Engineering and Technology Journal
Journal homepage: https://etj.uotechnology.edu.iq

Speech Recognition Algorithm in a Noisy Environment Based on Power Normalized Cepstral Coefficient and Modified Weighted-KNN

Mohammed E. Safi*, Eyad I. Abbas

Electrical Engineering Dept., University of Technology-Iraq, Alsina'a street, 10066 Baghdad, Iraq.
*Corresponding author Email: mohammed.e.safi@uotechnology.edu.iq

HIGHLIGHTS

- The proposed algorithm is based on PNCC feature extraction with a new classifier Weighted-KNN-DTW.
- Weighted-KNN-DTW classifier is a modification of Weighted KNN and DTW.
- The accuracy of the proposed algorithm was calculated with different levels of white noise (20dB, 15dB, 10dB, and 5dB).

ARTICLE INFO

Handling editor: Mohammed Y. Hassan

Keywords:
Speech Recognition
PNCC
MFCC
KNN
DTW
SVM

ABSTRACT

Speech recognition is widely used in robot control and automation. Nevertheless, the use of speech recognition in robots is limited due to its susceptibility to background noise. This paper proposes a speech recognition algorithm to control robots in noisy environments. The proposed algorithm is based on Perceptual Linear Predictive Cepstral Coefficients (PNCC), which is a noise-resistant feature extraction technique, and Modified K-Nearest Neighbors (KNN) with Dynamic Time Warping (DTW) as the classifier. A new KNN-DTW classifier is proposed, integrating weighted KNN and DTW. The proposed algorithm results from experiments comparing PNCC and Mel-frequency cepstral coefficients (MFCC) feature extraction techniques with different classifiers, namely KNN-DTW, two types of KNN (weighted KNN and Medium-KNN), and two types of Support Vector Machine SVM (Linear SVM and Quadratic SVM). The database used to investigate the accuracy was the audio-visual data corpus database UOTletters, which includes 30 speakers, 26 English letters, and 1560 utterances. The database is divided into 50% for training and 50% for testing purposes. In a noise-free environment, the accuracy of the proposed algorithm reached 100%. Moreover, the proposed algorithm demonstrates greater noise immunity across all five noise levels, with an average accuracy difference of 13.67% compared to baseline algorithms.

1. Introduction

One of the key techniques for controlling robotics, home appliances, security door, mobile applications, cars, etc., is speech recognition. However, as every user is fully cognizant, many features of this technology still need to be improved, such as the identification rate, noise resistance, and consistent accuracy [1,2]. PNCC feature extraction, an enhancement of the most popular MFCC feature extraction method, is a way to increase the accuracy of speech recognition systems in noisy environments [3-6].

Numerous studies on this topic have been conducted in recent years, in our previous work [7]. The proposed algorithm was used to accurately evaluate the recognition rate using large datasets (Speech Commands DataSet v0.01). This database contains Audio files with 64,727 different individuals (30 words, five repetitions each). This was the first study to integrate the Weighted KNN classifier with PNCC for speech recognition. Although KNN and SVM classifiers are frequently used in studies [8-13], it has been demonstrated that the two classifiers have a close accuracy. As a result, the best classifiers for this research were looked into for the two feature extraction techniques, PNCC and MFCC. The paper's findings demonstrate that the suggested technique, based on the PNCC-Weighted KNN algorithm, has higher accuracy and immunity to white noise. Ali et al. [14], proposed an MFCC and KNN classifier-based system for isolated word recognition in Pashtun—10 words (Pashton numerals 0 to 9, and 50 individuals make up the data collection). An accuracy of 76.8% was achieved for this investigation.

The researchers, Safi and Abbas [15], presented a microcontroller-driven Automatic Speech Recognition algorithm (ASR) to control the security door system. The ASR algorithm is based on DTW isolation word matching and MFCC feature extraction. The algorithm was tested with twenty-two individuals, and the database contains three passwords and three user authentications. 100% of accuracy was achieved.

http://doi.org/10.30684/etj.2023.140643.1469
Received 28 May 2023; Received in revised form 10 July 2023; Accepted 23 July 2023; Available online 22 August 2023
2412-0758/University of Technology-Iraq, Baghdad, Iraq

The paper's researchers, Imtiaz and Raja [16] 2017 suggested MFCC features extraction and DTW approach integrated by KNN classifier for isolated word automatic speech recognition. Two thousand audio tracks comprise the entire data set, comprising ten English words spoken by five individuals. 98.4% of the validation results are accurate according to this system.

AnggraeniIn et al. [17], proposed an ASR system to direct the movements of a 5 Degree of Freedom (DOF) robotic arm for picking and placing objects. The command recognition technique uses KNN classifiers in conjunction with MFCC for feature extraction. The data set consists of 20 commands that are repeated ten times for each word in the Indonesian language. The voice recognition rate for trained sets is 85%, compared to 80% for untrained respondents. Adiwijaya et al. [18], proposed an ASR system that pronounces the Arabic letters (Hijaiyah). Two feature extraction methods, MFCC and Linear Predictive Coding (LPC), are employed in this work as study cases. To estimate the best outcomes, it is advised to integrate KNN as a classifier in each method. Six speakers and 28 letter sounds make up the data set. The findings indicate that LPC has a 78.92F percent accuracy rate compared to MFCC's 59.87 percent.

She et al. [19], created a new feature extraction technique using the supplied blended features. The combination uses the cepstral coefficients as a foundation of the cochlear filter to maximize accuracy in noisy surroundings (CFCC). Three stages have been added to the feature extraction procedure. This was accomplished by getting the energy feature TEOCC and the feature TEOCC's compensatory effect on the auditory characteristics. Three stages have been added to the feature extraction procedure. This was accomplished by getting the energy feature TEOCC and the feature TEOCC's compensatory effect on the auditory characteristics. Last, Principal Component Analysis (PCA) is used for selection and optimization based on the fusion feature's feature redundancy. This technique uses the SVM classifier, and the database consists of ten and twenty-two Korean words from sixteen individuals, each repeated three times. For ten words, the accuracy is 92.79 percent, and for 20, it is 88.43 percent. Korkmaz et al. [20], suggested a system for classifying vowels of the Turkish language developed on a features vector created using the feature approaches Wavelet Decomposition Shannon Entropy, LPC, MFCC, Energy, and Zero Crossing Rate (ZCR). After optimization using a genetic algorithm, a 1-NN Cityblock classifier classified the characteristics vector. The database of this paper is 2762 total observations and 8 Turkish vowels spoken by ten individuals. The recognition rate of this work reached 100%. Alasadi et al. [21], different feature extraction methods have been proposed for the ASR system: the Modified Group Delay Function (ModGDF), the PNCC, and the MFCC. Forty speakers contributed 18 Arabic words to the data set. The findings of this paper have demonstrated that MFCC has a recognition rate of 97.5%, which is higher than ModGDF's recognition rate of 90.3%. Tuncer et al. [22], suggested a dynamic center mirror local binary pattern (DCMLBP), while the DWT was used for feature extraction. The identification of useful features is then accomplished using neighborhood component analysis (NCA). The decision tree (DT), KNN, SVM, bagged tree (BT), and linear discriminant analysis (LDA) are some of the classifiers employed in this work. There are 480 utterances in the database for this study, representing various ambient classes (eight classes by sixty utterances). SVM classifiers produced 99.97% of accuracy.

