Al-Rafidain Engineering Journal (AREJ) Vol.27, No.2, September 2022,pp.190-197

Comparison Between Features Extraction Techniques for Impairments Arabic Speech

Sura Ramzi Shareef* sura.ramzishareef@uomosul.edu.iq

Yusra Faisal Muhammad** yusrafaisalcs@uomosul.edu.iq

* Computer Engineering Department, College of Engineering, University of Mosul, Mosul, Iraq
** Computer Science Department, College of Computer Science and Mathematics, University of Mosul, Mosul, Iraq

Received: 7/2/2022 Accepted: 17/4/2022

ABSTRACT

Automatic Speech Recognition (ASR) is a tough task, with the existence of related noise and high unpredictability in a speech presenting the most severe problems. Especially with regard to the noise of speech impairments, whether due to disability or mispronunciation in children. Extraction of noise-resistant features to compensate for speech degradation due to noise impact has remained a difficult challenge in the last few years. This research investigated the impact of different wavelet generations for extracting speech features, then test the produced dataset from each technique with two types of deep learning techniquesdeep : long short-term memory (LSTM) and hyper deep learning model convolutional neural network with long short-term memory (CNN-LSTM). The result shows that the deep long short-term memory of MFCC has reached 93% as an accuracy while in the hyper deep learning model of CNN-LSTM the accuracy of MFCC was 91%, as the highest recorded accuracy which proves that MFCC would be the best feature extraction technique for our developed dataset.

Keywords:

Features; Impairments; Arabic; Speech; Extraction.

https://rengj.mosuljournals.com

==========

1. INTRODUCTION

Many researchers and engineers find speech to be a tough study issue since it is an effective form of communication between speech processing systems and humans [1]. ASR (Automatic Speech Recognition) is one of them. which is a research area on the reliability of speech processing and is a procedure of successively identifying numerous classes. The primary purpose is to turn the voice signal into a legible sequence of isolated Arabic words. Typically, recognition needs two major processes: the development of an auditory model and feature extraction. After many years of research, the speech recognition system still requires improvement and falls short of its purpose, as the computer cannot grasp all of the scenarios presented by a speaker in all surroundings [2]. Many commercially available Automatic Speech Recognition systems can function properly with clean speech. Nevertheless, if the noise is contaminated, the Automatic Speech Recognition system's performance might suffer considerably. As a result, developing a noise-resistant speech recognition system is critical. The speech waveform is employed in a speech recognition system to retrieve the discriminative feature vectors that indicate the speaker and the spectral information. The acoustic features are then employed for pattern detection by a speech recognizer [3]. The ASR system's recognition performance might suffer significantly without a suitable feature extractor. Relevant and excellent acoustic features may discern between speech classes and endure external noise, as well as the variation of various speakers [4]. As a result, noise resilience has been a critical issue in the field of voice processing. Based on the previous studies,

MFCC (Mel frequency cepstral coefficients) are remarkably often used speech parametrization approaches in speaker identification besides speech recognition. In the formulation of these coefficients, the features of the human auditory system are considered, which surely contribute to obtaining decent performance when used in a speech application. MFCC depends on cepstral area investigation, which imitates the human hearing procedure and could extract the significant features for every phoneme from discourse expressions. While, the primary drawback of this technology is that it has a low level of interference barrier, resulting in severe loss in speaker recognition [5]. To increase the noise resilience of ASR, a novel feature extraction based on a gammachirp filter bank is presented, wherein the operations of Discrete Cosine Transform (DCT) process, energy calculation, and mel-filter bank in MFCC[6] which is substituted by the operations of gammachirp filterbank production.

While MFCCs have received more attention in the context of speech emotion recognition in recent years, (GFCCs) Gammatone Frequency Cepstral Coefficients has continued to remain underestimated. GFCCs are occasionally employed for speaker recognition systems and speech [7-9]. In contrast to MFCCs, GFCCs are based on the Gammatone Filter Bank, where the filters imitate physiological aberrations in the external middle ear and inner ear [10]. They are more noise-resistant than MFCCs and are frequently employed in speaker recognition systems [11-13].

