214A Digital Speech Processing Final Project Report

Yuhang Li[1], *Yuheng He*[2], *Zilin Zeng*[3]

[1]yuhangli@g.ucla.edu, [2]yuheng98@g.ucla.edu, [3]zzeng12@g.ucla.edu

Abstract

In American English, there are multifarious accents, and one of the outstanding ones is African American English (AAE). Within African American English, there exist multiple regional accents. In this project, we explore the automatic classification of 5 regional AAE accents. Our goal is to sift out the most representing features of the utterances and use these features to train an XGBoost classifier. We investigated several acoustic features including Mel-frequency Cepstral Coefficients (MFCC), Linear Prediction Coding (LPC), and Power Normalized Cepstral Coefficients (PNCC). These acoustic features are evaluated both separately and collectively, in order for better classification performance. To ensure robustness, we test the classifier both on clean audio data and noisy ones. The features we selected show a strong representation of the regional accents and deliver a robust outcome.

Index Terms: Regional Accent Identification, Speaker Identification, Digital Speech Feature Processing

1. Introduction

As English becomes the lingua franca and social interactions continue to increase, the acquisition and processing of accented English language and the ability to understand speech are becoming increasingly vital. There are many aspects of speech that can provide information about a particular speaker's characteristics [1]. The accent is a linguistic trait that can provide insights into a speaker's identity based on the usage of language and it is a specific pattern of pronunciation of people belonging to a particular region or geographical location [2]. It also refers to different ways of pronouncing a language within a particular community [3]. Using only a brief speech sample, humans can determine various characteristics about an individual, such as their gender, age, native language, emotional or attentional state, and social background. There is widely available research focusing on the phonetic variation and specific accents of the English language. State-of-the-art speaker classification based on the acoustic-phonetic differences employs spectral features like Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Cepstral Coefficients (LPCC), and Power Normalized Cepstral Coefficients (PNCC) [4].

In this project, we are tasked to distinguish 5 regional accents from the Corpus of Regional African American Language (CORAAL) [5] dataset. Accent recognition is significant in speech processing tasks because it can alleviate misclassification and ensure fair-training of speech-language technologies [6]. The utterances are extracted from interview interactions with native English speakers from 5 different regions. As a relatively new dataset, modest research has been conducted. Alexander Johnson et al. focus on the dialect density estimation of the CORAAL dataset [6]. They have performed acoustic extraction from the data and have successfully estimated the dialect density. However, such an estimation also relies on the text transcript of the utterances. Accent classification, on the other hand, depends purely on acoustic features. As a result, in the accent classification task, we are expected to neglect grammar and within-language diction. It is also essential to test the classifier on noisy data in that in realistic implementation, the noisy background is the norm, and the extracted features should be robust enough to neglect the noise.

In our implementation, we experiment with the dataset using different feature sets including MFCC, LPC, and PNCC. In this work, we first inspect the dataset and analyze its properties (Sec. 2.1). We then extract features from the dataset and utilize a combination of the acoustic features to perform accent classification. The performance is evaluated using both clean test data and noisy test data to ensure system robustness (Sec. 2.2). We follow our results with a conclusion (Se. 3) and propose possible methods to enhance the performance (Sec. 4).

2. Project Description

2.1. Data

This project focuses on the speech samples from the Corpus of Regional African American Language (CORAAL) [5], which features speech recordings from regional varieties of African American English (or African American language). This dataset contains equal male and female identifying participants and the participants range from 20 to 80 of age [6]. The dataset provided is sampled at 44.1kHz in frequency. We are tasked to perform feature extraction from data in the following five US cities: speakers from Washington DC (DCB), speakers from Princeville, NC (PRV), speakers from Rochester, NY (ROC), speakers from Lower East Side Manhattan, NY (LES), and speakers from Valdosta, GA (VLD). This dataset includes audio files with clear utterances and low noise backgrounds and audio files with murmured utterances and noisy backgrounds. We extract the label from the file name and perform classification on both clear test data and noisy test data.

