## Introduction

Voice disorder can be caused by variety of disorders such as structural lesion of vocal cord, neurogenic disorder, or even psychological stressors. There might be a specific phonic pattern in different patho-etiology. I would like to applying ML/DL methods to classify voice disorder using voice recording. I found a dataset of voice recording from patients with vocal cord diseases and health control.

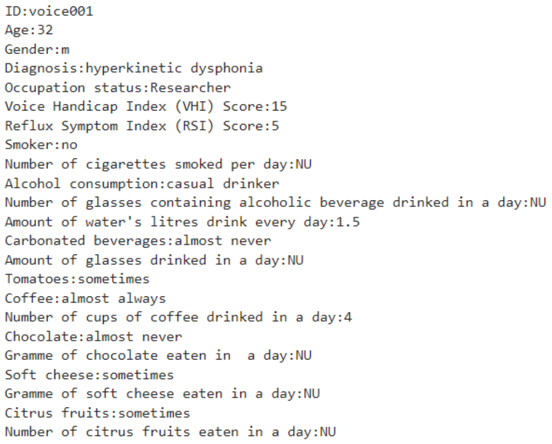
## Description of datasets

The dataset I used is *VOICED* (detailed description) from *Physionet*. It includes 208 voice samples (150 pathological [70 hyperkinetic dysphonia; 41 hypokinetic dysphonia; 39 reflux laryngitis]; 58 healthy) with expert-verified diagnosis. Subjects involved in the study aged between 18-70 years and diseases such as vocal folds or upper respiratory tract infections or with neurological disorders were excluded. The settings of voice recording as following:

- The acquired signals consist of a recording of a vocalization of the vowel "a" five seconds in length without any interruption of sound.
- All samples were recorded in a quiet room (< 30 dB of background noise) with humidity >30-40%.
- The voice recordings device was held at a distance of about 20 cm from the patient at an angle of about 45 degrees.
- All recordings were sampled at 8,000 Hz and their resolution was 32-bit. (So, a 5-second recording should contain 40,000 data point.)
- Each recording was filtered with an appropriate filter to remove any noise accidentally added during the acquisition.
- The participants were instructed to articulate the vocal sample, with a constant voice intensity, as they would during a normal conversation.
- For each subject certain training tests were performed about two/three times before the recording.

There are also demographic information including gender, age, pathology, lifestyle habits (e.g., smoking, alcohol and coffee consummation), occupational status, and the results of two specific medical questionnaires: the Voice Handicap Index (VHI) and Reflux Symptom Index (RSI).

*Figures: the left one is the content in info file; the right one is one of the voice recording (upper panel: whole recording, lower penal: 500-points recording).*



```
ID:voice001
Age:32
Gender:m
Diagnosis:hyperkinetic dysphonia
Occupation status:Researcher
Voice Handicap Index (VHI) Score:15
Reflux Symptom Index (RSI) Score:5
Smoker:no
Number of cigarettes smoked per day:NU
Alcohol consumption:casual drinker
Number of glasses containing alcoholic beverage drinked in a day:NU
Amount of water's litres drink every day:1.5
Carbonated beverages:almost never
Amount of glasses drinked in a day:NU
Tomatoes:sometimes
Coffee:almost always
Number of cups of coffee drinked in a day:4
Chocolate:almost never
Gramme of chocolate eaten in  a day:NU
Soft cheese:sometimes
Gramme of soft cheese eaten in a day:NU
Citrus fruits:sometimes
Number of citrus fruits eaten in a day:NU
```

