Mel Frequency Cepstral Coefficients Based Similar Albanian Phonemes Recognition

Bertan Karahoda[1], Krenare Pireva[1(✉)], and Ali Shariq Imran[2]

[1] Faculty of Computer Science and Engineering, UBT, Pristina, Kosovo
{bkarahoda,krenare.pireva}@ubt-uni.net
[2] Faculty of Computer Science and Media Technology,
Norwegian University of Science and Technology (NTNU), Trondheim, Norway
ali.imran@ntnu.no

Abstract. In Albanian language there are several phonemes that are similar in pronunciation like /q/ - /ç/, /rr/ - /r/, /th/ - /dh/ and /gj/ - /xh/. These phonemes are difficult to distinguish by human ear even for native speaking Albanians from different regions. The task becomes more challenging for automated speech systems, recognizing and classifying Albanian words and language due to the similar sounding phonemes. This paper proposes to use Mel Frequency Cepstral Coefficients (MFCC) based features to distinguish these phonemes correctly. The three layers back propagation neural network is used for classification. The experiments are performed on speech signals that are collected from different male and female native speakers. The speaker independent tests are performed for analyzing the performance of the classification. The obtained results show that the serial MFCC features can be used to classify the very similar speech phonemes with higher accuracy.

Keywords: Albanian phonemes · Similar phonemes classification · Serial MFCC features · Neural network classifier · Back propagation network

1 Introduction

Most of the current speech recognition techniques are designed to distinguish words. The word recognition is difficult to maintain since the system should be trained for each new word that is added in the database. A phoneme recognition approach could avoid this continual challenge, as a complete training could be made by using all possible phonemes of that particular language [1].

Although the Albanian language is widely spoken language, there are few works conducted to the Albanian phoneme recognition in some multilingual approaches [2]. Also, there is a lack of Albanian speech corpus which makes it difficult to conduct experiments for Albanian speech recognition systems. In Albanian language however, there is a problem when processing phoneme for recognition, as there are couples of phonemes that sound very similar but in fact are different, like the phonemes strong /rr/ and light /r/. Also the accent

S. Yamamoto (Ed.): HIMI 2016, Part I, LNCS 9734, pp. 491–500, 2016.
DOI: 10.1007/978-3-319-40349-6_47

of Albanian language differs from region to region. This further adds to the difficulty of correctly recognizing these phonemes.

Many speech phoneme recognition techniques are proposed by various researchers for several languages [1,3,4,7]. In speech recognition approaches, extracting the features that describe best the characteristics of the speech signal content is the most important process for further processing the speech signals. For feature extraction, the MFCC [3], the Linear Prediction based Cepstral Coefficients (LPCC) [5], and Wavelet Transform (WT) [6] based methods are widely used. Other techniques are also used for speech recognition. In [7,8], the MFCC is combined with wavelet transform to increase the speech recognition performance. In [9], the fuzzy modeling approach is proposed for speech phoneme recognition, whereas, in [10] the auditory based scale-rate filter selection method is presented. Researchers in [11] suggested a new feature extraction method called Fisher Weight Map for speaker independent phoneme recognition.

Furthermore, the neural networks and hidden Markov models are widely used for speech signal classification [12,13]. Each of the aforementioned methods describes the characteristics of the speech signal under consideration, and justify the need to use a specific method, as to increase the recognition rate in their particular contribution. Since there is no work conducted to the similar Albanian phonemes recognition, in this work the significance of MFCC features are investigated for Albanian language and its ability for classifying the similar phonemes. MFCC features are widely used in automatic speech and speaker recognition. They were introduced by Davis and Murmelstein in the 1980s, and have been used in many systems ever since [3,7,8].

The rest of the paper is organized as follows. Section 2 presents the proposed method describing feature extraction and classification process. Section 3 discusses the experimental results while the conclusion is provided in the last section.

2 Proposed Method

The proposed method consists of four major steps as shown in Fig. 1: (a) speech detection; to detect speech part within a signal, (b) noise removal; to remove any unwanted noise from the speech signal, (c) Mel Cepstrum analysis; to extract discriminative features useful for phenoms classification, (d) and artificial neural network; to classify amongst different Albanian phonemes from speech signals. Each of the steps are discussed further in the following subsection.

