Javanmardi, Farhad; Kadiri, Sudarsana; Kodali, Manila; Alku, Paavo

Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers

Published in:
INTERSPEECH 2022

DOI:
10.21437/Interspeech.2022-10420

Published: 01/09/2022

Document Version
Publisher's PDF, also known as Version of record

Interspeech 2022
18-22 September 2022, Incheon, Korea

Comparing 1-dimensional and 2-dimensional spectral feature representations in voice pathology detection using machine learning and deep learning classifiers

Farhad Javanmardi, Sudarsana Reddy Kadiri, Manila Kodali, Paavo Alku

Department of Signal Processing and Acoustics, Aalto University, Finland
firstname.lastname@aalto.fi

Abstract

The present study investigates the use of 1-dimensional (1-D) and 2-dimensional (2-D) spectral feature representations in voice pathology detection with several classical machine learning (ML) and recent deep learning (DL) classifiers. Four popularly used spectral feature representations (static mel-frequency cepstral coefficients (MFCCs), dynamic MFCCs, spectrogram and mel-spectrogram) are derived in both the 1-D and 2-D form from voice signals. Three widely used ML classifiers (support vector machine (SVM), random forest (RF) and Adaboost) and three DL classifiers (deep neural network (DNN), long short-term memory (LSTM) network, and convolutional neural network (CNN)) are used with the 1-D feature representations. In addition, CNN classifiers are built using the 2-D feature representations. The popularly used HUPA database is considered in the pathology detection experiments. Experimental results revealed that using the CNN classifier with the 2-D feature representations yielded better accuracy compared to using the ML and DL classifiers with the 1-D feature representations. The best performance was achieved using the 2-D CNN classifier based on dynamic MFCCs that showed a detection accuracy of 81%.

Index Terms: Voice pathology, Spectral features, Deep learning, CNNs, Mel-spectrogram

1. Introduction

Pathological voice may arise due to different factors such as diseases, infections as well as physiological and psychogenic problems that all impact the human voice production mechanism [1, 2]. Dysphonia, dysarthria and dyplophonia are examples of voice pathologies. The automatic detection of voice pathology refers to a binary classification task aiming at distinguishing pathological voices from healthy voices using machine learning (ML) or deep learning (DL). Clinicians can use the automatic detection of pathological voices as a potential medical tool for early diagnosis and treatment [3, 4]. The automatic voice pathology detection is beneficial because it enables an objective voice pathology assessment, which can be computed directly from the voice waveform. Hence, the detection can be conducted non-invasively and remotely away from hospital and therefore patients can avoid repeated visits to clinics for medical examinations [5].

Voice pathology detection has been subject to a significant progress in the past decade due to developments in signal processing as well as in ML and DL. Most of the existing studies have used a pipeline system consisting of two main stages (see Figure 1). The first stage is the feature extraction phase in which acoustic voice waveforms are represented in compressed forms using features. The second stage includes the classifier which predicts the output label (i.e., healthy vs. pathological) from the input features using a ML- or DL-based model. Several studies have investigated various feature extraction methods based on modeling the two main parts of voice production, the voice excitation and the vocal tract. In extracting voice excitation features, previous studies have investigated, for example, time-frequency aspects of the glottal source [6, 7, 8], prosody [9], and voice quality (represented by the harmonic-to-noise ratio [10], shimmer and jitter [11]). In computing vocal tract features, previous investigations have taken advantage of methods such as linear predictive cepstral coefficients (LPCCs) [12], perceptual linear prediction (PLP) [13] and mel-frequency cepstral coefficients (MFCCs) [14]. Regarding the classifier stage, several studies have explored conventional ML classifiers such as support vector machine (SVM) [4, 15, 16, 17], random forest (RF) [18] and decision trees [16, 19]. Due to recent advancements in deep learning, classical ML methods have been increasingly replaced by DL networks such as multilayer perceptron (MLP) [20], deep neural networks (DNNs) [21, 22], long short-term memory (LSTM) networks [23, 24], convolutional neural networks (CNNs) [25], combinations of CNN and MLP [4], and combinations of CNN and LSTM [26]. However, DL classifiers are known to be data hungry [27] and therefore they do not necessarily work well in detection tasks where the amount of training data is small.

