2nd International Conference on Problems of Logistics, Management and Operation in The East-West Transport Corridor (PLMO 2023) Baku, Azerbaijan
DOI:10.54381/plmo2023.31

Intelligent car interior components controlling by means of Azeri (Azerbaijani) spoken language words

Rustam MAMMADLI, Ali PARSAYAN*
Process Automation Engineering Department, Baku Higher Oil School, Azerbaijan
**Corresponding author*. E-mail: parsayan@bhos.edu.az

Abstract. This paper aims to design a control system for car interior components by means of some Azeri (Azerbaijani) spoken language selected command words. By using Google Speech API various existing languages can be integrated, however, different approach is followed to implement a system that recognizes sone command words in Azeri (Azerbaijani) spoken language. The proposed method consists of two MFCC and RASTA-PLP based feature extraction and a CNN based classification algorithms. Four command words from Azeri (Azerbaijani) spoken language are considered as "Aç", "Bağla", "Qoş", and "Söndür". A Raspberry-Pi based embedded system recognizes the command words and sends the decision commands to the STM32 based embedded system thorough the UART protocol for controlling. The simulations and implementation resulted 91.6% recognition accuracy in utilizing of the MFCC feature extractor.

***Keywords*:** Speech Recognition; MFCC; RASTA-PLP; CNN; Raspberry-Pi; STM32.; Intelligent Car Interior Components Controlling; Azeri (Azerbaijani) Spoken Language

1. INTRODUCTION

Speech signal is one of the most important communication signals between humans and machines. An advantageous of it is that transmission of spoken language signal digitally is faster than typing or writing by hand. Speech recognition can be defined as a process in which the human speech signal is received and converted into the text. Environment is a deciding factor in speech recognition process as it can affect the results adversely. Over the years, different controlling techniques have been introduced and as they are arising, it is a matter of task to make a controlling system more humane. Car interior components controlling is important as people spend more time on the cars while commuting, travelling, transporting etc. Speech recognition provides important features of the speech signal as [1]:

- Determination of the signal contents.
- Determination of the speaker identity and gender.
- Determination of health conditions.
- Recognition of the language, accent and emotional conditions.

As the microcontrollers based embedded systems are small, cost-saving and provide interoperability, they are the best choices for implementation of a speech recognition based (some Azeri (Azerbaijani) spoken language selected command words) Intelligent car interior components controller. A deep learning-based classifier can be implemented on a Raspberry-Pi based embedded system's python environment for recognition of the spoken language words, and then, the resulted commands can be sent to a STM32 based embedded system through the UART communication protocol for controlling. In this research, the MFCC and RASTA-PLP based feature extraction algorithms for the front-end part and a CNN based classification algorithm for the back-end part were implemented. According to the results of the simulations and implementation, the MFCC feature extractor performs well in terms of

recognition accuracy (91.6%) and computation time in comparison with the RASTA-PLP feature extractor.

