Denoising Esophageal Speech using Combination of Complex and Discrete Wavelet Transform with Wiener filter and Time Dilated Fourier Cepstra

Madiha Amarjouf[1*], *Fadoua* Bahja[2], *Joseph* Di Martino[3], *Mouhcine* Chami[1] and *El Hassan* Ibn Elhaj[1]

[1]Research laboratory in Telecommunications Systems: Networks and Services (STRS), Research team: Multimedia, Signal and Communications Systems (MUSICS), National Institute of Posts and Telecommunications (INPT), Av. Allal Al Fassi, Rabat, Morocco
[2]Laboratory of Innovation in Management and Engineering for Enterprise (LIMIE), Institut Supérieur d'Ingénierie et des Affaires (ISGA Rabat), 27 Avenue Oqba, Agdal, Rabat, Morocco
[3]Loria - Laboratoire Lorrain de Recherche en Informatique et ses Applications, B.P. 239 54506 Vandœuvre-lès-Nancy, France

Abstract. Esophageal speech is one of the pathological voices, which is known to be weak in intelligibility and hard to understand. Our approach's main idea is to reduce the esophageal speech noises using two-hybrid methods. This paper aims to merge the advantages of wavelet-based methods such as DWT and DTCWT, along with the standard methods such as the Wiener filter and the time dilated Fourier. The first hybrid method applies the filters on the vocal tract cepstrum, while the second one applies them at the synthesis stage. Two experiments were conducted as well to evaluate the results by objective analysis. The results obtained by the proposed hybrid methods gave good performances.

1 Introduction

Pathological voices are sounds produced by people suffering from a dysfunction or some voice trouble [1, 2]. This voice deterioration could be temporary or permanent [1]. Besides being hoarse, this voice undergoes changes in one or multiple acoustics parameters, e.g., intensity, pitch, frequency and timbre [1, 2, 3]. Vocal pathologies are generally grouped into three main categories of origins: Functional pathologies, organic and cancerous [1, 2]. Esophageal speech (ES) is the pathological voice that our work is about to study. It belongs to pathologies of cancerous origin. This voice disorder is due to total laryngectomy, which is a surgical operation consisting of complete removal of the larynx [1]. ES is a substitution voice that facilitates communication without the need of using a device or even the hand, which makes it the most used solution to communicate [1, 2]. Otherwise, the major problem with this voice is the fact that it is weak in intelligibility and difficult to understand [3]. Improving this voice through denoising methods is the aim of this study. The methods used are DWT, DTCWT, Wiener filter and the time dilated Fourier cepstra. In this study, we combined these techniques to attenuate the noise of esophageal speech and the results were noted in order to make an objective evaluation.

Many researchers have conducted speech denoising. [4] Have proposed a time-domain speech denoising method to reduce the hoarseness of dysphonic voices. This technique separates the signal and the noise subspaces based on singular value decomposition (SVD) of appropriate data matrices. Even the fact that the original signal's spectral characteristics were preserved while reconstructing the signal, the SNR was higher, especially for the high-frequency area. The same method was applied in [5] in order to implement a real-time hoarse denoising device. It was tested on speech signals degraded by white noise and applied to dysphonic voices. The same problem of high SNR was also found. A preprocessing technique using wavelet-based denoising of non-linear dynamic analysis for laryngeal paralysis patients has been studied in [6]; the work proved that nonlinear dynamic analysis under noisy conditions was improved efficiently using the proposed method. In [7] an improving technique for speech enhancement was introduced. This method consists of decomposing the noisy signal using wavelet packet decomposition, then using adaptive thresholding and cepstral subtraction for more denoising. It is based on principal component analysis (PCA) for speech denoising. [8] have introduced a preprocessing method for denoising speech. The procedure eliminates non-speech segments with a slight cost using speech segment detection before starting the denoising process. Also, this method has demonstrated that the denoising time is lower compared to the Wavenet-based denoising method.

