Perceptual Dissimilarity and Intra–Speaker Indication

Sidharth Gulati, Tushar Sudhakar Jee

University of California, Los Angeles

sidharth.gulati@ucla.edu, tjee@ucla.edu

Abstract

In this study, we examined the perceptual dissimilarity and a measure for intra - speaker identification. We examined 52 speech utterances by different speakers provided by the UCLA Speech Processing and Auditory Perception Laboratory(SPAPL). The UCLA medical school conducted a series of experiments to measure perceptual similarity (or dis- similarity) between and within speakers. The listeners are asked to listen to a pair of sentences and to tell if the sentences are from the same speaker or from two different speakers.The measure of uncertainty in their response is also taken into account. In this project, we estimated the acoustic characteristics such as F0, F1, F2, F3, F4, HNR, CPP and H1-H2 using VoiceSauce. MFCC coefficients and LPCC were computed to estimate the speech utterances. We found that the ensemble of these features provide an acceptable measure of perceptual dissimilarity and intra - speaker indication. We also found that the the variation in F1, F2 for the speakers examined doesn't vary considerably, thereby, providing a good estimate of the spectral features.

Keywords: Formants, HNR, CPP, H1-H2, MFCC, LPCC, Ensemble Methods, perceptual dissimilarity, intra-speaker indication.

1. Introduction

Speech recognition is a potential technology that will make our daily lives more secure. Apart from providing protection to humans computerized and electronic belongings, potential applications include command and control, dictation, transcription of recorded speech, searching audio documents and interactive spoken dialogues.It is one of the types of biometric used to identify and authenticate users on basis of their voice.The project uses set of acoustic features and algorithms to predict perceptual similarity or (dis)-similarity between and within speakers. Success in this task depends on extracting speaker-dependent features from the speech signal that can effectively distinguish one speaker from another.

Among features with reference to those employed in the past for speaker identification,the mel-frequency cepstral coefficients (MFCCs) rank quite high as they carry both speech and speaker information. Although MFCCs implicitly capture speaker-specific information, the paper aims at exploring other parameters to do the same explicitly.

Little is known about the extent to which voice pattern of an individual might vary with diverse speaking situations.This paper describes a preliminary investigation into speech variation in individual talkers and consequently the listeners perception of a voice sample coming from a particular speaker.It also covers extraction of speaker-specific features from the speech signal to distinguish one speaker from another.The feature set consists of four formants (F1, F2, F3, F4), the pitch frequency (F0),Harmonic to Noise Ratio(HNR), CPP and the difference between the strength of the first (H1) and second harmonics (H2).

Section 2 includes a brief summary of the Refered Literature.The System design that includes the Algorithm used is talked about in Section 3. In Section 4 the feature set employed and their computation is described. The database used and the experiments conducted are explained in Section 4. Section 5 discusses the given dataset and results obtained along with Discussions outlined in Section 6. Section 7 concludes our observations while trying to optimize the results. Section 8 deals with work we plan to do in the Future with an aim to explore a larger , diverse dataset interspersed with noise using different models. Section 9 and 10 comprise the Acknowledgment and References respectively.

2. Literature Review

- **The Relationship between Acoustic and Perceived Intraspeaker variability in Voice Quality Jody Kreiman1, Soo Jin Park, Patricia A. Keating and Abeer Alwan. [1]**

  The acoustic variations between and within 5 talkers and their effect on the likelihood of voice samples not being identified as coming from the same speaker were examined. Talkers were drawn from a large database recorded to capture everyday variations in vocal characteristics. In the paper, the discussion and results are aimed at answering whether talkers are consistently more similar acoustically and perceptually to themselves than other talkers.The results proved that both acoustically and perceptually talkers are more self similar than other talkers. It concluded that interspeaker variability in voice quality exceeded intraspeaker variability even though by a small margin.It also gives a brief outline of future work that would include a much larger data set with more talkers (including males) and various kinds of speech.