2. LITERATURE REVIEW

ASR (Automatic Speech Recognition) system itself contains both front-end and back-end parts. Front-end part extracts the features of a speech signal and represent them as a vector. Back-end part is used to make comparison between the vector of features and different reference model. Depending on the comparison, it gives an output and determines the words, sentences and etc. Different feature extraction methods have been implemented and developed for the front-end part as LPC (Linear Predictive Coding), MFCC (Mel-Frequency Cepstral Coefficient), PLP (Perceptual Linear Prediction), and Relative RASTA-PLP (Spectral Perceptual Linear Prediction Coefficients). MFCC feature extraction method have been developed by Davis and Mermelstein which consists of several steps such as selection of sampling rate, pre-emphasis, type of windowing technique and frequency wrapping [1-3]. In order to obtain better results, first and second order of derivative of static features coefficients can be taken. G. Zoric stated that the parameters which are extracted by using MFCC method are similar to the parameters that we use in understanding of a speech [4]. Rishiraj Mukherjee et al. stated that the perceived pitch and frequency of the tone can be measured in units of Mel which is not linearly dependent on the physical frequency, because the perceived pitch is not obtained linearly in our auditory system [5]. One of the reasons that MFCC method is considered as the best method for speech recognition is due to the fact that in using MFCC method, human perception sensitivity is defined depending on the frequencies. MFCC method is done by applying FFT (Fast Fourier Transform) on the pre-defined time frame and then power-spectrum is converted to Mel-Frequency Spectrum. After taking logarithm, we can apply IFFT (Inverse Fourier Transform) to get the results [6]. Another method is called Relative Spectral Perceptual Linear Predictive filtering (RASTA-PLP) [7]. It is the improved version of PLP method, although there are some differences. RASTA method contains a special band-pass filter, and it is used to smooth short-term noise problems in the log-spectral domain. For example, constant offset can be eliminated in the speech channel of the telephone. Vilas Thakare stated that RASTA removes offset which caused by the spectral coloration, wherefore a filter is applied to the energy in each frequency band [7]. The working principle of RASTA method depends on the slowly varying stimuli of the relative insensitivity of human auditory system. Suppression of the components which alternates with low rate improves the performance [8].

In recent years, machine learning techniques have become popular for several applications. Deep learning can be explained as a junction point for the graphical modelling, AI (Artificial Intelligence), ML (Machine Learning), optimization, and pattern recognition methods provide us processing huge chunks of data [9]. Dimitri Palaz et al. stated that DNN (Deep Neural Networks) method performs better than MLP (Multi-Layer Perceptron) with a single hidden layer [10]. DNN method can be utilized in context-dependent phonemes application, spectral features extraction application, Mel filter-bank energies with the basis of CNN (Convolutional Neural Networks) architecture application and Hybrid Principal application. Abdel-Hamid et al. investigated effectiveness of CNN methods in order to enhance performance of the hybrid model for the TIMIT phone recognition [11]. CNN method is able to learn about feature invariances and can be applied to both image analysis and recognition problems. Gain can be increased by using convolution and max polling over the frequency. Ali Bou Nassif et al. explained that convolutional layer has shared weights, whereas output is sub-sampled by the pooling layer, and it causes below layer data rate to be reduced [1]. There have been arguments about lack of invariance in CNN, which leads to ineffectiveness for pattern recognition problems. Nevertheless, CNN method has performed well in image recognition, computer vision, even with some modifications in speech recognition [11].

3. METHODOLOGY

Before analysing the technical details, we should understand how the speech signal is generated by a human. Speech can be categorized into different regions:

- Periodic signal of vocal folds contained part is called voiced region.
- Random signal part is called unvoiced region.
- No-excitation part refers to silence region.

In general, a person can hear a signal with frequency range (Bandwidth) of 20 Hz – 20 kHz, but this value can vary depending on the age. As shown figure 1, physiologically, vocal mechanism depends on the vocal tract, velum, nasal cavity, lips, tongue etc. [6]. So, basically, depending on the shape of the vocal tract, human can generate different sounds. The position of tongue and stretching/contracting of the throat matter when it comes to different phonemes. Overall, digital signal processing is analogous to this process. The vocal tract plays as a filter role. It can be stated as equation (1) that the speech signal can be described as a convolution of the vocal tract frequency response with glottal pulse [12].

$$X(t) = E(t) \cdot H(t) \quad (1)$$

By taking log of each side, it can be simplified the process mathematically as equations (2) and (3) respectively.

$$\log(X(t)) = \log(E(t) \cdot H(t)) \quad (2)$$

$$\log(X(t)) = \log(E(t)) + \log(H(t)) \quad (3)$$

So, with the help of logarithmic property, the glottal pulse and vocal tract frequency responses are separated from each other, where X(t) is the speech signal, E(t) is the glottal pulse signal, and H(t) is the vocal tract signal all in frequency domain. This separation is used in feature extraction stage of the speech recognition process. After feature extraction stage by means of MFCC or RASTA-PLP methods, the CNN classifier can classify those target Azeri (Azerbaijani) Spoken Language Words which will be sent for controlling. Therefore, the speech recognition, feature extraction, and classification processes of the proposed method can be described as the block diagram of figure 2.