This paper proposes a speech recognition algorithm based on PNCC as feature extraction and a new KNN-DTW classifier based on integrated weighted KNN and DTW to control robots in a noisy environment. The speech recognition algorithm interfaces with hardware to control the home appliances using MATLAB 2021a and a microcontroller. The database used to investigate the accuracy was the audio-visual data corpus database UOTletters Which has 30 speakers, 26 English letters, and 1560 utterances.

2. Research Components

In this section, the background of techniques of feature extraction and classifiers that are used and developed in this paper is illustrated.

2.1 PNCC Feature Extraction

The need for a trustworthy feature for ASR, which is high immunity in noisy environments in its structure with reasonable cost to compete with other feature extraction techniques, served as the primary motivation for the development of PNCC [6]. As shown in Figure 1, the structure of modified PNCC in [3], based on basic PNCC in [4], is the same as MFCC, PNCC starting with Pre-emphasis, as shown in Equation 1:

$$H(z) = 1 - 0.97Z^{-1} \tag{1}$$

The Short Time Fourier Transform is also carried out utilizing DFT after framing and windowing using Humming windows for 25.6 ms and a 10 ms cross-section between frames. Gammatone filter banks (40-channel) in place of MFCC's traditional triangle filter banks in the frequency range is the next step, which is the stage that sets all other varieties of PNCC apart from one another (200Hz to 8000Hz). This change was made in response to the finding in [3,4] that gamma-filter banks have higher ASR accuracy than triangle-filter banks. Getting spectral power in the short term, as shown in Equation 2:

$$P[m,l] = \sum_{k=0}^{(k/2)-1} \left|X\left[m, e^{j\omega k}\right] H_l(e^{j\omega k})\right|^2 \tag{2}$$

where: m is the frame number, l is the channel index, k is the DFT size, H_l is the response of lth of channels at frequency ωk.

Medium Time Power quantity ($\tilde{Q}$) alter information based on the Short Time power in MFCC is the final iteration of PNCC indicated in [3]. According to the following in Equation 3:

$$\tilde{Q}[m,l] = \frac{1}{2M+1}\sum_{\grave{m}=m-M}^{m+M} P\,[\grave{m},l] \quad (3)$$

where: M is the temporal integration factor.

According to [23], M=2 is advised (corresponding to five consecutive windows with 65.6ms of the total net). This step is based on studies showing that accuracy is improved in noisy environments with long-term processing e.g. [24-26]. This is true because the power related to noise changed more slowly than the power related to speech. Additionally, numerous research [27-29] have demonstrated that long-term processing with Gammatone channels yields more informational details beneficial for improving speech recognition.

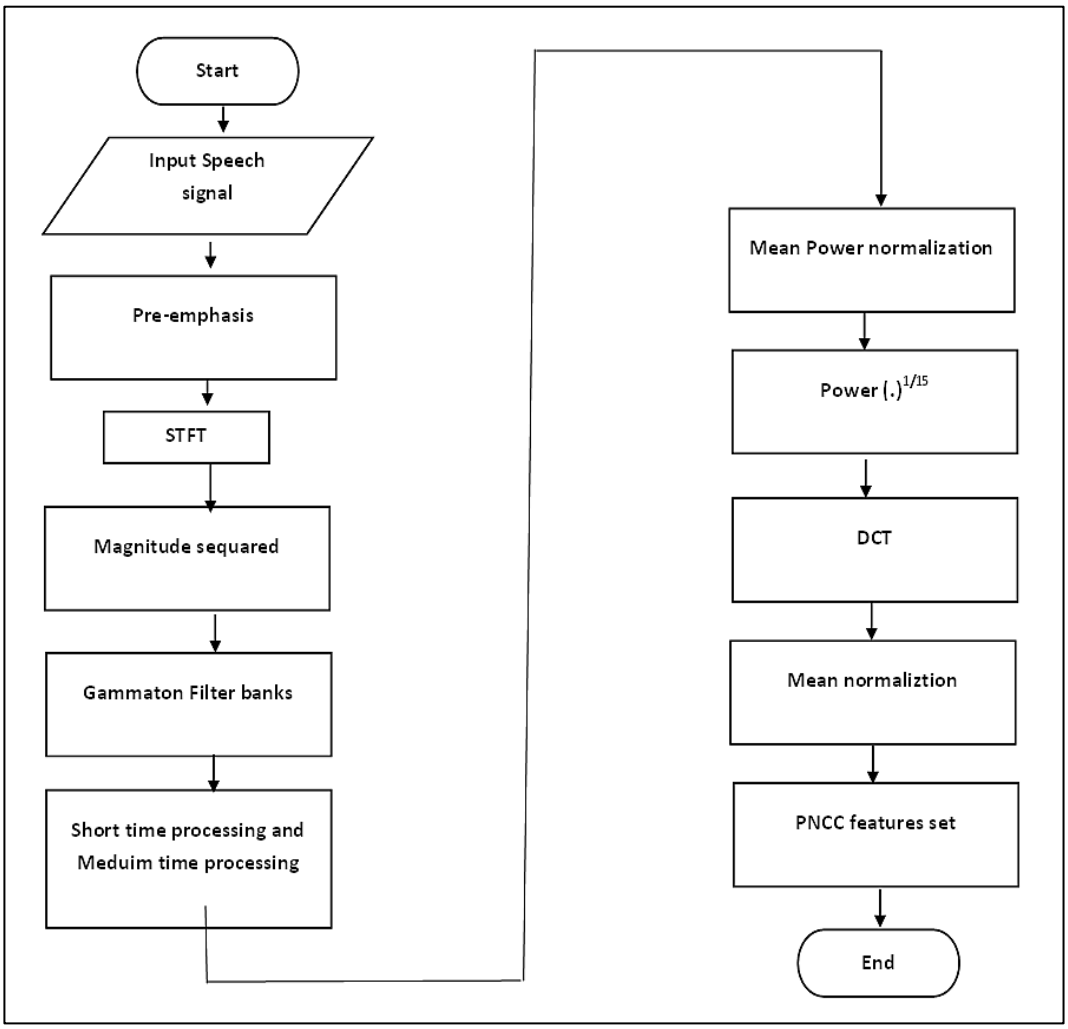

Figure 1: PNCC feature extraction algorithm

The following stage is asymmetric noise suppression (ANS), which involves using the information from the previous observations to select the flooring level and invest it in noise cancellation. The Asymmetric Lowpass Filter equation yields the lower envelope $\tilde{Q}_{le}[m,l]$, which denotes the average noise power, as shown in Equation 4:

$$\tilde{Q}_{le}[m,l] = \begin{cases} 0.999\,\tilde{Q}_{le}[m-1,l] + 0.001\,\tilde{Q}[m,l] \\ \quad if\ \ \tilde{Q}[m,l] \geq \tilde{Q}_{le}[m-1,l] \\ 0.5\,\tilde{Q}_{le}[m-1,l] + 0.5\,\tilde{Q}[m,l] \\ \quad if\ \ \tilde{Q}[m,l] < \tilde{Q}_{le}[m-1,l] \end{cases} \quad (4)$$

$\tilde{Q}_f[m,l]$ will be a floor level for $\tilde{Q}_1[m,l]$ concerning temporal masking $\tilde{Q}_{tm}[m,l]$, as shown in Equation 5:

$$\tilde{Q}_1[m,l] = \max\left(\tilde{Q}_{tm}[m,l], \tilde{Q}_f[m,l]\right) \quad (5)$$

Based on findings that the human auditory system places more emphasis on the rising edge of a power envelope than its onset (e.g. [22,23]), the temporal masking stage of PNCC was developed. Thus, $\tilde{Q}_p[m,l]$ is calculated from the maximum of $\tilde{Q}_p[m-1,l]$ and $\tilde{Q}_o[m,l]$) online peak power, as shown in Equation 6:

$$\tilde{Q}_p[m,l] = \max\left(0.85\ \tilde{Q}_p[m-1,l], \tilde{Q}_o[m,l]\right) \tag{6}$$

Then $\tilde{Q}_{tm}[m,l]$ obtained by Equation 7:

$$\tilde{Q}_{tm}[m,l] = \begin{cases} \tilde{Q}_o[m,l], & \tilde{Q}_o[m,l] \geq 0.85\ \tilde{Q}_p[m-1,l] \\ 0.2\ \tilde{Q}_p[m-1,l], & \tilde{Q}_o[m,l] < 0.85\ \tilde{Q}_p[m-1,l] \end{cases} \tag{7}$$

Following the development of $\tilde{Q}_1[m,l]$ via equation 5, If, $\tilde{Q}[m,l] \geq 2\ \tilde{Q}_{le}[m,l]$, then an "excitation segment" is necessary, and if, $\tilde{Q}[m,l] < 2\ \tilde{Q}_{le}[m,l]$ then a "non-excitation segment" is necessary.

The final output of ANS is given by Equation 8:

$$\tilde{R}[m,l] = \begin{cases} \tilde{Q}_1[m,l]\ , if\ excitation\ segment \\ \tilde{Q}_f[m,l], if\ non-excitation\ segment \end{cases} \tag{8}$$

To contribute noise compensation, smoothing is required [30] , as shown in Equation 9:

$$\tilde{S}[m,l] = \left(\frac{1}{l_2-l_1+1}\right)\sum_{\tilde{l}=l_1}^{l_2} \frac{\tilde{R}[m,\tilde{l}]}{\tilde{Q}[m,\tilde{l}]} \tag{9}$$

where: l_1= min $(l+N,1)$, $l_2 =$ min$(l+N,L)$, L is the total number of channels=40 (optimal value) based on [23], N = 4(optimal value) based on [23].

In the following step $\tilde{S}[m,l]$ is utilized to modify the short-time power $P[m,l]$, as shown in Equation 10:

$$T[m,l] = P[m,l]\tilde{S}[m,l] \tag{10}$$

Then mean power μ[m] is calculated, as shown in Equation 11:

$$\mu[\mathrm{m}] = \lambda_\mu\ \mu[\mathrm{m}-1] + \frac{(1-\lambda_\mu)}{L}\sum_{l=0}^{L-1} T[m,l] \tag{11}$$

The recommended value for the forgetting factor λμ is 0.999, according to [3]. Now it can obtain the normalized power $U[m,l]$ from the Equation 12 below:

$$U[m,l] = k\frac{T[m,l]}{\mu[\mathrm{m}]} \tag{12}$$

where: k is an arbitrary value.

The power-law nonlinearity is the final step that sets PNCC apart, as shown in Equation 13:

$$V[m,l] = U[m,l]^{1/15} \tag{13}$$

Finally, DCT is calculated to produce the cepstral coefficients, the same as the MFCC algorithm, as shown in Equation 14:

$$PNCC_n = \sum_{k=0}^{K} V_k \cos\left[n\left(k-\frac{1}{2}\right)\frac{\pi}{40}\right] \text{ for n=0,1, 2,...,M-1,} \tag{14}$$

where: M is the number of PNCC coefficients and V_k, k =0, 2..., K, represents the power function output for the kth filter.

2.2 Weighted-KNN Classifier

A simple-to-use supervised learning classifier is the KNN classifier [31,10]. It is predicated on the idea that neighboring samples of a similar nature exist [32]. The classifier investigates the distances of the k-objects closest to a sample to categorize

it. The sample is then classified under the class that occurs the most [33]. The size and kind of the dataset will determine how strictly the nearest neighbors (K) values should be chosen [14]. The Euclidean distance method is frequently used to calculate the distance (d) between the sample and the object [34] , as shown in Equation 15:

$$d_{(i,y_j)} = \sqrt[2]{\sum_{j=1}^{k}(S_j - y_j)^2} \tag{15}$$

where: S and y, for the jth of K-folded, are the sample features vectors and the object features vectors, respectively, i is the number of the cluster.

In Weighted KNN, the calculated distances are evaluated as weight [35] , as shown in Equation 16:

$$w_{(i,y_j)} = \frac{1}{d_{(i,y_j)}^{\,2}} \tag{16}$$

where: (w) is the weight for the k-objects most nearby to the tested sample.