- **A New Set of Features for Text-Independent Speaker Identification Carol Y. Espy-Wilson, Sandeep Manocha and Srikanth Vishnubhotla [2].**

  In addition to understanding the use of Gaussian Mixture Models(GMM) this paper discussed using small set of low level acoustic parameters to capture information about vocal tract shape and size of the speaker.It justifies using a set of eight acoustic parameters gives comparable performance to using 26 or 39 MFCC's for speaker identification.The speaker models were constructed using the Gaussian Mixture Models (GMM) and were trained using maximum-likelihood parameter estimation[3]

  The 32-mixture GMM gave the best performance for the feature set(F1-F4, H1-H2, Energy) and the MFCCs.It was highlighted that current work also dealt with the automatic extraction of creakiness and nasalization.Their

future work entailed including temporal information of the parameters,using a framework other than the GMMs and exploring a supervised approach to speaker identification. The changing relationship between acoustic properties and source/vocal tract information across classes of speech sounds justifies trying the supervised approach to speaker identification.

- **Some Acoustic Correlates Of Perceived (Dis) Similarity Between same-accent voices,Francis Nolan, Kirsty McDougall and Toby Hudson [4].**

  The paper involved constructing a model that would be able to predict how similar two voices are based on their acoustic and linguistic properties.Using Multidimensional scaling(MDS), pseudo perceptual dimensions were evaluated that helped quantify correlation between objects.First MDS dimension correlated with F0 confirming F0's key role in voice similarity followed in order by F3,F2 and F1.To disentangle the perception of personal voice quality from linguistic factors, this experiment used speakers carefully matched for accent. Foundation of the feature set can be further augmented considering further acoustic measures such as measures of rhythm, F0 dynamics, and other spectral parameters.

- **Speaker Identification Using GMM with MFCC Tahira Mahboob, Memoona Khanum, Malik Sikandar Hayat Khiyal, Ruqia Bibi [5].**

  They proposed a real time voice recognition system depending on Mel frequency Cepstral coefficient (MFCC) for feature extraction and on GMM for training.Subsequent to taking the speech sample as input,voice features are extracted by dividing the voice sample frame by frame.A Hamming window is applied to minimize discontinuities at edge of each frame. Vector quantization with K- means is used to find the clusters. A special case of ML (Maximum Likelihood) is used for estimation parameters. It resulted in the log probability of training sequence, which is further used to define the threshold and for verification of speaker. The speaker is recognized by comparing the log probability to the defined threshold in the system. The future work entailed further reduction in error rate.For the same instead of using K-means algorithm for clustering, Distribution based clustering may be tried. Maximum A Posteriori (MAP) estimation may be used to estimate parameters instead of estimating GMM parameters via the EM algorithm. In MAP, a GMM-UBM (universal background model) might be used to derive a speaker model.

- **Cepstral Peak Prominence: A comprehensive analysis Rub en Fraile,Juan Ignacio Godino-Llorentea [6].**

  The study of cepstral peak prominence (CPP) is presented, understanding its meaning and relation with voice perturbation parameters.The importance of CPP in understanding *Dysphonic Severity* is highlighted. CPP has been reported to be one of the most reliable and robust acoustic measures of dysphonia [7] .

  To a major extent the reason for the robustness is accorded to the fact that it does not need previous pitch detection and tracking [8]. Firstly, the meaning of the cepstral log-linear regression involved in the calculation of CPP following was analyzed based on the classical source-filter model of voice production. This helped conclude that subtraction of the log-linear regression from the first rahmonic value for CPP evaluation has little impact on the spectrum of the signal. Consequently, the conclusions of[9] regarding the relevance of the regression may well be more related to the relevance of calculating CPP following a systematic algorithm than to the effect of the regression itself. Further facets of the first rahmonic explained by [10] can be applied to CPP. The voice production is modelled using the traditional source-filter model and the first cepstral peak that is taken to have a Gaussian shape. Moreover, it is concluded that the meaning of CPP is very similar to that of the first rahmonic and some detailed insights are provided on its dependence with fundamental frequency and vocal tract resonances.

