Real Time Speech Recognition based on PWP Thresholding and MFCC using SVM

Wafa Helali
Faculty of Sciences of Tunis
University Tunis El-Manar
Tunis, Tunisia

Zied Hajaiej
Faculty of Sciences of Tunis
University Tunis El-Manar,
Tunis, Tunisia

Adnen Cherif
Faculty of Sciences of Tunis
University Tunis El-Manar
Tunis, Tunisia

***Abstract*—The real-time performance of Automatic Speech Recognition (ASR) is a big challenge and needs high computing capability and exhaustive memory consumption. Getting a robust performance against inevitable various difficult situations such as speaker variations, accents, and noise is a tedious task. It's crucial to expand new and efficient approaches for speech signal extraction features and pre-processing. In order to fix the high dependency issue related to processing succeeding steps in ARS and enhance the extracted features' quality, noise robustness can be solved within the ARS extraction block feature, removing implicitly the need for further additional specific compensation parameters or data collection. This paper proposes a new robust acoustic extraction approach development based on a hybrid technique consisting of Perceptual Wavelet Packet (PWP) and Mel Frequency Cepstral Coefficients (MFCCs). The proposed system was implemented on a Rasberry Pi board and its performance was checked in a clean environment, reaching 99% average accuracy. The recognition rate was improved (from 80% to 99%) for the majority of Signal-to-Noise Ratios (SNRs) under real noisy conditions for positive SNRs and considerably improved results especially for negative SNRs.**

Keywords-automatic speech recognition; perceptual wavelet packet transform; Mel frequency cestrum coefficients; SVM; Raspberry Pi 3

I. INTRODUCTION

Speech recognition technology is a widespread dynamic research area. Automatic Speech Recognition (ASR) has been vastly used in many human–machine interaction applications, such as electronics [1], mobile robots [2-4], car audio systems [5], manipulators in industrial assembly lines [6], and security systems [7]. Nonetheless, robust performance constitutes obviously a real concern for any real-time application due to various difficult conditions such as noisy background, accents, and speaker variations. As a result, the need for accuracy, high performance and fast embedded ASR are growing continuously. Many projects have been invested in ASR techniques in order to achieve proficient embedded systems that are able to imitate human behavior at all levels. The ASR accuracy obtained in laboratory environments is quite high, but once the recognition system is placed in a real background, the recognition rate gets roughly low. Several embedded voice recognition systems have been reported and some of them are implemented in Field Programmable Gate Arrays (FPGAs) [8-10] or in Digital Signal Processors (DSPs) [11, 12], all of them with a modest accuracy rate. ASR state-of-the-art systems are linking the performance to reasonable and controlled training conditions. Considering the noise impact, the system accuracy may become unacceptably low in some sensitive environments. Several researchers have shown their interests on speech feature extraction methods such as Linear Prediction Coefficients (LPC) [13], Perceptual Linear Predictive (PLP) [14] and Linear Predictive Cepstral Coefficients (LPCC) which are used due to their effectiveness and simplicity in speech/speaker recognition [15-16].

Mel Frequency Cepstral Coefficients (MFCCs) constitute feature parameters that present widely popular acoustic features mostly used in speech recognition [17]. In spite of its good performance achieved in clean background, the MFCCs feature extraction for speech recognition has been used to enhance speech recognition system performance in noisy environments. The most cited methods are the Cepstral Mean Subtraction (CMS) [18], the Power-Normalized Cepstral Coefficients (PNCCs) [19], and the Cepstral Mean Normalization (CMN) [20] which is a popular feature compensation method dealing with convolutional noise. In this same context, the majority of the published works demonstrated that the wavelet-based feature extraction [21-24] has better performance improvement than traditional Cepstral features in noisy environments. The already presented wavelet-based techniques rely on the multi-resolution PWP properties and combine the extracted MFCC features from various frequency sub bands to a unique feature vector.