2.2. Feature Extraction

The general methodology of audio classification depends heavily on extracting discriminatory features from the audio waveform and feeding these features to a classifier [7]. Audio data presents both in the time domain and transformation domain and using different approaches to extract the features provides varying performances. Several audio characteristics, such as MFCC, LPC, and PNCC, have been effectively utilized for audio classification. In this project, we are tasked to explore

regional accent classification performance using different features. We experiment on the aforementioned feature sets separately and collectively, and we also sought to find a feature combination that provides robust classification results.

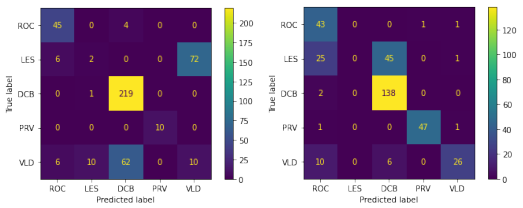
2.2.1. *MFCC*

The extraction and selection of the best parametric representation of acoustic signals are important tasks in the design of any speech recognition system. This is a critical process that has a significant impact on the performance of the speech recognition tasks [8]. A compact representation would be provided by a set of Mel-frequency Cepstral Coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale [7]. MFCC has 39 parameters and in this project while we select 13 parameters from the coefficients set. The first 12 parameters are related to the amplitude of frequencies which provides us with enough frequency channel to analyze the audio data. The thirteenth parameter is the energy in each frame that helps us to identify phones. The classification-based performance of MFCC features on clean test audio data and noisy test audio data is presented in Table 1. We also plot out the importance of different parameters of MFCC and confusion matrices for both clean test and noisy test in Fig. 1. It is noticeable in Fig. 1c that certain features play a more significant role in the classification task. In the MFCC coefficient set, feature 1, 3, 12, 4, 7 contribute more towards the accent classification. As shown in Fig. 1a and 1b, the DCB accent has higher classification accuracy as a result of more available training data. The LES accent gets misclassified severely due to the unsuccessful feature extraction of this audio dataset.

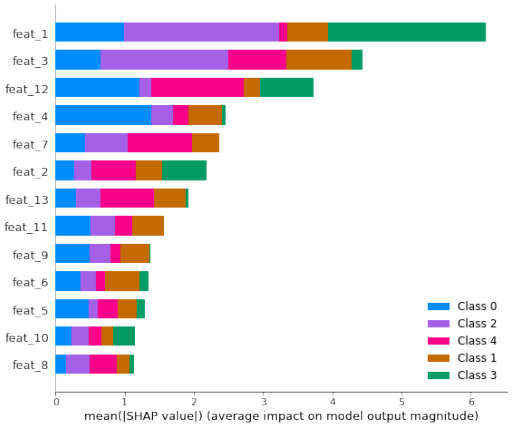
The result we obtain using MFCC feature library in `tochaudio` is different from the one provided by the class, which utilizes MFCC feature library in `librosa` even though both libraries select the first 13 features of MFCC. These feature differences are likely to happen on the mel-spectrogram level as the two libraries use different calculation approaches. We achieve a relatively high noisy test audio classification accuracy but a lousy rate for the clean data. In order to increase the accuracy rate for the clean test dataset, we need a more robust and effective feature extraction approach.

