Aalto University
School of Electrical
Engineering

ELEC-E5531 - Speech and Language Processing Seminar

Residual Neural Network precisely quantifies dysarthria severity-level based on short-duration speech segments

Xinran Li, Wilma Donner

31.01.2024

Introduction to Dysarthria

Dysarthria: A motor speech disorder affecting the muscles used in speech.

Impact: Makes it challenging for individuals to communicate effectively, affecting their quality of life.

Need for Technology: Limitations of current Intelligent Personal Assistants (IPAs).

Problem Statement

Shortcomings of traditional methods: Rely on a large "typical" speech dataset does not allow for the treatment of atypical patterns that occur in dysarthria.

Importance of severity-based classification: Dysarthria can vary significantly in severity

Problem Formulation

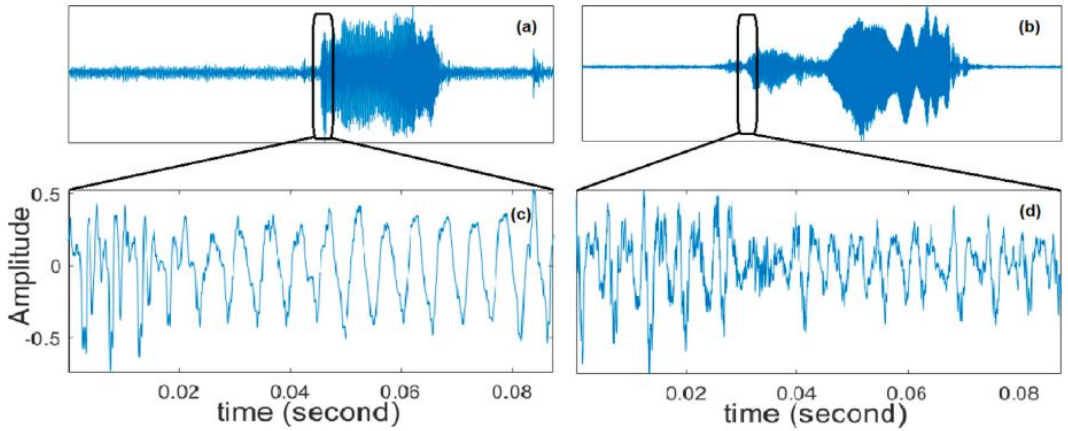
Challenge: Classifying dysarthric speech into four severity levels.

Intelligibility rating (%)	Severity-level
0–25	High
25–50	Medium
50–75	Low
75–100	Very low

Approach: Using transcriptions by listeners to categorize speech intelligibility.
Mathematical Model: Mapping speech features (X) to severity labels (Y).
Deep Learning Solution: ResNet allows us to process short-duration speech segments with greater precision.

Characterizing Dysarthria - Time-Domain Analysis

Normal(a) vs. **dysarthric(b)** speech waveform comparison.



Dysarthric speech: longer duration, pitch period, and production noise.

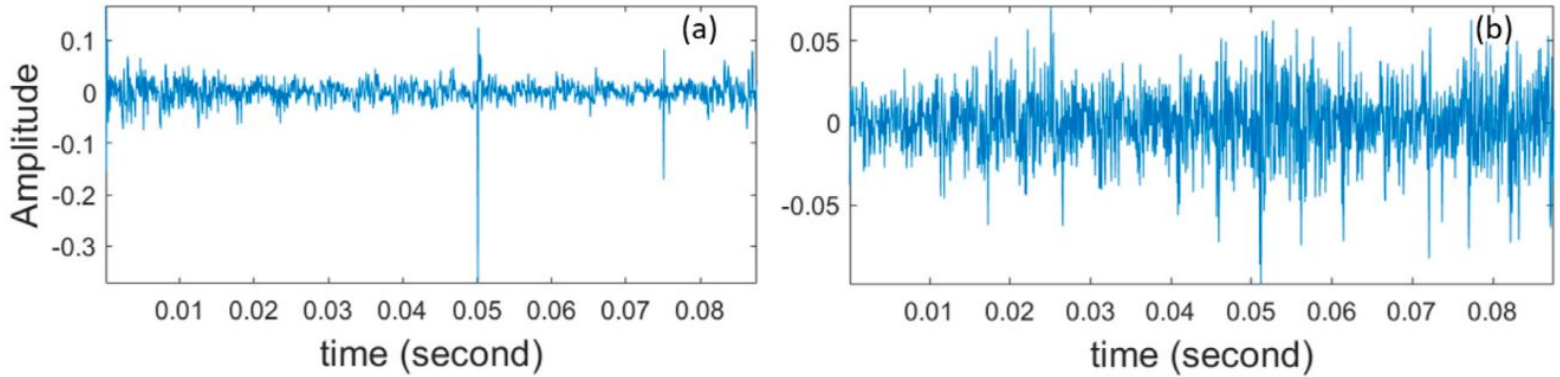
LP Residual Analysis:

LP Residual Analysis:

Equation $r(n) = s(n) - \bar{s}(n)$ where $\bar{s}(n) = \sum_{k=1}^{p} a_k \cdot s(n-k)$,

Reflects glottal closure instants (GCIs).

LP residual plot for **normal(a)** vs. **dysarthric(b)** speech.

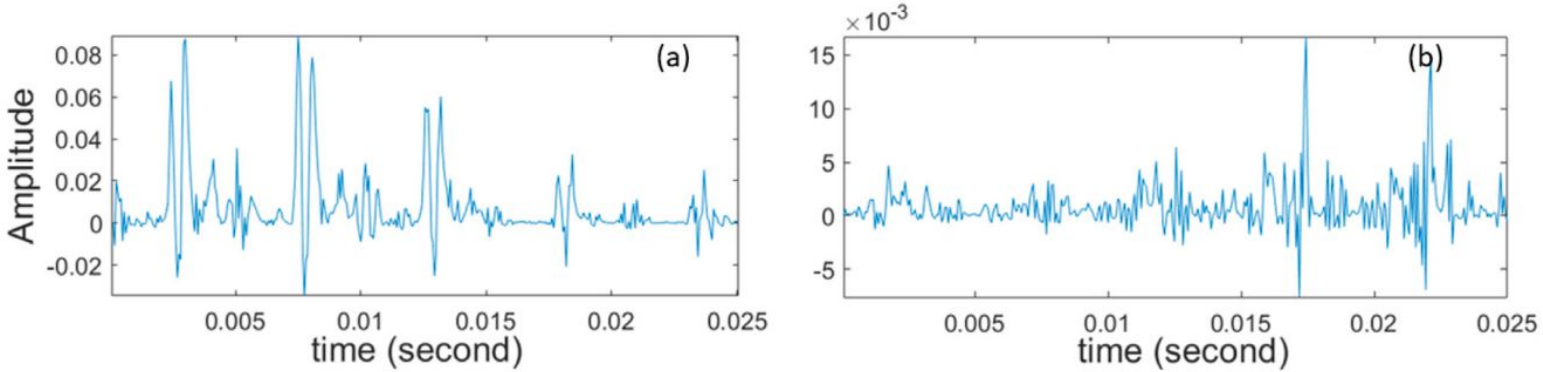


TEO Profile Analysis

Equation: *TEO{s(n)} = (s(n))^2 - s(n - 1) · s(n + 1).*

Captures signal energy and frequency.

TEO profile plot comparison for **normal(a)** vs. **dysarthric(b)** speech.



Nonlinearities in Dysarthric Speech Production

Linear vs. Nonlinear Speech Production:

Linear speech production would result in damped sinusoids and a decaying TEO profile.