The rest of this article is organized as follows: The second part introduces the processes and methods used in our work. Section three describes the denoising experiments. The results and discussion are given in the

* Corresponding author: amarjouf.madiha@inpt.ac.ma

fourth section. Finally, the fifth section gives a conclusion.

2 Methods

2.1 Extraction of the vocal tract cepstrum

Firstly, the short-time speech signal was multiplied by the normalised Hann window introduced by [9], to make the discontinuities in the speech signal less noticeable. Normalised Hann Window provided by formula (1) [9]:

$$W(n) = \begin{cases} \frac{2\sqrt{L}}{\sqrt{N}} * \left(\left(a - b\cos\left(\pi\frac{(2n+1)}{N}\right)\right)\right), & 0 \le n \le N \\ 0\ otherwise \end{cases} \quad (1)$$

Where:

- N: The length of the analysis window,
- L: The analysis step size
- a= b= 0.5

Then, a Fast Fourier Transform was applied to the windowed signal. After that, a log moderator was applied to the speech signal spectrum to separate the vocal tract spectrum and the excitation spectrum. The last step to obtain the cepstrum was applying the Inverse Fourier Transform (IFFT) to the logarithmic spectrum. Figure 1 resumes how the cepstrum is obtained.

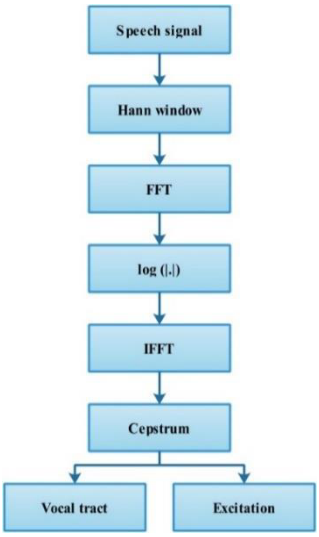

Fig. 1. The process of extracting the vocal tract and excitation cepstrum from the esophageal speech signal.

After calculating the cepstrum, it is easy to calculate the excitation data and the vocal tract data after separating them. Thus, the linguistic features of the speech signal are:

- C_0: The first coefficient
- $[C_1 \ldots C_p]$: The vocal tract cepstral vector
- $[C_{p+1} \ldots C_{N/2}]$: The excitation cepstral vector
- $[Ph_0 \ldots Ph_{N/2}]$: The phase coefficients.

Where N denotes the analysis window size, P represents the number of the vocal tract cepstrum, and $C_0 = 0$.

2.2 Denoising methods

2.2.1 Wavelet-based methods

- DWT

The Discrete Wavelet Transform gives a good localization in time and frequency [10]. It's used in several domains, especially in denoising images and speech signals [3, 11], and this is due to the multi-resolution analysis, which is the main characteristic of the wavelets. According to this analysis, the signal could be decomposed through a cascade of filters, associating a peer of filters to every resolution level: a low pass filter corresponding to low frequencies giving the approximation A_j, and a high pass filter corresponding to high frequencies giving the details D_j. This process is repeated to decompose the approximations [3, 10].

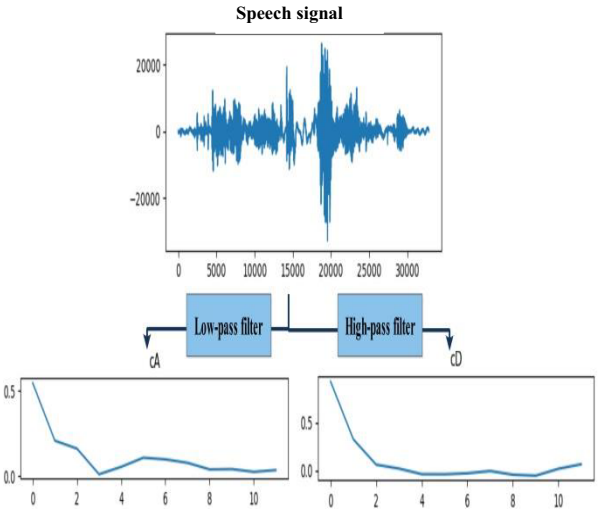

Fig. 2. The decomposition of the signal of an esophageal file sound in approximation A_j and details D_j by the DWT.