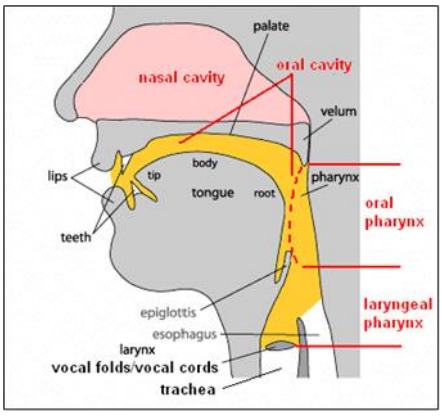

Fig. 1. Human speech generation vocal mechanism.

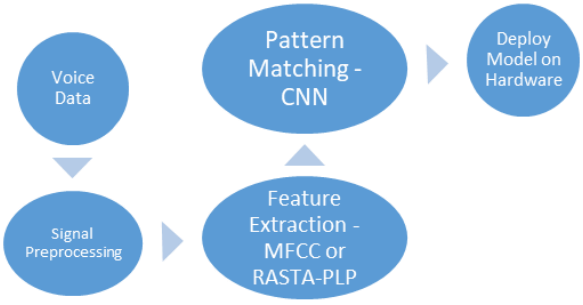

Fig. 2. The proposed method‘s block diagram

3.1 Speech Signal Pre-Processing

The structure of human vocal system has been created in a way that the sudden changes in the sound are not possible. Hence, in the digital processing of the sound signal, it can be divided into the short time frames because the change of the speech signal during short period is negligible. The short time frames construct the speech signal, and they overlap correspondingly to smooth the transition between the frames. Frame size should be chosen optimally, because if it to be short, enough samples will not available to get the spectral or if it to be long, the signal will be exposed to changes and disturb the stability. Figure 3 represents the basic steps for the pre-processing of the speech signal [6].

Fig. 3. Speech signal pre-processing steps

3.2 Mel-scale Frequency Cepstral Coefficients (MFCC)

Mel-scale is used for measurement of the pitches evaluated by the listeners who are equal equidistant from each other. Transition to the Mel-scale means that the signal frequency is transformed logarithmically. The main purpose of using Mel-scale transformation is relevant to the human hearing system. The human hearing system is less sensitive to the frequencies which have the value greater than 1000 Hz. So, the recipe to get Mel-spectrogram is applying STFT (Short Time Fourier Transform), converting amplitudes to DBs and then frequencies to the Mel-scale. Generally, the first 12-13 coefficients will be used while dealing with MFCC. To get the better results, first and second derivatives (Δ, $\Delta\Delta$) of the MFCCs can be taken. The first derivative can be obtained by subtracting the previous frame from the current frame. For obtaining the second derivate, the first derivate will be obtained and then the same process will be applied as explained previously. Overall, there will be 39 coefficients per each frame. Figure 4 sums up the steps for MFCC.

