INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

On the Use of High Frequency Information for Voice Pathology Classification

David Martínez[1], *Dayana Ribas*[2], *Eduardo Lleida*[2]

[1]LumenVox GmbH, Munich, Germany
[2]ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain

david.martinez@lumenvox.com, dribas@unizar.es, lleida@unizar.es

Abstract

We have observed significant differences in the high frequency content of the spectrum between healthy and pathological voices. Pathologies like larynx cancer, vocal fold lesions, and patients with larynx or vocal fold removal are examples of this. This finding invites to use high sampling frequencies in voice pathology classification systems to benefit from this high frequency information, which has been traditionally ignored. With a GMM classifier fed with MFCCs and a sampling frquency of 48 kHz we are able to improve AUC almost a 5% compared to a system using a sampling frequency of 8 kHz and more than 2% compared to a system using a sampling frequency of 16 kHz.
Index Terms: voice pathology classification, high frequency, sampling frequency, SVD, AVFAD

1. Introduction

Voice pathology detection is a binary classification problem with the goal of deciding if the voice uttered by a person is normal or pathological. Voice pathologies are numerous and diverse, and diagnosis rely on the subjective criteria of professional clinicians. Hence, automatic assessment tools are a great support to make more accurate and consistent diagnosis of pathologies.

Nonetheless, voice malfunction is a common problem in our society and at least 30% of us will suffer it at least once in his life according to [1]. Moreover, with the establishment of electronic devices and high-speed internet in most of the populated areas, such systems can be applied even in remote places without requiring the physical presence of a doctor. Potential savings are huge.

Different classification systems have been evaluated for voice pathology detection but the lack of a common setup makes it difficult to make fair comparisons [2]. Classifiers like Gaussian mixture models (GMM) [3], support vector machines (SVM) [4] and different types of neural networks [5, 6] present similar accuracies that range from 80% to 95% depending on the setup. In addition, a lot of investigations have focused on features that discriminate well between healthy and pathological voices, and we find a wide range of possibilities like jitter, shimmer, signal-to-noise ratios, Mel-frequency coefficients (MFCC), and combinations of them [7], to cite a few.

In this work, our main objectives are to share our findings about the utility of high-frequency content for the characterization of voice pathologies and to show how much voice pathology classification results improve when this information is used. We haven't found previous literature paying attention to this factor and most of the works use sampling frequencies of 8 kHz or 16 kHz [8]. For example, [9] makes a deep analysis on the influence of MFCCs in voice pathology detection but the audios are downsampled to 16 kHz. The authors in [10] do perform an analysis of the influence of each frequency band but their maximum sampling frequency is 25 kHz, which limits the maximum analysis frequency to 12.5 kHz. Interestingly, they find that the most relevant frequency band is the range 1-8 kHz, but their study is limited to three pathologies, namely vocal fold cysts, unilateral vocal fold paralysis and vocal fold polyps. In our work, we find also relevant even higher frequencies.

Even for characterization of certain impairments like laryngectomy [11] or dysphonia [12], the high frequency content has not been considered. However, we demonstrate that these standard sampling frequencies may not fully characterize certain pathologies.

Our work is organized as follows: Section 2 describes the audio material; in Section 3 we share the observations that motivated our study; Section 4 presents our classification system architecture; Section 5 shows the experiments that we have performend and obtained results; and in Section 6 conclusions are drawn.

2. Data Material

2.1. Saarbrücken Voice Database (SVD)

SVD [13] contains recordings of vowels /a/, /i/ and /u/, and the phrase in German "Guten Morgen, wie geht es Ihnen?". The sampling frequency is 50 kHz. It contains 687 healthy people and 1356 with pathologies, with a total of 71 pathologies including organic, functional and neurological impairments.

In our experiments we use all the pathologies and a 5-fold strategy following the same partition as in [2]. In this work all results are obtained using audio from the phrase "Guten Morgen, wie geht es Ihnen?". Previous work in the literature has shown that using phrases rather than sustained vowels increases discrimination between normal and pahtological voices [14].