- DTCWT

The Dual-Tree Complex Wavelet Transform; often confused with continuous wavelets (CWT); could give more suitable results than the DWT [12]. This method uses two real trees of DWT, the first DWT tree generates the real part and its coefficients, while the second one generates the imaginary part of the DTCWT. Two different sets of real filters are used in the two real trees of the DWT. So, in the final, the result is approximately an analytic wavelet [12, 13]. H_0 and L_0 are respectively the high-pass filter and the low-pass filter of the first tree, and H_1 and L_1 are respectively the low-pass filter and the high-pass filter of the second tree.

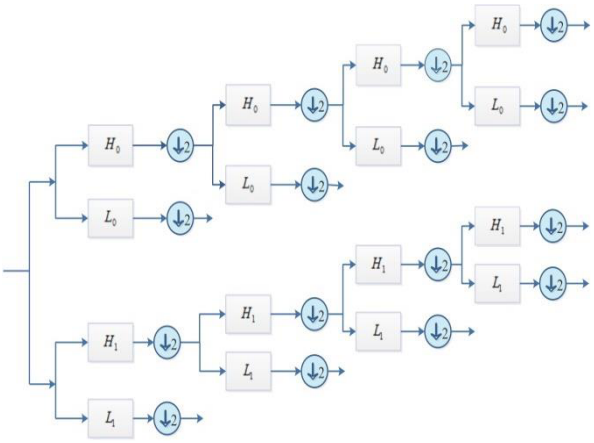

Fig. 3. Representation of the functioning of the DTCWT tree.

2.2.2 Wiener method

The Wiener method is a linear filter, belonging to the frequency filters. It is used to estimate the value of a noisy signal, by minimizing the mean squared error (MMSE) between the random estimated process and the desired process. It adapts the signal noise ratio for each treated part [14, 15].

2.2.3 The time Dilated Fourier Cepstra method

The dilation is a morphological operation, that aims to repair the interrupted signals by expanding them and making them clearer [2, 16, 17].

The time dilated Fourier Cepstra, as defined and detailed in [18], is used to enhance the esophageal speech in the frequency domain, by dilating the frequency axis of ratio 1/α. Thus, the frequency components will be changed, without corrupting the speech signal. The operator related to this operation is defined by formula (2) [18]:

$$D_\alpha[s](x) = s\left(\frac{x}{\alpha}\right), \alpha \in R^{+*} \quad (2)$$

Where α is called the α-rate homothety of a function $s \in L^2(R)$.

$D_\alpha^f = \alpha D_{1/\alpha}$ is the frequency dilated signal, it proceeds on the frequency support of a signal in the transposition direction:

- Towards the high frequencies for α >1
- Towards the low frequencies for α <1

The dilation algorithm used is detailed in [18].

2.2.4 Proposed hybrid methods

In this work, we propose two hybrid methods which aim to combine the advantages of the wavelet-based methods (DWT and DTCWT), with the standard methods (Wiener and Dilation). We proposed two models:

- Hybrid method 1

This proposed hybrid method denoised the extracted vocal tract cepstrum, using a combination of the methods described above. Figure 4 presents the schema of the first method.