In this paper, a new method for real-time speech recognition is proposed under both clean and noisy

Corresponding author: Wafa Helali (wafa.helali@yahoo.fr)

environments, and it is presented and implemented on a Raspberry Pi3 board. The proposed method is based on MFCC extraction from speech signal after applying wavelet thresholding. The main idea relies on obtaining coefficient exploitation which represents the wavelet transform decomposition after eliminating the small coefficients associated with the noise usually located in high frequencies. Then, the MFCCs method is applied to the signal. Finally, a feature vector is acquired by the obtained MFCC concatenation that constitutes one input parameters of the SVM used for classification.

Our main contribution resides in ensuring a good recognition rate, close to 100%, for positive SNRs. In real noisy areas, particularly within the range of [0, -10db], challenging results have been reached using the proposed real time approach. Obviously, real time implementation with Raspberry Pi gives excellent recognition performance in clean and noisy states.

II. FEATURE EXTRACTION

Feature extraction is the process of retaining useful information within a speech signal when rejecting the redundant and unwanted information. It represents merely the speech signal parameterization. This process includes:

- Segmentation of the speech signal into windows.
- Speech signal frequency decomposition into critical bands by transforming it into PWP.
- Parameter extraction.
- Coefficient calculation.

The feature extraction is mostly used, thanks to its better performance for ASR and low computational complexity under standard environment. MFCC and its hybrid feature extraction technique with PWP will be employed. A brief outline of the proposed method is described in Figure 1.

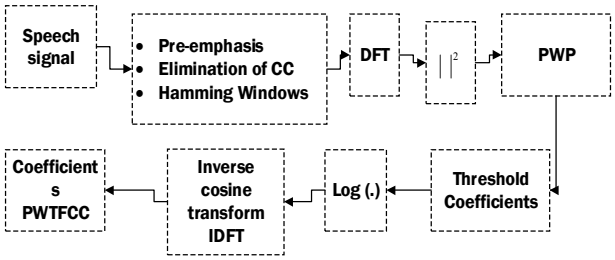

Fig. 1. PWTFCC algorithm.

A. Mel Frequency Cepstral Coefficients Meaning

MFCCs are frequency field features based on the human ear scale. The scale [25] is approximately linear until 1kHz and logarithmic at higher frequencies. These frequency domain features [26] offer more accuracy than time domain ones. In this technique, the same information can be incorporated in less coefficients, making it more compact. The calculation proceeds as described in our previous work [27]. Afterwards, FFT is computed for each speech frame so that signal frequency components could be extracted in time-domain. Then, the logarithmic Mel scaled filter bank is applied to the FFT frame. The log filter bank energies are calculated using the DCT. Only the first thirteen DCT coefficients are kept and the rest are discarded. These DCT coefficients decorrelate the features as well they arrange them in decreasing information order.

B. Perceptual Wavelet Packet

The wavelets offer a technique that represents the time-frequency domain. It has usually been used for signal decomposition into high and low frequency components. Its coefficients depict frequency content similarity measured between a chosen wavelet function and a given signal. These coefficients are calculated as a convolution of the signal and the scaled wavelet function, which can be explained as an expanded band-pass filter due to its band-pass spectrum [27-28]. Subsequently, the resulted wavelet transforms are exploited as a filter bank named Perceptual Wavelet Packet (PWP). The PWP results to a non-redundant restoration, which gives better spectral and spatial localization of signal configuration. Compared with other multi-scale representations such as Gaussian and Laplacian pyramid the PWP represents the privilege of multilevel decomposition, where the signal is decomposed in 'approximation' and 'detail' coefficients at each level [29], through an equivalent process to high-pass and low pass filtering components. As mentioned above, the wavelet transform was introduced for time and frequency analysis of transient signals and it was extended to multi-resolution wavelet transform theory via a Finite Impulse Response (FIR) filter approximation. The discrete wavelets used in multi-resolution analysis constitute an orthonormal basis. The PWP decomposition steps are explicated taking into account details and approximation coefficients.