Speech is always damaged by noises in real-world circumstances. In speaker recognition tasks, the authors presented new features, extracted by GFCC. The method surpasses the commonly used MFCC. The speaker verification system presented in [14] outperformed the GMM-UBM based one by merging GFCC and JFA. in [8, 15] discovered that GFCC increased MFCC considerably in noisy conditions, especially when the SNR is less than 10 dB. In the situation of high SNR, GFCC did not outperform MFCC. As a result, we have two choices for obtaining robust recognition: the first is to increase the noise robustness of GFCC by imposing various signal processing techniques, and the second is to treat noisy and clean speech separately. The research was published in [12, 16], where GFCC + i-vectors were used to address noisy conditions and session variability at the same time.

Additionally, Power Normalized Cepstral Coefficients (PNCC) features [17] replace the MFCC processing's log non-linearity with power-law nonlinearity and employ an asymmetric noise suppression approach to minimize background noise. In noisy circumstances, PNCC features significantly improved automated speech recognition accuracy compared to MFCC [17]. PNCC is a newly designed and very accurate feature that surpasses practically all other types of conventional features, even in extremely noisy circumstances [17]. The increased accuracy attained by PNCC is mostly due to essential features such as power-law nonlinearity, an asymmetric noise suppression module, and temporal masking. A bank of Gammatone filters represents the genuine human auditory filters, which have non-linearly rising bandwidths [18, 19]. The PNCC algorithm is used to extract features for voice recognition. This approach was tested for accuracy and complexity, which resulted in enhanced accuracy due to SNR having a higher value [20, 21].

This research developed a deep learning framework based on a Long Short-Term Memory network, in which we encoded each recorded voice using a deep network. LSTM has been applied successfully in context-dependent sequential classification tasks such as conversation modelling [22], dependency parsing [23], and voice recognition [24]. Based on our knowledge the first-ever attempt that an LSTM is applied to the collected dataset from children with impaired speech while pronouncing numbers and letters in the Arabic language, the built model will classify numbers and letters even with a pronunciation mistake. We analyzed the effectiveness of different wavelet families such as MFCC, PNCC, and GFCC in order to extract features from recorded sound. Furthermore, the suggested approach had no effect on processing performance and did not necessitate more computer resources than the previous version.

This paper's structure is arranged as follows. Section 1 describes the overview and literature of acoustic feature extraction. Section 2 goes into great depth on feature extraction strategies and the Deep-LSTM model. Section 3 then displays the experimental findings and discussion sections. Section 4 will bring our conclusion.

2. METHOD

This section explains the techniques that been used in this research and how these methods works.

2.1 Mel Frequency Cepstral Coefficients (MFCC)

The most often utilized feature extraction approach is MFCC. These coefficients are

obtained from the Mel Frequency Cepstrum and reflect audio depending on perception. This approach is regarded as the greatest possible approximation of the human ear. Figure 1 depicts a block diagram showing the construction of an MFCC processor.

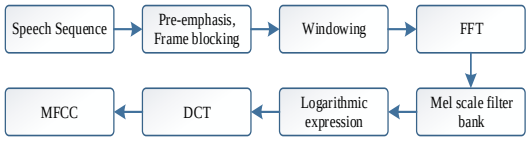

Fig. 1 Basic layout of MFCC

The input signal is fed directly via a first-order digital all-pole filter for pre-emphasis to spectrally flatten the signal, and the resulting signal is then passed through windowing, where it is split into frames using hamming windows. FFT is commonly used to transform time-domain speech signals to frequency-domain signals. Following windowing, FFT and Mel scale filter banks are used to obtain the Mel-spectrum. Mel-scale filter banks are made up of a succession of triangular band-pass filter banks that are built so that the lower limit of one filter is at the centre frequency of the previous filter and the higher limit of the same filter is at the centre frequency of the following filter [25]. The Mel scale is a logarithmic scale that approximates how the human ear hears audio signals. The Mel scale filter bank is used to transfer the powers of the previous spectrum onto the Mel scale utilizing triangular overlapping windows. The Mel scale formula is seen below:

$$Mel_f = 2595 \ln\left(1 + \frac{f}{700}\right) \quad (1)$$

Where Mel$f$ frequency is expressed in mel f and linear frequency is expressed in hertz. The log energy at the output of each filter bank is determined after the signal has been processed through the filter banks. To convert into the cepstral domain, the natural logarithm is used. Lastly, DCT is applied to each Mel spectrum (filter output) to return the data to time domain real values. This transformation decorrelates the features, and the first few coefficients are merged as a feature vector of a certain speech frame. Although DCT aggregates the majority of the information in the signal to its lower-order coefficients by eliminating the higher-order coefficients, a significant decrease in computing cost is achieved.

2.2 Gammatone Frequency Cepstral Coefficients (GFCC)

To improve identification performance, the optimal parametric representation of audio signals must be extracted. Because it influences the behaviour of the next phase, the efficiency of this process is critical. Figure 2 shows a block representation of the whole operation of the GFCC algorithm.

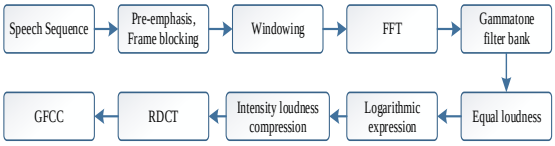

Fig. 2 The GFCC algorithm's block diagram

the FFT-based feature extraction approach is the GFCC algorithm. The technique is based on the GammaTone Filter Bank (GTFB), which attempts to imitate the human auditory system as a collection of overlapped band-pass filters. [26, 27]. The novel and robust GFCC technique, like the MFCC standard previously discussed in 2.1, calculates feature vectors from the spectra of a sequence of speech frames in a 32 ms window and superimposes them by 16 ms. The spectrum of a speech frame is first generated using the 512-point of (FFT) Fast Fourier Transformation. After that, the voice spectrum is routed via a bank of 20 gammatone filters GTFB. Based on the centre frequency of the filter, equal-loudness is applied to each of the filter outputs. After that, the logarithm function is applied to each of the filter outputs. Finally, to acquire the cepstral coefficients GFCC, the data should transition from the spectral to the cepstral domain. The Reverse Discrete Cosine Transform (RDCT) is applied to the filter outputs for this purpose.

2.3 Power-Normalized Cepstral Coefficients (PNCC)

Power- Normalized Cepstral Coefficients [17] was created to provide such qualities that may encourage strong identification when acoustic parameters changed. The goal was to achieve findings without altering the voice signals or increasing information loss. This technique's computational complexity was similar to that of Perceptual Linear Prediction (PLP) [28] and MFCC [29]. The method places a greater emphasis on auditory processing. In comparison to previous algorithms, PNCC processing contains several novel edges:

- To address environmental deterioration caused by shifting voice signals, short-time Fourier

analysis frames (20-30 ms) are combined with medium-time frames with durations of 70-120 ms.

- Before the creation of the PNCC effluent technology, none of the approaches supported an online real-time procedure.
- Power-law nonlinearity was used instead of traditional log nonlinearity in MFCC to establish a relative relationship between auditory-nerve finding signal intensity and intensity. Because it suppresses minor variant signals, nonlinearity is thought to give resilience.
- Asymmetric nonlinear low pass filters can estimate the percentage of trailing noise for all time frames and frequencies.

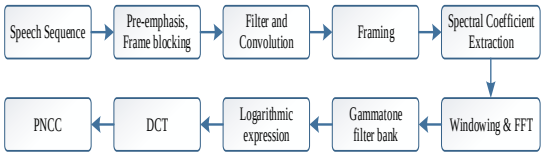

Fig. 3 The PNCC algorithm's block diagram

2.4 CNN-LSTM

The following describes the operation of each LSTM cell in our design. If x_t is one of the CNN models' outputs at each time step t, the LSTM model creates hidden activations at each time step, as indicated by Equation (2), which are then utilized to make the prediction. The LSTM model provides a transition relationship for the hidden representation at through an LSTM cell that receives the current time step's input x_t as well as the acquired information a^{t-1} from the previous step. As a result, when our LSTM network receives the CNN output of a speech sample as input, it analyses it and forwards the inherited information to the next phase. Each LSTM cell contains a cell state c^t, which can be determined using Equation 2, which works as memory and helps hidden units retain information from the past. The LSTM cell is seen in Figure 4. Using Equation 1, we generated a new candidate $c^{\sim t}$ to serve as a placeholder for c^t. Combining c^{t-1}, a^{t-1}, in addition to the input features at t yields the cell state c^t.

