International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 1, February 2022, pp. 376~382
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i1.pp376-382

Comparative study to realize an automatic speaker recognition system

Fadwa Abakarim, Abdenbi Abenaou
Research Team of Applied Mathematics and Intelligent Systems Engineering, National School of Applied Sciences, Ibn Zohr University, Agadir, Morocco

Article Info

Article history:

Received Jan 15, 2021
Revised Jul 13, 2021
Accepted Jul 26, 2021

Keywords:

Adaptive orthogonal transform
Automatic speech recognition
DTW
MFCCs
Speaker recognition system

ABSTRACT

In this research, we present an automatic speaker recognition system based on adaptive orthogonal transformations. To obtain the informative features with a minimum dimension from the input signals, we created an adaptive operator, which helped to identify the speaker's voice in a fast and efficient manner. We test the efficiency and the performance of our method by comparing it with another approach, mel-frequency cepstral coefficients (MFCCs), which is widely used by researchers as their feature extraction method. The experimental results show the importance of creating the adaptive operator, which gives added value to the proposed approach. The performance of the system achieved 96.8% accuracy using Fourier transform as a compression method and 98.1% using Correlation as a compression method.

Corresponding Author:

Fadwa Abakarim
Research Team of Applied Mathematics and Intelligent Systems Engineering, National School of Applied Sciences, Ibn Zohr University
B.P 1136, Agadir 80000, Morocco
Email: fadwa.abakarim@gmail.com

1. INTRODUCTION

Automatic speech recognition is a computer technique that is commonly used in several systems. In car systems, speech recognition identifies the driver's voice to start responding to their commands such as playing music, activating global positioning system (GPS), launching phone calls, and selecting radio stations. In the field of language education, speech recognition can teach proper pronunciation and help people to develop their oral expression, and also it facilitates education for blind students. In this context, the research proposed a method to identify the speaker's voice using adaptive orthogonal transformations [1] and comparing it with the method of mel-frequency cepstral coefficients (MFCCs) [2]-[4].

In order to identify the speaker's voice several methods are used to extract the special features of each voice, among them mel-frequency cepstral coefficients. Although numerous researchers chose it as their feature extraction method because of its several advantages [5], [6], it reaches its limit in the improvement of automatic speaker recognition system as described by references [7]-[9]. It needs a large voice training dataset and a long execution time to identify the voice of each speaker [10] and the same goes for other approaches such as principal component analysis (PCA), discrete wavelet transform (DWT) and empirical modal decomposition (EMD) as revealed by reference [11].

Janse *et al.* [12] presented a comparative study between mel-frequency cepstral coefficients and discrete wavelet transform, where it mentioned that MFCCs values are not very robust in the presence of additive noise and that DWT requires a longer compression time. Winursito *et al.* [13] combined MFCCs with data reduction methods with the aim of improving the accuracy and increasing the computational speed

of the classification process by decreasing the dimensions of the feature data. The data reduction process is designed in two versions: MFCC+SVD version 1 and MFCC+PCA version 2. The results showed a performance improvement for the proposed approach. Wang and Lawlor [14] proposed a method for a speaker recognition system by combining MFCCs with back-propagation neural networks. It revealed that this approach works successfully only when the number of unfamiliar speakers is not too large.

From these research studies, we deduced that many authors have used MFCCs as a feature extraction approach and to strengthen their methodologies, they have used other approaches in the classification process to obtain the desired results. In addition, other authors have developed new methods by addressing the limitations of MFCCs in order to obtain an improved algorithm, which is not sensitive to noise, and has a fast execution time. The goal of this study is to solve the problems mentioned above by developing a fast algorithm based on adaptive orthogonal transformations for the extraction of the informative features from the voice signal using the smallest possible training dataset, inspired by references [1], [15]. This paper is organized as follows: Section 2 describes the new approach of orthogonal operators, then the comparison results obtained between MFCCs and the proposed method are discussed in section 3, and finally section 4 concludes the paper.

2. RESEARCH METHOD

2.1. Pre-processing

Before starting to apply the proposed approach, it is first necessary to pre-process the signals as shown in Figure 1. This involves firstly removing silence, then secondly detecting the beginning and the end of the speech by using the zero-crossing rate (ZCR) [16]-[19]. The third step is making them equal in length by using zero padding [20], [21] because the training dataset can contain several signals that do not have the same length. The final step is compressing their size without losing quality to avoid the problem of system slowness by using Fourier transform [22], [23] or correlation [24], [25]. Figure 2 shows the input signal before and after removing silence with detection of the beginning and the end of speech. Figure 3 shows the speech signal after applying the Fourier transform method to detect the informative intervals. Figure 4 shows the speech signal after applying correlation to detect the informative intervals.