2.1 Speech Detection

The speech detection is the first major step. In speech detection part, the start time of the initial consonant is detected while looking for a sudden increase of amplitude and intensity. Also, from this point is the beginning of the processing phase. Once the speech is detected from the signal, the next step is to remove any unwanted noise from it.

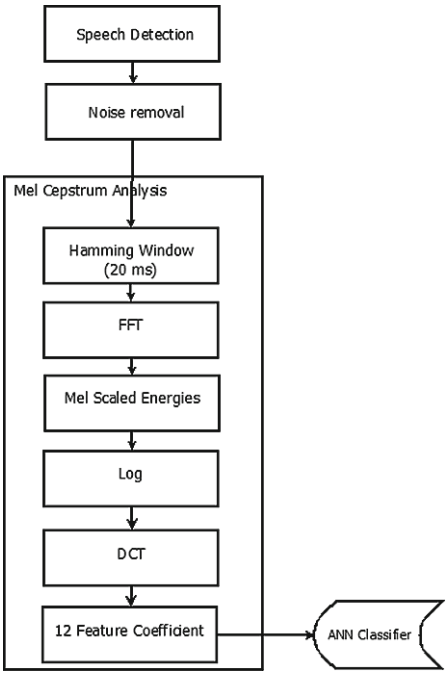

Fig. 1. Block diagram of proposed method

2.2 Noise Removal

The purpose of pre-processing steps is to remove any unwanted noise from the speech signal. As the Albanian phonemes sound very similar, therefore, it is important to obtain a better quality signal by removing unwanted noise. This will help obtain better features from the speech signal, ultimately improving classification results. The pre-emphasis filter is usually used as the first stage in processing the speech signals which improve the Signal to Noise Ratio (SNR). Furthermore, it is used to enhance specific speech information in higher frequencies and to calibrate the energy to analyse the wide spectrum of the speech signal. The pre-emphasis filter is expressed as follows [14]:

$$y(n) = x(n) - a * x(n-1) \tag{1}$$

The pre-emphasis filter coefficient used in our proposed method is set to 0.95, whereas, x(n) and y(n) are the values of input and output respectively.

2.3 Mel Cepstrum Analysis

The next step in the proposed method is to extract features for classification by performing Mel Cepstrum analysis on the speech signal.

The mel-scale frequencies reflect the human auditory system frequency response which are obtained from linear frequency f by using the Eq. (2):

$$Mel(f) = 2595log(1 + \frac{f}{700}) \tag{2}$$

For the various number of equally spaced mel frequencies the triangular filters are generated for linear frequencies obtained from Mel scaled frequency. Each filter is then multiplied with the Fourier spectrum of the original signal to obtain the log of the energies for Mel spaced filter banks. The discrete cosine transform (DCT) is used as the final step for obtaining the MFCC. The detailed steps are explained further in the following subsection.

2.4 Feature Extraction

The feature extraction process is as follows:

1. The detected speech signals are divided into 20 ms time intervals by using the hamming window. The overlap between the windows is set to 10 ms. The purpose of this overlap is to eliminate the spectral leakages. This is achieved because the time intervals of the hamming window is 20 ms and by keeping an overlap of half of the window size, spectral leakage can be avoided.
2. For each windowed signal the fast Fourier transform (FFT) is performed next to obtain the frequency amplitudes. The 20 Mel scaled filter banks are generated for the frequency range 300–5500 Hz, where most of the speech signals energies are concentrated. Fig. 2 shows the used mel filter banks and the frequency intervals.
3. Each Mel scaled filter bank is multiplied by windowed Fourier spectrum to obtain the energies of each filter bank.
4. Then DCT of the log of the energies obtained for 20 Mel scaled filter bank is computed and the first 12 coefficients of the DCT transform are used as feature vector for one windowed signal.
5. For approximately six-windowed signal the same procedure is repeated to obtain the 70 coefficients serial feature vector which corresponds to the approximately 70 ms time duration. Figure 3 shows the feature vectors obtained for the phoneme /rr/ from two different speakers.

