A Robust Feature Extraction Method for Real-Time Speech Recognition System on a Raspberry Pi 3 Board

Aymen Mnassri
Department of Physics,
Faculty of Sciences of Tunis,
Tunis-El Manar University, Tunisia
mnassri.aymen46@gmail.com

Mohamed Bennasr
Department of Physics,
Faculty of Sciences of Tunis,
Tunis-El Manar University, Tunisia
bennasr.mouhamed@gmail.com

Cherif Adnane
Department of Physics,
Faculty of Sciences of Tunis,
Tunis-El Manar University, Tunisia
adnene.cher@fst.rnu.tn

***Abstract*—The development of a real-time automatic speech recognition system (ASR) better adapted to environmental variabilities, such as noisy surroundings, speaker variations and accents has become a high priority. Robustness is required, and it can be performed at the feature extraction stage which avoids the need for other pre-processing steps. In this paper, a new robust feature extraction method for real-time ASR system is presented. A combination of Mel-frequency cepstral coefficients (MFCC) and discrete wavelet transform (DWT) is proposed. This hybrid system can conserve more extracted speech features which tend to be invariant to noise. The main idea is to extract MFCC features by denoising the obtained coefficients in the wavelet domain by using a median filter (MF). The proposed system has been implemented on Raspberry Pi 3 which is a suitable platform for real-time requirements. The experiments showed a high recognition rate (100%) in clean environment and satisfying results (ranging from 80% to 100%) in noisy environments at different signal to noise ratios (SNRs).**

Keywords-automatic speech recognition; discrete wavelet transform; Mel frequency cestrum coefficients; median filter; support vector machines; Raspberry Pi

I. INTRODUCTION

Speech recognition technology has been widely spread and has been applied in many research areas such as mobile robots [1-33], consumer electronics [4], car audio systems [5], security system manipulators [6], and manipulators in industrial assembly lines [7]. However, robust performance poses a problem for any real-time application due to various conditions such as noisy background, different accents, and speaker variations. Thus, high performance ASR systems are required. The accuracy of such systems obtained in laboratory environments is not sufficient. Moreover, it is practically lower in real conditions. Recently, several embedded ASR systems have been presented. Some of them have been implemented on digital signal processors (DSPs) [8-10], while others have been deployed on field programmable gate arrays (FPGAs) [11-13] without satisfactory enough accuracy rates. State-of-the-art on ASR provides good performance if the used and trained conditions are reasonably controlled and similar, and their performance is degraded in the case of mismatching conditions. In many recent works, the speech feature extraction methods have been discussed, such as the linear prediction coefficients (LPCs), the relative spectral-perceptual linear prediction (RASTA-PLP) [14], and the linear-predictive cepstral coefficients (LPCCs) which have been used because of their efficiency and simplicity in speech and speaker recognition [15]. Another technique that has been widely discussed in the speech recognition research area, is called the Mel-frequency cepstral coefficients (MFCC) [16]. In spite of its good performance in clean background condition, the MFCC's feature extraction for speech recognition is weak in noisy environments. In previous researches, a number of feature extraction methods have been proposed in order to enhance ASR performance under noisy conditions. The most cited ones are the power-normalized cepstral coefficients (PNCC) [17], the cepstral mean subtraction (CMS) [18], and the cepstral mean normalization (CMN) [19] which is considered as one of the most popular feature extraction techniques for dealing with convolutional noises. Furthermore, many works have highlighted the wavelet-based feature extraction method [20-23] which has led to reach better performance improvements in comparison to traditional cepstral features.

In this paper, a new robust feature extraction method for a real-time speech recognition system is presented and implemented on a Raspberry Pi 3 board. At first, the DWT transforms the received speech signal to the wavelet domain. Then, an iterative median filter is carried out over all output wavelet transformation coefficients to remove noisy samples. In the next state, the obtained coefficients are assembled into one vector and across the MFCC procedure to produce the final result. The proposed method provides excellent recognition rate under clean and noisy states. This can conserve more speech signal features which will be robust against noise effects This can conserve more speech signal features which will be robust against noisy effects. This method has a stronger ability in separating signal from noise compared to the threshold denoising method which preserves the details. The applied noise elimination method can improve the denoising effect which is one of the main speech recognition system objectives. Finally, a feature vector is carried out via the obtained MFCC

Corresponding author: Aymen Mnassri

coefficients concatenation. The performed vector will be used as an input parameter for the SVM multi-class. The proposed method has outstanding performance in clean and noisy environments.