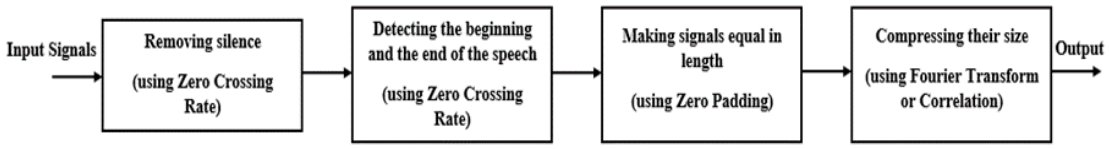

Figure 1. The pre-processing part

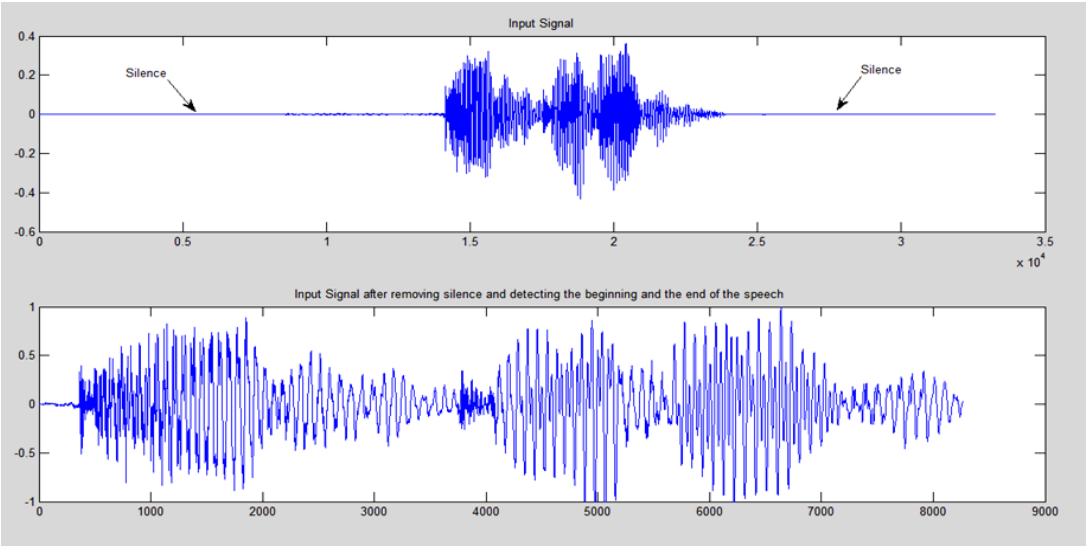

Figure 2. The input signal before and after removing silence with detection of the beginning and the end of speech