- **Speech Recognition System: A Review Nitin Washani,Sandeep Sharma [11].**

  This paper presents the advances made as well as highlights the pressing problems for a speech recognition system. Further the system is classified as Front and Back end for better clarity of Speech recognition system in each part. The Front end analysis included Speech Preprocessing and Feature Extraction. Speech Preprocessing including End Point Detection, Pre-emphasis Filtering, Noise Filtering, Framing, Windowing and Echo Cancelling. Feature extraction included elaboration on Cepstrum analysis, Spectrogram, MFCC (Mel Frequency Cepstrum Coefficient) and LPC (Linear Predictive Coefficient). The Back-End analysis comprises the Speech Classification block. The Speech Classification process classifies the extracted features and matches the input speech waveform to the best fitting sound on the test dataset, represents them as an output. The commonly reviewed techniques for Speech Classification are HMM (Hidden Markov Model), DTW (Dynamic Time Warping), VQ (Vector Quantization), ANN (Artificial Neural Network)[12]. Gaussian Mixture Models were found to give a poor Recognition Rate in Speech Recognition compared to other Classifiers but in Speaker Recognition the result were efficient. MFCC is a more preferred technique of Feature Extraction technique as it generates the training vectors by transforming speech signal into frequency domain, thus being less prone to noise interference.

3. System Design

A pictorial representation for the system flow is given in Figure 1. The algorithm implemented by our system is described below :

- Extract 71 features detailed subsequently forming the test and train data sets.
- The datasets are fed to the Ensemble model to predict Perceptual Dissimilarity indication value using Bag method for Regression and Intra-Speaker Indication Value using Ada-boost method for Classification.
- We trained the Statistical Model, here Bagged Regressor and AdaBoost Algorithm with 42 speech utterances pair and predicated the perceptual dissimilarity and class labels on the remaining 12 speech utterances pair.
- The predicted values from the model are compared to the actual values in the test data evaluating the Root Mean Square Error (RMSE) and Classification Error.

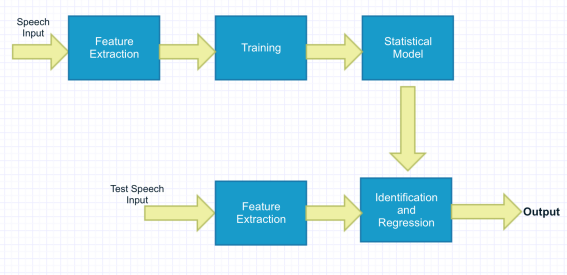

Figure 1: *Algorithm Design*

4. Features extraction

4.1. Vocal tract features

As mentioned in [2], the pitch frequencies and the formant frequencies of the voiced part of a speech provide significant amount of information regarding the shape of the vocal tract.

F_0, the pitch frequency, is the fundamental frequency of the glottal vibration. As different speakers have different size and shape of glottis, in general, it is usual for them to have different fundamental frequencies. In addition to the average of F_0, the temporal variation of the fundamental frequency has also been shown to be useful as an speaker cue [13]. The range of F_0 is usually used to capture such variation. In our experiment, the standard deviation of the F_0's across the frames are used as the description of such variation.

As for the formant frequencies, the higher order formants are more often used as features for the speaker since they tend to stay steady across time independent of the content of the speech. F3, F4 have been shown to contain information about the length and shape of the vocal tract. In our design, the $F_3, F4$ formants were extracted through taking average values across the time frames.

F_1, F_2 are more frequently used as indicator for different vowels. However, positive relations have been found between the closing of oral cavity and F1, the back(front) displacement of the tongue. So it is reasonable to use them as the feature for speaker since it is expected that different speaker will tend to have different behavior for these two characteristics. In order to capture such characteristics while keeping in mind that F_1, F_2 varies significantly depending on the text content, the standard deviations of F_1 and F_2 are taken in the in the hope that they would reflects the trajectories of the these formants across time.