Most of the previous voice pathology detection studies (including those referred to in the previous paragraph) have used 1-dimensional (1-D) features as input to ML and DL classifiers. For a given voice utterance, this corresponds to first extracting features frame-wise (e.g. in 20-ms frames), then merging the frame-wise features with statistical functionals (e.g., mean, max etc.) into a long 1-D feature vector to parameterize the entire utterance. Recently a few voice pathology detection studies have also investigated the usage of 2-dimensional (2-D) feature representations such as spectrograms and mel-spectrograms with CNN classifiers [10, 28, 29, 30, 31]. These 2-D features are "spectral images", which are computed by concatenating frame-wise features into 2-D feature matrices where the y-axis represents spectral amplitude in decibel (dB) and the x-axis represents time (in frames).

In this study, we compare 1-D and 2-D spectral feature representations in the detection of voice pathology using different ML and DL classifiers. More specifically, we explore four popular spectral feature representations (static MFCCs, dynamic MFCCs, spectrogram, and mel-spectrogram) by expressing them in the 1-D form (i.e., merging frame-wise features of each utterance with statistical functionals into a single feature vector per utterance) and in the 2-D form (i.e., concatenating all frame-wise features of each utterance into a feature matrix, that is, forming a "spectro-temporal image"). As classifiers, three

10.21437/Interspeech.2022-10420

ML classifiers (SVM, RF, Adaboost) and three DL classifiers (DNN, LSTM, and CNN) are investigated using the four spectral feature representations in the 1-D form. In addition, CNN classifiers are built using the 2-D form for the four spectral feature representations. To the best of our knowledge, this is the first systematic comparison between 1-D and 2-D spectral feature representations involving classifiers from the two classifiers camps (ML vs. DL).

The rest of the paper is organized as follows. In Section 2, a brief explanation of the proposed framework, feature extraction and the classifiers is provided. Section 3 describes the details of the experimental setup containing the database, training, validation and testing data used for the detection experiments, and evaluation metrics. In Section 4, the results and discussion of the detection experiments are presented. Finally, the summery of the paper is provided in Section 5.

2. Methods

Voice pathology detection systems are built using the popular two-stage architecture shown in Figure 1. In the feature extraction stage, four spectral feature representations are computed from the input voice signal. In the classifier stage, three ML and three DL classifiers based on the use of 1-D feature representations are built. In addition, one DL classifier utilising a 2-D spectral feature representation is built for the (binary) detection task.

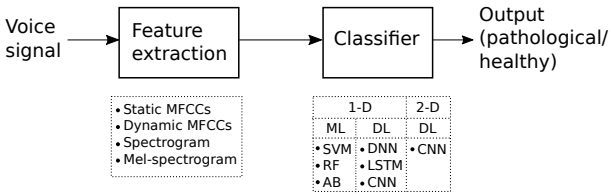

Figure 1: *The proposed framework for voice pathology detection.*

2.1. Feature Extraction

The feature extraction is computed in the current study by using four commonly used spectral features: MFCCs without and with the first and second derivatives (i.e., static MFCCs and dynamic MFCCs, respectively), spectrogram and mel-spectrogram. The computation of these spectral features will be described next. In order to keep the treatment straightforward, we start from spectrogram. The voice spectrum is computed with the short-time Fourier transform (STFT) in frames of 25-ms with a shift of 5 ms using the Hamming window. The spectrogram is computed by taking the logarithm of the amplitude spectrum. For the mel-spectrogram, a mel-filterbank with 80 filters is applied to convert the magnitude spectrum to the mel scale, and finally the resulting mel-spectrogram is mapped to a decibel scale through logarithm. To compute the MFCCs, the mel-scale of the spectrum is transformed via the discrete cosine transform (DCT) to compute mel-cepstral coefficients. From the entire mel-cepstrum, the first 13 coefficients (including the 0^{th} coefficient) are considered. The resulting cepstral coefficients are referred to as the static MFCCs. The first and second derivatives are also computed from the static MFCCs, and they are referred to as the dynamic MFCCs. All these feature representations are derived using the Librosa toolkit [32].