III. REAL TIME IMPLEMENTATION SLANT

The proposed speech recognition system's block diagram is illustrated in Figure 2. The various system steps are explained in this section. The microphone input speech is sampled at 16kHz. First of all, we mention that a Voice Activity Detector (VAD) is used as a noise estimator. The VAD's output presents the binary signal resulting of the comparison between the speech input signal and the threshold value. Thus, VAD value is either true (VAD=1) when the measured input is greater than the threshold and the signal is considered as a voiced frame, or the VAD value is false (VAD=0) and the signal frame is considered as a noisy frame. The second approach step consists on speech signal decomposition with the PWP. The PWP outcome is a multilevel decomposition, in which the signal is divided into 'approximation' and 'detail' coefficients at every stage. This process is similar to low-pass and high pass filtering. The simplest way to remove noise is by using the wavelet coefficients, which are the result of the wavelet transform decomposition. The small coefficients associated with the noise through the threshold step are eliminated. Indeed, the threshold purpose offers the ideal components from the noisy signal giving the noise level estimation. There are various threshold methods. Between the most commonly used are the hard and soft threshold. They are used and adopted in this work and modeled by:

$$y = sign(x)(|x| - \lambda) \quad (1)$$

where x, y and λ present respectively the input signal, the threshold signal, and the threshold value. The MFCCs are applied to the signal after the threshold and concatenation steps. The signal is filtered and windowed by the hamming window for FFT transformation. Next, the signal passes through a Mel-filter to obtain the twelve Cepstral coefficients. Finally, the resulted Cepstral coefficients are concatenated to construct the SVM classifier input. Similarly, this technique is applied also to our proper training speech database containing spoken words which are recorded by a mono-speaker.

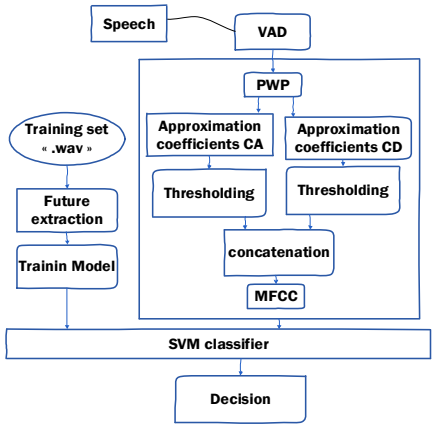

Fig. 2. The PWP decomposition steps of a 1D signal for three levels.

In order to increase the performance of our proposed speech recognition algorithm, a denoising module was added to the proposed system to enhance its robustness. This denoising module relies on Adaptive Median Filtering (AMF) [30] which is able to eliminate the data speckle noise without harming the embedded sharp contrasts. It's noticeable that the noise impact can be significantly reduced by applying the AMF to the temporal modulation spectrum, which is the Fourier transform for either real or imaginary acoustic spectrograms along the time axis. Thus, the resulting speech features can be more noise-robust and give better speech recognition performance. Figure 3 represents the modified speech recognition system in which the AMF is introduced as a speech denoising module.

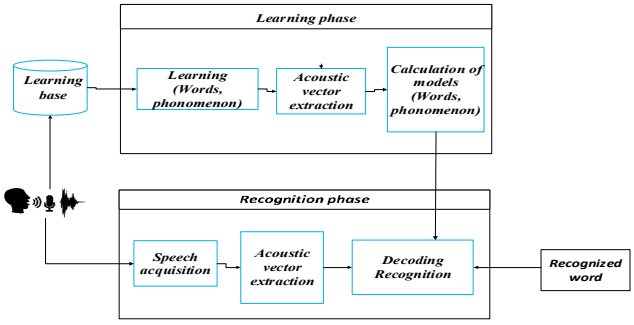

Fig. 3. The modified speech recognition system.