II. PROPOSED ALGORITHM METHODOLOGY

Figure 1 shows the procedure of the proposed speech recognition system. The main idea is to give more significant coefficients for the performance of the system in a clean and noisy environment. At first, the input speech from a microphone is sampled at 16KHz. Voice activity detector (VAD) is applied as silence detector to only keep signal parts where the speech is heard. The VAD output presents a binary activity founded on the comparison between the input speech signal and the threshold value. Thus, the VAD value is true (VAD=1) when the measured input is superior than the threshold. In this case, the signal is considered as a soundtrack. However, when the VAD value is false (VAD=0), the signal frame is considered as a silent frame.

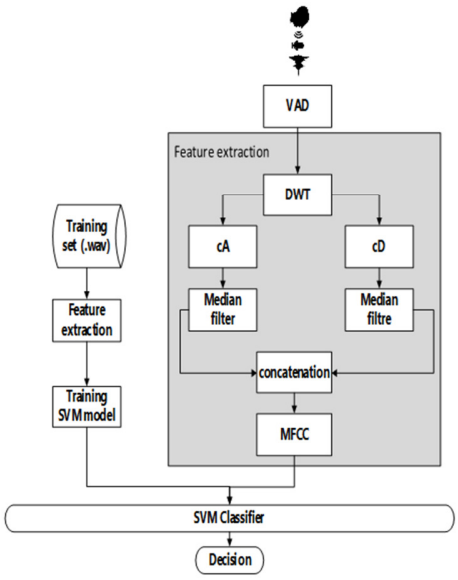

Fig. 1. Block diagram of the proposed speech recognition system

After receiving the speech signal, the next step consists of decomposing the speech signal using DWT. The DWT result is a multilevel decomposition, in which the signal is separated into an 'approximation' and a 'detail' coefficient at each level. This is fabricated from a similar process to low and high pass filtering, respectively. In this step, a median filter is performed to detail the approximation coefficients separately and to ensure a robust noisy speech removing process. The applied filter expression is indicated by (1):

$$\tilde{x}_t = med(x_{t-d},...,x_t,....,x_{t+d}) \quad (1)$$

where *med(..,..,..)* defines the median operator used to select elements within the sliding window and to take out the d^{th} largest one.

Median filtering (MF) is a rank-order filtering method which proves its ability to eliminate data speckle noise without harming the embedded sharp contrasts. The main concept of MF is the use of a $(2d+1)$ point sliding window to range through a signal, changing every signal point with a sliding window median value which is centered at that point. The MFCC function has then been used to extract the characteristics from the DWT output vector after the median filter application. Finally, the obtained cepstral coefficients are concatenated to construct one input for the SVM-classifier. This method is applied also to our proper training speech database containing Arabic speech words which are recorded by multiple speakers (women and men) for voice command. For the samples' classification, the "one against one" and "one against all" approaches were used.

A. Feature Extraction

The main objective of all feature extraction techniques is to select relevant and robust characteristic coefficients from the spoken utterance to improve the performance of the speech recognition system. In the present work, a new and robust hybrid feature extraction based on the MFCC and the DWT techniques was employed. The used experimental parameters of these techniques are presented below.

1) Discrete Wavelet Transform

DWT technique is a very popular tool used for the analysis of non-stationary signals. It can be thought of as an equivalent of filtering the speech signal with a bank of band pass filters, whose impulse responses are all approximately given by scaled versions of a mother wavelet [24]. In this work, the DWT will be used as a feature extraction tool because of its simplicity and reduced computation time. To compute the wavelet features, the decomposition is performed by one level, owing to the fact that the coefficients in this level include the essential frequency constituent of the speech signal. Various mother wavelets are used and validated for the decomposition such as Harr, Symlets and Daubechies. The latter presents the most suitable mother wavelet for our application, in particular to the first order dB1. After applying the Daubechies wavelets and extracting the features from each input speech command, the obtained coefficients were concatenated and used as the input vector for the next process.