The category to which the most K neighboring samples ($\hat{s_i}$) belonged is then specified. Each K-nearest to a certain cluster (i) of nearby samples ($\hat{s_i}$) as determined by the formula [36], as shown in Equation 17:

$$\hat{s_i} = \sum_{j=1}^{k}(w_{(i,y_j)} \tag{17}$$

2.3 Modified-KNN by DTW Classifier

The first proposed Modified KNN by DTW classifier was mentioned [15]. Instead of the Euclidean distance (d) only shown in Equation 15. DTW is frequently used to calculate the global distance (D) (similarity) between the sample and the object, as shown in Equation 18:

$$D(i,j) = d(S_i, y_j) + min\{D(i-1, j-1), D(i-1, j), D(i, j-1)\} \tag{18}$$

This paper proposes a new KNN-DTW classifier based on integrated weighted KNN in 2.2 and DTW.

3. The Proposed PNCC and Modified Weighted-KNN Algorithm

The proposed algorithm's general procedure is described in Figure 2, starting with the PNCC feature extracted to trained voices to store under its cluster label in the database. Similarly, in the test process, followed by the classification process using the proposed classifier.

The database used in this experiment is the audio of the UOT letters database, which consists of 30 speakers (10 to 62 years old), 26 English letters, two times of repetition, and the overall utterances are 1560. The database was created by researchers in [37]. They recorded movies of the database in their laboratory in real and noisy surroundings without filtering, depending on the camera microphone with a sampling rate of (11025 Hz), as shown in the sample of sound in Figure 3.

3.1 Features Extraction

PNCC feature extraction is mentioned in detail in section 2., but in this section, the parameters setup of PNCC of the proposed algorithm is illustrated. The sampling rate of the speech signal used in these experiments was 11025 samples/sec. The feature extraction started with Pre-emphasis using equation 1 to emphasize the high frequencies by using the common first-order FIR highpass filter to compensate for the high-frequency, low power. The Short Time Fourier Transform uses DFT after framing and windowing using Humming windows for 25.6 ms and a 10 ms cross-section between frames and getting spectral power in the short term for 40-channel of Gammatone filter banks as shown in Equation 2. The Medium Time Power quantity is calculated using equation 3. Setting the temporal integration factor M=2 corresponding to five consecutive windows with 65.6 ms of the total net. The Asymmetric Lowpass Filter equation yields the lower envelope $\tilde{Q}_{le}[m,l]$, which denotes the average noise power using equation 4. The temporal masking stage of PNCC was developed. Thus, $\tilde{Q}_p[m,l]$ is calculated from the maximum of $\tilde{Q}_p[m-1,l]$ and $\tilde{Q}_o[m,l]$) online peak power by using Equation 6. Following the development of $\tilde{Q}_1[m,l]$ via equation 5, If $\tilde{Q}[m,l] \geq 2\ \tilde{Q}_{le}[m,l]$, then an "excitation segment" is necessary, and if, $\tilde{Q}[m,l] < 2\ \tilde{Q}_{le}[m,l]$ then a "non-excitation segment" is necessary. The ANS is given by equation 8, and the contributed noise compensation smoothing is calculated by equation 9. In the following step $\tilde{S}[m,l]$ is utilized to modify the short-time power using equation 10. Then mean power μ[m] is calculated by equation 11 by setting the value for the forgetting factor λμ as 0.999. the normalized power $U[m,l]$ from equation 12 is obtained. The power-law nonlinearity is calculated by equation 13. The final step of feature extraction is The DCT by using equation 14. DCT is estimated to produce 13 cepstral coefficients for each frame, as shown in Figure 4.