2.2. Advanced Voice Function Assessment Database (AVFAD)

AVFAD [15] contains recordings of vowels /a/, /i/ and /u/, several phrases in Portuguese, a read text and a spontaneous speech recording. The sampling frequency is 48 kHz. It contains 363 healty people and 346 with pathologies, with a total of 26 pathologies including organic, functional and neurological impairments.

As in SVD, we use all pathologies and a 5-fold strategy, created using the modulus of the speaker id number. In this case, all our experiments have been done with phrase 5 of the database, "Sofia saiu cedo da sala".

10.21437/Interspeech.2023-2181

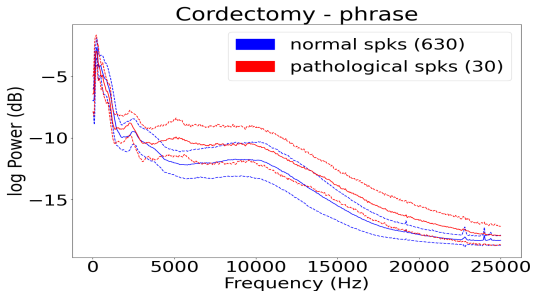
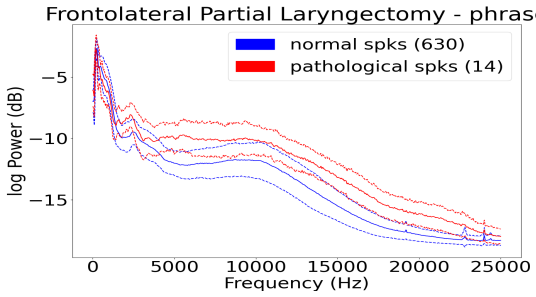
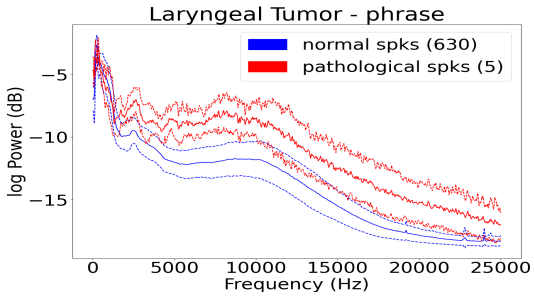
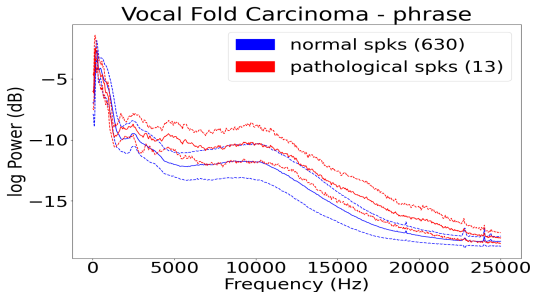
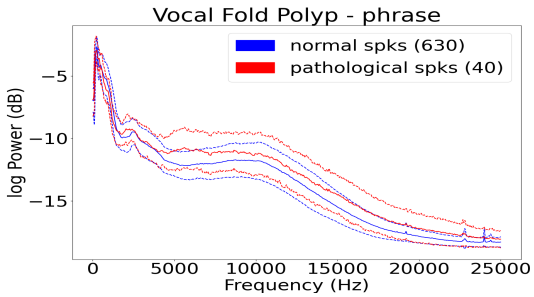
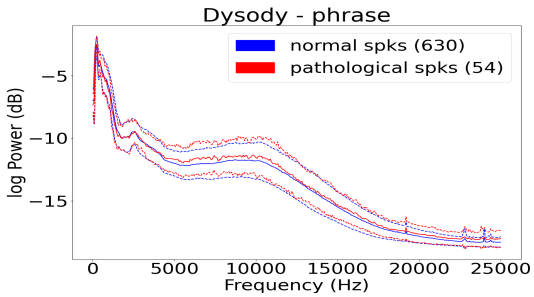
Nr.	a	b
1		
2		
3		

Table 1: *Figures with comparison between normal voice and pathological voices from SVD. 1a - Cordectomy; 1b - Frontolaterale partial laryngectomy; 2a - Laryngeal tumor; 2b - Vocal fold carcinoma; 3a - Vocal fold polyp; 3b - Dysodia.*

3. Observations

The main motivation for our work is the observation of wide-band spectrograms for several pathologies. In Table 1 we compare the average power spectrum for all the normal voices in the SVD with the average power spectrum for several pathologies $\pm$ 1 standard deviation interval. The figures are obtained on the phrase recording. We have selected the fiive pathologies where the differences in the higher part of the spectrum are more notable. To make the plots, we have used rectangular filter banks every 500 Hz.