4.2. Mel-Frequency Cepstrum Coefficients(MFCC)

MFCC was developed based upon frequency analysis to the ear, which aimed to model how human perceives sound. The cochlea inside the ear has the ability such that different parts of it vibrates on different groups of coherent incoming frequencies of the sound. According to the signals generating by these different portions of the cochleas, human brains are able to distinguish the different frequency components coupled in the sound. Human ears has also formed a well organized structure such that they are more sensitive to the the low frequencies compared to the high frequencies. MFCC is a feature takes all of the above characteristics of human ear into consideration.

MFCC of a given digital speech signal is computed in several steps:

- divided the signal into frames

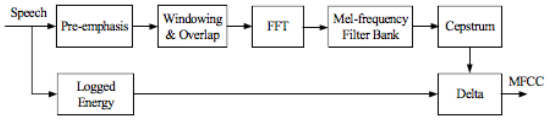

Figure 2: *Conventional Block Diagram for extracting MFCCs.*

- apply DFT to frames
- apply Mel-scale filterbank to the spectra and compute the energy in each filter
- take logarithm of the filterbank energies
- apply discrete cosine transform

Frequency is converted to Mel-scale using the following formula:

$$Mel(f) = 1127\ln(1 + \frac{700}{f}),$$

which means at high frequency, a larger range of frequency sweep will corresponds to the same Mel-scale interval of a smaller low frequency sweep. Converting into Mel-scale is really the part of MFCC that takes the sensitivity of the ears into account. Moreover, the use of filterbank to compute energies for groups of frequency effectively mimics the structure of the cochlea.

4.3. Cepstral Peak Prominence(CPP)

Cepstrum of a signal is the power spectrum of the logarithm of power spectrum of signal.With regard to computing the cepstrum, each short section of the acoustic signal is converted to the spectrum by a Fourier transform. The spectrum shows the intensity of each frequency in the signal. Applying a Logarithmic function and then taking Inverse Fourier transform, the spectrum is converted into the cepstrum. The frequency axis of the spectrum is converted into the quefrency axis which is in a time domain again. The cepstrum helps understand the harmonic structure of the spectrum as it describes the intensity of periodic patterns in the amplitude spectrum.

Cepstral Peak Prominence (CPP) is the difference in amplitude between the cepstral peak and the corresponding value on the regression line that is directly below the peak. It represents how far the cepstral peak emerges from the cepstrum background. It is an acoustic measure of voice quality that is a robust measure of Dysphonia severity[7] Widely used ,as Hartl et al proposed ,in combination with other parameters to assess the effects of surgical treatments [14].The consistency in performance of CPP to evaluate voice quality has inspired researchers to propose its application to purposes like assessment of speech intelligibility [15] and detection of cognitive load [16].

4.4. Harmonics-to-Noise Ratio (HNR)

A method to detect cycle to cycle perturbations ,it provides an indication of the overall periodicity of the voice signal by quantifying ratio between the voiced and unvoiced components. The Harmonic structure energy (periodic part) is evaluated from the speech signal, and the additive noise(aperiodic part) from different voice disorders.

The auditive impression of a voice with a low HNR is overwhelmingly hoarse, in contrast to that with hypertensive high HNR sounds.

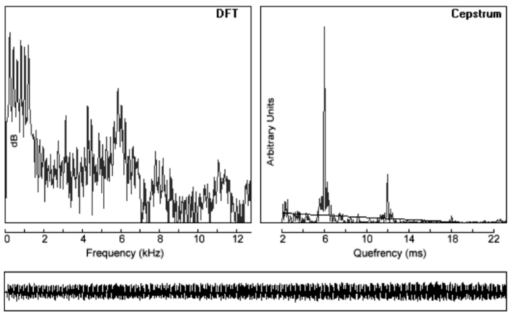

Figure 3: *DFT and Cepstral analysis for a normal female voice. The cepstral peak prominence in this sample corresponds to the fundamental frequency and is substantially greater than the average cepstral amplitude[17].*