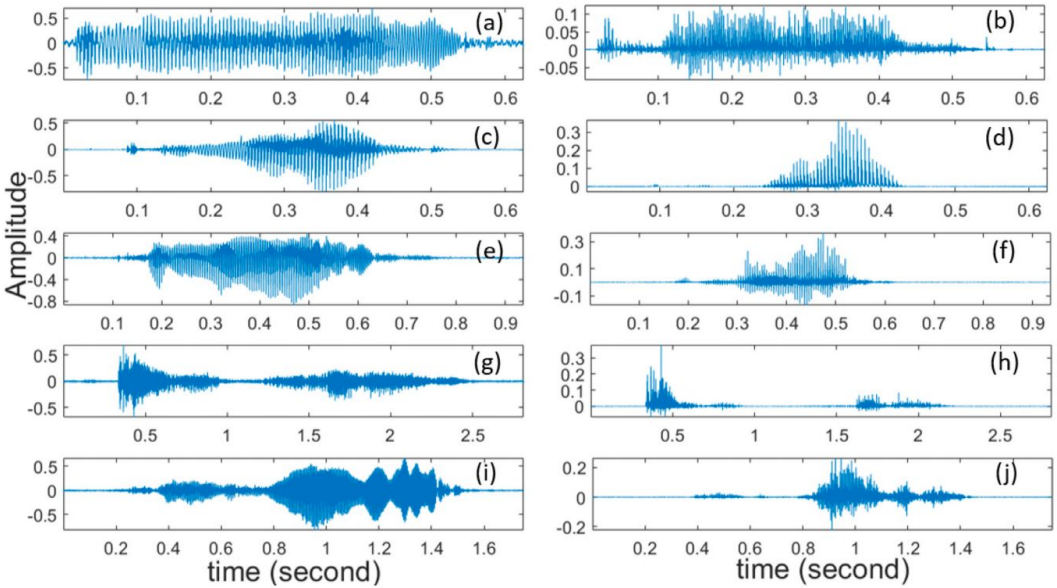
Nonlinear production involves additional factors like aeroacoustic mechanisms.

TEO Profile Observations:

In normal speech, TEO profiles display expected variations within glottal cycles. Dysarthric speech exhibits pronounced 'bumps' indicating significant nonlinear behavior.

As dysarthria severity increases, the Teager energy pulses and bumps become more pronounced and irregular.

TEO Profiles Across Severity Levels

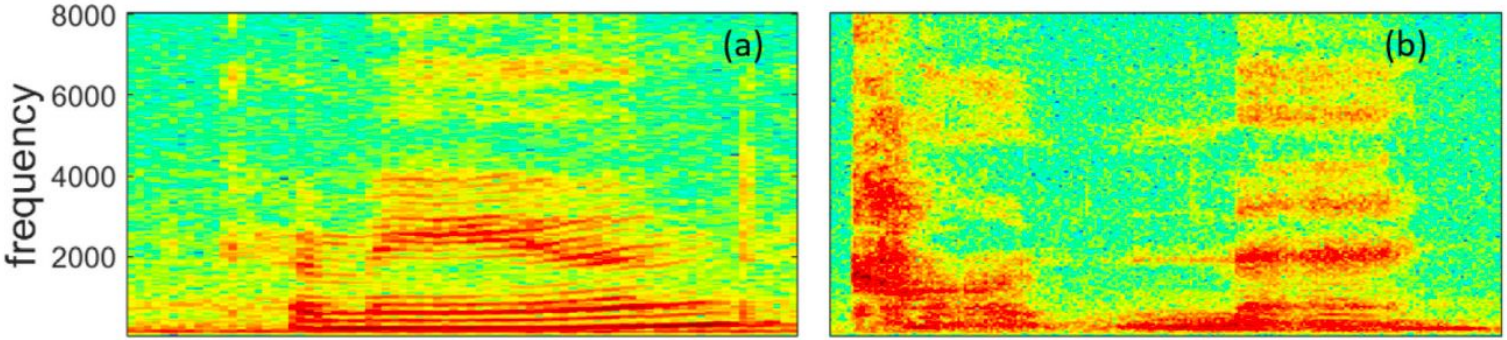


Analysis of nonlinearities via TEO profile (bumps in within GCIs):
(a) Normal Speech Waveform,
(b) TEO profile for normal speech signal.
Dysarthric Speech waveform for (c) severity-1, (e) severity-2, (g) severity-3, and (i) severity-4.
TEO Profile of Dysarthric Speech for (d) severity-1, (f) severity-2, (h) severity-3, and (j) severity-4.

Time-Frequency Analysis

Purpose: Examine spectral energy distribution in dysarthric speech.