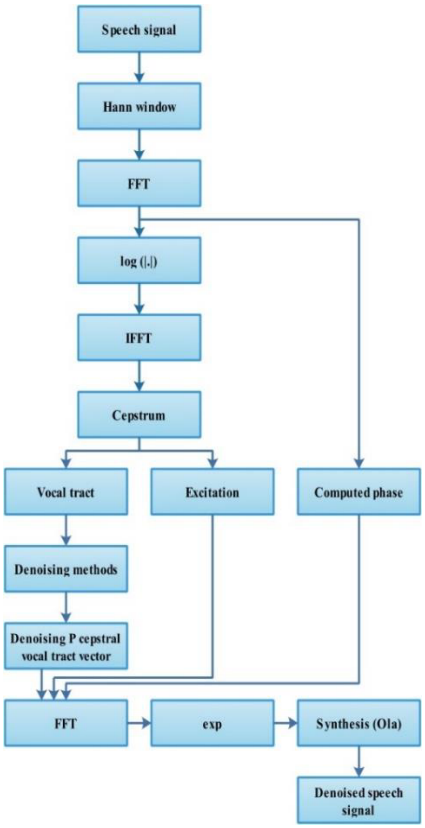

Fig. 4. Schema of hybrid method 1.

- Hybrid method 2

For hybrid method 2, the denoising methods were applied in the synthesis stage, on the Overlap-add method (Ola) without changing the vocal tract cepstrum. Figure 5 shows the schema of the given method.

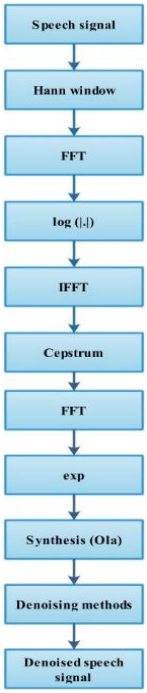

Fig. 5. Schema of hybrid method 2.

2.3 Synthesis stage

To reconstruct the speech signal and to return to the time domain [19], the IFFT operator was applied to the complex spectra, which were computed by multiplying the magnitude and the phase spectra. Then, the Overlap-add method (OLA) was used, it enables resynthesizing the original speech signal precisely from the windowed overlapping frames with low computation levels [19, 20, 21].

3 Experiments

3.1 Experiments

In this study, we conducted two experiments, the first one aimed to denoise esophageal speech signal using the proposed hybrid methods. This experiment compared the results of three cases, case 1 was about denoising the 25 vocal tract cepstrum and case 2 was about denoising the 33 vocal tract cepstrum (hybrid method 1). Case 3, used the second hybrid method. One esophageal speech file was used in experiment 1. The second experiment used the whole dataset. This experiment contained two cases; the first one was applying the first hybrid method on the 25 cepstral vocal tract vectors of each esophageal speech sound of the dataset. The choice of the number of cepstral vocal tract coefficients was made based on the first experiment. In the second case used hybrid method 2.

For experiments 1 and 2, the first cepstral coefficient C_0 was discarded. The $[C_{p+1}…C_{256}]$ cepstral coefficients left were used as the excitation features. Also, the 257 phase coefficients were used as well.

3.2 The dataset

The dataset used in this work is a French database, containing three parallel corpora, spoken by three French laryngectomees male speakers: PC, MH, and GM. Each corpus contains 289 phonetically balanced phrases. Table 1 presents the experimental settings of the experiments.

Table 1. The experimental settings.

Sampling rate	16 kHz
Step size	64ms
Analysis window size	512
P1	25
P2	33

4 Results and discussion

In this work, an objective evaluation has been done in order to evaluate the results of denoising the esophageal speech sounds. For experiment 1, the Signal-to-Error Ratio (SER) has been calculated. And for experiment 2, the SER has been calculated for each file sound of the three corpora, and then the average SER of each corpus has been computed. The Signal-to-Error-Ratio is given by formula (3) [2]:

$$SER = 10\,log_{10}\frac{\sum_k \|x_k - \hat{x}_k\|^2}{\sum_k \|x_k\|^2} \quad (3)$$

Where x_k and $\hat{x}_k$ are respectively the spectrum of the initial signal and the spectrum of the synthesized signal.

4.1 Results of Experiment 1

Table 2 resumes the results of experiment 1. For denoising the Ola method, either using 25 or 33 cepstra (which correspond to the number of vocal tract cepstral coefficients) have given the same SER results, thus the choice of cepstra does not affect the results since we are working on the whole signal.