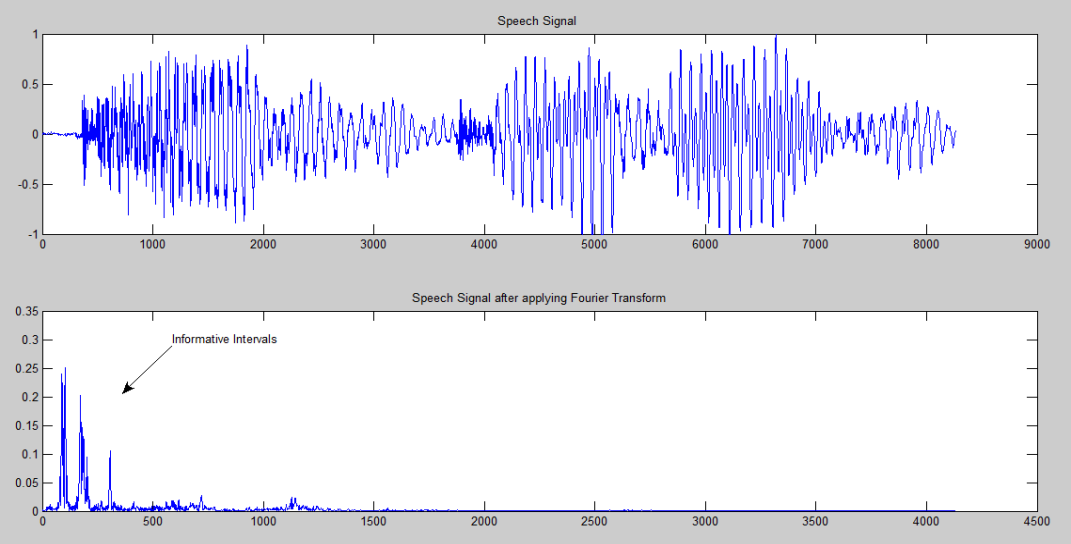

Figure 3. The speech signal after applying the Fourier transform

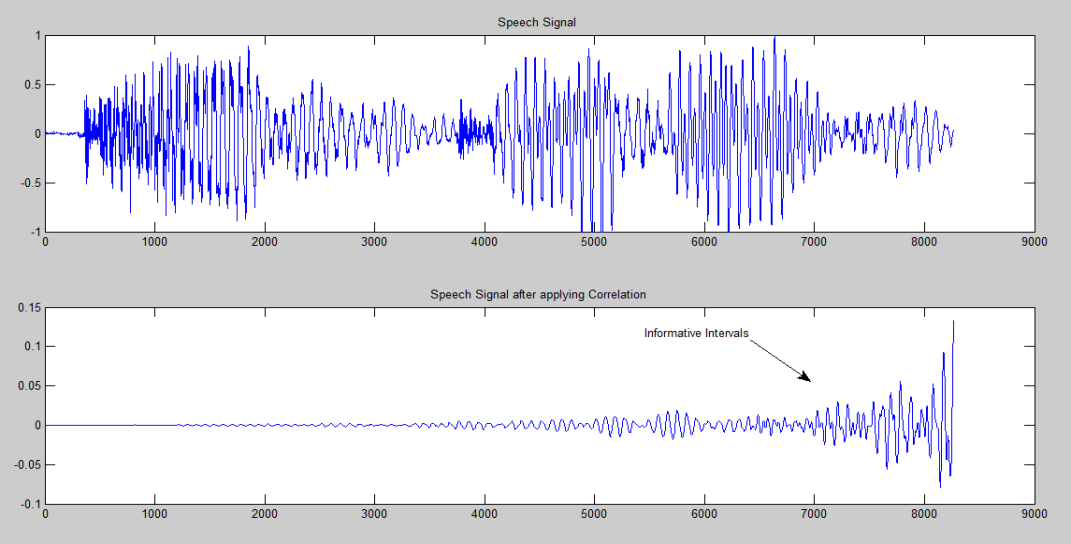

Figure 4. The speech signal after applying correlation

2.2. Theoretical background

Our approach consists in searching the informative features of the signal by using the operator H, which is a matrix operator of the transform (dimension N×N) whose number of rows corresponds to the number of basic functions. To decompose the vector X, the calculation of the discrete spectrum Y with the numerical methods can be represented by the following matrix [1], [15], [26]:

$$Y = \frac{1}{N} HX \quad (1)$$

where

$X = [x_1, x_2, \dots, x_N]^T$ is the initial signal to be transformed (size N = 2^n).

$Y = [y_1, y_2, \dots, y_N]^T$ is the vector of the spectral coefficients, calculated by the operator orthogonal H.

The calculation of the spectrum Y by using (1) requires N^2 multiplication and addition operations. The most efficient way to reduce the number of operations is to use a sparse matrix [27] where most of its elements are zero, which will make the calculation and execution time of the algorithm faster. The method of Good [28] which is used in the construction of fast transformation algorithms consists in expressing the

orthogonal spectral operator H as a product of sparse matrices G_i composed by minimum dimensional matrices called spectral kernels, where these matrices take the following form:

$$V_{i,j}(\alpha_{i,j}) = \begin{bmatrix} \cos(\alpha_{i,j}) & \sin(\alpha_{i,j}) \\ \sin(\alpha_{i,j}) & -\cos(\alpha_{i,j}) \end{bmatrix} \quad (2)$$

with $\alpha \in [0,2\pi]$.
Then G_i will be written as (3):

$$G_i = \begin{bmatrix} V_{i,1} & 0 & \dots & 0 \\ 0 & V_{i,2} & 0 & 0 \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & V_{i,N/2} \end{bmatrix} \quad (3)$$

with $i = 1 \dots n$, $n = \log_2 N$ which is the number of matrices G_i, then each matrix G_i contains $\frac{N}{2}$ spectral kernels of $V_{i,j}(\alpha_{i,j})$ of dimension 2×2.