IV. TEST PERFORMING AND OBTAINED RESULTS

In order to build the speech recognition system, voice commands and speech models have to be optimized based on a solid training database. In this experiment, the training database contained eleven commands recorded five times by a mono-speaker for a voice command under a silent environment. Each recorded data consisted of up to 4s of utterance. The speech recognition application needs more than just the simulation and the proposed algorithm was tested on a particularly suitable flexible platform. The complete setup has been implemented and tested on a Raspberry Pi 3 board.

A. Used Raspberry Pi Card Synopsis and System Pattern

The Raspberry Pi 3 is simply a performed sized card processor [31], containing a micro-controller and a CPU. The Raspberry Pi processor core system is a Broadcom BCM2837 System-on-Chip (SoC) multimedia processor, which has 64-bit quad-core ARMv8 Cortex A53 with 1GB of RAM. Besides, it's equipped with 16GB expandable to 128GB. An SD card slot ,1.2GHz SoC processor, Video Core IV GPU, 4 USB ports, 1 HDMI port, 40 GPIO pins which could be configured as digital output or input and a jack audio output. The Raspberry Pi is controlled by an amended version of Linux (Raspbian) optimized for the ARM architecture. As Raspbian is built based on Debian, it implicitly has all the compatibilities and features required for the program. Python 2.7 or 3.5 is already installed in the Raspbian operating system and therefore a new installation is not compulsory. Python 2.7 was selected because it owns more store community support accessible contrary to Python 3.5. The project requires some external Python packages that need to be separately installed. We have also installed some other measurement packages in order to evaluate the program performance. All specifications are mentioned in Table I.

TABLE I. RASPBERRY PI'S SOFTWARE SPECIFICATION FOR THE PROPOSED FRAMEWORK [30]

Name	Configuration
OS	Noobs (Rasbian)
Programming language	Python 2.7
Libraries	Numpy, SciPy, PyLab, Matplotlib, RPI.GPIO
Audio libraries	Pyaudio, Pydub, Wave
Performance Monitoring Utilities	BCMStat, TIME, htop

B. Real-time Performance

In order to validate the proposed speech recognition, based on MF-PWP/MFCC, algorithm's performance, a comparison was made of the improvements in speech recognition accuracy that can be obtained through the use of several types of features such as MFCC, PWP/MFCC and MF-MFCC. The recognition experiments were performed using noisy testing data with various noisy conditions: white Gaussian and babble noise, with a noise ratio (SNR) from -10db to +10db. Figure 4 compares the results for speech in white (Figure4(a)) and babble noise (Figure 4(b)) under different SNRs for several methods. Specifically, the recognition accuracy percent was compared for PWP/MFCC and MF-PWP/MFCC methods as described above, along with baseline MFCC and MF-MFCC. It

can be seen that the PWP/MFCC processing provides better accuracy than MFCC features for all the tested noises, although improvements are small in high SNRs. The lack of improvement observed for clean speech and high SNRs is a common observation for many approaches to robust speech recognition. It is also noted that the denoising module provides a trivial improvement in recognition accuracy expressly in lower SNRs.

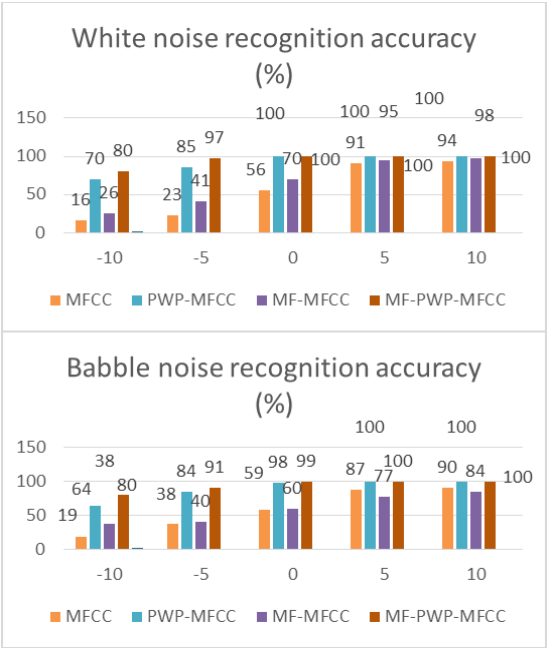

Fig. 4. Comparison of recognition accuracy in: (a) white noise and (b) babble noise for several feature extraction methods.