2.2. Classifier

Both ML- and DL-based classifiers are used for the detection of pathological voice. In the former group, the following three ML classifiers are used: SVM, RF and Adaboost. For the latter group, the following three DL classifiers are used: DNN, LSTM and CNN. Using all these six individual classifiers, pathology detection systems are built based on the 1-D spectral feature representations. To build systems using the 2-D feature representation, only the CNN classifier is considered because it is the only classifier architecture that can use 2-D features.

The three selected ML classifiers are popularly used to detect pathological voice and they have shown promising results [33]. In the current study, SVM with a radial basis function kernel and a regularization parameter of 1 is used. The number of trees used in RF is 100. In Adaboost, 50 estimators are used. All the ML classifiers were implemented using the Scikit-learn library [34].

The DL classifiers were built as follows:

- DNN: This classifier consists of three fully connected (dense) layers (D1, D2, D3) followed by the rectified linear unit (ReLU) activation function. The output dense layer (D3) performs the binary classification with the sigmoid activation function.
- LSTM: In this classifier, the input is passed through the LSTM layer and then the returned sequence output is fed to the first dense layer with the ReLU activation function. Finally the second dense layer with the sigmoid activation function predicts the output label.
- 1-D CNN: The input is fed to the two sequential convolutional layers, each followed by batch normalization and the ReLU activation function. The output is flattened and then delivered to a dense layer with the ReLU activation function. The resulting output is passed through the second dense layer with the sigmoid activation function to perform the binary classification task.
- 2-D CNN: In this classifier, two sequential convolutional layers are used and both layers are followed by batch normalization, the ReLU activation function and the max-pooling layer. The input is passed through these layers and then the output is flattened and delivered to the next two dense layers with the ReLU activation function. Dropout of 0.25 is used after the first dense layer. Finally, the resulting output is passed through the third dense layer with the sigmoid activation function to perform the binary classification. The size of the 2-D feature representation is fixed to 3 sec.

In the DL classifiers, the following hyper-parameters are used: a batch size of 32, 100 epochs with 20 as the early stopping, the Adam optimizer with a learning rate of 0.001, and the binary cross entropy-based error criterion. All the DL classifiers are implemented using the Pytorch library [35].

3. Experimental Setup

This section describes the voice pathology database, training, validation and testing data used for detection experiments, and the evaluation metrics.

3.1. Database

In this study, a dysphonia database known as Hospital Universitario Príncipe de Asturias (HUPA) is used [14, 36]. The

database consists of sustained phonations of the Spanish vowel /a/ by 200 patients (74 males, 126 females) and 239 healthy controls (101 males, 138 females). The pathological samples represent 15 different organic pathologies such as nodules, polyps, and Reinke's edema. The current binary detection study was carried out using balanced classes of 200 healthy and 200 pathological vowels of HUPA. Both classes consisted of vowels produced by 74 male and 126 female speakers.

3.2. Training, validation and testing

For the ML classifiers, detection experiments were conducted with stratified 10-fold cross-validation, where the data was split randomly into 10 equal portions (10 folds). Each fold consists of the same number of healthy and pathological voice samples. In each iteration, one fold was considered as testing data and the remaining nine folds were used for training the classifier. Both training and testing data were z-score normalized using the mean and standard deviation of the training data. The evaluation metrics were computed and saved in each iteration. This process was repeated for 10 iterations, and finally the evaluation metrics were computed by taking the average over the 10 iterations. The detection experiments with the DL classifiers were carried out in a similar manner as with the ML classifiers except that in the former, 10% of the training data was selected randomly as validation data in each iteration.

3.3. Evaluation metrics

The following five commonly used evaluation metrics are considered in this study [8, 37]: accuracy (ACC), sensitivity (SE), specificity (SP), F1-score (F1) and equal error rate (EER). For the first four metrics, a higher value indicates better performance. For EER, a lower value indicates better performance.

4. Results

The detection results given by the different ML and DL classifiers are reported in Table 1 for all the five metrics selected. We will next report our main observations by first discussing the results given by the classifiers that used the 1-D feature representations and then discussing the results given the 2-D CNN classifier.

From the results obtained using the 1-D spectral feature representations, it can be seen that the SVM classifier based on mel-spectrogram showed the highest detection accuracy (75%) among the ML classifiers. Among the DL classifiers, the 1-D CNN classifier based on spectrogram showed the highest accuracy (75.75%). The results also show that the differences in accuracy between the classifiers using the four 1-D spectral feature representations were smaller for the ML systems compared to the DL systems.