The top 2 figures in the table correspond to two surgical procedures: cordectomy, removal of one or two vocal folds, and laryngectomy, removal of part or all of the larynx. In the two cases the differences with normal voices start around 5 kHz and the difference remains up to 24 kHz. It is interesting to see that in laryngectomy, the average curves $\pm$ 1 std do not overlap from around 15 kHz to 22 kHz, frequencies not possible to be analyzed even for a sampling frequency of 30 kHz.

Figure 2a shows patients with laryngeal tumor. The difference with normal is significant from 5 kHz up to 24 kHz. In Figure 2b, with focal fold carcinoma, we see a similar effect, not as prominent as in the previous case, although from 15 kHz to 22 kHz the overlap of the $\pm$1 std regions is very small. In Figure 3a, vocal fold polyp, has a similar pattern to vocal fold carcinoma.

In Figure 3b we have dysodia, which is included for reference to compare with a case that doesn't have a visually different spectrogram from the normal class.

One point in common in 2 of the 6 selected pathologies is that there is an injury in the vocal cords, so, it seems that this type of damages provoke higher excitement of high frequencies than normal.

Most of these cases contain differences not only in the high part of the spectrum, but also in the middle, mainly down to 5 kHz, therefore we don't expect these pathologies not to be well classified with lower sampling frequencies but to have a more robust behaviour or even a boost in accuracy when the higher sampling frequencies are used.

Out of the scope of this work, although a highly interested information, is the reason why the high frequencies are altered compared to normal voices. In the case of lesions in the vocal cords, one plausible hypothesis is that polyps or other protuberances provoke a perturbartion in the air flow that excites this wide range of frequencies.

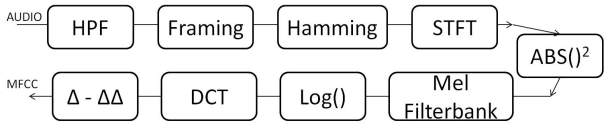


Figure 1: *MFCC computation diagram.*

4. System Description

4.1. Features

We use MFCCs. We need to include information of the whole frequency range in the voice pathology classication task and MFFCs are features that work on the frequency domain, efficiently compressing the information in the whole spectrum. MFCCs have been extensively used in speech technology [16]. The main steps to compute them are depicted in Figure 1.

We first apply a high pass filtering. Then the signal is framed in chunks of 25 ms every 10 ms. A Hamming window is applied in each frame. Then, the short time Fourier transform (STFT) is computed at each frame with as many points as the next power of two to the number of samples in the frame. The square of the absolute value of the STFT is computed to get the power spectrum, and then a Mel filterbank is applied. Next we apply a log function over the Mel-filtered values and finally a discrete Fourier transform (DCT) is applied, which returns a pre-selected number of coeffcientes. Finally, we also obtain the first and second derivatives of the MFCCs.

4.2. Classifier

We use a GMM classifier [17]. The GMM classifier learns two generative models: one for the normal class and one for the pathological class, and at inference time the score to make decisions is the difference between the logarithm of the score obtained by the normal model minus the logarithm of the score obtained by the pathological model.

The goal of each GMM model is to learn a frequency representation (captured by the mean of each Gaussian) for each part of the evaluated sentence, and depending on the length of the sentence we need more or less Gaussian units.

This classifier has shown state-of-the-art performance during the last decade [8]. Other typical choices are SVMs or neural networks. It is interesting to note that the advent of deep neural networks has not shown convincing improvements yet in voice pathology classification tasks, probably due to the lack of huge databases, as it has happened in other related areas like speech or image recognition where the size of available databases is light-years larger. Only recently, some approaches have tried to make benefit of self-supervised models pre-trained with huge amounts of data with succes [18].