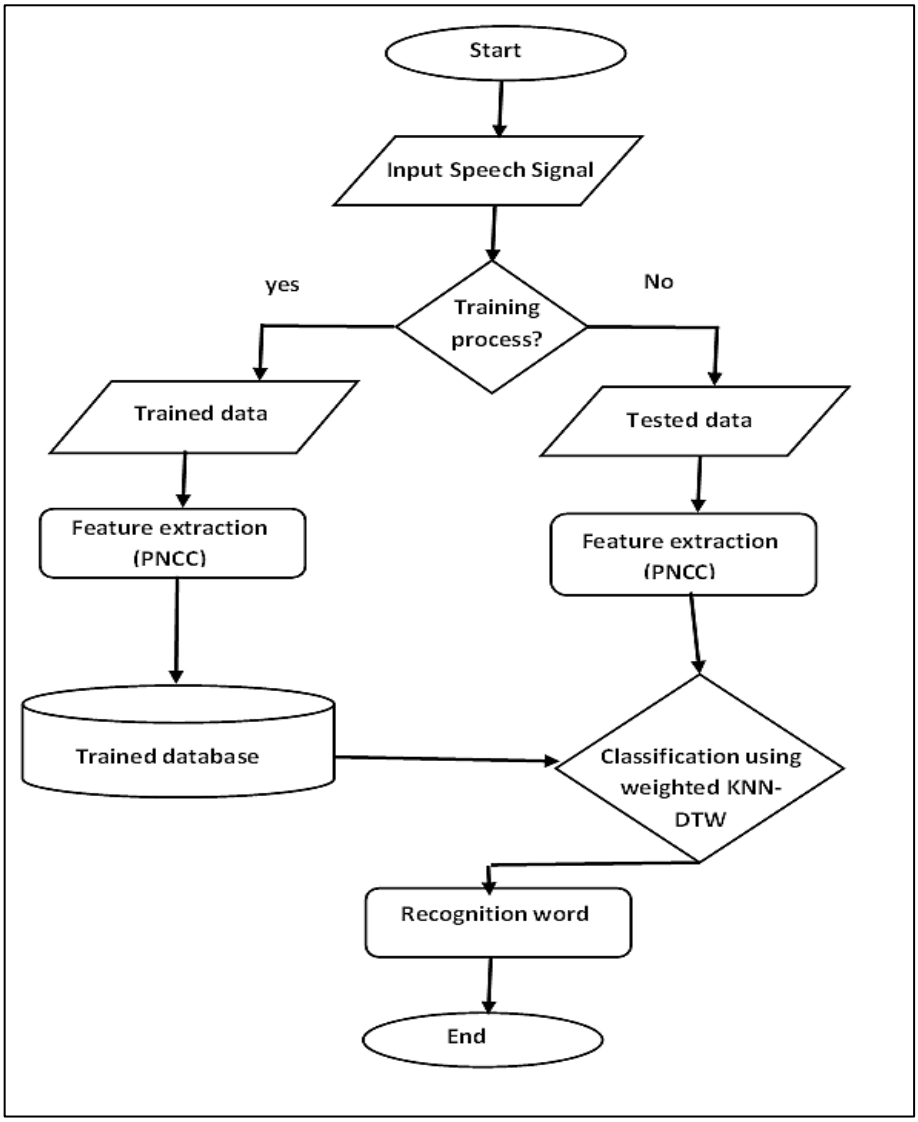

Figure 2: The proposed algorithm

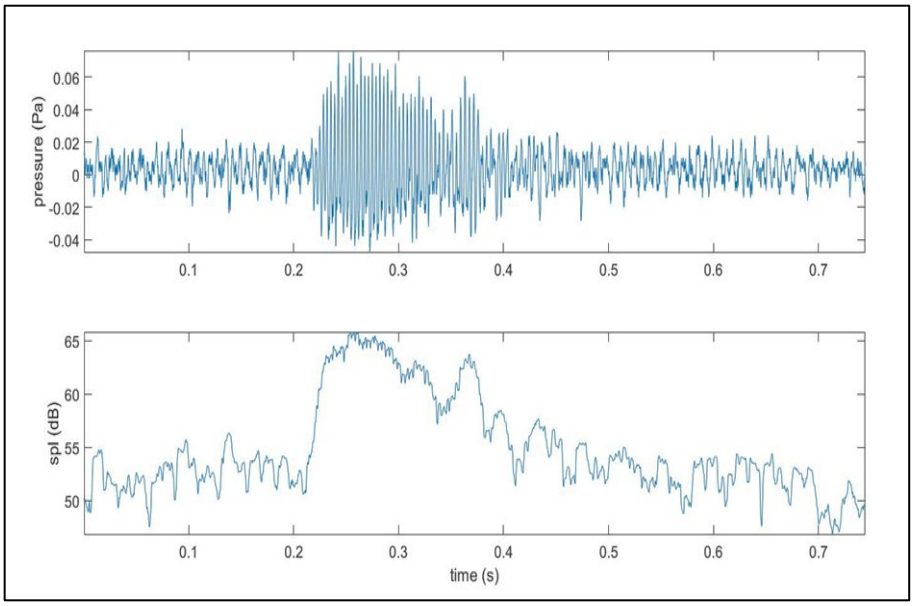

Figure 3: Sample of sound (letter 'B') for UOTletters database

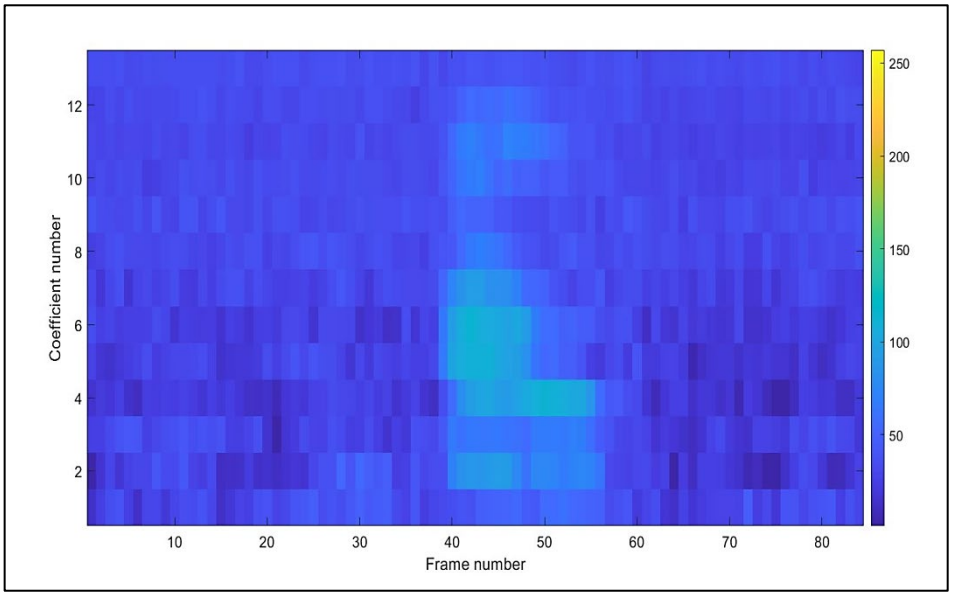

Figure 4: PNCC features extraction for a sample speaker pronounced the letter "A"

3.2 Classification

The proposed classifier (Weighted-K-NN-DTW) is a modification of the K-NN classifier explained in section 2.3. The proposed classifier procedure steps are shown below:

1) Set the K of the neighbor features.
2) Find the similarity between the test and trained features using DTW, as shown in Equation18.
3) Calculate the K nearest neighbors per each class's DTW, as shown in Equation 19:

$$m_{i,k} = Min(D_{(i,1)}, D_{(i,2)}, \dots \dots D_{(i,j)}) \quad (19)$$

where: D is the similarity based on DTW.

4) Calculate the weight (w) for each set of K nearest neighbors using Equation 16.
5) Compute the sum of weights for each class (M_i) using Equation 17.
6) Composed the weight matrix of the set of K nearest neighbors based on each class index (the index represents the class label) , as shown in Equation 20:

$$M = \{M_1, M_2, \dots M_i \dots \dots M_I\} \quad (20)$$

7) Select the maximum weight of the set of K nearest neighbors, as shown in Equation 21:

$$M_{max.} = Max(M_i) \quad (21)$$

where: $M_{max.}$ is the maximum weight in the (M) matrix.

8) The recognized class represents the spatial position represented by the (M) matrix index, as shown in Equation 22:

$$i_{max.} = index(M_{max.}) \quad (22)$$

where: $i_{max.}$ is the maximum index in the (M) matrix.

4. Results and Discussion

In this paper, the accuracy of the proposed algorithm was calculated with different levels of white noise (noise-free, 20dB, 15dB, 10dB, and 5dB). The accuracy of the proposed algorithm in a noise-free environment reached 100%, as shown in Figure 5. The k-nearest number is set to 5 according to experiments with respect to the accuracy, as shown in Figure 6, which has the higher accuracy at the k-folded point with an accuracy of 100%.