So, the formula of Y will be:

$$Y = \frac{1}{N} HX = \frac{1}{N} G_1 G_2 \dots G_n X = \frac{1}{N} \prod_{i=1}^{n} G_i X \quad (4)$$

The algorithm goes through a procedure of adaptation of the operator H to a class of input signals. It consists in calculating the average of the statistical features at the pre-processing part (section 2 part 1) to form the standard vector $\hat{R}_{sd}$. We can say that the operator H is adapted to a class of signals represented by a standard vector $\hat{R}_{sd}$ if it verifies the following condition:

$$\frac{1}{N} H_a \hat{R}_{sd} = Y_t = [y_{t1}, 0, \dots, 0]^T \text{ with } y_{t1} \neq 0 \quad (5)$$

where Y_t is the target vector that constructs the adaptation criterion of the operator H_a to $\hat{R}_{sd}$.
The target vector Y_t is calculated as (6):

$$Y_i = G_i Y_{i-1} \quad (6)$$

with $i = 1 \dots \log_2 N$ and $Y_0 = \hat{R}_{sd}$.

In a simplified way, the synthesis procedure of the operator of orthogonal transformation is as follows:
For $i = 1$, $Y_1 = G_1 \hat{R}_{sd}$ with Y_1 contains $\frac{N}{2^1}$ non-zero number of elements.
For $i = 2$, $Y_2 = G_2 Y_1$ with Y_2 contains $\frac{N}{2^2}$ non-zero number of elements.
For $i = n$, $Y_n = Y_t = G_n Y_{n-1}$ with Y_t contains $\frac{N}{2^n}$ non-zero number of elements.
Then the calculation of the orthogonal spectral operator is:

$$H_a = G_n G_{n-1} \dots G_1 \quad (7)$$

Figure 5 shows the overall process to extract the informative features from the input signals by using our approach:
As shown in Figure 5, the extraction of the informative features consists of 7 steps:

- Step 1: Input signals go through a pre-processing process (section 2 part 1).
- Step 2: We calculate the average of the statistical features obtained during the pre-processing part (using Fourier transform and correlation).
- Step 3: The operator synthesis algorithm is applied to the average of the statistical features obtained from the previous step.
- Step 4: The output of the algorithm is the adaptive operator H.
- Step 5: The projection multiplication is applied between the operator H and the rest of the statistical features.
- Step 6: The result of the previous operation is a set of informative features that characterize each signal of the class.
- Step 7: The average of the feature vectors is calculated. The result is an informative feature vector with a minimum dimension that characterizes the whole class.