In any case, since our objective is not to obtain the best possible results, but to share our findings about the importance of high-frequency content for voice pathology detection, the GMM classifier is better suited for our problem than the other choices because it makes easier to analyse what the system is learning.

4.3. Tunable Parameters

To avoid overfitting in our experiments, we use a held-oput dataset (AVFAD) to tune the system parameters. The tunable parameters are:

- Number of Mel filter banks: to control de resolution of the spectrum.
- Number of MFCC coefficients: to control the compression rate of the spectrum information.
- Number of Gaussians: to control de resolution of the parts of the sentence where a different spectrum is computed.

5. Experiments

5.1. Metrics

To report results, we focus on uncalibrated metrics like area under the cureve (AUC) and equal error rate (EER) because we have not optimized the threshold of the system to make decisions. However, to enable comparisons with other works in the literature, we use a threshold of 0 to make decisions and report unweighted avearage recall (UAR).

We also compute the 95% confidence interval on the mean of the reported metrics computed over the 5 folds.

5.2. System Development on AVFAD

AVFAD database is used as development data to tune the system parameters and select the best configuration for each sampling frequency. We use AUC as optimization metric. In Figure 2 we show the 5-fold average AUC for a sweep of the tunable parameters as described in Section 4.3 (number of Mel filters, number of MFCC coefficients and number of Gaussians). In Table 2 we show the optimal configuration for each sampling frequency with the corresponding obtained AUC.

Table 2: *Optimal configurations and results in AVFAD*

Sampling Freq.	8 kHz	16 kHz	22 kHz	48 kHz
#filters	40	40	20	120
#MFCC coefficients	30	30	30	25
#Gaussians	16	16	16	16
AUC (%)	89.32	89.91	90.20	90.23

In Figure 2 we can observe several things:

- Best results are obtained with a sampling frequency of 48 kHz in most points.
- The difference between 48 kHz and the rest is largest for 3 Gaussians and is reduced as we increase the number of Gaussians. This indicates that for simpler systems, the high frequency information is more relevant to differentiate between normal and pathological voices, and as we introduce more parameters we are able to include more information from the low-frequency band to improve the differentiation between the two classes.
- For the 48 kHz system, the addition of more Gaussians does not offer a big improvement in terms of AUC. This indicates that by using high frequency content we do not need to compare a big diversity of frequency patterns (or in other words, only a few modes in the data distribution are relevant for the classification task).
- Higher number of MFCC coefficients help.
- For 8 kHz and 16 kHz the optimal number of filter is 40 and for 48 kHz it is 120. This looks reasonable since this way the frequency band covered by each filter is similar in all cases. For the lower sampling frequency 120 filters may provoke overfitting. For 22 kHz we see that the optimal is 20 filters but the result for 120 filters is similar.