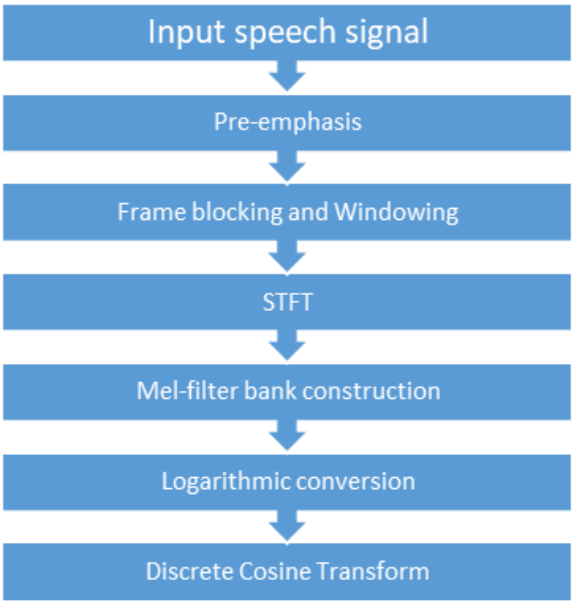

Fig. 4. The MFCC feature extraction method's processing steps

3.3 Relative Spectral Perceptual Linear Prediction (RASTA-PLP)

The alternative method which used for feature extraction is RASTA-PLP. To understand its working principle, first the PLP (Perceptual Linear Prediction) feature extraction method should be analysed, because RASTA-PLP and PLP methods are similar to each other, except that the RASTA-PLP method uses a band-pass filter for each spectral element. There are three main steps which constitutes basis for the PLP-related analysis. They are critical band frequency, curve of equal-loudness and power law between intensity and loudness processing steps [6]. Figure 5 represents the block diagram of the RASTA-PLP method processing steps.

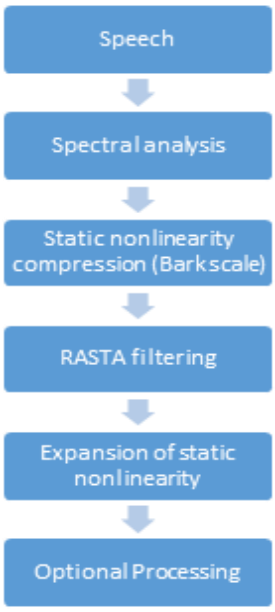

Fig. 5. The RASTA-PLP feature extraction method's processing steps

3.4 Convolutional Neural Network (CNN)

The CNN classifier is an algorithm which is able to learn and differentiate features in the image by processing them with weights and biases. It is mainly used for image processing and is more effective than MLP and contains less parameters than dense layers. The CNN classifier aims emulation of the human vision system. Components of a CNN is subjected to learn extraction of different features. Main CNN components are Convolution and Pooling. Convolution contains kernel, also known as grid of weights [12]. Kernel is applied over image by sliding till the end. During feature extraction, spectrogram of a sound signal is obtained. So, its image too. Only the mapping of the parameters of the sound signal with convolution/pooling process parameters must be done. For example, the amplitude parameter represents the pixel value, and the time and frequency parameters depict the size.

3.5 System Hardware Architecture and Flowchart

A Raspberry-Pi based embedded system is used for recognizing of those target Azeri (Azerbaijani) spoken language words and sends the decision commands to the used STM32 based embedded system thorough the UART protocol for controlling. Figure 6 shows the connection between the hardware components. Therefore, flowchart of the proposed method can be given as figure 7.

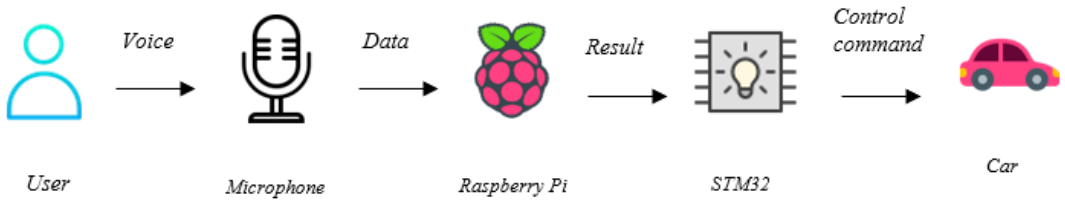

Fig. 6. Hardware architecture of the proposed method

4. SIMULATIONS, IMPLEMENTATION AND DISSCUSION ON THE RESULTS

Database was collected from 40 different people. Each of them recorded 1 second of audio file for four Azeri (Azerbaijani) Spoken Language command words. Audio files were saved as .wav files. However, this data can be insufficient for training. Therefore, data augmentation techniques were applied as the pitch shifting technique. Python program's librosa.effects.pitch_shift function was used to adjust of the pitch of the sound and the results be saved as the new audio files. In this function, N_steps parameter shows the steps to shift the sound. In conclusion, data was increased to 80 samples for each command. Also, other pre-processing algorithms as pre-emphasizing were applied on all content of the database. Figure 8 shows the difference between original and pre-emphasized versions of "Bağla" command word implemented with Python program's librosa.effects.preemphasis function.

