Acquiring Knowledge for Mimicking Dysarthric Speech by Incorporating Its Features Into Synthetic Speech

Tomasz Piernicki1, Grazina Korvel2, Bozena Kostek1
1 Gdansk University of Technology, Poland
2 Institute of Data Science and Digital Technologies, Vilnius University

Introduction

The purpose of this study is two-fold. First of all, it is to conduct in-depth analyses, allowing the extraction of features associated with dysfunctional speech, and in particular, dysarthria.
The methods of speech signal analysis, such as temporal, spectral, time-frequency, etc. are to be used.
The outcome of this analytical approach is a set of features that corresponds best to dysarthria.
Hence, the second purpose of this study is to propose techniques that may be employed to synthesize normal speech patterns with the most relevant dysarthria features to create dysfunctional speech.

Background

Since existing datasets containing dysarthric speech are very small, this study is to create a dataset for machine learning purposes.

UA SPEECH

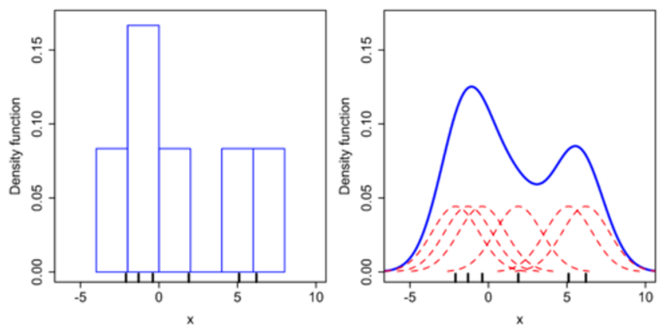
UA Speech is a database containing dysarthric speech samples from 19 individuals diagnosed with cerebral palsy. The speech materials encompass 765 isolated English words from each participant, comprising 300 distinct uncommon words, along with three repetitions of digits, computer commands, radio alphabet, and common words.
Data acquisition involved utilizing an 8-microphone array and one digital video camera to capture the speech recordings. For the purpose of statistical analyses, recordings from microphone no. 8 were used.

TORGO

The TORGO database comprises speech recordings from seven individuals exhibiting speech impediments due to cerebral palsy or amyotrophic lateral sclerosis. Additionally, age- and gender-matched control subjects are included for comparison. The stimuli used in the study are sourced from diverse origins, including the TIMIT database, lists of identified phonetic contrasts, and assessments of speech intelligibility.
Speech is captured using a head-mounted and a directional microphone. For articulatory data, the researchers utilize electromagnetic articulography, enabling precise measurements of tongue and other articulator movements during speech, along with 3D reconstruction based on binocular video sequences.

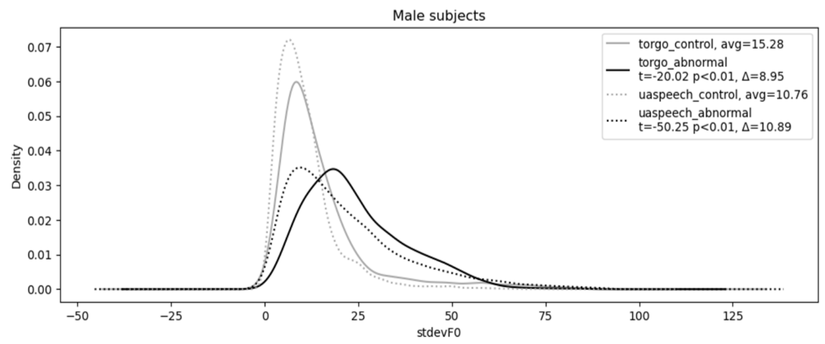
Kernel Density Estimation (KDE)

- non-parametric statistical technique employed to estimate the probability density function (PDF) of a random variable
- offering a smoothed representation that aids in visualizing and analyzing the data
- the fundamental idea is to place a kernel, typically a smooth function such as a Gaussian, at each data point and sum these kernels to generate a continuous estimate



The blue histogram on the left illustrates the raw data, while the blue curve on the right depicts the Kernel Density Estimate for the reaction times. On the plot on the right individual Gaussian kernels are depicted with a pink dashed line.
[image source: Wikipedia]

