Automatic Classification of Speech Dysfluencies in Continuous Speech

1928
K. N. Toosi University of Technology

Mostafa Kavarizadeh Dr. Mansour Vali

Speech & Sound Processing Lab

Introduction

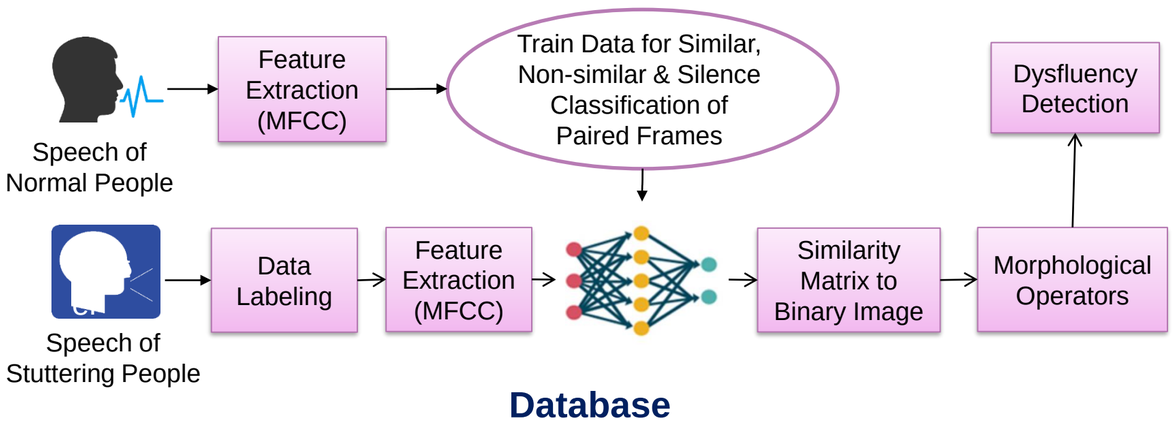
- **Chain of speech is full of unwanted stops between words.**
- **Fluent speech is proper use of pauses and silences between words.**
- **Three types of dysfluencies are more common in the speech of the stuttering.**

Description of 3 types of dysfluency

Type of dysfluency	Description	Example
Part-word repetition	To repeat a sound or syllable	W-W-where is she going?
word repetition	To repeat a whole word	WHERE-where is she going?
Prolongation	Sustain a sound for a long	where is SHSHshe going?

Stages of Dysfluency Detection

- **Extracting MFCC features from successive frames of the normal speech.**
- **Training NN to detect similar, dissimilar frames and silent segments.**
- **MFCC feature extraction from successive frames of the stuttered speech.**
- **Classification of the similarity between frames by the NN.**
- **Output of NN is the similarity matrix between pairs of frames.**
- **Extraction of similar components from similarity matrix by morphological image processing tools.**



Database

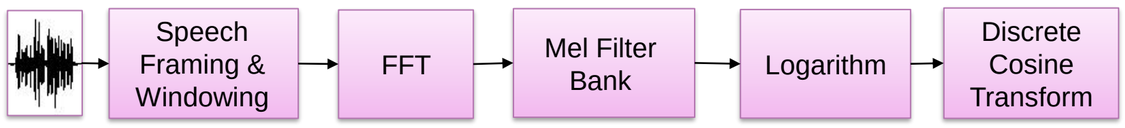
- **FarsDat data**: Fluent speech data base
- **University College London Archive of Stuttered Speech (UCLASS) data**: dysfluent speech data base

Number of fluent and dysfluent words in UCLASS database

Text name	Fluent words	Prolongation	Syllables/ Words repetition
Arthur the rat	8514	281	85
One more week to Easter	4277	131	29
Total	12791	412	114

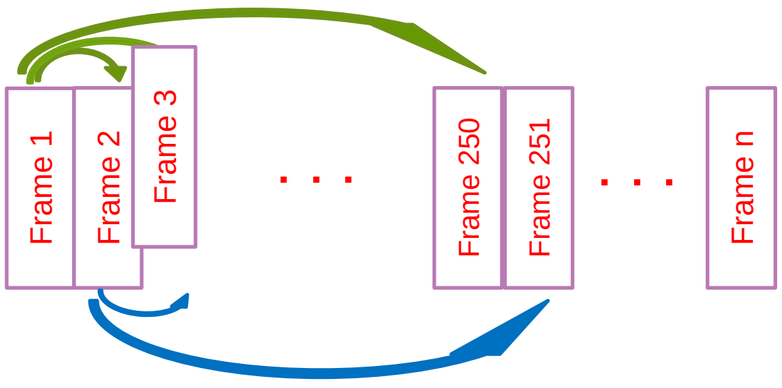
Feature Extraction

13 Mel Frequency Cepstral Coefficients (MFCC) along with the first derivative (total of 26 features) are extracted from each frame.