After applying the signal pre-processing steps on all content of the collected database as "Bağla" command word, all MFCC coefficients were obtained directly using Python program's librosa.feature.mfcc function. The delta MFCCs also were obtained by using delta function and changing the order to relevant value. Figure 9 represents all MFCC coefficients for "Bağla" command word pictorially. Pre-processing of the same command word using the RASTA-PLP relevant function resulted the spectral Features of figure 10 and the cepstral Features of figure 11.

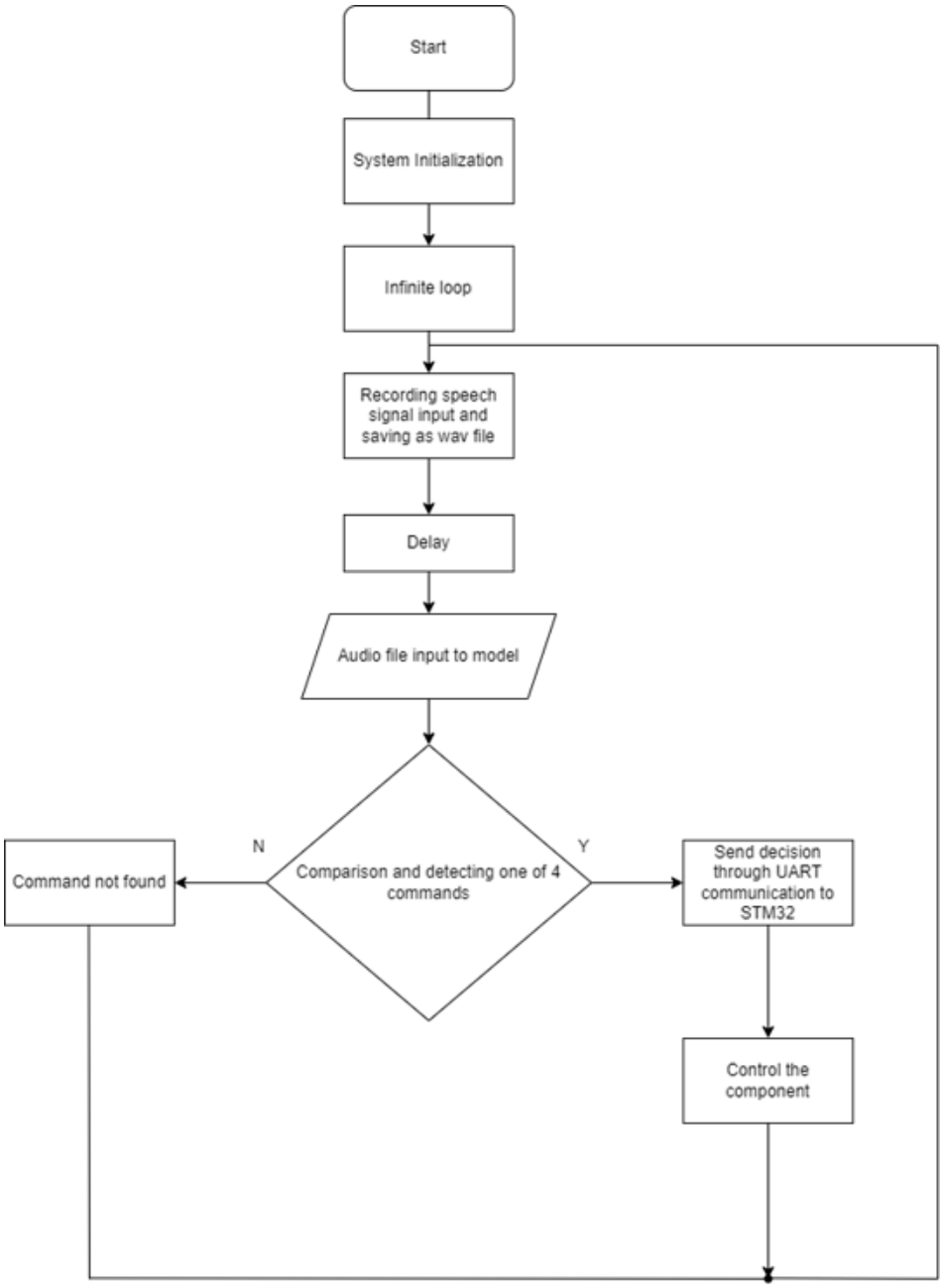

Fig. 7. Flowchart of the of the proposed method

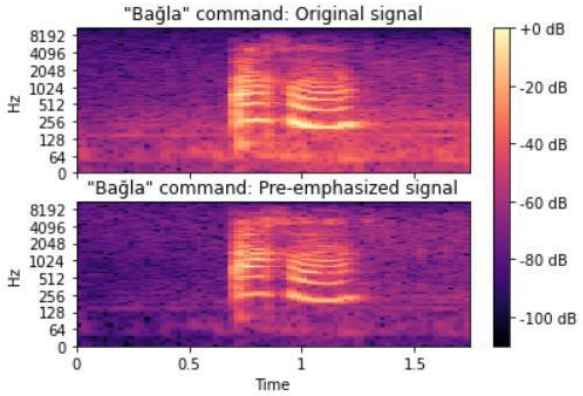

Fig. 8. Difference between original and pre-emphasized versions of "Bağla" command word

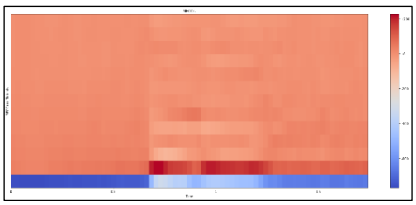

Fig. 9. All MFCC coefficients for “Bağla” command word

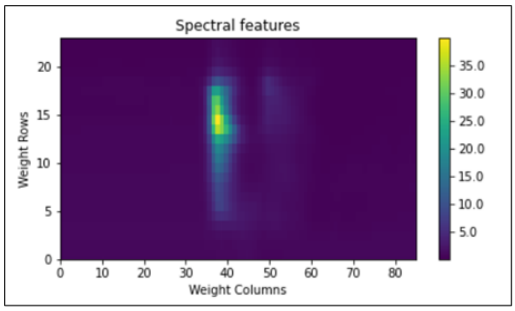