Table 1: *MFCC Features Classification*

Dataset	Classification Accuracy
Clean Train	100%
Clean Test	63.98%
Noisy Test	73.20%



(a) *Clean Confusion Matrix.* (b) *Noisy Confusion Matrix.*



(c) *Most Important MFCC Features.*

Figure 1: *MFCC Results.*

2.2.2. *LPC, PNCC, Prosody, and Phonation*

Linear Prediction Coding (LPC) models the system output as a linear combination of past outputs and present input. It is the process of predicting a signal sample based on the past p samples [9]. Viewing the LPC analysis as an AR model (All-pole model) in the frequency domain serves as a powerful tool for estimating the filter function (i.e.vocal tract envelope spectrum, formant frequencies) for a speech signal [10]. The order of LPC (p) is the number of past samples considered for prediction, which is equivalent to the number of LPC coefficients (filter poles) returned from the analysis. Each of the p coefficients would be tantamount to one of the first p formants (peaks) showing up on the signal's frequency spectrum. Empirically, the LPC order for a speech signal ranges from 15 to 20. The classification based on the performance of order-20 LPC features on clean test audio data and noisy test audio data is presented in Table 2.

Table 2: *order-20 LPC Features Classification*

Dataset	Classification Accuracy
Clean Train	100%
Clean Test	51.67%
Noisy Test	67.43%

We realize that the LPC feature is low in accuracy for both clean test audio data and noisy test audio data compared to MFCC. To improve the overall performance of the LPC feature, we increase its order from 20 to 48 by referring to the formula [10]:

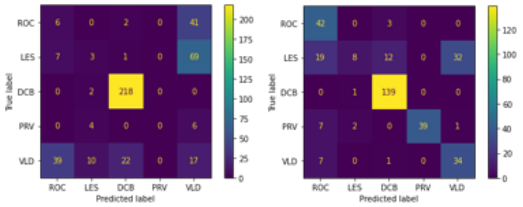
$$p = 2F_{max}\ (\text{in kHz}) + [2 - 4]$$

Moreover, instead of doing the LPC analysis on a full-length audio signal, we calculate the LPC for every 40-second window frame and average over the LPC coefficient arrays to retrieve an averaged LPC feature. The aforementioned modifications boost accuracy for both clean and noisy test sets as indicated by Table 3.

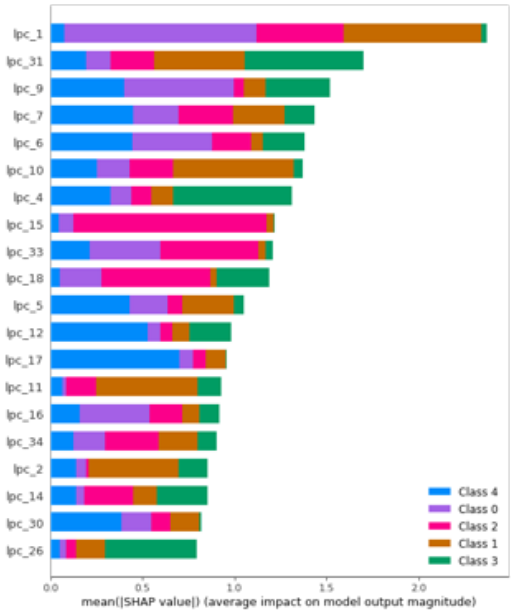
Making limited progress, the averaged LPC remains an inferior feature set for classifying clean test audio data. Neverthe-

Table 3: *order-48 averaged LPC Features Classification*

Dataset	Classification Accuracy
Clean Train	100%
Clean Test	54.59%
Noisy Test	75.50%



(a) *Clean Confusion Matrix.* (b) *Noisy Confusion Matrix.*



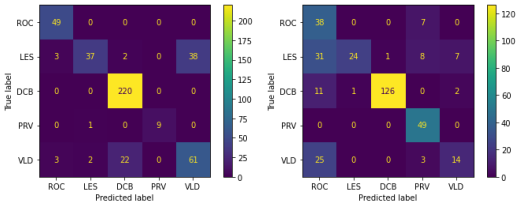
(c) *Most Important order-48 averaged LPC Features.*