The proposed algorithm was a result of experiments of comparison on PNCC and MFCC features extraction techniques with different classifiers (proposed KNN-DTW classifier, weighted-KNN, Medium KNN, Linear SVM, and Quadratic SVM), as shown in Table 1. The best accuracy results were for the proposed KNN-DTW classifier in both techniques (PNCC and MFCC), 100% and 98.72%, respectively. At the same time, the Weighted KNN was 99.97% and 96.67% for the PNCC and MFCC, respectively. In the case of Medium-KNN, accuracy had a drop in its values of 89.54% for PNCC and 86.55 for MFCC. Nevertheless, a significant drop in accuracy was shown in SVM classifiers: Quadratic SVM has an accuracy of 76.97% and 74.41% for the PNCC and MFCC, respectively. In comparison, the worse accuracy was for Linear SVM of 60.27% and 40.45% for the PNCC and MFCC, respectively. Consequently, modifying the weighted KNN classifier by DTW improved the accuracy due to the accuracy of DTW in show similarity rather than the Euclidian distance. Moreover, integrating the proposed classifier with PNCC feature extraction improves the accuracy in a noise-free environment.

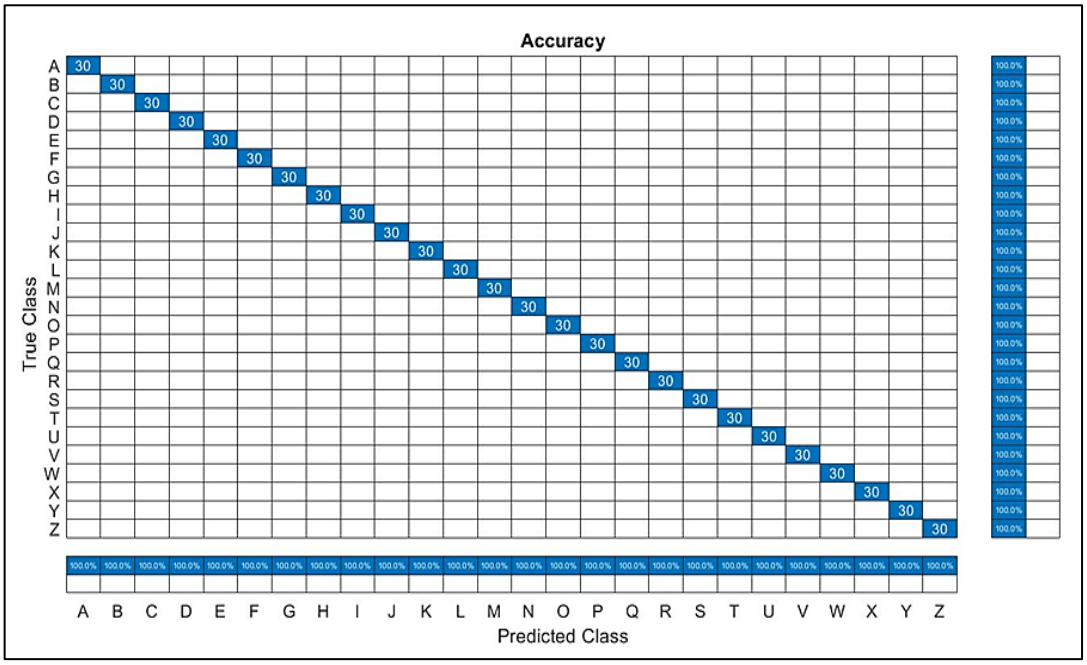

Figure 5: Accuracy of the proposed algorithm in a noise-free environment

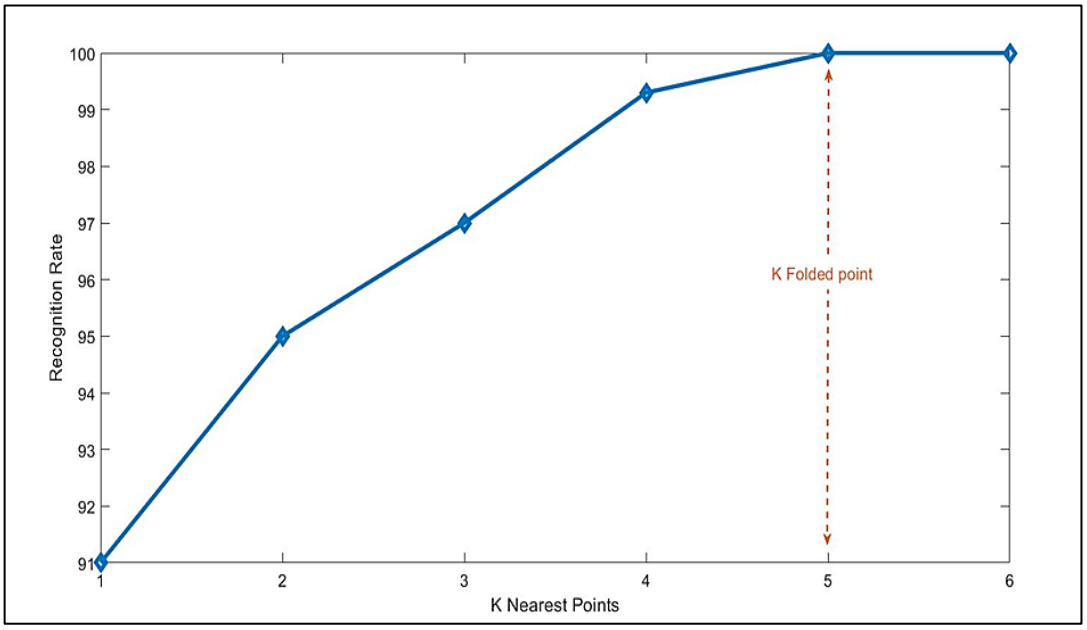

Figure 6: The accuracy of the proposed algorithm increases with respect K-nearest number

For the noisy environments, the two approaches, PNCC features extraction with the Proposed KNN-DTW classifier (proposed algorithm), and MFCC features extraction with the KNN-DTW classifier, were selected to investigate the most immunity approach to noise. As shown in Table 2, the approach based on MFCC features extraction with the Proposed KNN-DTW classifier has given recognition rates of 98.72, 72.19, 18.75, 9.15, and 3.16 for noise levels (20, 15,10, and 5) dB, respectively. On the other hand, the approach based on PNCC features extraction with the Proposed KNN-DTW classifier has given recognition rates of 82.05, 50.90, 25.77, and 12.05 for noise levels (20, 15,10, and 5) dB, respectively. As results have shown, the proposed algorithm based on PNCC has more immunity to noise in all five noise levels with an average accuracy difference of 13.67% with respect algorithm based on MFCC.

As a practical comparison between the proposed algorithm and related works, the proposed algorithm has higher accuracy than the algorithms based on (Weighted K-NN and PNCC) in [7] (KNN and MFCC) in [14,17], and (SVM and PNCC) in [21], as shown in Table 1.