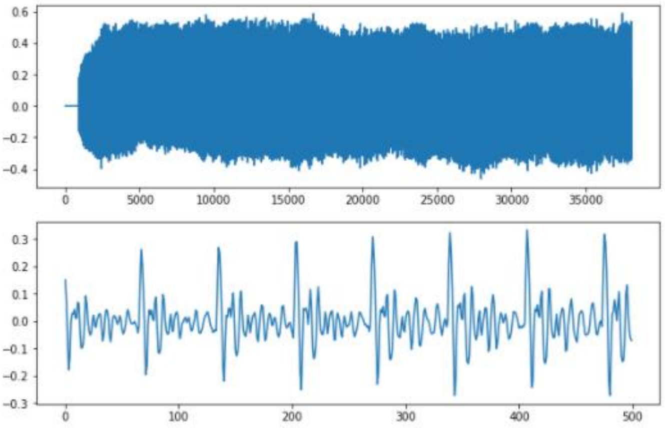
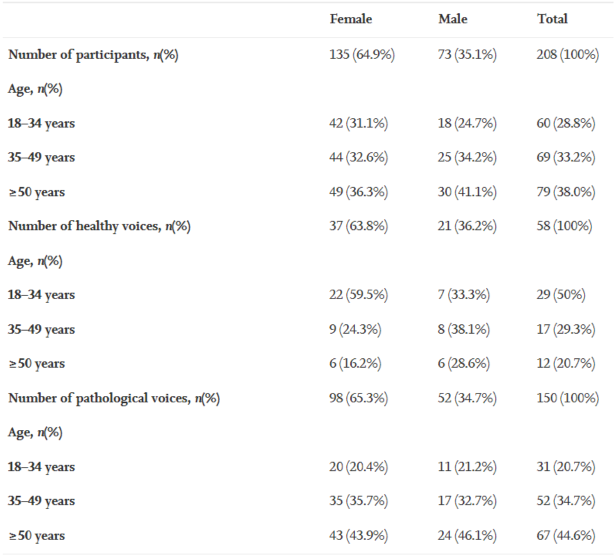
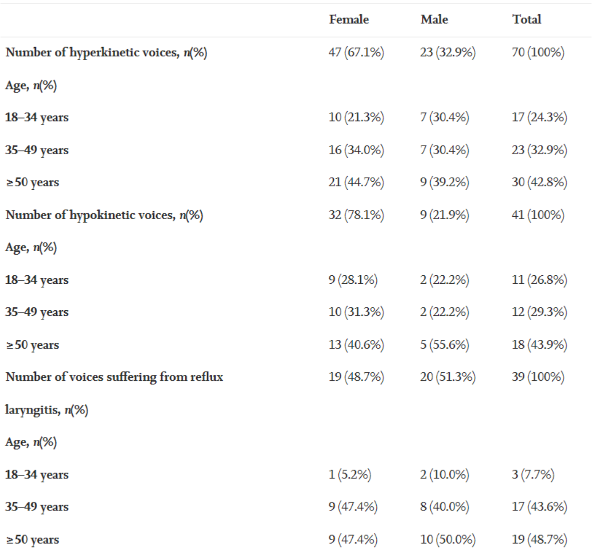


*Table: demographic information*



| | Female | Male | Total |
|---|---|---|---|
| **Number of participants, *n*(%)** | 135 (64.9%) | 73 (35.1%) | 208 (100%) |
| **Age, *n*(%)** | | | |
| **18–34 years** | 42 (31.1%) | 18 (24.7%) | 60 (28.8%) |
| **35–49 years** | 44 (32.6%) | 25 (34.2%) | 69 (33.2%) |
| **≥50 years** | 49 (36.3%) | 30 (41.1%) | 79 (38.0%) |
| **Number of healthy voices, *n*(%)** | 37 (63.8%) | 21 (36.2%) | 58 (100%) |
| **Age, *n*(%)** | | | |
| **18–34 years** | 22 (59.5%) | 7 (33.3%) | 29 (50%) |
| **35–49 years** | 9 (24.3%) | 8 (38.1%) | 17 (29.3%) |
| **≥50 years** | 6 (16.2%) | 6 (28.6%) | 12 (20.7%) |
| **Number of pathological voices, *n*(%)** | 98 (65.3%) | 52 (34.7%) | 150 (100%) |
| **Age, *n*(%)** | | | |
| **18–34 years** | 20 (20.4%) | 11 (21.2%) | 31 (20.7%) |
| **35–49 years** | 35 (35.7%) | 17 (32.7%) | 52 (34.7%) |
| **≥50 years** | 43 (43.9%) | 24 (46.1%) | 67 (44.6%) |