2.5 Artificial Neural Network Classifier

The last step in the proposed method is the classification of Albanian phonemes using the extracted feature vectors. The three-layer backpropagation neural network is used for phonemes classification. The 70 coefficients feature vector obtained as a result of feature extraction process is fed to the 70 neurons in input layer. The 30 neurons are used in hidden layer and 8 neurons in output layer to classify the 8 different phonemes.

The speech phonemes from 4 male and 4 female speakers are used for training the network which created a total of 64 training samples for 8 phonemes. The structure of the used neural network classifier is given in Fig. 4.

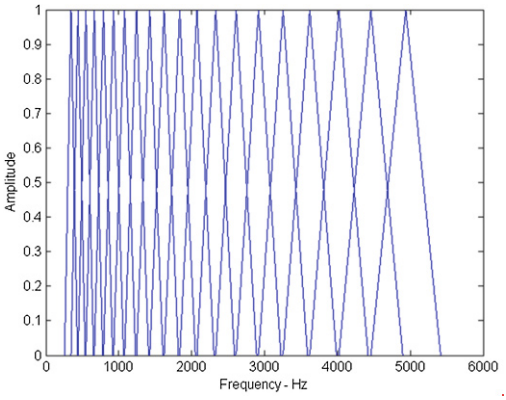

Fig. 2. Mel scaled filter banks

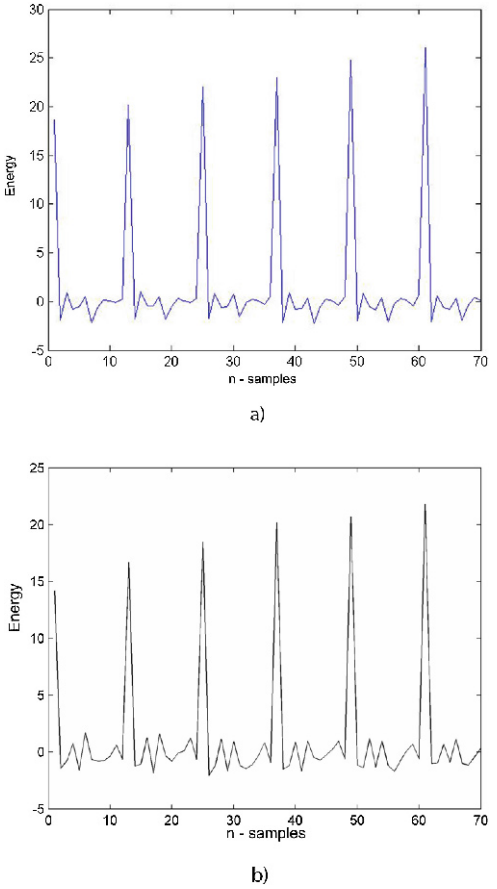

Fig. 3. The n-sample feature vectors; (a) Phoneme /rr/ female speaker, (b) phoneme /rr/ male speaker

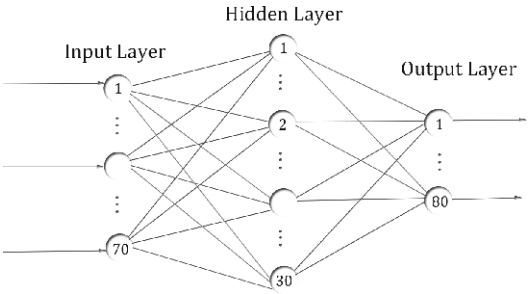

Fig. 4. Neural network structure (70 neurons in input layer, 30 neurons in hidden layer and 8 neurons in output layer)

3 Results

Due to the unavailability of Albanian language database, the speech samples are collected manually from people originally from different regions with diverse ascent. The speech signals are collected from 23 male and female speakers in silent environment with 48000 sample rate, which makes a total of 184 isolated phoneme signals. From those 184 phonemes, 64 of them are used for training the backpropagation neural network, and the rest 120 phonemes are used for testing.

All the speech signals contain the initial consonant followed by the vowel. The signal energies are computed for 20 ms durations, to detect the beginning of the initial consonants, the signal is subjected to the threshold. The threshold level is set to 0.2 in our case which is calculated experimentally. 120 phonemes from 15 male and female participants were used for testing which were not included in training set. The trained backpropagation network is tested with all 120 test phonemes, the results of which are discussed further.

