaiwongyen, S. Duangpummet, J. Karnjana, W. Kongprawechnon and M. Unoki, "Potential of Speech-Pathological Features for Deepfake Speech Detection," in IEEE Access, vol. 12, pp. 121958-121970, 2024, doi: 10.1109/ACCESS.2024.3447582.

International Conference

1. Anuwat Chaiwongyen., Duangpummet, S., Karnjana, J., Kongprawechnon, W., and Unoki, M. (2023, November). “Deepfake-speech Detection with Pathological Features and Neural Networks”. In 2023 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC).

Thank you for your kind attention

NECTEC
a member of NSTDA

SIIT
Sirindhorn International Institute of Technology

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY
1990

NSTDA