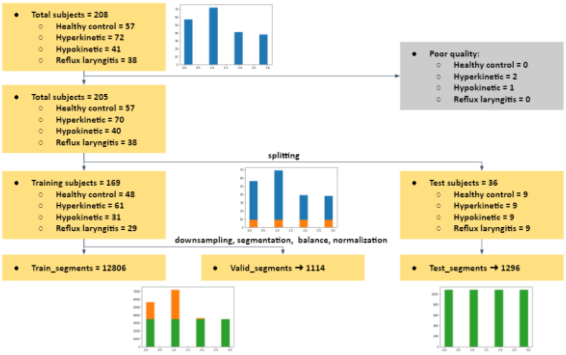


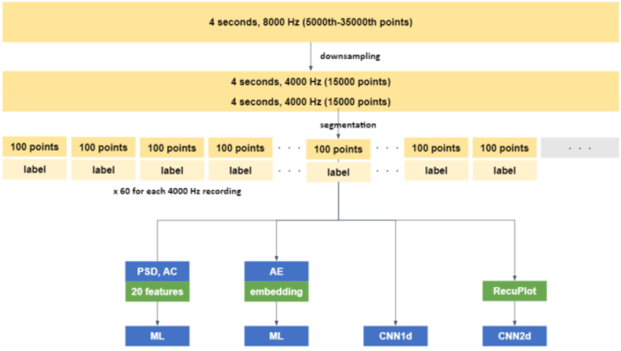
| | Female | Male | Total |
|---|---|---|---|
| **Number of hyperkinetic voices, *n*(%)** | 47 (67.1%) | 23 (32.9%) | 70 (100%) |
| **Age, *n*(%)** | | | |
| **18–34 years** | 10 (21.3%) | 7 (30.4%) | 17 (24.3%) |
| **35–49 years** | 16 (34.0%) | 7 (30.4%) | 23 (32.9%) |
| **≥50 years** | 21 (44.7%) | 9 (39.2%) | 30 (42.8%) |
| **Number of hypokinetic voices, *n*(%)** | 32 (78.1%) | 9 (21.9%) | 41 (100%) |
| **Age, *n*(%)** | | | |
| **18–34 years** | 9 (28.1%) | 2 (22.2%) | 11 (26.8%) |
| **35–49 years** | 10 (31.3%) | 2 (22.2%) | 12 (29.3%) |
| **≥50 years** | 13 (40.6%) | 5 (55.6%) | 18 (43.9%) |
| **Number of voices suffering from reflux laryngitis, *n*(%)** | 19 (48.7%) | 20 (51.3%) | 39 (100%) |
| **Age, *n*(%)** | | | |
| **18–34 years** | 1 (5.2%) | 2 (10.0%) | 3 (7.7%) |
| **35–49 years** | 9 (47.4%) | 8 (40.0%) | 17 (43.6%) |
| **≥50 years** | 9 (47.4%) | 10 (50.0%) | 19 (48.7%) |

## Description of all methods

Three recordings are discarded because of poor quality. Total 205 recordings are split into training set and test set. Then the recordings in both data set are underwent down-sampling to signal with sampling rate of 4000 Hz and then dividing into 100-point segments. Supervised learning will be used since the dataset contained labels (hyperkinetic, hypokinetic, reflux laryngitis, and healthy). A testing dataset containing 9 subjects per classes is split. And the other subjects will be the training and validation dataset.

*Figures: the left one shows the way to split dataset; the right one shows the methods used here.*