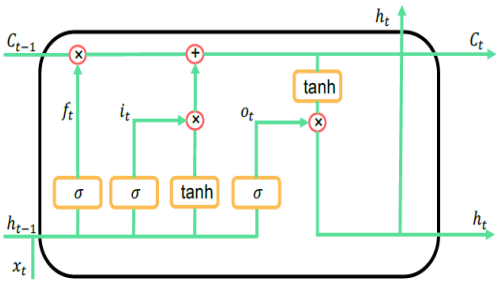

Fig.4 The Architecture of The Lstm Cell.

$$c^{\sim t} = \tanh(W_a^c\, a^{t-1} + W_x^c\, x_t) \quad (2)$$
$$c^t = f^t \otimes c^{t-1} + u^t \otimes c^{\sim t} \quad (3)$$
$$a^t = o^t \otimes \tanh(c^t) \quad (4)$$

Here w_a^c and w_x^c signify the weight parameters used to produce a candidate cell state. Hereafter we will ignore the bias terms in the following since they can be absorbed into weight matrices. Then, as shown in the diagram, we build a forget gate layer f^t, an update gate layer ut, and an output gate layer o^t.

$$f^t = \sigma\,(W_a^f\, a^{t-1} + W_x^f\, x_t) \quad (5)$$
$$u^t = \sigma\,(W_a^u\, a^{t-1} + W_x^u\, x_t) \quad (6)$$
$$o^t = \sigma\,(W_a^o\, a^{t-1} + W_x^o\, x_t) \quad (7)$$

At every step t, the hidden representation at is an accumulation of information from previously processed features, and so impacts the development of the final output. The alteration in cell state generates a memory flow, which enables the modelling of long-term spatial and temporal dependency. The abovementioned processing approach is done to all sample data input features. The first three natural frequencies are the final output. Figure 5 depicts the architecture of the outlined CNN-LSTM model.

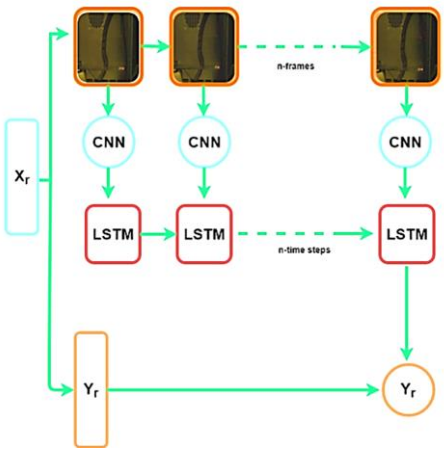

Fig.5 Depicts The CNN-LSTM Model's Architecture.

3. EXPERIMENTAL RESULTS AND DISCUSSION

The speech dataset has been collected through the participation of 38 impaired students from schools for people with special needs in Mosul-Iraq. The dataset contains 770 speeches recorded. All students who participated in the research ranged from 7 to 11 years old, they have some commonalities related to speech impairment. They suffer from motor speech disorders.

The specialist advised the youngsters to pronounce letters and numerals in Arabic language carefully. For each letter and each number, each kid was instructed to repeat the pronunciation 10 times. The number used in the dataset were from 0 to 9, letters are “Aleef”, “Baa”, “Taa”, “Thaa”, “Geem”, “Haa”, “Khaa”, “daal”,and“Tufaha”, [30].

This research attempts the experiments in two directions, first, Highlight the use of popular feature extraction techniques in ARS with a problem of understanding speech impairments and measuring their capacity to recognize words correctly.