Finally, it can be observed that the proposed features based on MF-PWP/MFCC perform better than other features under all test conditions. With clean and noisy data testing, we can obtain a great and expectant recognition rate with the MF-PWP/MFCC for real-time speech recognition system. Aiming to validate the proposed speech recognition algorithm performance with several feature extractions we have measured the memory use, the CPU use, and the execution time. Table II presents the CPU and memory use. This verification is obtained using htop, a popular Linux text mode utility, which is ideal for monitoring system processes and performance metrics.

TABLE II. COMPARISON OF RESOURCE CONSUMPTION AND EXECUTION TIME FOR DIFFERENT FEATURE EXTRACTION METHODS

Algorithm	Memory consumption (bytes)	CPU usage (%)	Memory usage (%)	Extraction time (ms)
MFCC	8460	9.2	5.8	670
PWP/MFCC	8486	9.2	5.8	675
MF-MFCC	8500	10.6	6.4	678
MF-PWP/MFCC	8580	10.9	6.4	680

Table II shows that the average CPU usage is 10.9% in MF-PWP/MFCC. On the other hand, in MFCC, PWP/MFCC and MF-MFCC, it is around 9.2%, 10.6% and 10.9% respectively. In addition, the maximum time execution difference of the proposed algorithm to the other algorithms doesn't exceed 15ms. It was noticed that this low difference in time execution and resources consumption did not affect the proposed algorithm's robustness.

C. Recognition Rate Comparison

The negative recognition rate part was given much attention and it represents the main contribution of this study. The comparison with the work in [30]is shown in Table III:

TABLE III. RECOGNITION RATE RESULT COMPARISON IN BABBLE NOISE

SNR (db)	MFCC	MFCC [30]	PWP-MFCC	PWP-MFCC [30]	MF-MFCC	MF-MFCC [30]	MF-PWP-MFCC	MF-PWP-MFCC [30]
-10	19	10.9	64	61.81	38	10.09	80	65.45
-5	38	30.9	84	78.18	40	37.27	91	80.09
0	59	59.09	98	85.45	60	68.15	99	97.27
5	87	68.15	100	90	77	70.9	100	100
10	90	76.36	100	93.65	84	85.45	100	100

Generally, the published works do not take into account the range [0,-10db]. Recognized words in this noisy area are very hard to extract. Although, the recognition rate within the range [0, 10db] reaches nearly 100% which is also the current case.

V. CONCLUSION

A new real-time speech recognition algorithm has been proposed in this paper. The proposed algorithm exploits the PWP combined with MFCC in order to match speech features in addition to the SVM classification block. The proposed method proves its effectiveness to pick up an ideal recognition rate of about 100% in clean environment. The recognition rate ranges from 98.18% to 100%, even in noisy environments from 0db to10db with the use of adaptive median filter as a denoising module. In the real noisy part, principally inside the range [0, -10db], good results have been reached with the proposed real time method. For real-time experimentation a Raspberry Pi has been used as the hardware platform. The proposed system's performance was sufficient for a wide range of speech-controlled applications. As future work, resource consumption and its impact of speech embedded applications in addition to accuracy and timing will be investigated.

REFERENCES

[1] D. Karaboga and E. Kaya, "Adaptive network based fuzzy inference system (ANFIS) training approaches: a comprehensive survey," *Artificial Intelligence Review*, vol. 52, no. 4, pp. 2263–2293, Dec. 2019, doi: 10.1007/s10462-017-9610-2.

[2] H. A. Yanco, A. Norton, W. Ober, D. Shane, A. Skinner, and J. Vice, "Analysis of Human-robot Interaction at the DARPA Robotics Challenge Trials," *Journal of Field Robotics*, vol. 32, no. 3, pp. 420–444, May 2015, doi: 10.1002/rob.21568.