Fig. 10. All RASTA-PLP method based spectral features for “Bağla” command word

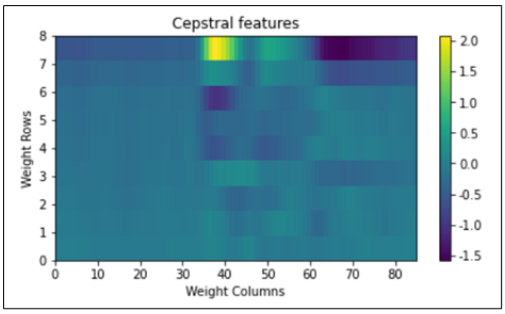

Fig. 11. All RASTA-PLP method based cepstral features for “Bağla” command word

After obtaining the pre-processed database, the simulated CNN was trained by 80% of the database for both MFCC and RASTA-PLP methods based extracted features. Figures 12 and 13 show the training and validation accuracies and figures 14 and 15 shows the training and validation losses for both methods. After training process, it was observed that the trained CNN by the MFCC method based extracted features outperformed on the one by RASTA-PLP. Then, the trained CNN evaluated by the remined 20% content of the database as the test part. Table 1 represents the results which show more 91% of recognition accuracy of the CNN which is trained by MFCC method based extracted features for the test data. Hence, the implemented CNN model was trained for the MFCC method based extracted features.