Table 2. The SER Results of the two hybrid methods for the operations performed on one sound file

Parameters / Operations	SER for 25 cepstra	SER for 33 cepstra	SER For Ola Function
Synthesis of the ES sound	16,94	16,94	16,94
Dilation	7,63	7,83	3,35
Wiener	10,42	9,81	16.94
DWT+Dilation (of cA)	11,59	10,01	5,83
DWT+Dilation (of cA and cD)	10,34	9,28	2,89
DWT+Wiener (of cA And cD=0)	10,36	10,20	7,23
DWT+Wiener (of cA And cD≠0)	13,01	13,27	14,81
DWT+Wiener (of cA and cD)	9,73	9,81	12,52
DTCWT+Dilation level 1	7,66	7,81	5,43
DTCWT+Dilation level 2	6,86	6,62	8,91
DTCWT+Dilation level 3	12,12	11,66	13,76
DTCWT+Dilation level 4	8,36	7,89	10,09
DTCWT+Dilation level 5	5,70	5,93	6,86
DTCWT+Wiener level 1	7,26	7,21	13,25
DTCWT+Wiener level 2	10,17	10,03	15,33
DTCWT+Wiener level 3	14,19	14,30	16,13
DTCWT+Wiener level 4	10,77	10,66	15,31
DTCWT+Wiener level 5	7,08	7,13	12,74
DTCWT+Wiener level 6	10,14	5,43	14,07
DTCWT+Wiener level 7	15,20	11,36	15,34

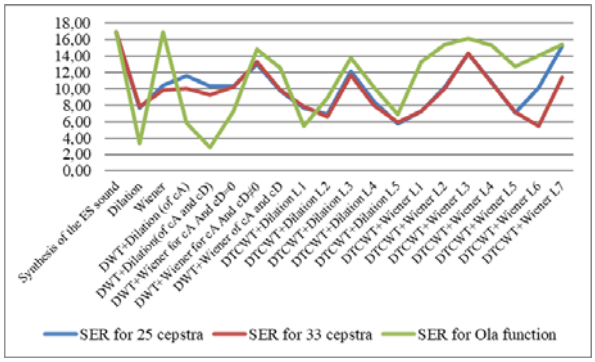

Fig. 6. Variation curves of the SER of hybrid method 1 (25/33 cepstra) and hybrid method 2 (Ola function) according to the operations performed in experiment 1.

Figure 6 shows that the Signal-to-Error Ratio curves for 25 and 33 cepstra are almost identical for all operations performed except for a few operations where the SER of the 25 cepstra exceeds the curve of the 33 cepstra. For both 25 and 33 cepstra, the SER of the combination of the 3rd level of the DTCWT with the Wiener filter gives the highest value compared to the other operations. Even though the last operation that designates the combination of the 7th level of DTCWT with the Wiener filter shows the highest peak of the results for the 25 cepstra, the successive division of the 25 cepstra by the DTCWT to obtain the 7th level decreases the number of cepstra to 1 for each tree, which means that the number of cepstra that underwent a change is almost negligible.

Also, the following operations have the highest SER values: Combining DWT with the Wiener filter applied on the approximation cA and the details cD ≠ 0, as well as combining the 3rd level of the DTCWT with the dilation.

Concerning the operations performed at the level of the Ola function, its Signal-to-Error Ratio curve shows several significant peaks than the SER curves of 25 and 33 cepstra, precisely for the combination of the DTCWT (from level 1 to 7) with the Wiener filter, and the combination of the 3rd level of the DTCWT with the dilation as well as the combination of the DWT with the Wiener filter applied to the cA approximation and the cD ≠ 0 details.

Furthermore, we notice that the application of the Wiener filter alone in the Ola function, without combination with other methods gives the highest peak, which value is identical to the value of the synthesis operation of the original speech sound, which means that the Wiener filter alone could not make any changes to the speech signal due to the nature of the esophageal noise. The combination of the 3rd level of the DTCWT with the Wiener filter shows the highest value and the closest to the synthesis operation value. Similarly, the two combinations of level 2 and level 4 of the DTCWT with the Wiener filter give almost the same high result.