[3] A. Pereira, C. Oertel, L. Fermoselle, J. Mendelson, and J. Gustafson, "Responsive Joint Attention in Human-Robot Interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2019, pp. 1080–1087, doi: 10.1109/IROS40897.2019.8968130.

[4] I. Tiddi, E. Bastianelli, E. Daga, M. d'Aquin, and E. Motta, "Robot–City Interaction: Mapping the Research Landscape—A Survey of the

Interactions Between Robots and Modern Cities," *International Journal of Social Robotics*, vol. 12, no. 2, pp. 299–324, May 2020, doi: 10.1007/s12369-019-00534-x.

[5] Y. Zheng, Y. Liu, and J. H. L. Hansen, "Navigation-orientated natural spoken language understanding for intelligent vehicle dialogue," in *2017 IEEE Intelligent Vehicles Symposium (IV)*, Jun. 2017, pp. 559–564, doi: 10.1109/IVS.2017.7995777.

[6] T. Hino, S. Ito, T. Liu, and M. Maeda, "Set-based particle swarm optimization with status memory for knapsack problem," *Artificial Life and Robotics*, vol. 21, no. 1, pp. 98–105, Mar. 2016, doi: 10.1007/s10015-015-0253-6.

[7] A. Koduru, H. B. Valiveti, and A. K. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology*, vol. 23, no. 1, pp. 45–55, Mar. 2020, doi: 10.1007/s10772-020-09672-4.

[8] S. Zhu, C. Xu, J. Wang, Y. Xiao, and F. Ma, "Research and application of combined kernel SVM in dynamic voiceprint password authentication system," in *2017 IEEE 9th International Conference on Communication Software and Networks (ICCSN)*, May 2017, pp. 1052–1055, doi: 10.1109/ICCSN.2017.8230271.

[9] E. Rodríguez-Orozco *et al.*, "FPGA-based Chaotic Cryptosystem by Using Voice Recognition as Access Key," *Electronics*, vol. 7, no. 12, p. 414, Dec. 2018, doi: 10.3390/electronics7120414.

[10] Q. Li *et al.*, "MSP-MFCC: Energy-Efficient MFCC Feature Extraction Method With Mixed-Signal Processing Architecture for Wearable Speech Recognition Applications," *IEEE Access*, vol. 8, pp. 48720–48730, 2020, doi: 10.1109/ACCESS.2020.2979799.

[11] P. J. Dugan, H. Klinck, J. A. Zollweg, and C. W. Clark, "Data Mining Sound Archives: A New Scalable Algorithm for Parallel-Distributing Processing," in *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 768–772, doi: 10.1109/ICDMW.2015.235.

[12] K. Gupta and D. Gupta, "An analysis on LPC, RASTA and MFCC techniques in Automatic Speech recognition system," in *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, Jan. 2016, pp. 493–497, doi: 10.1109/CONFLUENCE.2016.7508170.

[13] S. P. Panda, A. K. Nayak, and S. C. Rai, "A survey on speech synthesis techniques in Indian languages," *Multimedia Systems*, vol. 26, no. 4, pp. 453–478, Aug. 2020, doi: 10.1007/s00530-020-00659-4.

[14] V. M. Patel, N. K. Ratha, and R. Chellappa, "Cancelable Biometrics: A review," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 54–65, Sep. 2015, doi: 10.1109/MSP.2015.2434151.

[15] V. M. Patel, N. K. Ratha, and R. Chellappa, "Cancelable Biometrics: A review," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 54–65, Sep. 2015, doi: 10.1109/MSP.2015.2434151.

[16] L. Jiao *et al.*, "A Survey of Deep Learning-Based Object Detection," *IEEE Access*, vol. 7, pp. 128837–128868, 2019, doi: 10.1109/ACCESS.2019.2939201.