Figure 2: *LPC Results.*

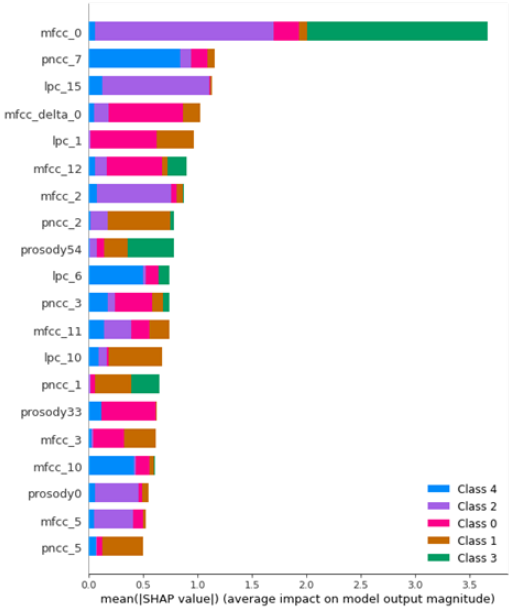
less, just like `torchaudio` MFCC, the averaged LPC manifests a robust performance for noisy test audio data. Looking into the shap explainer (Fig. 2a), we also notice that higher-ordered coefficients have considerable impacts on speaker predictions, which explains why higher-ordered LPC analysis gains better accuracy for the noisy test set. By analyzing the confusion matrices (Fig. 2b and Fig. 2c), we realize accents from ROC, LES, and VLD can be hard to discern using merely LPC features. Attempting to get better accuracy for these regions, we choose three additional feature sets: Prosody [11] [12], Phonation [13], and Power Normalized Cepstral Coefficients (PNCC) [14]. We used `spafe` [15] for PNCC feature extractions, and we installed `DisVoice` [16] for extracting the prosody features and the phonation features. PNCC is very similar to MFCC except that it uses a power-law nonlinearity rather than a log nonlinearity and deploys a Gammatone Filter Bank rather than a Mel Filter bank. The most important difference comes from that PNCC adopts a noise-suppression algorithm based on asymmetric filtering that suppresses background excitation. The prosody features represent the pitch contour and energy contour for voiced speech segments, whereas the phonation features indicate jitters and shimmers over the voiced segments as well as the derivatives of fundamental frequencies. We expect to further fortify our feature set's noise resilience with PNCC and mitigate the deficiency on the clean test set with features pertaining to the fundamental frequencies. The classification accuracies for noisy test sets remain steady, yet we achieve a more reliable classification for the clean test set as Table 4 denotes.

Table 4: *MFCC-Prosody-PNCC-LPC-Phonation*

Dataset	Classification Accuracy
Clean Train	100%
Clean Test	84.12%
Noisy Test	72.33%



(a) *Clean Confusion Matrix.* (b) *Noisy Confusion Matrix.*



(c) *Most Important features among the MFCC-Prosody-PNCC-LPC-Phonation Feature Set.*

Figure 3: *MFCC-Prosody-PNCC-LPC-Phonation Results.*

2.2.3. Scale is all you need

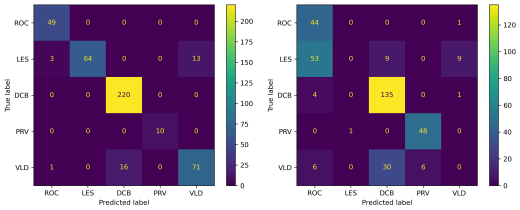
Theoretically speaking, more features can represent the data better, and help increase the approximation capability of a classifier. To validate whether more features can improve the classifier performance in African American English Classification task, inspired by [6], we use the `OpenSmile` toolkit [17] to