Features: Spectrogram(time–frequency representation of speech signal)

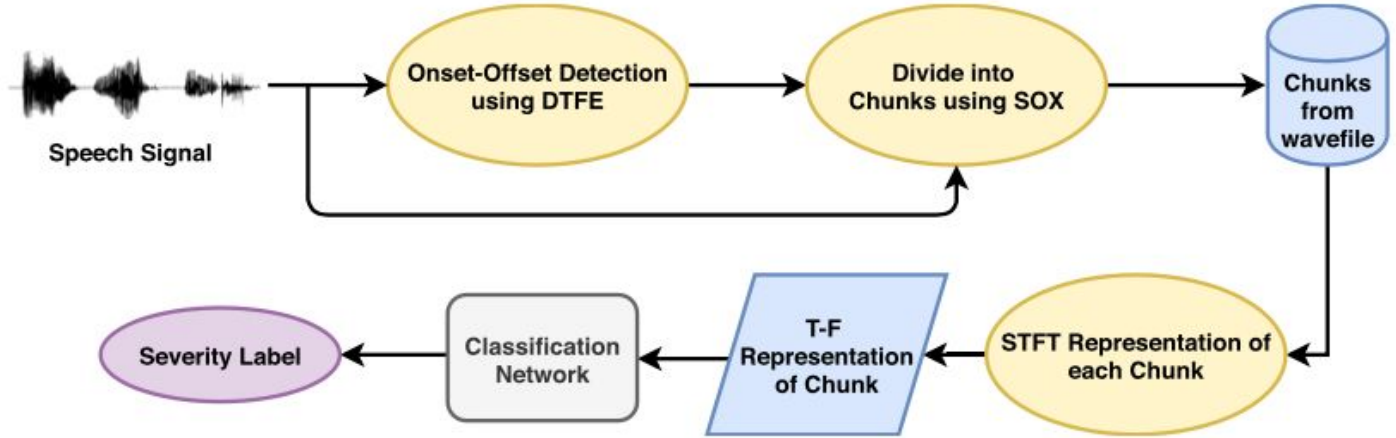


(a) represents normal speech, while (b) represents dysarthric speech.

Approach

These components are needed to solve the problem:

1. onset–offset detection;
2. Time–Frequency (T–F) representation of selected short-duration speech segments;
3. mapping technique for utilizing features to do efficient classification.



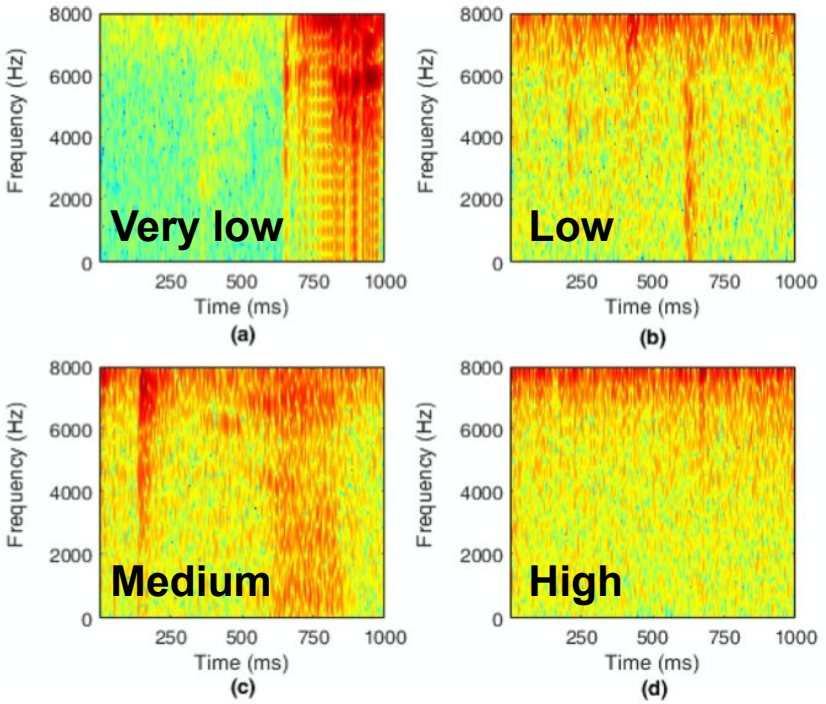
Onset-offset detection

- Detect onset-offset of speech signals using DFTE (Direct Time Fundamental Frequency Estimation) based on frequency
- Time stamps of all onsets and offsets are extracted
- Taking 100 ms to either side of each timestamp, forming 200ms long chunks