[17] R. Chakroun and M. Frikha, "Efficient text-independent speaker recognition with short utterances in both clean and uncontrolled environments," *Multimedia Tools and Applications*, vol. 79, no. 29, pp. 21279–21298, Aug. 2020, doi: 10.1007/s11042-020-08824-7.

[18] C. Kim and R. M. Stern, "Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 7, pp. 1315–1329, Jul. 2016, doi: 10.1109/TASLP.2016.2545928.

[19] S.-S. Wang, P. Lin, Y. Tsao, J.-W. Hung, and B. Su, "Suppression by Selecting Wavelets for Feature Compression in Distributed Speech Recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 3, pp. 564–579, Mar. 2018, doi: 10.1109/TASLP.2017.2779787.

[20] M. A. Islam, W. A. Jassim, N. S. Cheok, and M. S. A. Zilany, "A Robust Speaker Identification System Using the Responses from a Model of the Auditory Periphery," *PLoS One*, vol. 11, no. 7, p. e0158520, Jul. 2016, doi: /10.1371/journal.pone.0158520.

[21] N. Das, S. Chakraborty, J. Chaki, N. Padhy, and N. Dey, "Fundamentals, present and future perspectives of speech enhancement," *International Journal of Speech Technology*, Jan. 2020, doi: 10.1007/s10772-020-09674-2.

[22] C. Jiang, L. Ba, X. Tang, and D. Wen, "Speaker Verification Using IMNMF and MFCC with Feature Warping Under Noisy Environment," in *2018 Chinese Automation Congress (CAC)*, Nov. 2018, pp. 2583–2588, doi: 10.1109/CAC.2018.8623278.

[23] A. K. H. Al-Ali, V. Chandran, and G. R. Naik, "Enhanced forensic speaker verification performance using the ICA-EBM algorithm under noisy and reverberant environments," *Evolutionary Intelligence*, May 2020, doi: 10.1007/s12065-020-00406-8.

[24] O. Mamyrbayev, A. Toleu, G. Tolegen, and N. Mekebayev, "Neural architectures for gender detection and speaker identification," *Cogent Engineering*, vol. 7, no. 1, p. 1727168, Jan. 2020, doi: 10.1080/23311916.2020.1727168.

[25] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, 1 edition. Englewood Cliffs, N.J: Pearson, 1993.

[26] N. Holighaus, G. Koliander, Z. Průša, and L. D. Abreu, "Characterization of Analytic Wavelet Transforms and a New Phaseless Reconstruction Algorithm," *IEEE Transactions on Signal Processing*, vol. 67, no. 15, pp. 3894–3908, Aug. 2019, doi: 10.1109/TSP.2019.2920611.

[27] W. Helali, Z. Hajaiej, and A. Cherif, "Automatic Speech Recognition System Based on Hybrid Feature Extraction Techniques Using TEO-PWP for in Real Noisy Environment," *IJCSNS - International Journal of Computer Science and Network Security*, vol. 19, no. 10, pp. 118–124, Oct. 2019.

[28] A. Rinoshika and H. Rinoshika, "Application of multi-dimensional wavelet transform to fluid mechanics," *Theoretical and Applied Mechanics Letters*, vol. 10, no. 2, pp. 98–115, Jan. 2020, doi: 10.1016/j.taml.2020.01.017.

[29] D. G. Manolakis and V. K. Ingle, *Applied Digital Signal Processing: Theory and Practice*, 1 edition. New York: Cambridge University Press, 2011.

[30] A. Mnassri, M. Bennasr, and C. Adnane, "A Robust Feature Extraction Method for Real-Time Speech Recognition System on a Raspberry Pi 3 Board," *Engineering, Technology & Applied Science Research*, vol. 9, no. 2, pp. 4066–4070, Apr. 2019.

[31] S. N. Truong, "A Low-cost Artificial Neural Network Model for Raspberry Pi," *Engineering, Technology & Applied Science Research*, vol. 10, no. 2, pp. 5466–5469, Apr. 2020.