2) Mel Frequency Cepstral Coefficients

The analysis of the MFCC coefficients is the most popular and powerful in terms of demonstration that aims to present the data by coefficient vectors.

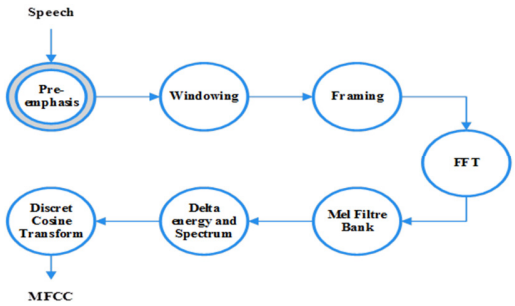

Fig. 2. Calculation of the MFCC coefficients

The cepstral coefficients (MFCC) are frequently used in the ASR domain due to their low complexity of the estimation algorithm and their high performance. In addition, the MFCC representation can describe the structure of the human auditory system better than the traditional linear predictive features [25]. The calculation process can be implemented as shown in Figure 2. Table I illustrates the experimental parameters used for extracting the MFCC coefficients.

TABLE I. EXPERIMENTAL PARAMETERS USED FOR EXTRACTING THE MFCC COEFFICIENTS

Parameters	Value
Coefficient of preemphasis filter	0, 9
Ccepstrum number	12 coefficients
Analysis window length	20ms
Analysis window	Hamming
Recovery (%)	50 %

B. Classification

Regarding classification, our choice is the SVM method, which is a very interesting progress in terms of implementation and extension to multi-class problems. The SVM method is applied with success to applications such as facial recognition and handwriting recognition. Furthermore, the SVM is a kernel technique, which means that the hyperplane is found in a feature space using a non-linear transformation that aims to transform the input space in a feature space which has a much bigger dimension. Machine learning methods (SVMs) are basically based on binary classifiers, whereas the acoustic classifier step in the ASR can be fixed as a multiclass problem. In order to extend the binary algorithm to multiclass decision, we used two approaches which are: the "one-against-one" (OAO) [26] and the "one-against-all" (OAA) [27]. The OAA approach compares each class with the other classes in order to categorize k classes and to construct k binary SVM classifiers. In contrast, the OAO approach consists in employing a classifier for each pair of classes. For the k-classes problem, the "OAO" method discriminates the samples of one class from the samples of another class, so that k(k−1)/2 SVMs are constructed. In this work, the classification accuracy is evaluated with two approaches of multi-class SVM's by using the Gaussian kernel. The parameters of this kernel (c, σ) are fixed according to the optimization obtained by the genetic algorithm described in [28].

III. REAL TIME IMPLEMENTATION

This part is dedicated to the hardware and software description used for the real-time implementation. To prove the reliability of the proposed speech recognition method, a real-time simulation is required. For that, the adopted algorithm has been tested on a flexible embedded Raspberry Pi 3 board which seems to be convenient with the particularity of our application. The Raspberry Pi system is a Broadcom BCM2837 system-on-chip (SoC) multimedia processor, which has 64-bit quad-core ARMv8 Cortex A53 with 1GB of RAM, 16GB (expandable to 128GB) SD card slot, 1.2GHz processor (SoC). Table II represents the Raspberry Pi software specifications for the proposed framework. To test the implemented Raspberry Pi algorithm, the experiment data acquired in real-time environment are given to calculate response time. For that, a prototype robot similar to a wheelchair has been designed and tested using the Raspberry Pi board (Figure 1). For the control of the moving robot, Arabic voice commands have been set up to regarding direction and robot speed. Two male and two female speakers were requested to issue voice commands under silent environment (Table III).