Two types of tests are carried out, the first one is performed for each phoneme individually in separate classes as shown in Table 2, whereas, in the second test we have divided each pair of similar phonemes into four classes as shown in Table 1. The recognition tests is performed for each of these classes, as each class has different characteristics from other phonemes pairs. The tests are performed without using any filter for noise reduction.

Table 1 shows the correct recognition rates for each class which is calculated based on the number of correct recognitions. The highest recognition rate was achieved in the third class /rr-r/, whereas, the lowest recognition rate were in /dh-th/ class. The average recognition rate for all classes combined together is 72 %.

Table 2 shows the correct recognition rates for each phoneme in individual classes.

Table 1. Recognition performance for 4 phonmes classes

Phoneme classes	ç-q	dh-th	rr-r	gj-xh
Correct classification rate (%)	82	50	95	60
Overall performance (%)	72			

Table 2. Recognition performance for each phonmes classes

Phoneme classes	ç	q	dh	th	r	rr	gj	xh
Correct classification rate(%)	60	66.7	46.7	53.3	86.7	93.3	73.3	66.6
Overall performance(%)	68.3							

As it is seen in Table 2, the best recognition rate was obtained for the phoneme /rr/ followed by /r/, and the worst recognition rate was for phoneme /dh/. The overall recognition rate achieved is 68.3 % as depicted in Table 2, which makes the difference between the performances of two tests to 3.7 %, as can be compared from the overall performance from two tables (Tables 1 and 2).

For showing the number of correct and incorrect prediction compared to the actual outcomes, we used the confusion matrix. In order to compute the sensitivity, precision and accuracy rates, the confusion matrix for a two class classifier has been used [15]. Table 3 summarizes different attributes of the confusion matrix.

Table 3. Confusion matrix 2x2, for two classes (positive and negative)

		Prediction	
		P	N
Actual	P ´	TP(true positive)	FN (false negative)
	N´	FP(false positive)	TN (true negative)

- TP (true positive) is the number of correct predictions that an instance is positive,
- FN (false negative) is the number of incorrect of predictions that an instance negative,
- FP (false positive) is the number of incorrect predictions that an instance is positive,
- TN (true negative) is the number of correct predictions that an instance is negative,

The sensitivity or Recall rate (TPR) is the proportion of actual positive cases which are correctly identified. It is expressed as the proportion of number that are TP of all the numbers that are actual positive (TP+FN):

$$TPR = \frac{TP}{TP + FN} \tag{3}$$

Precision or Positive Predictive Value (PPV), is the proportion of positive cases that were correctly identified. It is expressed as the proportion of the number that are TP, of all the numbers that outcome positive (TP+FP):

$$PPV = \frac{TP}{TP + FP} \tag{4}$$

And, the accuracy (ACC) is the total number of predictions that were correct, calculated with Eq. (5), as follows:

$$ACC = \frac{TP + TN}{P + N} \tag{5}$$

The sensitivity, precision and accuracy rates for each phoneme class are represented in Table 4.

Table 4. Confusion matrix representing the sensitivity, precision and accuracy

Phonemes	ç	q	dh	th	r	rr	gj	xh
Sensitivity (%)	27	40	40	50	80	33	27	33
Precision (%)	33	40	67	50	43	50	31	33
Accuracy (%)	84	85	90	88	84	88	83	83

The accuracy of the classification for the individual 8 classes test is above 80 % for all the phoneme classes, whereas, the precision is above 60 % in the phoneme /dh/, and above 50 % in /th/ and /r/. The average accuracy rate is 86 %.

4 Conclusions and Future Work

The Albanian phonemes /ç/ - /q/, /rr/ - /r/, /th/ - /dh/ and /gj/ - /xh/ are very similar in pronunciation. They are also difficult to distinguish by human ear in a spoken Albanian language. This paper proposed a method for accurately recognizing Albanian phonemes using the MFCC based features and the three layers back propagation neural network. The combination of MFCC features and neural network classification model was able to classify the similar phonemes with acceptable recognition rate.

