ECE 443: Speech Signal Processing

Project Report

Classification of Pathological Data

using MFCC & LP Residual features

I. Understanding Spasmodic Dysphonia

The human standard of life can be severely affected by their individual pathological voice condition. Some common impairments to the voice are structural lesions, neoplasms, and neurogenic disorders. Spasmodic dysphonia, or laryngeal dystonia, is one such. In spasmodic dysphonia, the muscles inside the vocal folds spasm (make sudden, involuntary movements), interfering with vocal fold vibrations. Spasmodic dysphonia causes voice breaks during speaking and can make the voice sound tight, strained, or breathy. Symptoms of spasmodic dysphonia can come on suddenly or gradually appear over the span of years. Spasmodic dysphonia is of three kinds i.e

- **Adductor spasmodic dysphonia** - This is the most common form of spasmodic dysphonia. In this disorder, spasms cause the vocal folds to slam together and stiffen. These spasms make it difficult for the vocal folds to vibrate and produce sounds.
- **Abductor spasmodic dysphonia** - This is less common. In this disorder, spasms cause the vocal folds to remain open. The vocal folds cannot vibrate when they are open too far. The open position also allows air to escape from the lungs during the speech. As a result, the voice often sounds weak and breathy.
- **Mixed spasmodic dysphonia**. - This is a combination of the above two types and is very rare. Because the muscles that open *and* the muscles that close the vocal folds are not working properly, it has features of both adductor and abductor spasmodic dysphonia.

Datasets

The dataset contains samples of healthy people and people suffering from Spasmodic dysphonia. It contains 42 samples of healthy people and 30 samples of Spasmodic Dysphonia. Since the data is skewed, we can try modifying our loss function. The main idea is to have more penalties for misclassification in the minority class and lesser penalties for the majority class misclassification.

Pre-Processing Block

The preprocessing block is crucial since the data is noisy and contains background speakers (whispers).

- **Pre-emphasis Filter**: The use of this block is to highlight the high-frequency components of the voice sample

Feature Engineering

We have decided to use both local and global features for our model.The local features include - jitter,shimmer,fundamental frequency ,etc . For the global features, we have MFCC, LPC, Cepstral Coefficients, etc.

- **MFCC:** Mel-frequency Cepstrum is a representation of a sound signal, based on the linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. The coefficients that collectively comprise the Mel-frequency Cep-strum are called MFCC features. In contrast to the linearly-spaced frequency bands obtained from the cepstrum of a sound signal, in a Mel-frequency cepstrum, the frequency bands are uniformly spaced on the mel scale.
- **MFCC on LP residual:** LPC is a spectral estimation technique because it provides an estimate of the poles of the vocal tract transfer function. LP Residual gives us the speaker information --as it characterizes the excitation features. We compute MFCCs on this vector to extract vocal tract information.
- **The first and second derivatives of cepstral coefficients:** These are useful to investigate the properties of the dynamic behavior of the speech signal.Delta features were first introduced to include the

temporal dynamics along with the static cepstral features. The cepstral feature vector carries spectral static information i.e., filterbank power spectral envelope of an individual speech frame. However speech is a dynamic signal, meaning that it changes over time, and framing the time signal into a smaller segment destroys the temporal dynamics and trajectories between the successive time frames.

- **Jitter:** Jitter is defined as the parameter of frequency variation from cycle to cycle. This describes the instabilities of the oscillating pattern of the vocal folds, quantifying the cycle-to-cycle changes in fundamental frequency. The jitter is affected mainly by the lack of control of vibration of the cords; the voices of patients with pathologies often have a higher percentage of jitter. Most researchers considered typical value variation between 0.5 and 1.0% for the sustained phonation in young adults.
  To determine this parameter, which reflects the variation of the successive periods, the algorithm has to implement a function that detects the timing of the fundamental period. The steps for this are :.

  i.)Remove the linear trend of the signal
  ii.)Take modulus
  iii.)Take a moving average with a length corresponding to about 10ms (length similar to a glottal period). Then peak is searched as the maximum of the acoustic signal under a window of 15 samples before and 15 samples after the index of the maximum of the moving average.

- Finally, we get a vector that contains the peak levels corresponding to the beginning of the glottal pulse signal.

After the determination of the onset time of the glottal pulses, the jitter can be determined by the formula given below

$$Jitter(\%) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}(|T_i - T_{i+1}|)}{\frac{1}{N}\sum_{i=1}^{N}(T_i)}$$

Where Ti is the duration in seconds of each period and N is the number of periods. The threshold limit for detecting pathologies is 1.04%.

- **Shimmer:** This relates to the amplitude variation of the sound wave.