TABLE II. RASPBERRY PI SOFTWARE SPECIFICATIONS

Name	Configuration
Operating system	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

TABLE III. VOICE COMMANDS

Arabic commands	Robot action
تقدم	Go forward
امام	Go forward with speed
أسرع	Speed
يسار	Go left
يمين	Go right
استدر	Turn
توقف	Stop
تراجع	Go back with speed
خلف	Go back
واصل	Keep going

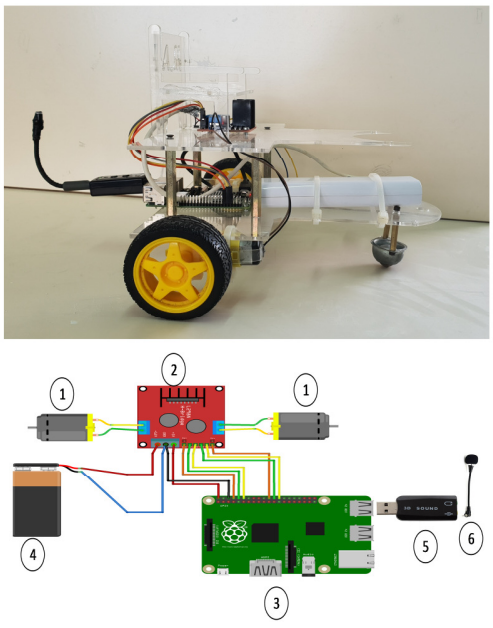

Fig. 3. Hardware architecture of wheelchair prototype

The wheelchair prototype (Figure 1) contains two motors (1) used for the left and the right wheels. These motors are controlled separately by an H-Bridge Motor based on the L298N component (2). The CPU board can dictate the rotation speeds of the motors individually using pulse-width modulation (PWM). The H-bridge motor and the Raspberry Pi board (3) are directly powered by the power bank battery (4). The Raspberry Pi 3 requires a 5V DC power source with typical

bare-board active current consumption of 400mA and the H-bridge motor requires approximately +6V of minimum power supply for operation. The voice acquisition command is done using a microphone (6) with an external USB sound-card (5).

IV. TESTS AND RESULTS

To prove the performance of our proposed speech recognition algorithm based on MF-DWT/MFCC, we compared it through the use of several types of features such as MFCC, DWT/MFCC and MF-MFCC based on the two multiclass approaches, OAA and OAO The recognition experiments were performed using noisy testing data with different various noisy conditions: white Gaussian and babble noise, with a noise ratio (SNR) ranging from -10db to 10db. A comparatives study between feature extraction (MFCC, MFMCC, DWT-MFCC and MFDWT-MFCC) methods in babble and white noise states under different SNRs is summarized in Tables IV and V. This comparison is obtained according to the classification rate based on the OAO and OAA approaches. It can be observed that the performance of MFCC, MFMCC and DWT-MFCC methods degrades considerably in the presence of noise for SNRs less than 0dB. On the contrary, our proposed method based on MFDWT-MFCC gives good results even for SNRs lower than 0dB. It can increase the classification rate up to 23% in comparison with DWT-MFCC and more than 50% compared to MFMFCC for white noise with a SNR of -5dB. The proposed feature extraction gives the best results for the two multi-class SVM based methods (OAO, OAA). These results justify the importance of the use of the denoising module with median filter in wavelet domain which provides a significant improvement in recognition accuracy, especially in lower SNRs. We can conclude that the proposed feature extraction based on MFDWT-MFCC performs better than the other ones all test conditions, with noisy and clean testing data. Therefore, we can obtain a perfect and encouraged reconnaissance rate with the MFDWT-MFCC for a real-time voice command system.

TABLE IV. CLASSIFICATION ACCURACY IN BABBLE NOISE FOR FEATURE EXTRACTION METHODS USING OAA AND OAO APPROCHES

	Classification rate (%)									
	OAA					**OAO**				
SNR (dB) Features	-10	-5	0	5	10	-10	-5	0	5	10
MFCC	10.9	30.90	59.09	68.13	76.36	27.27	34.54	54.54	74.54	87.27
DWT-MFCC	61.81	78.18	85.45	90	93.63	60	83.63	94.54	97.27	100
MFMFCC	10.9	37.27	68.13	70.90	85.45	33.63	41.81	60	75.45	87.27
MFDWT-MFCC	65.45	80.09	97.27	100	100	80.09	88.18	97.27	99.09	100