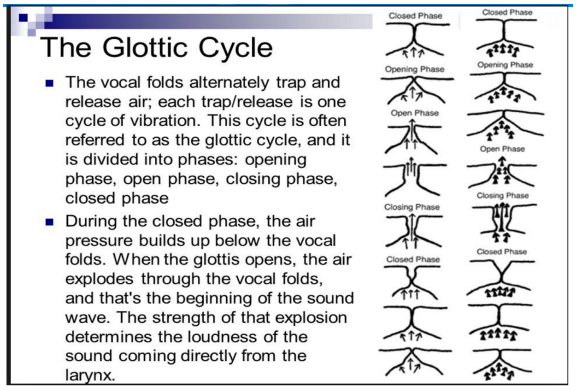

Figure 4: *Glottic Cycle [18]*

$$HNR = 10\log(1 + \frac{E[v^2]}{E[u^2]})(dB)$$

4.5. H1-H2

The difference in amplitudes of first and second harmonics(H1,H2) are useful for quantifying the degree of Glottal Adduction in different voices perceived. The Airway at the vocal cord level is called Glottis and the vocal folds coming together to close the glottis is Adduction. H1-H2 indicates the relative length of the opening phase of the Glottal pulse as is described in the figure titled Glottic cycle.

4.6. Linear Predictive Cepstral Coding(LPCC)

A cepstrum has a number of advantages (source-filter separation, compactness, orthogonality) contrary to the Linear Prediction Coefficients that are too sensitive to numerical precision. Thus often desirable to transform LP coefficients into Cepstral coefficients.The Linear Prediction Coefficients are represented by a_k and cepstral coefficient by c_n.

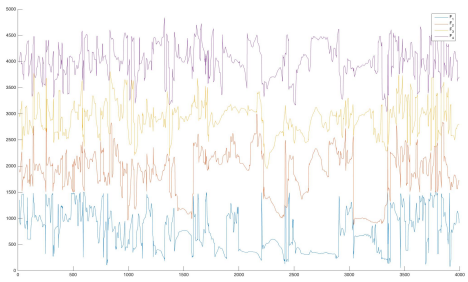

Figure 5: *Formant (F0-F4) Plot for Speaker 51A_2a.*

$$c(n) = \begin{cases} 0 & \text{if } n < 0 \\ log(G) & \text{if } n = 0 \\ a_n + \sum_{k=1}^{n-1} \frac{k}{n} c_k a_{n-k} & 0 < n \leq p \end{cases}$$

5. Results

5.1. Feature Dimension

We used 71 features for training and testing in this project. The break up of features is given below :

Table 1: *Features Used*

Feature Name	Number
$F0$	1
$F1$	1
$F2$	1
$F3$	1
$F4$	1
HNR	1
CPP	1
$H1-H2$	1
$MFCC$	13
$LPCC$	50
$Total$	71

We have trained our system on 40 speech utterances and tested on 12 speech utterances, provided by UCLA Speech Processing and Auditory Perception Laboratory(SPAPL). Figure 5 shows a sample formant plot for Speaker 51A_2a.

Figure 6 shows a sample H1-H2, CPP, HNR, MFCC and LPCC plots for Speaker 51A_2a.

5.2. Perceptual Dissimilarity

We used Bagged Trees Algorithm of Ensemble method for predicting Perceptual Dissimilarity. The minimum Average Root Mean Square Error for Perceptual Dissimilarity was found to be 1.88. We also, tested our algorithm on the test data posted on EEWeb and obtained a RMS error of 2.1.