Second direction, apply different approaches based on deep learning to know the advantages of classifiers, with various features type. In addition, to observe the performance of feature extraction techniques, which one yields more effective features? to resistant speech impairments problem in ARS with a different approaches' environment.

Experiments were conducted by using python language to test the performance of the system in two steps. First, extracted a set of 39 features from each frame of the audio signal using the three feature extraction techniques, explained earlier (MFCC, PNCC and GFCC). While the second step, exploits the approaches of Deep learning, specifically, LSTM and hybrid CNN-LSTM approach to creating a model for ARS. Both approaches are applied using the set of features result from the first step separately for each set of features.

Table 1: LSTM Model

Features	Features Numbers	Accuracy
MFCC	13+Delta1(13) +Delta2(13)	93.0%
PNCC	13+Delta1(13) +Delta2(13)	80%
GFCC	13+Delta1(13) +Delta2(13)	78%

Table 2: CNN-LSTM Model

Features	Features Numbers	Accuracy
MFCC	13+Delta1(13) +Delta2(13)	91.00%
PNCC	13+Delta1(13) +Delta2(13	64.67%
GFCC	13+Delta1(13) +Delta2(13	61.67%

The results show the LSTM approach with MFCCs gave better performance .In Table (1) and Table (2) we notice that MFCC is the best technique than the other two techniques dealing with our dataset. Accordingly, in deep long short-term memory the accuracy of MFCC has reached 93% while in the hyper deep learning model CNN-LSTM the accuracy of MFCC was 91%, moreover PNCC with Lstm has achieved accuracy 80% while in CNN-LSTM has achieved 64.67%. lastly, GFCC reached 78% accuracy in LSTM while the accuracy in CNN-LSTM reached 61.67% which is the lowest accuracy. From the explanation above we realize that LSTM model is better than CNN-LSTM by dealing with audio dataset as in figure 6.

The experiments results reveal using MFCC features in establishing an ARS can be able to recognize words, that children with speech impairments are saying, helps them facilitate talk or used speech recognition commands technology.

The confusion in context or sequence of phonemes in an audio signal, that is resulting from motor speech disorders in speech impairment, could be the main reason for the poor performance of both PNCC and GFCC feature extraction techniques. The GammaTone Filter Bank produces robust features against noise that affect quality audio rather than recognize phonemes. The PNCC and GFCC outperform MFCCs in speech emotion recognition.

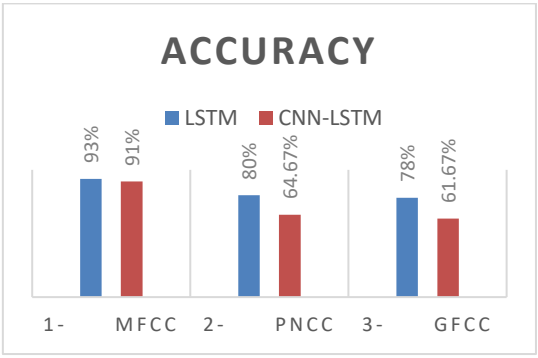

Fig. 6 Models Results

4. CONCLUSION

Type of features and the techniques are adopted in extracting it are known a difficult challenge for tasks ASR from a performance aspect. This paper investigates the ability of which audio signal features extraction techniques are more resistant to Impairments Arabic speech in children for automatic speech recognition systems by comparing three common techniques MFCC, PNCC and GFCC. The competence of these feature extraction techniques is assessed based on their ability to recognize words correctly. In fact,

the goodness of these features reflected the accuracy of prediction results for deep learning approaches adopted in this work. Although this research did not investigate the analysis of the differences between the audio signal of speech impairment problem and normal precisely. However, the findings of the research can suggest general conclusions.

First, several of the technical aspects of the sophisticated feature extraction framework referenced in the literature were confirmed. Where, the Gammatone Frequency Cepstral Coefficients (GFCC), power normalized cepstral coefficient (PNCC), and Mel Frequency Cepstral Coefficients (MFCC) are presented.