TABLE V. CLASSIFICATION ACCURACY IN WHITE NOISE FOR FEATURE EXTRACTION METHOD USING OAA AND OAO APPROCHES

	Classification rate (%)									
	OAA					**OAO**				
SNR (dB) Features	-10	-5	0	5	10	-10	-5	0	5	10
MFCC	10.09	27.27	43.63	90.90	90.90	10,09	30	54	83.63	90.90
DWT-MFCC	69	70.90	85.45	98.18	100	69	92.72	97.27	100	100
MFMFCC	27.27	43.63	69.09	92.72	94.54	27.27	41.81	69.09	92.72	94.54
MFDWT-MFCC	80	94.54	98.18	100	100	80	96.36	100	100	100

In Table VI, the results of our tests regarding CPU usage, memory performance, and execution time required to run the proposed real-time speech recognition system, are summarized. To validate performances, our choice was HTOP which is an interactive process for Linux and a very reliable monitoring tool for performance metrics. The results show the average usage of our Raspberry pi CPU which is 10.8% when tested with the proposed algorithm. When using MFCC, DWT-MFCC and MFMFCC, the CPU usage is at 9.1%, 9.5% and 10.3% respectively. Also, the recorded maximum execution time does not exceed 15ms when compared with the other algorithms. According to these low differences in resource consumption and execution time, we conclude that they will have no effect on our algorithm quality in real-time usage.

TABLE VI. RESOURSE CONSUMPTION AND EXECUTION TIME COMPARISON

Algorithm	**Memory consumption (Byte)**	**Memory usage (%)**	**CPU usage (%)**	**Running time (ms)**
MFCC	8325	9,1	5,9	645
DWT/MFCC	8376	9,1	5,9	650
MF-MFCC	8450	10,3	6,2	652
MF- DWT/MFCC	8499	10,8	6,2	660

V. CONCLUSION

In this paper, a new robust feature extraction system for real-time speech recognition has been presented. This technique is based on a hybridization between DWT and MFCC. The proposed system proves its reliability by having 100% recognition rate when tested in clean environment. The recognition rate ranges from 80% to 100% in noisy environments from -10 dB to 10dB. Regarding the real-time problem the performances of the proposed methodology were evaluated by implementing a voice command application using a wheel chair prototype based on a Raspberry Pi card. Results show that the proposed methodology has significantly sufficient performance in terms of recognition rate, recognition processing time and resource consumption.

REFERENCES

[1] M. Kos, M. Rojc, A. Zgank, Z. Kacic, D. Vlaj, "A speech-based distributed architecture platform for an intelligent ambience", Computers & Electrical Engineering, Vol. 71, pp. 818-832, 2018

[2] K. Bader, B. Lussier, W. Schon, "A fault tolerant architecture for data fusion: A real application of Kalman filters for mobile robot localization", Robotics and Autonomous Systems, Vol. 88, pp. 11-23, 2017

[3] B. Jensen, N. Tomatis, L. Mayor, A. Drygajlo, R. Siegwart, "Robots meet humans-Interaction in public spaces", IEEE Transactions on Industrial Electronics, Vol. 52, No. 6, pp. 1530–1546, 2005

[4] H. K. Lam, F. H. Leung, "Design and training for combinational neurallogic systems", IEEE Transactions on Industrial Electronics, Vol. 54, No. 1, pp. 612-619, 2007

[5] N. Hataoka, Y. Obuchi, T. Mitamura, E. Nyberg, "Robust speech dialog interface for car telematics service", First IEEE Consumer Communications and Networking Conference, Las Vegas, USA, January 5-8, 2004

[6] K. Saeed, M. Nammous, "A speech-and-speaker identification system: Feature extraction, description, and classification of speech-signal image", IEEE Transactions on Industrial Electronics, Vol. 54, No. 2, pp. 887–897, 2007

[7] D. Yongda, L. Fang, X. Huang, "Research on multimodal human-robot interaction based on speech and gesture", Computers & Electrical Engineering, Vol. 72, pp. 443-454, 2018