4.2 Results of Experiment 2

Table 2 resumes the results of the average SER of the two hybrid methods for the 289 esophageal speech files from the three corpora GM, HM, and PC.

Table 2. The average of the SER of the operations performed on the three corpora, for the two hybrid methods

Parameters / Operations	The mean SER of 289 (GM corpus)	The mean SER of 289 (MH corpus)	The mean SER of 289 (PC corpus)
Synthesis of the esophageal speech sound	24,03	25,51	33,77
Dilation (Cepstra)	11,85	10,59	9,92
DWT+Dilatation of cA and cD=0	10,23	9,794	8,39
DWT+Dilatation of cA and cD	11,50	11,01	10,34
DWT+Wiener of cA and cD=0	15,15	16,04	12,51
DWT+Wiener (of cA and cD)	17,37	17,88	13,42
DWT+Wiener(cA andcD) (Ola)	19,02	19,02	13,85
DTCWT+Dilation Level 3 (Cepstra)	5,95	8,56	6,38
DTCWT+Dilation Level 3 (Ola)	9,92	11,10	7,99
DTCWT+Wiener Level 3 (Cepstra)	9,76	11,93	9,64
DTCWT+Wiener Level 3 (Ola)	16,82	17,93	16,87

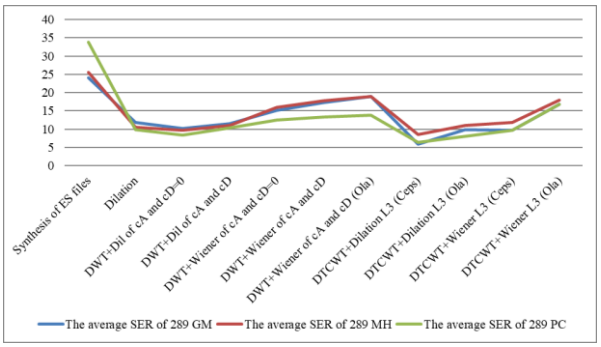

Fig. 7. The average of SER curves of the two hybrid methods for the 289 ES sound files obtained from the PC, MH, and GM corpora

Figure 7 shows that the three SER average curves of the 289 sound files obtained from the three corpora PC, MH, and GM, have the same pattern for all the operations performed. The highest mean values are those of the DWT operations combined with the Wiener filter applied on the approximation cA, and on details cD at the level of the Ola function. In addition to that, the combination of DTCWT level 3 with the Wiener filter at the level of the Ola function and the DWT combined with the Wiener filter applied on the approximation cA with cD = 0 at the level of the cepstra show remarkable results.

5 Conclusion

This study proposes two hybrid methods for reducing the noise of esophageal speech. From the first and second experiment, we can conclude that the application of the Wiener filter or the dilation without combining them with the DTCWT or the DWT (either for the 25

cepstra, the 33 cepstra, or the Ola function) do not show good results. Otherwise, the two proposed hybrid methods which consist of combining the wavelet-based methods along with the Wiener and Dilation method show better results. Also, we notice that discrete and complex wavelets (DWT and DTCWT) show significant denoising results. Furthermore, the second hybrid method shows an important improvement in the results and more precisely the combinations with the Wiener filter. The outcomes of the techniques used in this paper are still to be improved in future work, using new combinations with other denoising techniques and even using neural networks.