The overall performance for four phoneme classes was 72 %, where for each phoneme class the overall performance was 68.3 %. If we compare the recognition rate of other languages that used the same feature extraction method MFCC, for example, in Thai speech [1], the vowel was recognized with 67.71 %, whereas,

the initial consonant with 63.82 %. Comparing to that, the results of this paper are satisfactory, even though the classification accuracy is still very low.

In general, the difference between the overall performances of four classes of similar phonemes and the eight classes of individual phonemes are found to be miniscule, which shows that the MFCC based features can be used to classify speech signals which are very similar in pronunciation, with acceptable results. The accuracy of recognition rate was calculated using confusion matrix which resulted with 86 % of accuracy.

As a future work, the paper can be extended further by applying other feature extraction algorithms for speech recognition and evaluating the classification results using Bayes classifier and SVM. Further comparisons can also be made from the results obtained by other classifiers.

References

1. Theera Umpon, N., Chansareewittaya, S., Auephanwiriyakul, S.: Phoneme and tonal accent recognition for THAI speech. Exp. Syst. Appl. **38**(10), 13254–13259 (2011)
2. Caranica, A., Buzo, A., Cucu, H., Burileanu, C.: Speed@ mediaeval 2015: Multilingual phone recognition approach to query by example STD (2015)
3. Dabbaghchian, S., Sameti, H., Ghaemmaghami, M., BabaAli, B.: Robust phoneme recognition using MLP neural networks in various domains of MFCC features. In: 2010 5th International Symposium on Telecommunications (IST), pp. 755–759, December 2010
4. Sharifzadeh, S., Serrano, J., Carrabina, J.: Spectro-temporal analysis of speech for spanish phoneme recognition. In: 2012 19th International Conference on Systems, Signals and Image Processing (IWSSIP), pp. 548–551, April 2012
5. Zbancioc, M., Costin, M.: Using neural networks and LPCC to improve speech recognition. In: 2003 International Symposium on Signals, Circuits and Systems SCS 2003, vol. 2, pp. 445–448 (2003)
6. Sahu, P., Biswas, A., Bhowmick, A., Chandra, M.: Auditory ERB like admissible wavelet packet features for timit phoneme recognition. Int. J. Eng. Sci. Technol. **17**(3), 145–151 (2014)
7. Tavanaei, A., Manzuri, M., Sameti, H.: Mel-scaled discrete wavelet transform and dynamic features for the persian phoneme recognition. In: 2011 International Symposium on Artificial Intelligence and Signal Processing (AISP), pp. 138–140, June 2011
8. Xue-ying Zhang, X., Bai, J., zhou Liang, W.: The speech recognition system based on bark wavelet MFCC. In: 2006 8th International Conference on Signal Processing, vol. 1 (2006)
9. Halavati, R., Shouraki, S.B., Zadeh, S.H.: Recognition of human speech phonemes using a novel fuzzy approach. Appl. Soft Comput. **7**(3), 828–839 (2007)
10. Fartash, M., Setayeshi, S., Razzazi, F.: A scale-rate filter selection method in the spectro-temporal domain for phoneme classification. Comput. Electr. Eng. **39**(5), 1537–1548 (2013)
11. Muroi, T., Takiguchi, T., Ariki, Y.: Speaker independent phoneme recognition based on fisher weight map. In: 2008 International Conference on Multimedia and Ubiquitous Engineering MUE 2008, pp. 253–257. IEEE (2008)

12. Rahman, M., Islam, M.: Performance evaluation of MLPC and MFCC for HMM based noisy speech recognition. In: 2010 13th International Conference on Computer and Information Technology (ICCIT), pp. 273–276, December 2010
13. Paulraj, M., Bin Yaacob, S., Nazri, A., Kumar, S.: Classification of vowel sounds using MFCC and feed forward neural network. In: 2009 5th International Colloquium on Signal Processing Its Applications CSPA 2009, pp. 59–62, March 2009
14. Muda, L., Begam, M., Elamvazuthi, I.: Voice recognition algorithms using MEL frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. arXiv preprint (2010). arXiv:1003.4083
15. Visa, S., Ramsay, B., Ralescu, A.L., Van Der Knaap, E.: Confusion matrix-based feature selection. In: MAICS, pp. 120–127 (2011)