This indicates the instabilities of the oscillating pattern of the vocal folds, quantifying the cycle-to-cycle changes in amplitude. The shimmer changes with the reduction of glottal resistance and mass lesions on the vocal cords and is correlated with the presence of noise emission and breathiness. The shimmer is estimated as the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20. It is expressed in decibels. To determine the Shimmer parameters the methods used for the jitter are followed. The algorithm began by determining the onset time of the glottal pulses of the signal and the respective magnitude of the signal at that sample. Then we use the formula given below :

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20log\frac{A_{i+1}}{A_i}\right|$$

In the formula below Ai are the extracted peak-to-peak amplitudes and N is the number of extracted F0 periods.The limit to detect pathologies is 0.350 dB.

- **Harmonic to noise ratio (HNR):** Defined as the log ratio of the energies of the periodic and the aperiodic components. This quantifies the ratio of signal information over noise due to turbulent airflow, resulting from an incomplete vocal fold closure in speech pathologies. The first component arises from the vibration of the vocal cords and the second follows from the glottal noise, expressed in dB. The evaluation between the two components reflects the efficiency of speech, i.e., the greater the flow of air expelled from the lungs into the energy of vibration of vocal cords. In these cases, the HNR will be greater. A voiced sound is thus characterized by a high HNR. A low HNR denotes an asthenic voice and dysphonia. That is, a value of less than 7 dB in HNR is considered pathological.

Machine Learning Models

- Since the problem involves classification, we can use techniques such as Support Vector Machine, Logistic regression, GMMs, KNN, Random forest, XGBoost, etc.
- We have experimented with the RBF kernel in the case of SVM.
- As explained in the previous slides, we have experimented with two modes of loss calculation: balancing the loss value across classes to handle the class imbalance and the normal mode. We call the former mode, “class balanced mode”.
- SVM deals with hinge loss and is a max margin classifier, while logistic is based on cross-entropy loss. GMM is based on the EM algorithm. Random forest and XGBoost have similar roots, they are tree-based ensembling methods. KNN is based on simple majority voting amongst the K nearest feature vectors. We have tried to accommodate different types of classification algorithms.
- XGBoost is a special tree boosting based algorithm which is designed to take a lesser amount of resources to train on, but produces good results. It is a gradient boosting based tree ensemble model. Reference (For more details) : https://arxiv.org/pdf/1603.02754.pdf
- The evaluation metrics for our models have been mentioned below.

Evaluation metrics :

- **Precision** -This tells us how many, out of all instances that were predicted to belong to class X, actually belonged to class X. The precision for class X is calculated as

$$TP/(TP + FP)$$

TP = the number of true positives for class X
FP = the number of false positives for class X

- **Recall -** This expresses how many instances of class X were predicted correctly. The recall is calculated as

$$TP/(TP + FN)$$

TP = the number of true positives for class X
FN = the number of false negatives for class X

- **F1 score -** This is the harmonic mean of precision and recall.
- **Accuracy score:** This denotes the number of samples predicted correctly divided by the total number of samples.

Results

Things we tried as part of this project:

1. Concatenating all per frame MFCC feature vectors and feeding them to classifiers
2. Building a classifier which predicts at frame level and taking a majority vote of all frames during inference with MFCC features.
3. Varying the number of MFCC feature values from 13-20 and studying its effect.
4. Adding jitta, shimmer features and studying their effect.
5. Adding HNR feature value and studying its effect.
6. Trying out different classifiers and studying how various algorithms are able to capture the nuances in speech signals. Trying out two modes : class balanced, normal modes wherever relevant

We have the detailed list of results here:
https://docs.google.com/spreadsheets/d/13xdr6rXQnn15iVOiMZEflnoxoT0dqECf8gPF75C4G3g/edit?usp=sharing

General observations

a. Concatenation vs Frame by frame : There is almost 27% difference in accuracy between frame by frame analysis and simple concatenation (90%, 63% accuracy respectively). Huge difference might be because the number of samples is quite low when we approach at signal level. (There are only 72 files), whereas there are around 55K frames in total.
b. There is not much of a difference in end result when we vary the number of mfcc feature values from 13-20. There is one peculiar case though, of SVM, for which 15 feature values gave around 86% accuracy but 13,19 feature values gave 90% accuracy. In all other cases there is no

notable change.

c. Adding Jitta, shimmer, HRM feature values indeed helped improve the accuracy by 0.5-1.5%. Since the feature vector size is very small, adding rich features might have helped.
d. Out of all classifiers, class balanced versions of Random forest and SVM performed well on the whole. We can clearly observe that the class balanced mode is outperforming the normal mode in all the cases. This is expected as the dataset is imbalanced.

Paper References

1. Voice Disorder Identification by using Machine Learning Techniques
2. Vocal Acoustic Analysis – Jitter, Shimmer and HNR Parameters
3. Linear prediction residual features for automatic speaker verification anti-spoofing
4. http://vlab.amrita.edu/?sub=3&brch=164&sim=616&cnt=1108