In the noise effect experiments shown in Table 2. the proposed classifier is a modification of the KNN, and it can be taken as the optimal alternative classifier in noisy environments instead of the classifiers [14,17] for MFCC and classifiers in [7,21] for PNCC. Under this assumption, the proposed algorithm has higher accuracy than the related works in noisy environments, as shown in Table 2.

Moreover, the proposed algorithm is compared with related works based on the accuracy and the size of the database of studies, as shown in Table 3. The results have shown that the proposed algorithm is higher accuracy than the baseline algorithms in [14, 7,16-19,21,22], and less cost than the algorithm in [20,22].

Table 1: Accuracy of PNCC and MFCC with Different Classifiers

Classifier	PNCC Accuracy %	MFCC Accuracy %
Proposed WKNN-DTW	100	98.72
Weighted KNN	99.97	96.67
Medium KNN	89.54	86.55
Linear SVM	60.27	40.45
Quadratic SVM	76.97	74.41

Table 2: Accuracy of the PNCC and MFCC for the proposed classifier

Noise Level	PNCC Accuracy %	MFCC Accuracy %
Noise-free	100	98.72
20 dB	82.05	72.19
15 dB	50.90	18.75
10 dB	25.77	9.15
5 dB	12.05	3.61
Average	54.15	40.48

Table 3: Comparison between the proposed algorithm and baseline algorithms

Features	Classifier	Size of database (Words x Speakers)	Accuracy %	Ref.
MFCC	KNN	10x50	76.80	[14]
MFCC	KNN-DTW	10x5	98.40	[16]
MFCC	KNN	20x1	85.00	[17]
LPC	KNN	28x6	78.92	[18]
CFCC+PCA	SVM	20x16	88.43	[19]
LPC+MFCC+Energy+ZCR	1-NN	8x10	100	[20]
PNCC	SVM	18x40	97.50	[21]
DCMLBP+DWT+NCA	SVM	8x60	99.97	[22]
PNCC	WKNN	26x30	99.97	[7]
PNCC	Proposed WKNN-DTW	26x30	100	The proposed algorithm

5. Conclusion and Future Works

This paper proposed a speech recognition algorithm for controlling robots in a noisy environment. The outcomes demonstrated that the suggested method is more accurate than the baseline and related work. Moreover, it has more immunity to noisy environments by 13.67% with respect baseline algorithm. Nevertheless, the proposed system needs to be improved in the high noise level. One of the future works suggestions is to integrate the proposed audio speech recognition with a visual lip-reading algorithm to improve accuracy in noisy environments, which is the reason for using the audio of the UOT letters database. Another future work is implementing a hardware interface between the proposed algorithm and a mobile robot.

Author contributions

Conceptualization, M. Safi and E. Abbas; methodology, M. Safi.; software, M. Safi validation, M. Safi and E. Abbas; formal analysis, M. Safi; investigation, M. Safi; resources, M. Safi; data curation, M. Safi; writing—original draft preparation, M. Safi; writing—review and editing, M. Safi; visualization, M. Safi; supervision, E. Abbas; project administration, M. Safi and E. Abbas. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author.

Conflicts of interest

The authors declare that there is no conflict of interest.

References

[1] C. G. Le Prell, O. H. Clavier, Effects of noise on speech recognition: Challenges for communication by service members, Hear. Res., 349 (2017) 76–89. https://doi.org/10.1016/j.heares.2016.10.004

[2] E. I. Abass, M. E. Safi, Speech Recognition Based Microcontroller for Wheelchair Movement, Eng. Tech. J., 32 (2014)

[3] C. Kim, R. M. Stern, Power-Normalized Cepstral Coefficients (PNCC) for Robust Speech Recognition," IEEE/ACM Trans. Audio Speech Lang. Process., 24 (2016) 1315–1329. https://doi.org/10.1109/TASLP.2016.2545928

[4] C. Kim, R. M. Stern, Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction, Proc. Annu. Conf. Int. Speech Commun. Assoc. Brighton, UK, September (2009) 28–31.

[5] F. De-La-Calle-Silos, R. M. Stern, Synchrony-Based Feature Extraction for Robust Automatic Speech Recognition, IEEE Signal Process. Lett., 24 (2017) 1158–1162. https://doi.org/10.1109/LSP.2017.2714192

[6] T. Fux, D. Jouvet, Evaluation of PNCC and extended spectral subtraction methods for robust speech recognition, 23rd Eur. Signal Process. Conf. (2015) 1416–1420. https://doi.org/10.1109/EUSIPCO.2015.7362617

[7] M. E. Safi, E. I. Abbas, Isolated word recognition based on PNCC with different classifiers in a noisy environment, Appl. Acoust., 195 (2022) 108848. https://doi.org/10.1016/j.apacoust.2022.108848

[8] M. Khan, T. Goskula, M. Nasiruddin, R. Quazi, Comparison between k-nn and SVM method for speech emotion recognition, Int. J. Comput. Sci. Eng., 3 (2011) 607–611.

[9] R. Amami, D. B. Ayed, N. Ellouze, An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel Recognition, Int. J. Intell. Inf. Process., 3 (2012). https://doi.org/10.4156/IJIIP.vol3.issue1.6

[10] K. Chaka, N. Le Thanh, R. Flamary, C. Belleudy, Performance Comparison of the KNN and SVM Classification Algorithms in the Emotion Detection System EMOTICA, Int. J. Sens. Net. Data Commun., 7 (2018) 1–9. https://doi.org/10.4172/2090-4886.1000153

[11] S. Prabavathy, V. Rathikarani, P. Dhanalakshmi, Classification of Musical Instruments using SVM and KNN, Int. J. Innov. Technol. Explor. Eng., 9 (2020) 1186–1190, https://doi.org/10.35940/ijitee.G5836.059720

[12] N. A. J. Gnamele, Y. B. Ouattara, T. A. Kobea, G. Baudoin, J. M. Laheurte, KNN and SVM classification for chainsaw sound identification in the forest areas, Int. J. Adv. Comput. Sci. Appl., 10 (2019) 531–536. https://doi.org/10.14569/ijacsa.2019.0101270

[13] L Chen, S Gunduz, M. T. Ozsu, Mixed Type Audio Classification with Support Vector Machine, 2006 IEEE international conference on multimedia and expo. IEEE, (2006) 781–784. https://doi.org/10.1109/ICME.2006.262954