Neural Network Training

NN inputs: The 26 features of each frame are paired together with 26 features of the next 250 frames, forming 52-component representation vectors.



Feature Label

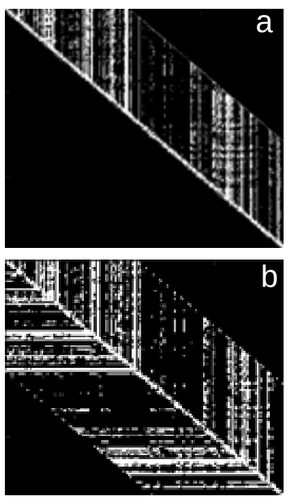
$$\begin{bmatrix} F_1 & F_1 & 1 & 0 & 0 \\ F_1 & F_2 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_1 & F_{250} & 0 & 1 & 0 \\ F_2 & F_2 & 0 & 0 & 1 \\ F_2 & F_3 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ F_2 & F_{251} & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}$$

Similarity Matrix

•Output of the NN is a similarity matrix which is transformed into a binary image.

•The similarity between *i*th and *j*th frames is the same as the similarity between *j*th and *i*th frames.

•Binary image of similarity matrix(a) could be transform to a symmetric image(b).

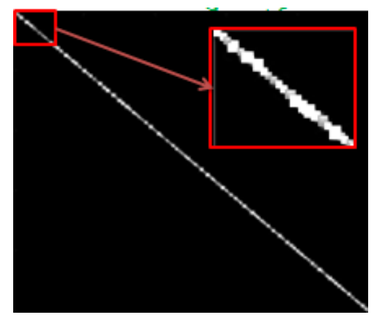


Speech Rate Estimation

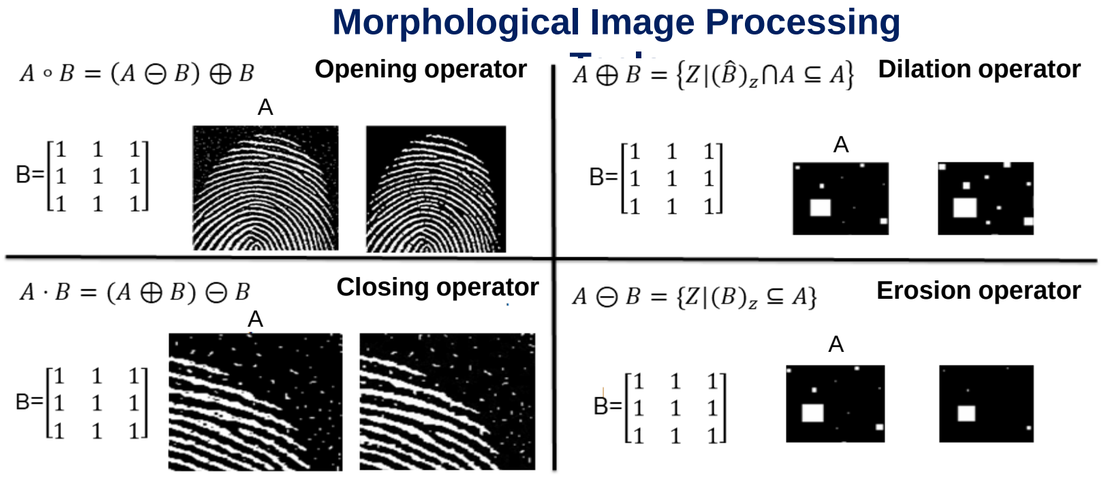
- **Speech rate is the number of phonemes spoken per unit of time.**

$$Sr = \frac{Nv}{T}$$

- **Where *Sr* represents the speech rate, *Nv* denotes the number of phonemes and *T* is the total time of the speech signal.**