Second, the research presents a type of comparison between normal and impaired speech from the aspect of the possibility of using the same feature extraction techniques and deducing which of them is can be extracting more robust speech features, without affecting system performance for impaired speech.

Third, research results behold that Deep learning approaches based on MFCC features of the audio signals show the best accuracy in classifying Arabic characters and numerals according to experimental data. So, it can be said the MFCC techniques are more resistant to Impairments Arabic speech in children comparable to other techniques PNCC and GFCC.

Although the highest predictive accuracy is yield with MFCC features, it requires a variety of modifications, tests, and experiments with a more in-depth investigation. Especially with analysis speech impairment problem to avoid accuracy deterioration that might be expected from an increase in vocabulary and take the user independence into account. Furthermore, noisier situations, such as mixes of ambient noises.

REFERENCES

[1] X. Huang, A. Acero, H.-W. Hon, and R. Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*: Prentice hall PTR, 2001.

[2] D. Jurafsky and J. H. Martin, "Speech and Language Processing: International Version: an Introduction to Natural Language Processing," *Computational Linguistics, and Speech Recognition, Pearson*, 2008.

[3] A. Kuamr, M. Dua, and T. Choudhary, "Continuous Hindi speech recognition using Gaussian mixture HMM," in *2014 IEEE Students' Conference on Electrical, Electronics and Computer Science*, 2014, pp. 1-5.

[4] S. Sadhu, R. Li, and H. Hermansky, "M-vectors: sub-band based energy modulation features for multi-stream automatic speech recognition," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6545-6549.

[5] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, "Minimum mean-squared error estimation of mel-frequency cepstral coefficients using a novel distortion model," *IEEE transactions on audio, speech, and language processing*, vol. 16, pp. 1654-1661, 2008.

[6] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," in *2006 IEEE international symposium on circuits and systems*, 2006, p. 4 pp.

[7] W. Burgos, "Gammatone and MFCC features in speaker recognition," 2014.

[8] X. Shi, H. Yang, and P. Zhou, "Robust speaker recognition based on improved GFCC," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 1927-1931.

[9] T. L. Nwe and H. Li, "On fusion of timbre-motivated features for singing voice detection and singer identification," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 2225-2228.

[10] A. G. Katsiamis, E. M. Drakakis, and R. F. Lyon, "Practical gammatone-like filters for auditory processing," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, pp. 1-15, 2007.

[11] X. Zhao and D. Wang, "Analyzing noise robustness of MFCC and GFCC features in speaker identification," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 7204-7208.

[12] M. Jeevan, A. Dhingra, M. Hanmandlu, and B. Panigrahi, "Robust speaker verification using GFCC based i-vectors," in *Proceedings of the International Conference on Signal, Networks, Computing, and Systems*, 2017, pp. 85-91.

[13] P. S. R. Singh, N. Kaur, and P. Singh, "Speech Based Biometric System Using GFCC Features," *Imperial Journal of Interdisciplinary Research*, vol. 3, pp. 1156-1160, 2017.

[14] P. Das and U. Bhattacharjee, "Robust speaker verification using GFCC and joint factor analysis," in *Fifth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, 2014, pp. 1-4.

[15] G. K. Liu, "Evaluating gammatone frequency cepstral coefficients with neural networks for emotion recognition from speech," *arXiv preprint arXiv:1806.09010*, 2018.

[16] X. Zhang, X. Zou, M. Sun, and P. Wu, "Robust Speaker Recognition Using Improved GFCC and Adaptive Feature Selection," in *International Conference on Security with Intelligent Computing and Big-data Services*, 2018, pp. 159-169.

[17] 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, pp. 1315-1329, 2016.

[18] M. Slaney, "An efficient implementation of the Patterson-Holdsworth auditory filter bank," *Apple*

Computer, Perception Group, Tech. Rep, vol. 35, 1993.

[19] A. Badi, K. Ko, and H. Ko, "Bird sounds classification by combining PNCC and robust Mel-log filter bank features," *The Journal of the Acoustical Society of Korea,* vol. 38, pp. 39-46, 2019.