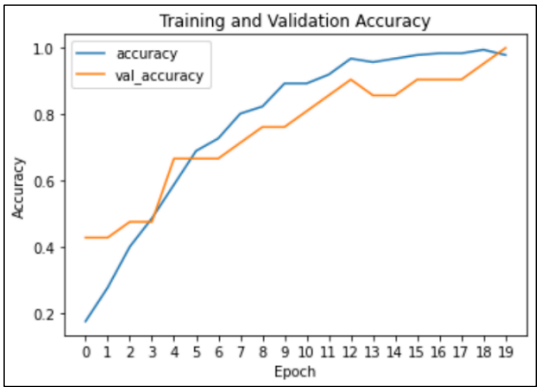

Fig. 12. The training and validation accuracies of CNN for MFCC method based extracted features

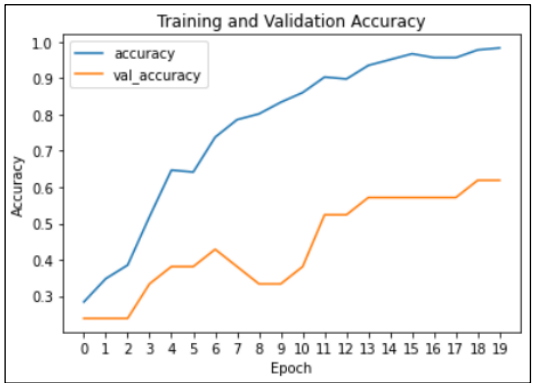

Fig. 13. The training and validation accuracies of CNN for RASTA-PLP method based extracted features

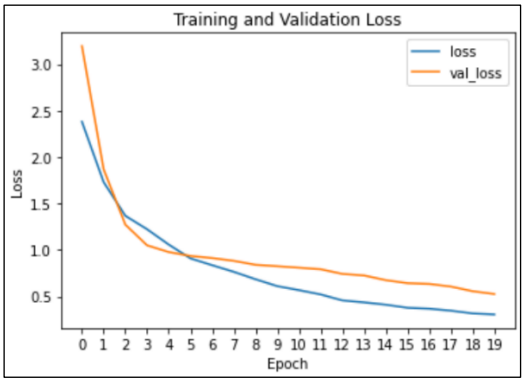

Fig. 14. The training and validation losses of CNN for MFCC method based extracted features

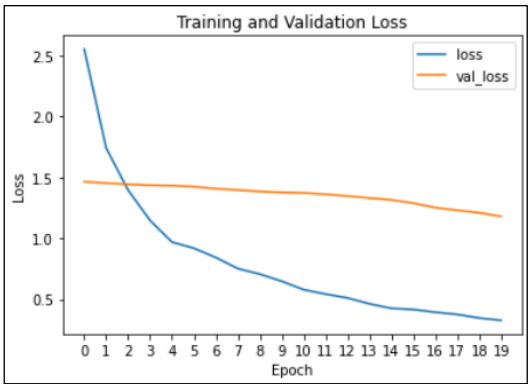

Fig. 15. The training and validation losses of CNN for RASTA-PLP method based extracted features

Table 1. Comparison table for the trained CNN.

	Training Accuracy	Validation Accuracy	Test Accuracy	Feature extraction Computation time
MFCC	1.0	0.97	91.6%	7.31 seconds
RASTA-PLP	0.98	0.81	75%	11.02 seconds

Finally, the hardware was implemented using a Raspberry-Pi based embedded system and its python environment for recognition of the Azeri (Azerbaijani) spoken language selected command words and sending the commands, and a STM32 based embedded system with its UART based connection with the used Raspberry-Pi for controlling. Figures 16 and 17 show the UART based connection between the used embedded systems.