extract `CompParE16` features [18] from an audio file. Here we use the Low-Level Descriptors (LLD) to extract 65 features, which include several feature groups such as pitch, FFT spectrum, signal energy, cepstral, and so on. It is worth mentioning that the `OpenSmile` toolkit can also extract Functionals level features with a size of 6373. Limited by time and available memory, we choose the LLD, whose number of features is still much larger than single features like MFCC. The test classification accuracies of the xgboost classifier trained based on `OpenSmile` LLD features are listed in Table 5. Compared to the results using MFCC only, the classification accuracy on the clean test set increases a lot while an accuracy drop appears on the noisy test set. To shed more light on the contribution of each feature, We also plot a beeswarm figure to explain the output of the model in Fig. 4c. It's obvious that MFCC-SMA features, which refer to MFCC features of the signal processed by simple moving average (SMA) low-pass filtering, play a significant role in the classifier. Beyond that, the magnitude spectrum field (PCM-FFTMag) also matters in the trained classifier, which may also help increase the classification accuracy on the clean test set. The improved classification accuracy on the clean test set demonstrates the benefit of using more features in this task. Although the classification accuracy is not as good as using the MFCC feature only reported above, the accuracy is still higher than the results of using many other single features like LFCC or LPC.

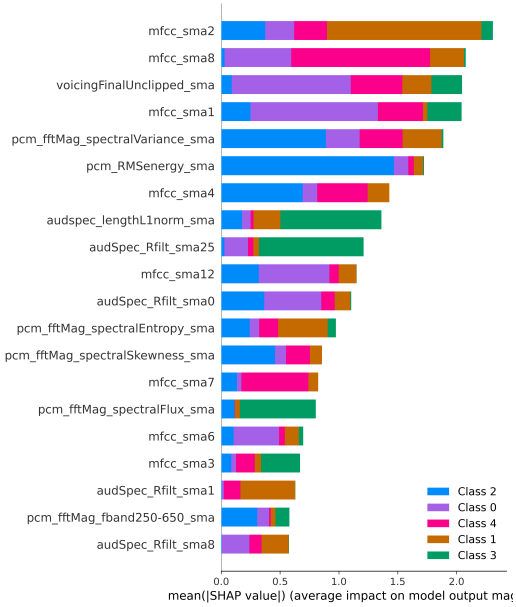
Table 5: *ComParE16 Features Classification Accuracy*

Dataset	Classification Accuracy
Clean Train	100 %
Clean Test	92.62%
Noisy Test	65.41%

The confusion matrices of the clean test set and the noisy test set are also plotted in Fig. 4a and 4c. We notice that on the clean test set, the LES samples are sometimes misclassified to the ROC or the VLD while all the wrong classification results of the VLD are the DCB. Therefore, we hypothesize that the ROC accent is similar to the VLD and the DCB, and the VLD resembles the DCB. This phenomenon is more obvious in the classification results of the noisy test set that the LES are mostly classified as the ROC while all the VLD are regarded as the DCB. Intuitively, the noise would affect the feature extraction negatively, making it more difficult to distinguish two similar accents as the key features may be lost. We conjecture that the imbalance of the training dataset causes the misclassification of the VLD to the DCB as the DCB samples dominate the dataset. To prove this guess, we randomly remove 2300 DCB samples and keep 157 DCB samples, which are fewer than the VLD samples with a size of 567. Based on the clipped new dataset, we retrain the classifier and plot the confusion matrices of the clean test set and the noisy test set in Fig. 5. We can find that most clean DCB samples are misclassified to the VLD as the new dataset has more VLD samples right now. In the noisy test set, the VLD, the DCB and the VLD are likely to be classified as the ROC which had the most samples in the new dataset. These observations further confirm our guess that the imbalance of the training dataset will cause the classification tendency of the model.