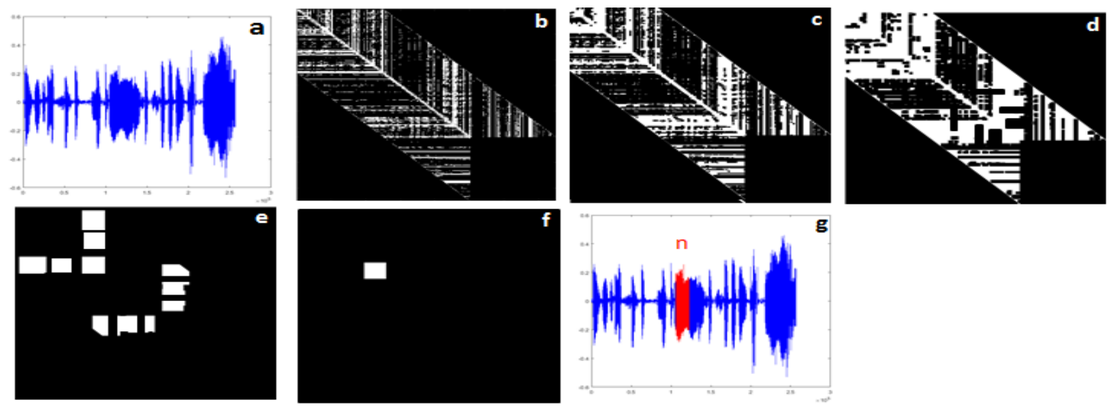


Morphological Image Processing



Prolongation Detection

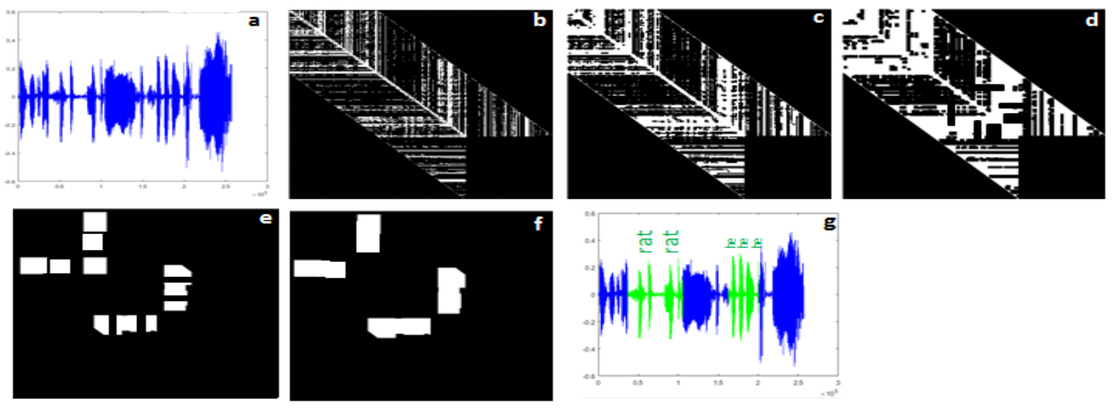
Example : «*There was once a young rat nnnnamed Arthur. he could*»



(a) A few seconds of a UCLASS sample with one prolongation in words "nnnnnamed"
(b) Visualized similarity matrix for MFCC features
(c) Resulted image of applying dilation operator with the 3 × 3 square structural element
(d) Resulted image of applying closing operator with the 10 × 10 square
(e) Results of applying opening operator with the 55 × 55 square structural element
(f) Connected-components which are connected to diagonal line.
(g) Indicate the location of prolonged dysfluency in the speech signal

Phoneme / Word Repetition

Example : «*There was once a young rat rat named Arthur. He he he could*»



(a) A few seconds of a UCLASS sample with two repetitions in words "rat" & "he"
(b) Visualized similarity matrix for MFCC features
(c) Resulted image of applying dilation operator with the 3 × 3 square structural element
(d) Resulted image of applying closing operator with the 10 × 10 square
(e) Results of applying opening operator with the 55 × 55 square structural element
(f) Deletion connected-components which are connected to diagonal line.
(g) Indicate the location of repetition dysfluencies in the speech signal

Results

Accuracy of Edge Detection

Read Text	Arthur the rat	One more week to Easter	Total
Prolongation	98.38%	97.92%	98.23%
Repetition	96.57%	96.42%	96.52%
Total	97.78%	96.89%	97.48%

Accuracy of Dysfluency Detection

Read Text	Arthur the rat	One more week to Easter	Total
Prolongation	98.58%	95.42%	97.57%
Repetition	97.65%	93.10%	96.49%
Total	98.36%	95.00%	97.34%