5.3. Intra Speaker Indication

We used AdaBoost Algorithm of Ensemble Trees for Intra Speaker Indication and obtained a minimum classification Er-

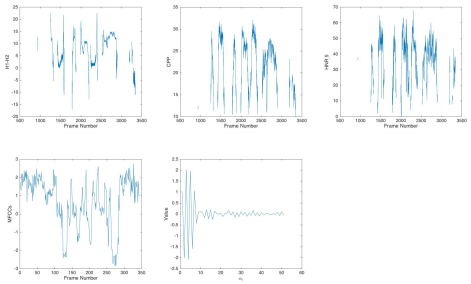

Figure 6: *Feature Plot for Speaker 51A_2a.*

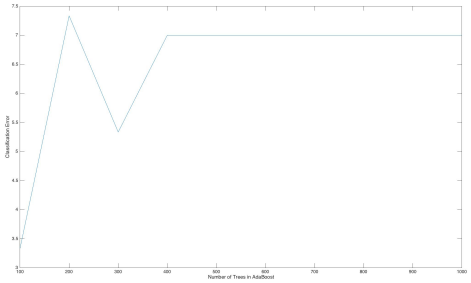

Figure 7: *Classification Error Rate vs Number of Trees*

ror Rate was 3.33%. The classification error rate on tested data was 5.33%.

Apart from the Ensemble Tree method, we used Support Vector Machine and Gaussian Naive Bayes for classification. The classification error rate for corresponding classifiers is given below.

Table 2: *Classification Error Rate for different Classifiers.*

Classifier	Classification Error
$AdaBoost$	3.33
SVM	27.33
$NaiveBayes$	$\approx$ 11.67

As it is evident from the above table, minimum classification error was achieved using Ensemble (AdaBoost) Classifier. Also, the experiment reported a very high classification error for SVM. This is due to the fact that there is not enough data to train high dimension support vectors and the corresponding margins between the support vector is low.

Figure 7 shows the variation of classification error rate as a function of number of trees used in the AdaBoost Algorithm of Ensemble Method. As we can see that as the number of trees increases, the classification error rate also increases until it becomes constant at 7%. It is also evident that the minimum classification error rate is obtained with 100 trees in the AdaBoost algorithm of Ensemble methods.

Figure 8 shows the variation of Root Mean Square error as a function of number of trees used in the Bagged Trees Algorithm of Ensemble Method. As we can see that as the number of trees increases, the RMS error also increases. It is also evident that

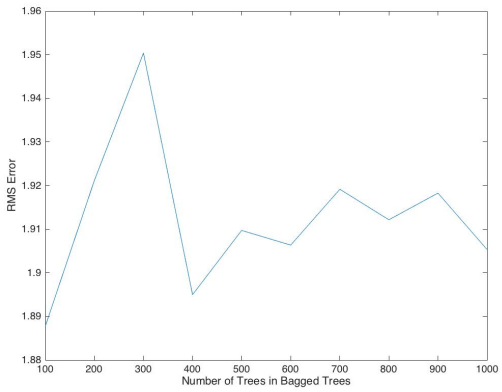

Figure 8: *Root Mean Square Error vs Number of Trees*

the minimum Root Mean Square error of 1.88 is obtained with 100 trees in the Bagged Tree algorithm of Ensemble methods.

6. Discussion

For using Formants(F1 - F4) as features, we tried modeling them using Gaussian Mixture Model, but the classification error rate was high that is, $\approx 9\%$. Also, the RMS error was 2.43, which is quite high. But, when we used the average value of corresponding formants as features, classification error considerably dropped to 3.33% and a RMS error of 1.88. Same, technique was done for using MFCCs as features, but the classification rate was quite high($\approx 15\%$). But, with mean value of MFCC it dropped considerably to 3.33

We also tried to use 60 LPCC coefficients but the classification error rate and RMS error dropped considerably. A optimum value of 1.88 of RMS in case of perceptual dissimilarity and 3.33% for intra speaker indication was obtained with 50 LPCC features.

The performance for LPC coefficients was also not good as we obtained a classification error rate of $\approx 10\%$ and an RMS error of 2.1. We then, did the cepstral analysis on LPC coefficients and obtained a classification error rate of 3.33% and RMS error of 1.88.