From the results obtained using the 2-D spectral feature representations (i.e., with the 2-D CNN), it can be seen that the MFCC-based features (both static and dynamic) performed better than the two spectrogram-based feature representations. The best performance was obtained using dynamic MFCCs that yielded an accuracy of 81%. Compared to the two best performing classifiers (in ML and DL) that use 1-D spectral feature representations, i.e., 1-D CNN using spectrogram and SVM using mel-spectrogram, 2-D CNN using dynamic MFCCs gave an absolute improvement in accuracy that was 5.25% and 6%, respectively. It is also worth noting that the 2-D CNN classifier based on dynamic MFCCs gave the best performance in all the metrics (except for sensitivity) when all the classifiers and spectral representations are compared. We argue that the improved performance that was obtained by using dynamic MFCCs as the 2-D feature representation with the CNN classifier is due to the capability of dynamic MFCCs to effectively reflect temporal irregularities (e.g. jitter) that occur during phonation and during onsets and offsets in pathological voice.

Table 1: *The results of ML and DL classifiers with four 1-D and 2-D feature representations. ACC refers to accuracy, SE refers to sensitivity, SP refers to specificity , F1 refers to F1-score, and EER refers to equal error rate.*

Classifier	Feature	ACC	SE	SP	F1	EER
ML classifiers with the 1-D feature representations						
SVM	Static MFCCs	72.00	0.68	0.75	0.71	0.292
	Dynamic MFCCs	72.00	0.75	0.68	0.72	0.290
	Spectrogram	73.50	0.69	0.77	0.72	0.260
	Mel-spectrogram	75.00	0.71	0.78	0.74	0.250
RF	Static MFCCs	68.50	0.67	0.69	0.68	0.323
	Dynamic MFCCs	71.00	0.69	0.72	0.70	0.278
	Spectrogram	74.25	0.69	0.79	0.72	0.277
	Mel-spectrogram	74.00	0.71	0.77	0.72	0.264
Adaboost	Static MFCCs	64.00	0.61	0.67	0.62	0.355
	Dynamic MFCCs	63.50	0.65	0.61	0.64	0.365
	Spectrogram	68.75	0.68	0.69	0.68	0.320
	Mel-spectrogram	64.75	0.64	0.65	0.64	0.345
DL classifiers with the 1-D feature representations						
DNN	Static MFCCs	60.50	**0.81**	0.39	0.67	0.375
	Dynamic MFCCs	63.00	0.79	0.47	0.68	0.375
	Spectrogram	72.75	0.75	0.70	0.73	0.265
	Mel-spectrogram	69.50	0.64	0.75	0.67	0.310
LSTM	Static MFCCs	65.00	0.64	0.66	0.64	0.340
	Dynamic MFCCs	69.25	0.69	0.69	0.68	0.295
	Spectrogram	70.50	0.70	0.70	0.70	0.300
	Mel-spectrogram	69.25	0.60	0.78	0.66	0.295
1-D CNN	Static MFCCs	59.25	**0.81**	0.37	0.66	0.390
	Dynamic MFCCs	69.00	0.72	0.65	0.69	0.330
	Spectrogram	75.75	0.78	0.73	0.76	0.230
	Mel-spectrogram	74.50	0.70	0.78	0.73	0.230
DL classifier with the 2-D feature representations						
2-D CNN	Static MFCCs	78.00	0.80	0.75	0.78	0.215
	Dynamic MFCCs	**81.00**	0.77	**0.85**	**0.79**	**0.170**
	Spectrogram	77.75	0.80	0.75	0.78	0.220
	Mel-spectrogram	75.25	0.75	0.75	0.75	0.235

5. Conclusion

The current work is the first systematic comparison of 1-D and 2-D spectral feature representations with ML- and DL-based classifiers in voice pathology detection. The detection experiments were carried out with the HUPA database. The experimental results revealed that using 2-D spectral feature representations with the CNN classifier achieved better detection accuracy compared to using 1-D spectral feature representations with ML and DL classifiers. This finding indicates that 2-D spectral feature representations are spectro-temporal "images" that contain voice pathology cues (e.g., related to temporal irregularities) that can be effectively used by the CNN classifier. It was found that the CNN classifier using dynamic MFCCs as the 2-D representation performed best showing an accuracy of 81%. Contrary to what might have been expected, the CNN classifier did not suffer from the smallish amount of training data available in the study. This may be due to the fact that the data consisted of sustained phonations of a single vowel (the vowel /a/), which might have reduced the problem of data scarcity. Further investigations with pathological

databases representing continuous speech are needed to better understand how the 2-D CNN classifier based on the dynamic MFCC feature representation works in other pathology voice detection scenarios.