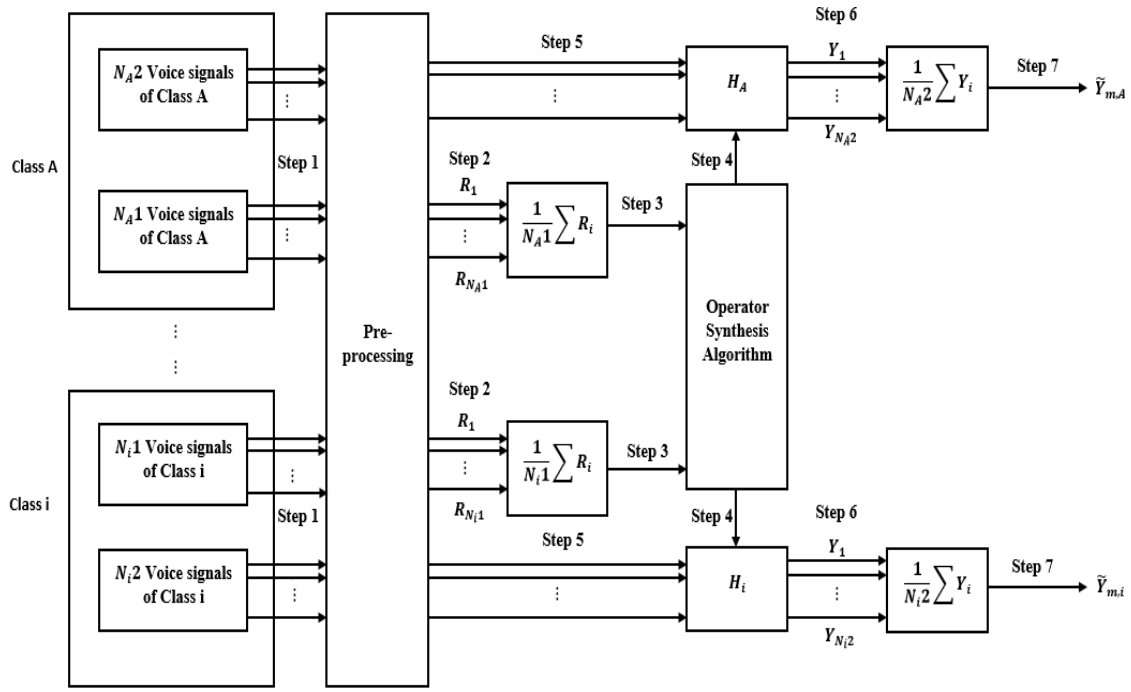

Figure 5. The overall process to extract the informative features from the input signals by using the adaptive orthogonal transform method

2.3. Datasets

The speech signals are recorded and filtered by the Audacity program. Each signal is recorded by default at 22 kHz with a duration between 1 s and 2 s. The training dataset contains 10 classes, where each class contains 100 voice recordings of the speaker (Class 1 of Speaker A contains 100 of his voice recordings, the same applies to Class 2 of Speaker B up to Class 10 of Speaker J).

The test dataset contains 6000 voice recordings of speaker A, B, …, J and other unfamiliar speakers. To test the similarity between the speaker's voice in the training dataset and the speaker's voice in the test dataset, dynamic time wrapping (DTW) is used. Dynamic time wrapping or DTW [29]-[32] consists in comparing two voice signals by considering the Euclidean distance between the two vectors obtained by the applied method, which is defined by (8):

$$D_i = \sqrt{\sum_{i=1}^{n}(a_i - b_i)^2} \quad (8)$$

with

D_i : The distance between the i vector of the spectrum a and the i vector of the spectrum b.

n : Dimension of a and b spectrum.

Therefore, the vector a_i will correspond to class i if $D_i = \min(D_{i=1\ldots C})$ where C is the number of classes.

3. RESULTS AND DISCUSSION

The quality of recognition is measured by calculating the recognition rate which is defined as (9).

$$Rate = \frac{the\ recognized\ speakers\ count}{the\ test\ dataset\ size} * 100 \quad (9)$$

Tables 1 and 2 show the voice recognition rate according to the size of the interval of the analysis. From these tables, we observe that the adaptive orthogonal transform method gives good results compared to the MFCCs approach. As mentioned in section 2 part 1, we used correlation and Fourier transform to work only with the informative intervals of the signal instead of working with the whole signal. As we can see there is a 47.5% difference in voice identification rates with Fourier transform intervals between using our approach (96.8%) and MFCCs (49.3%). On the other hand, we found a 45.0% difference in voice identification rates with correlation intervals between using our approach (98.1%) and the MFCCs (53.1%). Correlation intervals

give us better results than Fourier transform intervals, either for our approach or for the MFCCs. The proposed method has succeeded in identifying 5886 voice recordings among 6000 voice recordings of the test dataset (rate 98.1%) compared to MFCCs that identified only 3186 voice recordings (rate 53.1%), and these results show the efficiency of our algorithm.

Table 1. The voice recognition rate according to the size of the interval using Fourier transform with MFCCs and the adaptive orthogonal transform method.

Size of interval using Fourier Transform	The voice recognition rate with MFCCs (%)	The voice recognition rate with the adaptive orthogonal transform method (%)
128	23.8	68.9
256	32.8	75.3
512	35.6	82.3
1024	38.5	91.5
2048	45.6	93.1
4096	49.3	96.8

Table 2. The voice recognition rate according to the size of the interval using correlation with MFCCs and the adaptive orthogonal transform method

Size of interval using Correlation	The voice recognition rate with MFCCs (%)	The voice recognition rate with the adaptive orthogonal transform method (%)
128	25.6	65.8
256	30.3	73.2
512	37.3	81.7
1024	43.2	90.2
2048	49.1	95.6
4096	53.1	98.1

4. CONCLUSION

MFCCs is one of the most usable and well-known methods in the field of signal processing. However, it needs a large training dataset and a long execution time to extract the important features if the number of test dataset unfamiliar speakers is large, so for these reasons we developed a new method based on the creation of the operator H which is adaptable to any input signal. Even though its creation goes through several iterations $\log_2 N$ iterations where N is the length of the signal, an advantage of working with a sparse matrix where most of its elements are zero is that it makes the calculation and execution time of the algorithm faster. Our future goal is to increase the number of voice recordings in the test dataset and to decrease the number of voice recordings in the training dataset to see if the method continues to give successful results or not. In addition, we will combine it with other methods that are commonly used as classification methods such as hidden markov model (HMM) and artificial neural networks (ANN).