[8] J. Manikandan, B. Venkataramani, K. Girish, H. Karthic, V. Siddharth, "Hardware implementation of real-time speech recognition system using TMS320C6713 DSP", 24th Internatioal Conference on VLSI Design, Chennai, India, January 2-7, 2011

[9] C. C. Shen, W. Plishker, S. S. Bhattacharyya, "Design and optimization of a distributed, embedded speech recognition system", IEEE International Symposium on Parallel and Distributed Processing, Miami, USA, April 14-18, 2008

[10] B. Kamdar, B. Mirchandani, D. Shah, Y. S. Rao, "Real time speech recognition using IIR digital filters implemented on an embedded system", International Conference on Communication, Information & Computing Technology, Mumbai, India, October 19-20, 2012

[11] M. M. Da Silva, D. A. Evin, S. Verrastro, "Speaker-independent embedded speech recognition using Hidden Markov Models", IEEE Conference on Computer Sciences, Buenos Aires, Argentina, November 30-December 2, 2016

[12] J. Li, D. An, L. Lang, D. Yang, "Embedded speaker recognition system design and implementation based on FPGA", Procedia Engineering, Vol. 29, pp. 2633-2637, 2012

[13] G. Tamulevicius, V. Arminas, E. Ivanovas, D. Navakauskas, "Hardware Accelerated FPGA Implementation of Lithuanian Isolated Word Recognition System", Elektronika Ir Elektrotechnika, Vol. 99, No. 3, pp. 57-62, 2010

[14] M. A. A. Zulkifly,N. Yahya, "Relative spectral-perceptual linear prediction (RASTA-PLP) speech signals analysis using singular value decomposition (SVD)", IEEE 3rd International Symposium in Robotics and Manufacturing Automation, Kuala Lumpur, Malaysia, September 19-21, 2017

[15] H. Gupta, D. Gupta, "LPC and LPCC method of feature extraction in Speech Recognition System", 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, India, January 14-15, 2016

[16] D. O'Shaugnessy, "Automatic speech recognition: History, methods and challenges", Pattern Recognition, Vol. 41, No. 10, pp. 2965-2979, 2008

[17] C. Kim, 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, 2016

[18] S. Furui, "Cepstral analysis technique for automatic speaker verification", IEEE Transactions On Acoustics, Speech, and Signal Processing, Vol. 29, No. 2, pp. 254-272, 1981

[19] O. Viikki, K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition", Speech Communication, Vol. 25, No. 1-3, pp. 133-147, 1998

[20] S. Kim, M. Ji, H. Kim, "Noise-robust speaker recognition using subband likelihoods and reliable-feature selection", ETRI Journal, Vol. 30, pp. 89-100, 2008

[21] S. Okawa, E. Bocchieri, A. Potamianos, "Multi-band speech recognition in noisy environments", IEEE International Conference on Acoustics, Speech and Signal Processing, Seattle, USA, May 15, 1998

[22] W. C. Chen, C. T. Hsieh, E. Lai, "Multiband approach to robust text independent speaker identification", Computational Linguistics and Chinese Language Processing, Vol. 9, No. 2, pp. 63-76, 2004

[23] M. I. Abdalla, H. M. Abobakr, T. S. Gaafar, "DWT and MFCCs based Feature Extraction Methods for Isolated Word Recognition", International Journal of Computer Applications, Vol. 69, No. 20, pp. 21-26, 2013

[24] R. X. Gao, R. Yan, Wavelets Theory and Applications for Manufacturing, Springer, 2010

[25] L. R. Rabiner, B. H. Juang, Fundamentals of Speech Recognition, Prentice Hall, 1993

[26] C. Hsu and C. Lin, "A comparison of methods for multiclass support vector machines", IEEE Transactions on Neural Networks, Vol. 13, No. 2, pp. 415-425, 2001

[27] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and other Kernel-based Learning Methods, Cambridge University Press, 2000

[28] A. Mnassri, M. Bennasr, A. Cherif "GA Algorithm Optimizing SVM Multi-Class Kernel Parameters Applied in Arabic Speech Recognition", Indian Journal of Science and Technology, Vol. 10, No 27, pp. 1-9, 2017