[14] Z. Ali, A. W. Abbas, T. M. Thasleema, B. Uddin, T. Raaz, S. A. R. Abid, Database development and automatic speech recognition of isolated Pashto spoken digits using MFCC and K-NN, Int. J. Speech Technol., 18 (2015) 271–275. https://doi.org/10.1007/s10772-014-9267-z

[15] M. E. Safi, E. I. Abbas, Microcontroller - Controlled security door based on speech recognition, Al-Sadiq Int. Conf. Multidisciplinary in IT and Comm. Sci. Appl., (2016) 1-6. https://doi.org/10.1109/AIC-MITCSA.2016.7759909

[16] M. A. Imtiaz, G. Raja, Isolated word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN, Asia Pacific Conf. on Multimedia and Broadcasting (APMediaCast), Bali, Indonesia, (2016) 106-110. https://doi.org/10.1109/APMediaCast.2016.7878163

[17] D. Anggraeni, W. S. M. Sanjaya, M. Munawwaroh, M. Y. S. Nurasyidiek, I. P. Santika, Control of robot arm based on speech recognition using Mel-Frequency Cepstrum Coefficients (MFCC) and K-Nearest Neighbors (KNN) method, Int. Conf. Advan. Mechatronics, Intelligent Manufacture, and Industrial Automation, Surabaya, Indonesia, (2017) 217-222. https://doi.org/10.1109/ICAMIMIA.2017.8387590

[18] Adiwijaya, M. N. Aulia, M. S. Mubarok, W. Untari Novia, F. Nhita, A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters Pronunciation classification system, 5th International Conference on Information and Communication Technology, Melaka, Malaysia, (2017) 1-5. https://doi.org/10.1109/ICoICT.2017.8074689

[19] Y. Shi, J. Bai, P. Xue, D. Shi, Fusion Feature Extraction Based on Auditory and Energy for Noise-Robust Speech Recognition, IEEE Access, 7 (2019) 81911–81922. https://doi.org/10.1109/ACCESS.2019.2918147

[20] Y. Korkmaz, A. Boyacı, T. Tuncer, Turkish vowel classification based on acoustical and decompositional features optimized by Genetic Algorithm, Appl. Acoust., 154 (2019) 28–35. https://doi.org/10.1016/j.apacoust.2019.04.027

[21] A. A. Alasadi, T. H. Aldhayni, R. R. Deshmukh, A. H. Alahmadi, A. S. Alshebami, Efficient Feature Extraction Algorithms to Develop an Arabic Speech Recognition System, Eng. Technol. Appl. Sci. Res., 10 (2020) 5547–5553. https://doi.org/10.48084/etasr.3465

[22] T. Tuncer, E. Aydemir, S. Dogan, Automated ambient recognition method based on dynamic center mirror local binary pattern : DCMLBP, Appl. Acoust., 161 (2020) 107165. https://doi.org/10.1016/j.apacoust.2019.107165

[23] C. Kim, Signal Processing for Robust Speech Recognition Motivated By Auditory Processing, Diss. Johns Hopkins University, 2010.

[24] C. Kim, R. M. Stern, Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring, IEEE Int. Conf. on Acoustics, Speech and Signal Process., Dallas, TX, USA, (2010) 4574-4577. https://doi.org/10.1109/ICASSP.2010.5495570

[25] H. Hermansky, N. Morgan, RASTA Processing of Speech, IEEE Trans. Speech Audio Process., 2 (1994) 578–589. https://doi.org/10.1109/89.326616

[26] D. Gelbart, N. Morgan, Evaluating long-term spectral subtraction for reverberant ASR, IEEE Work. Autom. Speech Recognit. Understanding, Madonna di Campiglio, Italy, (2001) 103-106. https://doi.org/10.1109/ASRU.2001.1034598

[27] H. Hermansky, S. Sharma, TempoRAl Patterns (TRAPs) in ASR of noisy speech, IEEE Int. Conf. Acoust. Speech Signal Process., Phoenix, AZ, USA, 1 (1999) 289-292 . https://doi.org/10.1109/ICASSP.1999.758119

[28] S. Thomas, S. Ganapathy, H. Hermansky, Recognition of Reverberant Speech Using Frequency Domain Linear Prediction, IEEE Signal Process. Lett., 15 (2008) 681–684. https://doi.org/10.1109/LSP.2008.2002708

[29] S. P. Rath, D. Povey, K. Veselý, J. H. Černocký, Improved feature processing for deep neural networks, Proc. Annu. Conf. Int. Speech Commun. Assoc., (2013) 109–113. https://doi.org/10.21437/interspeech.2013-48

[30] C. Kim, R. M. Stern, Feature extraction for robust speech recognition based on maximizing the sharpness of the power distribution and on power flooring, IEEE Int. Conf. Acoust. Speech Signal Process., Dallas, TX, USA, (2010) 4574-4577. https://doi.org/10.1109/ICASSP.2010.5495570

[31] Ranny, Voice recognition using k nearest neighbor and double distance method, Int. Conf. Ind. Eng. Manag. Sci. Appl., Jeju, Korea (South), (2016) 1-5. https://doi.org/10.1109/ICIMSA.2016.7504045

[32] T. Cover, P. Hart, Nearest Neighbor Pattern Classification, IEEE Trans. Inf. Theory, 13 (1967) 21-27. https://doi.org/10.1109/TIT.1967.1053964

[33] H. Bhavsar, A. Ganatra, A Comparative Study of Training Algorithms for Supervised Machine Learning, Int. J. Soft Comput. Eng., 2 (2012) 74–81.

[34] Z. Jan, M. Abrar, S. Bashir, A. M. Mirza, Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique, Springer-Verlag Berlin Heidelb., (2008) 40–51.

[35] J. E. S. Macleod, A. Luk, D. M. Titterington, A Re-Examination of the Distance-Weighted k-Nearest Neighbor Classification Rule, IEEE Trans. Syst. Man. Cybern., 17 (1987) 689–696. https://doi.org/10.1109/TSMC.1987.289362

[36] G. Fan, Y. Guo, J. Zheng, W. Hong, Application of the Weighted K-Nearest Neighbor Algorithm for Short-Term Load Forecasting, energies, 12 (2019). https://doi.org/10.3390/en12050916

[37] W. H. Ali, T. R. Saeed, M. H. Al-Muifraje, FPGA Implementation of Visual Speech Recognition System based on NVGRAM-WNN, Int. Conf. Comput. Sci. Software Eng., Duhok, Iraq, (2020) 132-137. https://doi.org/10.1109/CSASE48920.2020.9142095