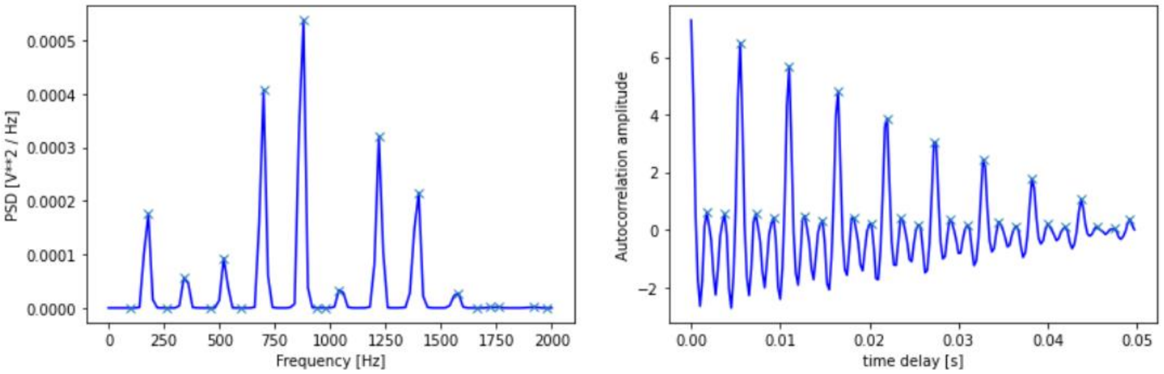


4 seconds, 8000 Hz (5000th-35000th points)
downsampling
4 seconds, 4000 Hz (15000 points)
4 seconds, 4000 Hz (15000 points)
segmentation
100 points
label
x 60 for each 4000 Hz recording
PSD, AC
20 features
ML
AE
embedding
ML
CNN1d
RecuPlot
CNN2d

The algorithms used here include:

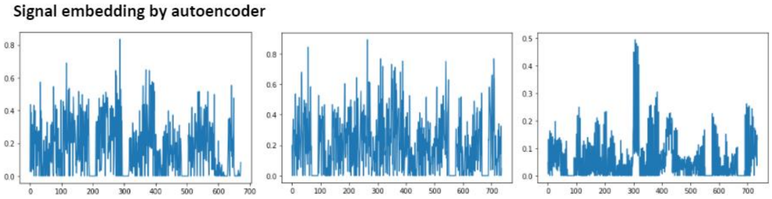
A. ML methods (SVM, decision tree, random forest) with selected features which are extracted from the segments using power spectral density (PSD) and autocorrelation (AC). I choose the top five frequencies and their power values, and the top one time-delay and their similarity values to be the input features.

*Figures: the left one is the PSD of a segment; the right one is the AC of a segment.*



B. Hybrid method using autoencoder-1dCNN to get low-dimensional features and then input the features to ML model (SVM, decision tree, random forest).

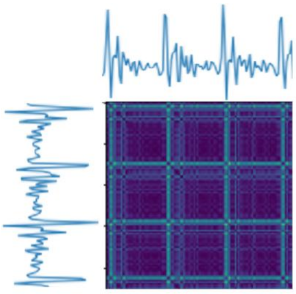
*Figures: three examples of signal embeddings*



C. 1-demensional CNN (dilated and non-dilated) using original signal

D. 2-demensional CNN (dilated and non-dilated) using signal-derived recurrence plot

*Figures: an examples of recurrence plot*



The performance of different algorithms will be compared.

Data pre-processing:

- Down-sampling—Each 5-s 8000-Hz recording will be down-sampled to two 5-s 4000-Hz recordings by re-sampling the point alternatively.
- Segmentation—Each 4000-Hz recordings will be segmented into 100-point segment.
- Balance—because the segment number of normal control and hyperkinetic voice are more, some segment would be discarded to make them equal to the number of hypokinetic voice and reflux laryngitis.
- Normalization—for (A) and (B), the selected features are normalized to the range of [0, 1]

**Justification for all methods**

Time series / signal is an interesting material that variety of methods can be used to analyze it:

(A) There are known methods to extract frequency domain and time domain feature from signals, and machine learning method can be applied. In this way, feature importance can be calculated, and also the model and the result are easy to be interpreted and visualized.

(B) Autoencoder can self-learn the features from input data and also reduce the dimension. Then the extracted features can be used for further analysis. It is really helpful when data analytic skill or tool is limited.

(C) I used 1-D CNN instead of RNN which is typically used to solve problems with sequential data because I want to compare 2-D CNN and 1-D CNN with similar architecture. And the time series can be treated as 1-dimension image and be processed by 1-D convolution.

(D) Time series can be transformed to a 2-D array by some visualization algorithms. Recurrence plot is a matrix whose element represent the distance between each point and the other one in the signal and it is used for analysis of recurring state of a dynamical system recurs. And it can be an input to CNN, just like an image. Dilated convolution is also applied in 1-D and 2-D CNN, which have a broader receptive field and can improve the model performance.