[20] A. A. Alasadi, R. R. Deshmukh, and S. D. Waghmare, "Review of Modgdf & PNCC Techniques for Features Extraction in Speech Recognition," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, 2019, pp. 1-7.

[21] H. M. S. Naing, R. Hidayat, R. Hartanto, and Y. Miyanaga, "Discrete wavelet denoising into mfcc

[25] O. Cheng, W. Abdulla, and Z. Salcic, "Performance evaluation of front-end processing for speech recognition systems," *School of Engineering Report. The University of Auckland, Electrical and Computer Engineering*, 2005.

[26] W. H. Abdulla, "Auditory based feature vectors for speech recognition systems," *Advances in Communications and Software Technologies,* pp. 231-236, 2002.

[27] M. Kleinschmidt, J. Tchorz, and B. Kollmeier, "Combining speech enhancement and auditory feature extraction for robust speech recognition," *Speech Communication,* vol. 34, pp. 75-91, 2001.

[28] E. Yücesoy and V. V. Nabiyev, "Comparison of MFCC, LPCC and PLP features for the determination of a speaker's gender," in *2014 22nd Signal Processing and Communications Applications Conference (SIU)*, 2014, pp. 321-324.

[29] Z. M. Dan and F. S. Monica, "A study about MFCC relevance in emotion classification for SRoL database," in *2013 4th International Symposium on Electrical and Electronics Engineering (ISEEE)*, 2013, pp. 1-4.

[30] Sura Ramzi Shareef, Y.F.Al-I.,"Towards developing impairments arabic speech dataset using deep learning," *Indonesian Journal of Electrical Engineering and Computer Science*(*ijeecs*),*Vol.25,No.3, March , pp.2502-4752,*2022*DOI: 10.11591*

for noise suppressive in automatic speech recognition system," *International Journal of Intelligent Engineering and Systems,* vol. 13, pp. 74-82, 2020.

[22] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *arXiv preprint arXiv:1503.02364,* 2015.

[23] C. Dyer, M. Ballesteros, W. Ling, A. Matthews, and N. A. Smith, "Transition-based dependency parsing with stack long short-term memory," *arXiv preprint arXiv:1505.08075*, 2015.

[24] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *2013 IEEE international conference on acoustics, speech and signal processing*, 2013, pp. 6645-6649.

مقارنة بين تقنيات استخلاص الميزات عند ضعف الكلام العربي

سرى رمزي شريف* **يسرى فيصل محمد****
sura.ramzishareef@uomosul.edu.iq yusrafaisalcs@uomosul.edu.iq

*جامعة الموصل - كلية الهندسة – قسم هندسة الحاسوب
**جامعة الموصل - كلية العلوم والرياضيات – قسم علوم الحاسوب

الملخص

تعد عملية التعرف على الكلام التلقائي واحدة من المهام الصعبة، مع وجود ضوضاء مصاحبة في اغلب الاحيان للكلام وعدم القدرة على التنبؤ في الكلام المنطوق يؤدي الى مشاكل حادة في عملية تحويل الكلمات المنطوقة إلى نص. استخراج ميزات مقاومة للضوضاء لتعويض هذا التراجع في الاداء هو الاخر تحدياحتى السنوات القليلة الماضية. هذا البحث يحقق في تأثير الميزات المختلفة المستخرجة من موجات الكلام. ثم اختبرت هذه الميزات مع نوعين من تقنيات التعلم العميق هما الذاكرة طويلة المدى LSTM النموذج التقليدية ونموذج هجين يتضمن الشبكة العصبية التلافيفية ذات الذاكرة طويلة LSTM-CNN.أظهرت نتائج هذا البحث أن ميزات MFCC أكثر مقاومة الضوضاء ، حيث حققت أعلى دقة مع نموذج LSTM بنسبة 93 ٪ومع الموديل الهجين LSTM-CNN كانت الدقة 91.٪.

الكلمات داله :
سمات؛ ضعف؛ عربى؛ خطاب؛ استخلاص