7. Conclusions

The goal of the project was to build a system that could discern the perceptual properties of different speaker and thereby, predicting the measure of intra speaker indication. This project implemented 2 well known features namely, MFCC and LPCC along with formants, pitch period, CPP, HNR and the difference between first and second harmonics. As an ensemble of these features, we obtained a minimum classification error rate of 3.33% and a minimum root mean square error of 1.88. We also observed slight variations from the well established literature regarding the performance of F1 and F2 over the whole speech utterance. Though, in general, F1 and F2 vary considerably, but the finding of this project suggest that F1 and F2 not vary considerably, thereby giving a high accuracy rate for perceptual dissimilarity and intra speaker indication.

8. Future Work

As a next step to this project, and a part of our curriculum we would like to build a system to identify speakers using noisy conversational speech data. As mentioned in the discussion, we tried using GMM with M =4,8,16 but the efficiency was not good, due to small dataset. We would model the voice features using GMM with a large dataset and analyze the perceptual dissimilarity and intra speaker indication.In another attempt, we would like to analyze voiced and unvoiced segments of the speech separately and test on the same dataset to have a better estimate of the spectral features of the speaker examined in this project.

9. Acknowledgments

We would like to thank the UCLA Speech Processing and Auditory Perception Laboratory(SPAPL) for their constant support and help in this project. We would also like to thank Professor Abeer Alwan for her guidance in technical concepts and continuous motivation to complete the project in a timely manner.

10. References

[1] J. Kreiman, S.Park, P. Keating, and A. Alwan, "The relationship between acoustic and perceived intraspeaker variability in voice quality."

[2] C. Y. Espy-wilson, E. Manocha, and S. Vishnubhotla, "A new set of features for textindependent speaker identification," pp. 1475–1478, 2006.

[3] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models."

[4] F. Nolan, K. McDougall, and T.Hudson, "Some acoustic correlates of perceived (dis) similarity between same-accent voices," 2011.

[5] T. Mahboob, M. Khanum, M. Khiyal, and R. Bibi, "Speaker identification using gmm with mfcc," 2015.

[6] R. Fraile, J. Ignacio, and Godino-Llorentea, "Cepstral peak prominence: A comprehensive analysis," 2014.

[7] Y. Maryn, N. Roy, M. DeBodt, P. V. Cauwenberge, and P. Corthals, "Acoustic measurement of overall voice quality: A meta-analysis," pp. 2619–2634, 2009.

[8] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," pp. 769–778, 1994.

[9] Y. D. Heman-Ackah, "Reliability of calculating the cepstral peak without linear regression analysis," pp. 203–208, 2004.

[10] P. J. Murphy, "On first rahmonic amplitude in the analysis of synthesized aperiodic voice signals," pp. 2896–2907, 2006.

[11] N. Washani and S. Sharma, "Speech recognition system: A review," 2015.

[12] L. Rabiner and R. Schafer, "Digital processing of speech signals."

[13] J. Edlund and M. Heldner, "Speaker classification ii," C. Müller, Ed. Berlin, Heidelberg: Springer-Verlag, 2007, ch. Underpinning /Nailon/: Automatic Estimation of Pitch Range and Speaker Relative Pitch, pp. 229–242.

[14] D. Hartl, O. Laccourreye, J. Vaissiere, and D. Brasnu, "Acoustic analysis of autologous fat injection versus thyroplasty in the same patient," pp. 987–992, 2003.

[15] T. Haderlein, C. Moers, B. Mobius, F. Rosanowski, and E. Noth, "Intelligibility rating with automatic speech recognition, prosodic, and cepstral evaluation," Berlin Heidelberg, pp. 195–202, 2011.

[16] T.F.Yap, J.Epps, and E.Ambikairajah, "Voice source features for cognitive load classification," Prague, pp. 5700–5703, 2011.

[17] S. N. Awan and N. Roy, "Toward the development of an objective index of dysphonia severity: A four-factor acoustic model," 2006.

[18] E. Ramos, "Physiology of larynx."