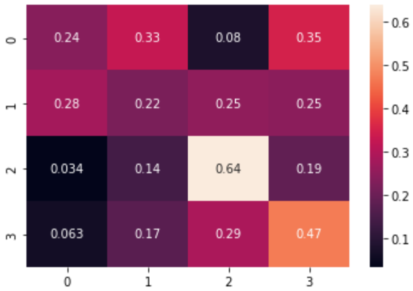
**Results**

1. For validation dataset which is split from training dataset, the autoencoder combined with machine learning has best performance up to accuracy of 0.79 and dilated convolution do improve the performance (0.83). And the 2D-CNN failed to distinguish different pathological voice.

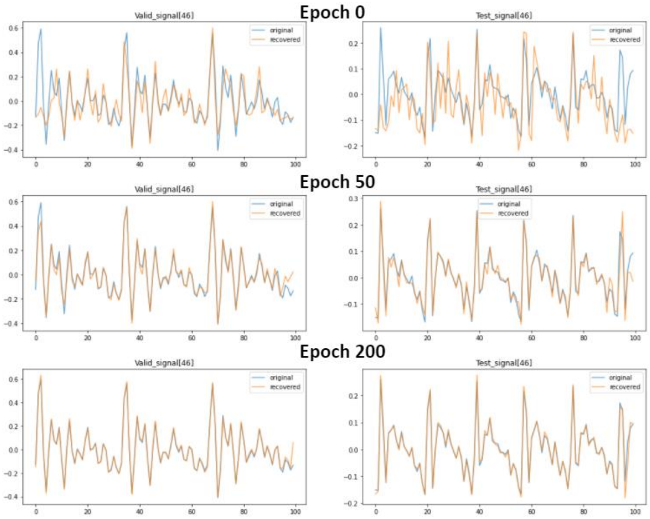
| | kernel | channels | dilation | epoch | ML | Train (acc) | Valid (acc) | Test (acc) |
|---|---|---|---|---|---|---|---|---|
| AE+ML | 1*3 | [16, 32] | [1, 1] | 50 | SVM(rbf) | 0.8402 | 0.7944 | 0.4313 |
| AE+ML | 1*3 | [16, 32] | [3, 2] | 50 | SVM(rbf) | 0.8721 | 0.8285 | 0.3920 |
| AE+ML | 1*3 | [16, 32] | [3, 2] | 50 | SVM(poly, 3) | 0.8716 | 0.7998 | 0.3920 |
| ML | 12 features | | | | SVM(rbf) | 0.6729 | 0.6499 | 0.3931 |
| ML | 12 features | | | | SVM(poly, 3) | 0.6517 | 0.6274 | 0.4046 |

| | kernel | channels | dilation | FC | epoch | Train (acc) | Valid (acc) | Test (acc) |
|---|---|---|---|---|---|---|---|---|
| CNN1d | 1*3 | [16, 32] | [1,1] | [512,256,128] | 50 | 0.5307 | 0.5676 | 0.4334 |
| CNN1d | 1*3 | [16, 32] | [3,2] | [512,256,128] | 50 | 0.5562 | 0.5339 | 0.3547 |
| CNN2d | 3*3 | [16, 32] | [1,1] | [512,256,128] | 50 | 0.2501 | 0.2487 | 0.25 |
| CNN2d | 3*3 | [16, 32] | [3,2] | [512,256,128] | 50 | 0.2502 | 0.2451 | 0.2508 |

2. When applied on the test dataset, the accuracy became poor (0.43); the performance on the class 2 (hypokinetic dysphonia) is best and followed by class 3 (reflux laryngitis).



3. The performance of signal regeneration of the autoencoder is really good and the embedding provide 736 features for each signal segment.



## Conclusion / Future work

1. The dataset is not sufficient to be representative and may have bias so the model is not robust. Data augmentation to increase and balance the samples can be helpful. However, augmentation method I used here is probably not appropriate for this dataset. The voice recording is short and relatively stable so that the segments from same recording are too similar to be a unique sample to improve the model performance. Other data augmentation technique such as SMOTE should be tried.
2. Neural network using the recurrence plot which is actually a dimension-increasing method that causes lots of computation and is really time-consuming. Simpler methods such as machine learning with feature extraction or a “shallow” neural network should be considered first which can be also working.
3. There are other methods can be tried including adding residual connections, multi-task learning, recurrent neural network, and even combination of self-learned features and selected features. And transfer learning with pre-trained model for signal analysis is also a good option.