Standard Deviation of Fundamental Frequency



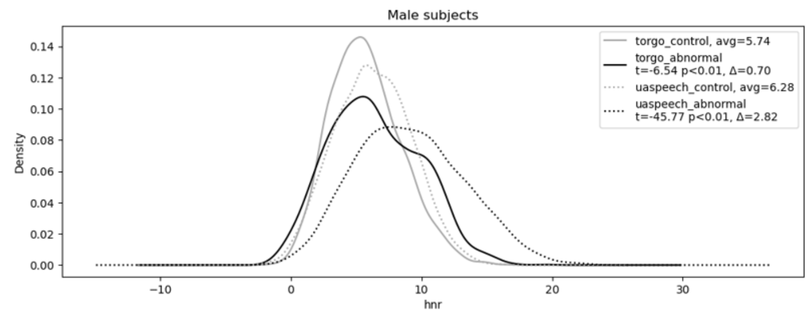
The distribution of the stdevF0 in the TORGO dataset, has a higher mean for abnormal speech compared to the control datasets. Same behavior is observed for the UA SPEECH dataset. Flaccid, spastic and hypokinetic dysarthria are charag tcterised by monopitch. Mapping this phenomena to acoustic features the standard deviation of F0 should be smaller for dysarthric speakers.

Harmonics-to-Noise Ratio (HNR)

The Harmonics-to-Noise Ratio (HNR) is a measure used to assess the purity of a signal by comparing harmonic components to noise components.

HNR quantifies the proportion of energy in a signal that is attributable to harmonics, which are multiples of the fundamental frequency, as opposed to noise.

A higher HNR value indicates a cleaner and more harmonically rich signal, while a lower HNR suggests a higher level of noise or non-harmonic interference.



Based on figure above for the male subject of UaSpeech dataset, the average HNR is significantly higher for abnormal speakers compared to the control group. Similar behavior was observed for the male subjects of the Torgo database.

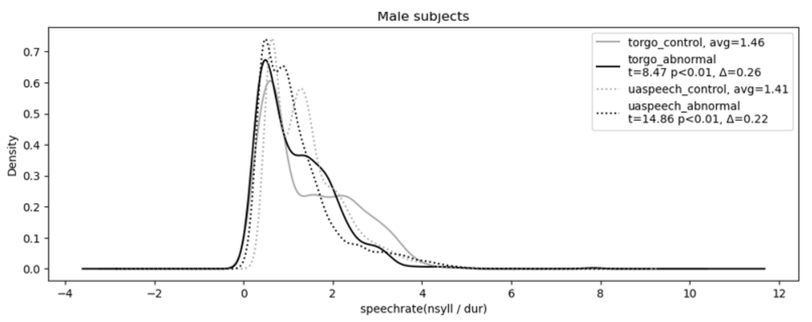
Speech Rate

Speech rate refers to the speed at which a person speaks, measured in terms of the number of syllables uttered per unit of time.

Variations in speech rate can convey different emotions, intentions, or levels of engagement.

A higher speech rate is associated with faster delivery, while a slower rate may indicate more deliberate or contemplative speech.

Speech rate plays a role in determining the intelligibility and comprehensibility of spoken language.



Based on figure above the speech rates for both control groups are significantly higher compared to the corresponding abnormal speaker groups. For Torgo Dataset the left side of the distributions overlap in a great manner, and the right tale of the distribution diverges. Similar, yet not so distinguishable, phenomena can be observed for the UaSpeech. Abnormal speaker groups contain patients of varying severity of illness and this may be observed directly in the speech rate feature.

General features of dysarthric speech

Fundamental Frequency:

- In general, dysarthric speakers are identified with mono pitch.
- Dysarthric speakers have breaking voice and locally “harsh” voices.

HNR:

- For dysarthric speakers, their HNR is higher by approx .5 dB.
- HNR for severe dysarthria may differ by > 5 dB between healthy, and abnormal speakers.

Speech Rate:

- Average speech rate of dysarthric speaker is 0.24 lower when compared to control group.

Expanding the study

Several speech synthesis techniques are applied to enlarge existing sets of dysarthric speech for use with a deep model approach. However, before that, the synthesized dysarthric speech outcome will be examined both by objective measures and subjective tests. The results will be compared in the form of stationary analyses and regarding a sequence of dysarthric utterances to highlight changes detected changes.

Conclusions

- The analytical approach allowed for discerning features corresponding to dysarthric speech.
- This means that they allow for differentiating between ‚normal’ and dysfunctional speech.
- On the basis of these features, synthesis will be performed with changes applied (with differentiated percentages).
- This is to enlarge existing datasets (mostly in English) for preparing an automated system for helping phonitricians dealing with the examination and treatment of voice, speech, language.