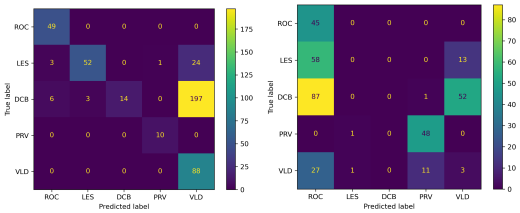


(a) *Clean Confusion Matrix.* (b) *Noisy Confusion Matrix.*



(c) *Most Important LLD Features.*

Figure 4: *ComParE16 LLD Results.*



(a) *Clean Confusion Matrix.* (b) *Noisy Confusion Matrix.*

Figure 5: *Clipped Dataset Results.*

2.3. Attempts to increase the accuracy on noisy test data

With 65 `CompParE16` features, although the classification accuracy on the clean test set has reached over 90%, the accuracy on the noisy test set is still not good enough. Aiming to improve the performance of the classifier on the noisy samples, we look for features that are stable to the noise or the denoise methods. We notice that our MFCC method reported before achieves 73.2% accuracy on the noisy test set, so we concatenate it with the `CompParE16` features. Though `CompParE16` features already included the MFCC features, the feature values may vary due to the different choice of the parameters, as we see the difference between MFCC value extracted from `torchauio` and `librosa`. However, the combined features do not help improve the performance but instead smear the per-

formance on both clean and noisy test sets, as shown in Table 6. Aside from selecting the features carefully, we also try the denoise algorithm using the `noisereduce` package and custom `Audio-Denoising` [19]. However, the accuracy is even worse.

Table 6: *MFCC + ComParE16 Features Classification Accuracy*

Dataset	Classification Accuracy
Clean Train	100 %
Clean Test	82.32%
Noisy Test	64.84%

Data augmentation is commonly used in machine learning to enrich the amount and diversity of the existing training data and help improve model accuracy. Therefore, we want to inject the noise into the original audio segments to see whether it can help improve the performance. Although there are lots of datasets including all kinds of noise, they overkill the classification task here. Without loss of generality, we randomly select a certain number of noisy test data and added them to the clean train dataset. And then we test the trained classifier on the clean train set and the remaining noisy test set. We plot the accuracy curves of the test clean set and test noisy set including both average and standard deviation values based on 30 times repetitions, as shown in Fig 6. We can see the accuracy on the noisy test set can be improved a lot to 80% even though we solely include 10 additional noisy samples in the training. As we further increase the number of noisy samples used for the training, the classification accuracy on the test noisy set surpasses the one on the test clean set, not to mention that the number of noisy samples used for the training is much smaller than the number of clean training samples. The performance on the clean test set degrades merely slightly. This indicates the effectiveness of including noisy samples in the training process.

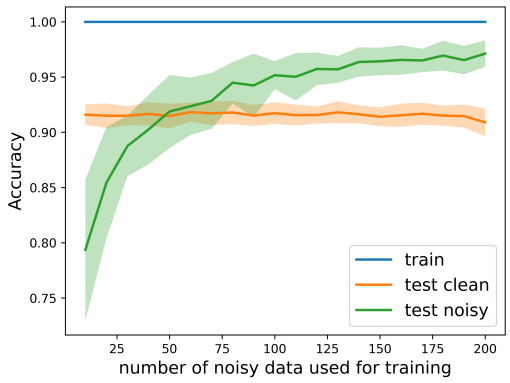


Figure 6: *Impact of adding noisy samples into the training dataset*

3. Conclusion: Best model achieved

To further improve the classifier performance on both clean and noisy test sets, we concatenate `Librosa-MFCC`, `torchaudio-MFCC`, `PNCC`, `LPC-40s window`, `Prosody`, `Phonation` and `ComParE` features. We notice the order of features matters and the final performance varies in the range of 3%. We shuffle the order of features and use the shuffled features to train the classifier. We repeat 40 times to get the best model, whose performance on listed in Table 7. Although the accuracy on the clean test set or noisy test set is not the best compared to previous results, which may be due to the interaction between different features, this model achieves the best performance balance on the two test sets.