6. Acknowledgements

This work was supported by the Academy of Finland (grant number 313390). The computational resources were provided by Aalto ScienceIT.

7. References

[1] A. Aronson, *Clinical Voice Disorders; An Interdisciplinary Approach*. Thieme Inc, 1985.

[2] N. R. Williams, "Occupational groups at risk of voice disorders: a review of the literature," *Occupational Medicine*, vol. 53, no. 7, pp. 456–460, 2003.

[3] F. Amato, M. Cannataro, C. Cosentino, A. Garozzo, N. Lombardo, C. Manfredi, F. Montefusco, G. Tradigo, and P. Veltri, "Early detection of voice diseases via a web-based system," *Biomedical Signal Processing and Control*, vol. 4, no. 3, pp. 206–211, 2009.

[4] N. Narendra and P. Alku, "Glottal source information for pathological voice detection," *IEEE Access*, vol. 8, pp. 67 745–67 755, 2020.

[5] C. G. Goetz, G. T. Stebbins, D. Wolff, W. DeLeeuw, H. Bronte-Stewart, R. Elble, M. Hallett, J. Nutt, L. Ramig, T. Sanger *et al.*, "Testing objective measures of motor impairment in early parkinson's disease: Feasibility study of an at-home testing device," *Movement Disorders*, vol. 24, no. 4, pp. 551–556, 2009.

[6] A. A. Dibazar, S. Narayanan, and T. W. Berger, "Feature analysis for automatic detection of pathological speech," in *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, vol. 1, Oct 2002, pp. 182–183.

[7] F. Rudzicz, "Phonological features in discriminative classification of dysarthric speech," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4605–4608.

[8] S. R. Kadiri and P. Alku, "Analysis and detection of pathological voice using glottal source features," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 367–379, 2019.

[9] C. Manfredi, M. D'Aniello, P. Bruscaglioni, and A. Ismaelli, "A comparative analysis of fundamental frequency estimation methods with application to pathological voices," *Medical Engineering & Physics*, vol. 22, no. 2, pp. 135 – 147, 2000.

[10] H. Kim, J. Jeon, Y. J. Han, Y. Joo, J. Lee, S. Lee, and S. Im, "Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy," *Journal of Clinical Medicine*, vol. 9, no. 11, p. 3415, 2020.

[11] M. Brockmann, M. J. Drinnan, C. Storck, and P. N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *Journal of Voice*, vol. 25, no. 1, pp. 44 – 53, 2011.

[12] V. Parsa and D. G. Jamieson, "Identification of pathological voices using glottal noise measures," *Journal of Speech, Language, and Hearing Research*, vol. 43, no. 2, pp. 469–485, 2000.

[13] M. A. Little, D. A. Costello, and M. L. Harries, "Objective dysphonia quantification in vocal fold paralysis: Comparing nonlinear with classical measures," *Journal of Voice*, vol. 25, no. 1, pp. 21 – 31, 2011.

[14] J. D. Arias-Londoño, J. I. Godino-Llorente, M. Markaki, and Y. Stylianou, "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," *Logopedics Phoniatrics Vocology*, vol. 36, no. 2, pp. 60–69, 2011.

[15] M. Pishgar, F. Karim, S. Majumdar, and H. Darabi, "Pathological voice classification using mel-cepstrum vectors and support vector machine," *arXiv preprint arXiv:1812.07729*, 2018.

[16] Z. Fan, J. Qian, B. Sun, D. Wu, Y. Xu, and Z. Tao, "Modeling voice pathology detection using imbalanced learning," in *2020 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*. IEEE, 2020, pp. 330–334.