T-F representation

- STFT is applied to each chunk
- Differences in spectograms can be seen even in very short segments

  -> Hypothesis that short segments are enough to classify severity

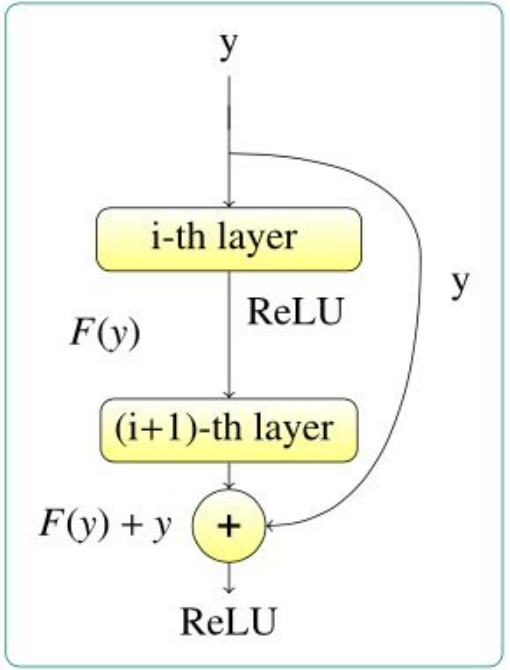


Mapping technique - CNN vs. ResNet

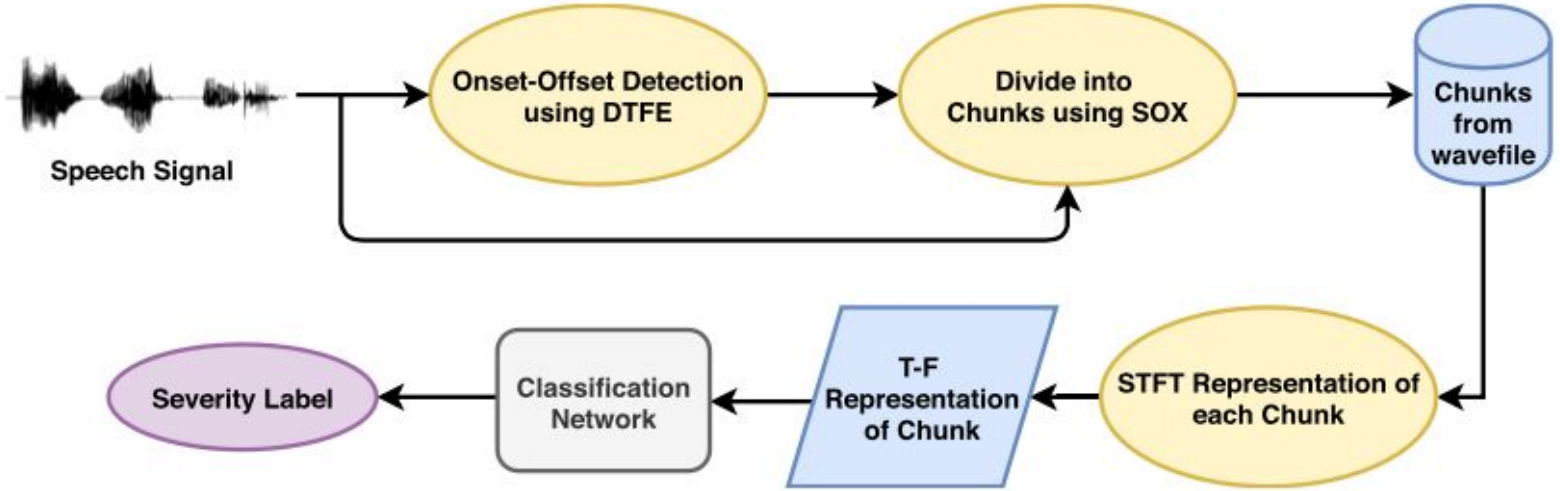
- In many cases, having many stacked layers in neural networks result in better classification
- Adding layers can be problematic
  - Accuracy degrades because of training error
  - Deep neural networks are more difficult to train
- Instead of making CNN models deeper, ResNet (Residual Networks) are used
- ResNet is more effective than CNN in image classification