References

1. O. Lachhab, *Reconnaissance Statistique de la Parole Continue pour Voix Laryngée et Alaryngée*, tel.archives-ouvertes.fr, (2017) https://tel.archives-ouvertes.fr/tel-01563766/
2. I. Ben Othmane, *Conversion de la voix: Approches et applications*, *tel.archives-ouvertes.fr*, (2019). https://tel.archives-ouvertes.fr/tel-02276259.
3. F. Bahja, *Détection du fondamental de la parole en temps réel: application aux voix pathologiques*, tel.archives-ouvertes.fr, (2013). https://tel.archives-ouvertes.fr/tel-00927147
4. C. Manfredi, M. D'aniello, and P. Bruscaglioni, *Comparison between AR and SVD approaches for speech denoising*, (2001).
5. C. Manfredi, L. Landini, F. Faita, and V. *Gemignani, SVD-based portable device for real-time hoarse voice denoising*, IEEE Xplore, (Jul. 01, 2002)
6. Y. Zhang, J. J. Jiang, and F. A. Feroze, *Wavelet-based denoising for improving nonlinear dynamic analysis of pathological voices*, *ur.booksc.eu*, (2005)
7. M. Shafieian and M. Rahmanian, *An unsupervised approach for improving speech enhancement using wavelet packet transform and adaptive thresholding*, Bdigital2.ula.ve, **26**, no. 3, (2019), doi: pp 92.0200.
8. S.-J. Lee and H.-Y. Kwon, *A Preprocessing Strategy for Denoising of Speech Data Based on Speech Segment Detection*, Applied Sciences, **10**, no. 20, p. 7385, (2020), doi: 10.3390/app10207385.
9. D. W. Griffin and J. S. Lim, *Signal estimation from modified short-time Fourier transform*, IEEE Transactions on Acoustics, Speech, and Signal Processing, **32**, no. 2, pp. 236–243, (1984), doi: 10.1109/TASSP.1984.1164317.
10. Z. Khawaja, *Analyse des états de surface en science des matériaux: caractérisation multi-échelles par ondelette et détermination de l'anisotropie des surfaces*, HAL Archives Ouvertes, (2014). https://hal.archives-ouvertes.fr/tel-01081204/
11. A. Lallouani, *Débruitage d'un signal de la parole corrompu par un bruit coloré en utilisant la transformée en ondelettes et implantation sur un processeur de traitement numérique des signaux*, espace.etsmtl.ca, (2004).
12. I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, *The dual-tree complex wavelet transform*, IEEE Signal Processing Magazine, **22**, no. 6, pp. 123–151, (2005), doi: 10.1109/msp.2005.1550194.
13. P. Loiseau, *Ondelettes complexes pour l'analyse des lois d'échelles*, (2006).
14. K. Nabgha, M. Khannoussi, and A. Tazi, *Bruit et filtrage*, dspace.univ-adrar.edu.dz, (2018).
15. A. Jeanvoine, *Intérêt des algorithmes de réduction de bruit dans l'implant cochléaire : Application à la binauralité,* tel.archives-ouvertes.fr, (2012).
16. J. Balado, P. van Oosterom, L. Díaz-Vilariño, and M. Meijers, *Mathematical morphology directly applied to point cloud data*, ISPRS Journal of Photogrammetry and Remote Sensing, **168**, no. 168, pp. 208–220, (2020), doi: 10.1016/j.isprsjprs.2020.08.011.
17. A. Soni and A. P. Singh, *Automatic Pulmonary Cancer Detection using Prewitt & Morphological Dilation,* 2nd International Conference on Data, Engineering and Applications (IDEA), (Feb. 2020), doi: 10.1109/idea49133.2020.9170680.
18. I. Ben Othmane, J. Di Martino, and K. Ouni, *Enhancement of esophageal speech obtained by a voice conversion technique using time dilated Fourier cepstra*, International Journal of Speech Technology, **22,** no. 1, pp. 99–110, (2018), doi: 10.1007/s10772-018-09579-1.
19. W. Verhelst, *Overlap-add methods for time-scaling of speech*, Speech Communication, **30**, no. 4, pp. 207–221, (2000), doi: 10.1016/s0167-6393(99)00051-5.
20. M. Bahoura, *Efficient FPGA-Based Architecture of the Overlap-Add Method for Short-Time Fourier Analysis/Synthesis*, Electronics, **8**, no. 12, p. 1533, (2019), doi: 10.3390/electronics8121533.
21. E. B. George and M. J. T. Smith, *Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model*, IEEE Transactions on Speech and Audio Processing, **5**, no. 5, pp. 389–406, (1997), doi: 10.1109/89.622558.