[17] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, and V. Saferis, "Categorizing normal and pathological voices: Automated and perceptual categorization," *Journal of Voice*, vol. 25, no. 6, pp. 700 – 708, 2011.

[18] Z. Dankovičová, D. Sovák, P. Drotár, and L. Vokorokos, "Machine learning approach to dysphonia detection," *Applied Sciences*, vol. 8, no. 10, p. 1927, 2018.

[19] W. Yuanbo, Z. Changwei, F. Ziqi, Z. Yihua, Z. Xiaojun, and T. Zhi, "Voice pathology detection and multi-classification using machine learning classifiers," in *2020 International Conference on Sensing, Measurement & Data Analytics in the era of Artificial Intelligence (ICSMD)*. IEEE, 2020, pp. 319–324.

[20] K. Ezzine and M. Frikha, "Investigation of glottal flow parameters for voice pathology detection on svd and meei databases," in *2018 4th International conference on advanced technologies for signal and image processing (ATSIP)*. IEEE, 2018, pp. 1–6.

[21] Z.-Y. Chuang, X.-T. Yu, J.-Y. Chen, Y.-T. Hsu, Z.-Z. Xu, C.-T. Wang, F.-C. Lin, and S.-H. Fang, "Dnn-based approach to detect and classify pathological voice," in *2018 IEEE international conference on big data (big data)*. IEEE, 2018, pp. 5238–5241.

[22] L. Chen and J. Chen, "Deep neural network for automatic classification of pathological voice signals," *Journal of Voice*, 2020.

[23] V. Gupta, "Voice disorder detection using long short term memory (lstm) model," *arXiv preprint arXiv:1812.01779*, 2018.

[24] S. A. Syed, M. Rashid, S. Hussain, and H. Zahid, "Comparative analysis of cnn and rnn for voice pathology detection," *BioMed Research International*, vol. 2021, 2021.

[25] M. K. Reddy and P. Alku, "A comparison of cepstral features in the detection of pathological voices by varying the input and filterbank of the cepstrum computation," *IEEE Access*, vol. 9, pp. 135 953–135 963, 2021.

[26] P. Harar, J. B. Alonso-Hernandezy, J. Mekyska, Z. Galaz, R. Burget, and Z. Smekal, "Voice pathology detection using deep learning: a preliminary study," in *2017 international conference and workshop on bioinspired intelligence (IWOBI)*. IEEE, 2017, pp. 1–4.

[27] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.

[28] H. Wu, J. Soraghan, A. Lowit, and G. Di Caterina, "A deep learning method for pathological voice detection using convolutional deep belief networks," *Interspeech 2018*, 2018.

[29] ——, "Convolutional neural networks for pathological voice detection," in *2018 40th annual international conference of the ieee engineering in medicine and biology society (EMBC)*. IEEE, 2018, pp. 1–4.

[30] H. Ding, Z. Gu, P. Dai, Z. Zhou, L. Wang, and X. Wu, "Deep connected attention (dca) resnet for robust voice pathology detection and classification," *Biomedical Signal Processing and Control*, vol. 70, p. 102973, 2021.

[31] T. Arias-Vergara, P. Klumpp, J. C. Vasquez-Correa, E. Noeth, J. R. Orozco-Arroyave, and M. Schuster, "Multi-channel spectrograms for speech processing applications using deep learning methods," *Pattern Analysis and Applications*, vol. 24, no. 2, pp. 423–431, 2021.

[32] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8. Citeseer, 2015, pp. 18–25.

[33] S. Hegde, S. Shetty, S. Rai, and T. Dodderi, "A survey on machine learning approaches for automatic detection of voice disorders," *Journal of Voice*, vol. 33, no. 6, pp. 947–e11, 2019.

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[35] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035.

[36] L. Moro-Velázquez, J. Gómez-García, J. I. Godino-Llorente, and G. Andrade-Miranda, "Modulation spectra morphological parameters: A new method to assess voice pathologies according to the grbas scale," in *BioMed research international*, 2015.

[37] J. A. G. García, L. Moro-Velázquez, and J. I. Godino-Llorente, "On the design of automatic voice condition analysis systems. part I: review of concepts and an insight to the state of the art," *Biomedical Signal Processing and Control*, vol. 51, pp. 181 – 199, 2019.