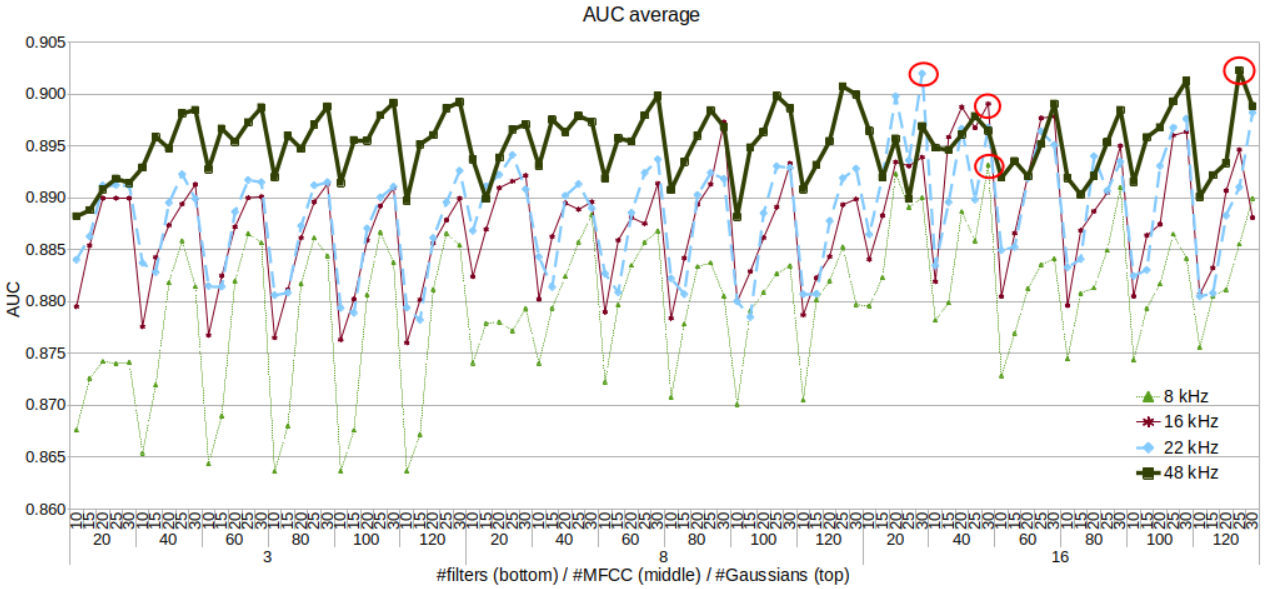


Figure 2: *AVFAD average AUC results for sampling frequencies of 8 kHz (green-thin-dotted-triangle), 16 kHz (red-thin-continuous-star), 22 kHz (blue-thick-dashed-diamond), 48 kHz (black-thick-continuous-square). In the x-axis we show a sweep for number of Mel filters (20, 40, 60, 80, 100), number of MFCC coefficients (10, 15, 20, 25, 30) and number of Gaussians (8, 16, 32). Red circles indicate maximum value of each line.*

5.3. Evaluation on SVD

In Table 3 we compare the results for SVD using the best configuration obtained with AVFAD database for the four sampling frequencies studied.

Table 3: *Results for SVD using AVFAD's best configuration*

Sampling Freq.	8 kHz	16 kHz	22 kHz	48 kHz
AUC avg (%)	84.38	86.70	86.99	88.52
CI(95%)	±2.82	±2.73	±2.09	±2.15
EER avg (%)	23.17	20.95	21.43	19.33
CI(95%)	±2.73	±2.72	±2.34	±1.67
UAR avg (%)	76.29	78.63	78.83	80.52
CI(95%)	±2.17	±2.73	±1.71	±1.60

We can see a robust and progressive improvement of results in all metrics from 8 kHz to 48 kHz, with some overlaps in the confidence intervals. Focusing on AUC, we see relative improvements with 48 kHz of 4.90% over 8 kHz, 2.10% over 16 kHz, and 1.76% over 22 kHz. We also see that the difference between 22 kHz and 16 kHz is very small.

6. Conclusions

We have based our work in the observation of the wide-band power spectrums of pathological voices, which indicates that high frequency content can be a potential good information to be used in voice pathological classifiers. We have seen that certain pathologies like cordectomy, laryngectomy, and several lesions in the vocal cords and larynx, including larynx cancer, have a clear distinct spectrum pattern than normal voices.

The hypothesis has been tested on SVD database using a GMM architecture tuned on AVFAD database. The results show that by using a sampling frequency of 48 kHz we can obtain an average relative improvement of 2.10% over a sampling frequency of 16 kHz and 4.90% over 8 kHz.

In future work we plan to study if the high frequency information can also help us to better distinguish across pathologies in a multiclass classification problem. We think that voice classification can benefit from hand-crafted setups and continuing this analysis on other data and other part of speech like sustained vowels will help us to improve the classification rates.

7. Acknowledgements

This work was supported in part by the European Union's Horizon 2020 Research and Innovation Programme under Marie Skłodowska-Curie under Grant 101007666, in part by Grant MCIN/AEI/10.13039/501100011033, in part by the European Union NextGenerationEU/PRTR under Grant PDC2021-120846-C41 and Grant PID2021-126061OB-C44, and in part by the Government of Aragon under Grant T36_20R.