Residual Network (ResNet)

- Identity shortcut connections are used to skip some layers
- Addresses problem with vanishing gradients and efficiency



Method


Speech Signal
Onset-Offset Detection using DTFE
Divide into Chunks using SOX
Chunks from wavefile
STFT Representation of each Chunk
T-F Representation of Chunk
Classification Network
Severity Label

Dataset

- Source: Universal Access (UA) corpus (Kim et al., 2008)
- Severity level of each dysarthric speaker
- 8 speakers used in experiments
- Speakers said 455 distinct words
- 90% of data was used for training
- 10% used for testing

Comparison methods

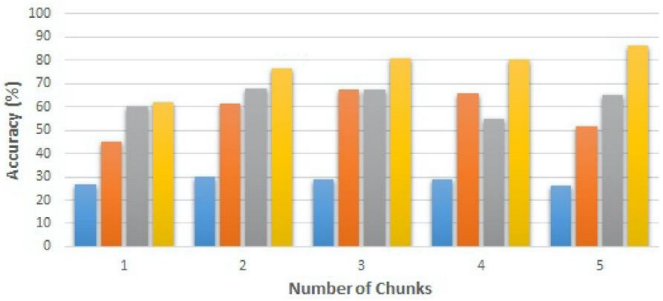
- Gaussian Mixture Model (GMM) used as a baseline
- CNN
- Light CNN (LCNN)
    - Has performed well before

- System was assessed for different number of chunks to prove hypothesis

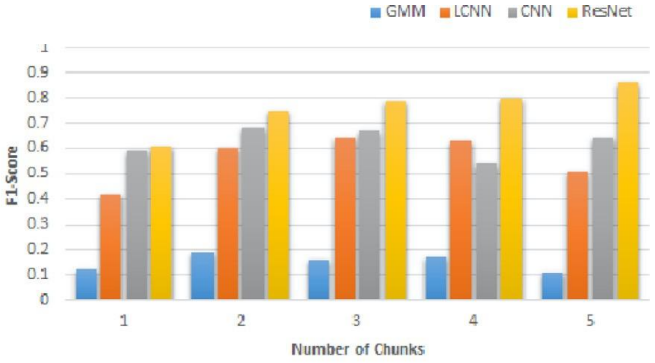
Performance evaluation

- Accuracy and F1 score was used to evaluate
- Accuracy: Number of correctly predicted out of all inputs
- F1:

$$\textbf{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



(a) accuracy



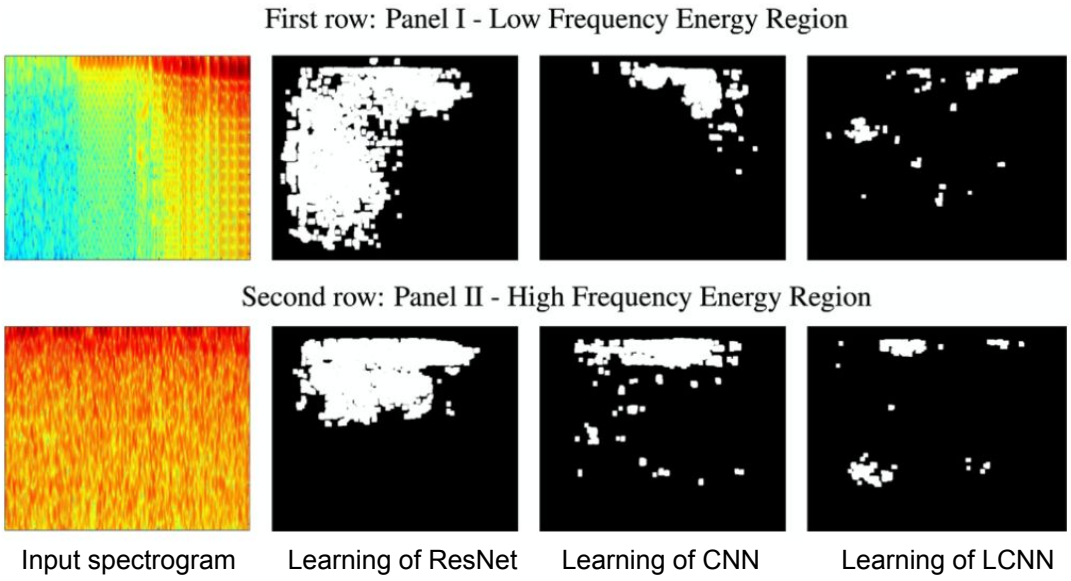
(b) F1-score

Analysing results - Learning performance

How effective is the methodology in terms of

- Learning performance
- Amount of training data

- ResNet captures both energy regions efficiently

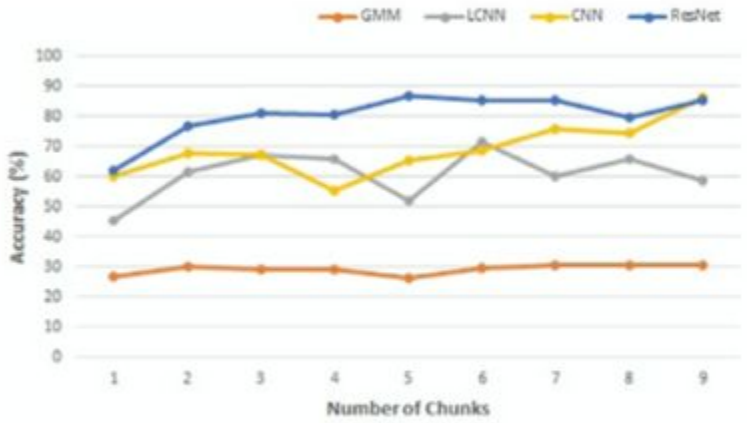


Analysing results - Amount of training data

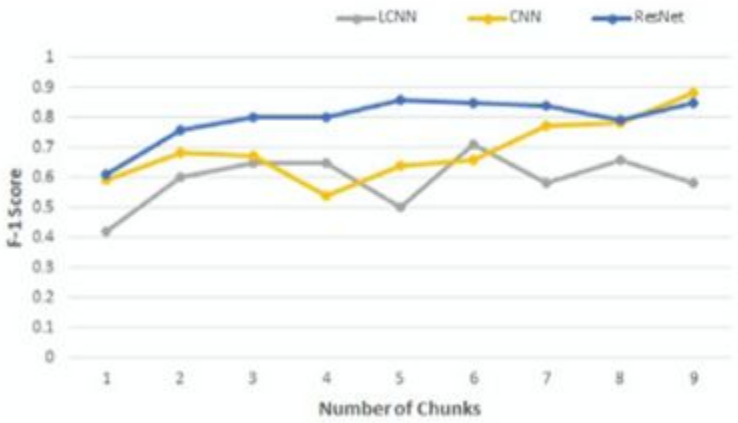
- What happens when more training data is available

Evaluation of baseline CNN *vs.* ResNet when entire speech utterance is available for training.

Systems	Accuracy (%)	F1-Score
ResNet	**98.90**	**0.98**
CNN	91.76	0.91



(a) accuracy



(b) F1-score

Conclusions

- GMM performs poorly in comparison
- Deep-learning and particularly ResNet is more accurate and efficient
- Resnet outperforms CNN both with short and long segments
- Only ResNet was successful with short segments

- Onset-offset detection and spectrogram creation are time-consuming

  -> Real-time implementation is tricky
- Study opens up many possibilities with ResNet
- In the future modified versions of ResNet that are suitable for real-time application could be utilized

Paper

Gupta, S., Patil, A. T., Purohit, M., Parmar, M., Patel, M., Patil, H. A., & Guido, R. C. (2021). Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, *139*, 105-117.

Assignments

1. Describe how the irregularities observed in the TEO energy pulses correlate with the severity of dysarthria and how these findings might affect the classification of dysarthric speech severity.
2. Shortly describe how the performance (not efficiency) of the methodologies was measured.