Table 7: *Final Model Classification Accuracy*

Dataset	Classification Accuracy
Clean Train	100 %
Clean Test	91.28%
Noisy Test	72.91%

To prove the reliability of our feature set concatenated, we test the trained classifier using a hidden clean set and a hidden clean set, which are never used during the training and validation stage. The new result shows as follows in Table 8:

Table 8: *Final Model Classification Accuracy on Hidden Test Set*

Dataset	Classification Accuracy
Clean Test Hidden	87.98%
Noisy Test Hidden	74.83%

The classification accuracy looks decent and relatively steady on both testing sets. We manage to achieve a classification accuracy of approximately 90 percent on clean testing audio and a classification accuracy of above 70 percent on noisy testing audio.

4. Future Work

We believe that we can achieve better classification accuracy if we were given more freedom for the datasets. We have proven that including a moderate amount of noisy data in the training induces a promising accuracy for the noisy test set. We think the classification can be more reasonable if we are capable of using different sets of features for classification depending on the Signal to Noise Ratio of the signal. Features have trade-offs, thus it is almost impossible to come up with a set of features that perform well for testing under both clean and noisy situations. The classification can be even more accurate if we tailor the feature set applied to the classifier depending on the noisy level of the testing signal.

5. References

[1] F. William, A. Sangwan, and J. H. Hansen, "Using human perception for automatic accent assessment," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[2] D. Crystal, *A dictionary of linguistics and phonetics*. John Wiley & Sons, 2011.

[3] H. Behravan, V. Hautamäki, and T. Kinnunen, "Foreign accent detection from spoken finnish using i-vectors." in *Interspeech*, vol. 2013, 2013, p. 14th.

[4] N. Zheng, T. Lee, and P.-C. Ching, "Integration of complementary acoustic features for speaker recognition," *IEEE Signal Processing Letters*, vol. 14, no. 3, pp. 181–184, 2007.

[5] T. Kendall and C. Farrington., "The corpus of regional african american language." 2021, the Online Resources for African American Language Project. [Online]. Available: http://oraal.uoregon.edu/coraal.

[6] A. Johnson, K. Everson, V. Ravi, A. Gladney, M. Ostendorf, and A. Alwan, "Automatic dialect density estimation for african american english," *arXiv preprint arXiv:2204.00967*, 2022.

[7] V. Tiwari, "Mfcc and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[8] F. Zheng, G. Zhang, and Z. Song, "Comparison of different implementations of mfcc," *Journal of Computer science and Technology*, vol. 16, pp. 582–589, 2001.

[9] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[10] L. R. Rabiner and R. W. Schafer, *Theory and applications of Digital Speech Processing*. Pearson Higher Education, Inc, 2011, p. 473–504.

[11] N. Dehak, P. Dumouchel, and P. Kenny, "Modeling prosodic features with joint factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2095–2103, 2007.

[12] J. Vasquez, J. R. Orozco, T. Bocklet, and E. Noeth, "Towards an automatic evaluation of the dysarthria level of patients with parkinson's disease," *Journal of Communication Disorders*, vol. 76, pp. 21–36, 08 2018.

[13] J. Vásquez-Correa, P. Klumpp, J. R. Orozco-Arroyave, and E. Nöth, "Phonet: A Tool Based on Gated Recurrent Neural Networks to Extract Phonological Posteriors from Speech," in *Proc. Interspeech 2019*, 2019, pp. 549–553.

[14] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (pncc) for robust speech recognition," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4101–4104.

[15] A. Malek, "Spafe: Simplified python audio features extraction," *Journal of Open Source Software*, vol. 8, p. 4739, 01 2023.

[16] C. Vasquez, "Disvoice," https://github.com/jcvasquezc/DisVoice.

[17] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[18] B. Schuller, S. Steidl, A. Batliner, J. Hirschberg, J. K. Burgoon, A. Baird, A. Elkins, Y. Zhang, E. Coutinho, and K. Evanini, "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language," in *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*, vol. 8. ISCA, 2016, pp. 2001–2005.

[19] A. Patare, "Audio-denoising," https://github.com/AP-Atul/Audio-Denoising.