8. References

[1] M. S. Benninger, C. E. Holy, P. C. Bryson, and C. F. Milstein, "Prevalence and occupation of patients presenting with dysphonia in the united states," *Journal of Voice*, vol. 31, no. 5, pp. 594–600, 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0892199716304416

[2] M. Huckvale and C. Buciuleac, "Automated Detection of Voice Disorder in the Saarbrücken Voice Database: Effects of Pathology Subset and Audio Materials," in *Proc. Interspeech 2021*, 2021, pp. 1399–1403.

[3] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006, voice Models and Analysis for Biomedical Applications. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809406000267

[4] Y. Niu, J. Cao, F. Shen, and P. Ren, "The study of voice pathology detection based on mfcc and svm," in *2020 7th International Conference on Biomedical and Bioinformatics Engineering*, ser. ICBBE '20. New York, NY, USA: Association for Computing Machinery, 2021, p. 27–30. [Online]. Available: https://doi.org/10.1145/3444884.3444890

[5] M. A. Mohammed, K. H. Abdulkareem, S. A. Mostafa, M. Khanapi Abd Ghani, M. S. Maashi, B. Garcia-Zapirain, I. Oleagordia, H. Alhakami, and F. T. AL-Dhief, "Voice pathology detection and classification using convolutional neural network model," *Applied Sciences*, vol. 10, no. 11, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/11/3723

[6] R. Islam, E. Abdel-Raheem, and M. Tarique, "Voice pathology detection using convolutional neural networks with electroglottographic (egg) and speech signals," *Computer Methods and Programs in Biomedicine Update*, vol. 2, p. 100074, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666990022000258

[7] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit," in *Advances in Speech and Language Technologies for Iberian Languages*, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández, and D. Ramos Castro, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 99–109.

[8] R. Islam, M. Tarique, and E. Abdel-Raheem, "A survey on signal processing based pathological voice detection techniques," *IEEE Access*, vol. 8, pp. 66 749–66 776, 2020.

[9] S. Tirronen, S. R. Kadiri, and P. Alku, "The effect of the mfcc frame length in automatic voice pathology detection," *Journal of Voice*, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S089219972200087X

[10] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. Farahat Ibrahim, "Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions," *IEEE Access*, vol. 6, pp. 6961–6974, 2018.

[11] D. Globlek, B. Simunjak, M. Ivkic, and A. Bonetti, "Some characteristics of voice in near-total laryngectomy," *Logopedics Phoniatrics Vocology*, vol. 30, no. 2, pp. 94–96, 2005. [Online]. Available: https://doi.org/10.1080/14015430500246130

[12] G. Pouchoulin, C. Fredouille, J.-F. Bonastre, A. Ghio, and A. Giovanni, "Frequency study for the characterization of the dysphonic voices," *Proceedings of the 8th INTERSPEECH Conference (INTERSPEECH 2007)*, 08 2007.

[13] M. Pützer and W. Barry. Saarbrücken voice database. [Online]. Available: http://www.stimmdatenbank.coli.uni-saarland.de/

[14] J. Gómez-García, L. Moro-Velázquez, and J. Godino-Llorente, "On the design of automatic voice condition analysis systems. part ii: Review of speaker recognition techniques and study on the effects of different variability factors," *Biomedical Signal Processing and Control*, vol. 48, pp. 128–143, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1746809418302416

[15] L. M. Jesus, I. Belo, J. Machado, and A. Hall, "The advanced voice function assessment databases (avfad): Tools for voice clinicians and speech research," in *Advances in Speech-language Pathology*, F. D. M. Fernandes, Ed. Rijeka: IntechOpen, 2017, ch. 14. [Online]. Available: https://doi.org/10.5772/intechopen.69643

[16] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[17] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[18] D. Ribas, M. A. Pastor, A. Miguel, D. Martínez, A. Ortega, and E. Lleida, "Automatic voice disorder detection using self-supervised representations," *IEEE Access*, vol. 11, pp. 14 915–14 927, 2023.