ACKNOWLEDGEMENTS

A special thanks to Mr. T. Hobson from Anglosphere English Center for reviewing for spelling and grammatical mistakes, and to all the participants who recorded their voices for this research.

REFERENCES

[1] A. Abenaou, F. Ataa Allah, and B. Nsiri, "Towards an automatic speech recognition system in amazigh based on orthogonal transformations," *Asinag*, pp. 133-145, 2014.
[2] N. Easwari and P. Ponmuthuramalingam, "A comparative study on feature extraction technique for isolated word speech recognition," *International Journal of Engineering and Techniques*, vol. 1, no. 6, pp. 108-115, 2015.
[3] L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques," *Journal of Computing*, vol. 2, no. 3, 2010.
[4] P. P. Singh and P. Rani., "An approach to extract feature using MFCC," *IOSR Journal of Engineering*, vol. 4, no. 8, pp. 21-25, 2014, doi: 10.9790/3021-04812125.
[5] X. Wu, H. Gong, P. Chen, Z. Zhong, and Y. Xu, "Surveillance robot utilizing video and audio information," *Journal of Intelligent and Robotic Systems*, vol. 55, no. 4, pp. 403-421, 2009, doi: 10.1007/s10846-008-9297-3.
[6] A. N. A. Kumar and S. A. Muthukumaraswamy, "Text dependent voice recognition system using MFCC and VQ for security applications," *2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA*), vol. 2, 2017, pp. 130-136, doi: 10.1109/ICECA.2017.8212779.
[7] W. Zunjing and C. Zhigang, "Improved MFCC-based feature for robust speaker identification," *Tsinghua Science and Technology*, vol. 10, no. 2, pp. 158-161, 2005, doi: 10.1016/S1007-0214(05)70048-1.
[8] W. Chen, M. Zhenjiang, and M. Xiao, "Differential MFCC and vector quantization used for real-time speaker recognition system," *2008 Congress on Image and Signal Processing*, vol. 5, pp. 319-323, 2008, doi: 10.1109/CISP.2008.492.