Fig. 16. UART Communication

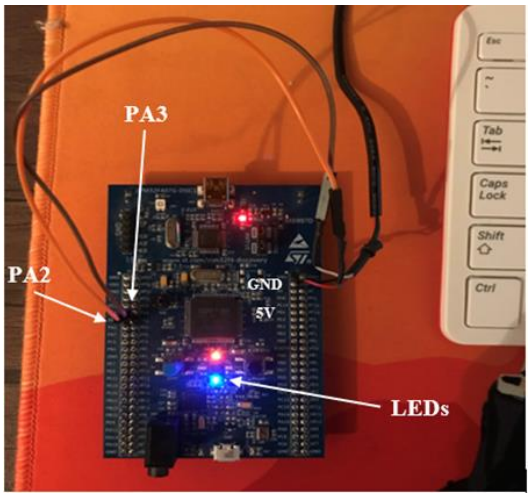

Fig. 17. PA2 and PA3 pins as Transmitter and Receiver

5. Conclusion

In this research, the MFCC feature extraction method were used for training of the classifier as the front-end part, and a CNN as the classifier for the back-end part. Also, a hardware was implemented using a Raspberry-Pi based embedded system and its python environment for recognition of the Azeri (Azerbaijani) spoken language selected command words as the feature extractor and classifier hardware, then sending the commands to the controller, and a STM32 based embedded system with its UART based connection with the used Raspberry-Pi as the controller for controlling of a car interior components intelligently. The simulations and implementation resulted 91.6% recognition accuracy. For improvement of the implemented system, increasing the size of the database, utilizing high speed and high-performance processing hardware, and be re-designing it for processing of the continuous speech signals rather than the separated command words can be followed.

References

1. Nassif, Ali Bou, et al. "Speech recognition using deep neural networks: A systematic review." IEEE access 7 (2019): 19143-19165..
2. Ursin, Markku. "Triphone clustering in Finnish continuous speech recognition." Diplomityö, Teknillinen korkeakoulu 2002..
3. Ali Parsayan, Mensur Gulami, Javid Mahmudov, Yusif Aliyev, Rovshan Akberov, “The First Azeri (Azerbaijani) Language Next Word Predictor.” Information Systems and Signal Processing Journal (2020) 5: 1-4.
4. Zorić, Goranka. “Automatic lip synchronization by speech signal analysis.” Diss. University of Zagreb. Faculty of Electrical Engineering and Computing. Department of Telecommunications, 2005.

5. Mukherjee, Rishiraj, Tanmoy Islam, and Ravi Sankar. "Text dependent speaker recognition using shifted MFCC." 2013 Proceedings of IEEE Southeastcon. IEEE, 2013.
6. Këpuska, Veton Z., and Hussien A. Elharati. "Robust speech recognition system using conventional and hybrid features of MFCC, LPCC, PLP, RASTA-PLP and hidden Markov model classifier in noisy conditions." Journal of Computer and Communications 3.06 (2015): 1.
7. Shrawankar, Urmila, and Vilas Thakare. "Feature extraction for a speech recognition system in noisy environment: A study." 2010 second international conference on computer engineering and applications. Vol. 1. IEEE, 2010.
8. Nayana, P. K., Dominic Mathew, and Abraham Thomas. "Performance comparison of speaker recognition systems using GMM and i-vector methods with PNCC and RASTA PLP features." 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT). IEEE, 2017.
9. Cho, Youngmin, and Lawrence Saul. "Kernel methods for deep learning." Advances in neural information processing systems 22 (2009)..
10. Palaz, Dimitri, and Ronan Collobert. “Analysis of CNN-based speech recognition system using raw speech as input.” No. REP_WORK. Idiap, 2015.
11. Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." IEEE/ACM Transactions on audio, speech, and language processing 22.10 (2014): 1533-1545.
12. Valerio Velardo, “The Sound of AI”, 2020.