[9] F. Z. Chelali and A. Djeradi, "Text dependant speaker recognition using MFCC, LPC and DWT," *International Journal of Speech Technology*, vol. 20, no. 3, pp. 725-740, 2017, doi: 10.1007/s10772-017-9441-1.
[10] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef, "Comparative study of automatic speech recognition techniques," *IET Signal Processing*, vol. 7, no. 1, pp. 25-46, 2013, doi: 10.1049/iet-spr.2012.0151.
[11] F. Abakarim and A. Abenaou, "Amazigh isolated word speech recognition system using the adaptive orthogonal transform method," *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, 2020, pp. 1-6, doi: 10.1109/ISCV49265.2020.9204291.
[12] P. V. Janse *et al.*, "A comparative study between MFCC and DWT feature extraction technique," *International Journal of Engineering Research and Technology*, vol. 3, no. 1, pp. 3124-3127, 2014.
[13] A. Winursito, R. Hidayat, A. Bejo, and M. N. Y. Utomo, "Feature data reduction of MFCC using PCA and SVD in speech recognition system," *2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE)*, pp. 1-6, 2018, doi: 10.1109/ICSCEE.2018.8538414.
[14] Y. Wang and B. Lawlor, "Speaker recognition based on MFCC and BP neural networks," *2017 28th Irish Signals and Systems Conference (ISSC)* , 2017, pp. 1-4, doi: 10.1109/ISSC.2017.7983644.
[15] M. Azergui, A. Abenaou, and H. Bouzahir, "Bearing fault classification based on the adaptive orthogonal transform method," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 9, no. 1, pp. 375-380, 2018, doi: 10.14569/IJACSA.2018.090151.
[16] D. S. Sheteb and P. S. B. Patil, "Zero crossing rate and energy of the speech signal of devanagari script," *IOSR Journal of VLSI and Signal Processing*, vol. 4, no. 1, pp. 1-5, 2014, doi: 10.9790/4200-04110105.
[17] R. G. Bachu, S. Kopparthi, B. Adapa, and B. D. Barkana, "Voiced/unvoiced decision for speech signals based on zero-crossing rate and energy," *Advanced Techniques in Computing Sciences and Software Engineering*, 2010, pp. 279-282, doi: 10.1007/978-90-481-3660-5_47.
[18] H. Aouani and Y. Ben Ayed, "Speech emotion recognition with deep learning," *Procedia Computer Science*, vol. 176, pp. 251-260, 2020, doi: 10.1016/j.procs.2020.08.027.
[19] T. Ijitona, H. Yue, J. Soraghan, and A. Lowit, "Improved silence-unvoiced-voiced (SUV) segmentation for dysarthric speech signals using linear prediction error variance," *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, 2020, pp. 685-690, doi: 10.1109/ICCCS49078.2020.9118462.
[20] J. Luo, Z. Xie, and M. Xie, "Interpolated DFT algorithms with zero padding for classic windows," *Mechanical Systems and Signal Processing*, vol. 70, pp. 1011-1025, 2016, doi: 10.1016/j.ymssp.2015.09.045.
[21] B. Zhang, T. Xiao, and J. Zhong, "A simple determination approach for zero-padding of FFT method in focal spot calculation," *Optics Communications*, vol. 451, pp. 260-264, 2019, doi: 10.1016/j.optcom.2019.06.065.
[22] J. Obuchowski, A. Wyłomańska, and R. Zimroz, "Selection of informative frequency band in local damage detection in rotating machinery," *Mechanical Systems and Signal Processing*, vol. 48, no. 1-2, pp. 138-152, 2014, doi: 10.1016/j.ymssp.2014.03.011.
[23] R. Mankar, M. J. Walsh, R. Bhargava, S. Prasad, and D. Mayerich, "Selecting optimal features from fourier transform infrared spectroscopy for discrete-frequency imaging," *The Analyst*, vol. 143, no. 5, pp. 1147-1156, 2018, doi: 10.1039/c7an01888f.
[24] C. Charayaphan, A. E. Marble, S. T. Nugent, and D. Swingler, "Correlation algorithm and sampling techniques for estimating the signal-to-noise ratio of the electrocardiogram," *Journal of biomedical engineering*, vol. 14, no. 6, pp. 516-520, 1992, doi: 10.1016/0141-5425(92)90106-U.
[25] J. Zhou and A. Qu, "Informative estimation and selection of correlation structure for longitudinal data," *Journal of the American statistical Association*, vol. 107, no. 498, pp. 701-710, 2012, doi: 10.1080/01621459.2012.682534.
[26] A. Abenaou, "Development of a method for the synthesis of adaptive spectral operators for the analysis of random," *International Journal of Computer Theory and Engineering*, vol. 9, no. 4, pp. 273-276, 2017, doi: 10.7763/IJCTE.2017.V9.1150.
[27] S. Wang, J. Liu, and N. Shroff, "Coded sparse matrix multiplication," *2018 International Conference on Machine Learning (ICML)*, 2018, pp. 5152-5160.
[28] I. J. Good, "The interaction algorithm and practical fourier analysis," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 20, no. 2, pp. 361-372, 1958, doi: 10.1111/j.2517-6161.1958.tb00300.x.
[29] P. Senin, "Dynamic time warping algorithm review," *Information and Computer Science Department University of Hawaii at Manoa Honolulu USA*, vol. 855, no. 1-23, pp. 40, 2008.
[30] A. Ismail, S. Abdlerazek, and I. M. El-Henawy, "Development of smart healthcare system based on speech recognition using support vector machine and dynamic time warping," *Sustainability*, vol. 12, no. 6, pp. 2403, 2020, doi: 10.3390/su12062403.
[31] I. D. G. Y. A. Wibawa and I. D. M. B. A. Darmawan, "Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini," *Journal of Physics: Conference Series*, vol. 1722, no. 1, 2021, doi: 10.1088/1742-6596/1722/1/012014.
[32] S. Kinkiri and S. Keates, "Speaker identification: Variations of a human voice," in *2020 International Conference on Advances in Computing and Communication Engineering (ICACCE)*, 2020, pp. 1-4, doi:10.1109/ICACCE49060.2020.9154998.