Identification of Letters Hijaiyah Pronunciation Using Neural Network (Backpropagation) and Pre-Processing of Mel-Frequency Cepstral Coefficient

1st Wafira Rahmania[1], 2nd Arini[1], 3rd Anif Hanifa Setyaningrum[1], 4th Arie Purnomosidi[2], 5th Muhammad Taufik Rusydi[2]
{wafirarahmania13@mhs.uinjkt.ac.id[1], arini@uinjkt.ac.id[1], anif.hanifa@uinjkt.ac.id[1], purnomosidi@unsa.ac.id[2], taufik@unsa.ac.id[2]}

UIN Syarif Hidayatullah, Department of Informatics Engineering, Jakarta, Indonesia[1], University of Surakarta, Faculty of Law, Indonesia[2]

Abstract. To avoid mistakes in pronouncing hijaiyah letters. The writer applies mel-frequency cepstral coefficient to extract and will yield characteristic value of voice signal. Implementation of Artificial Neural Networks (Backpropagation) is used for classification on the identification of 8 letters of hijaiyah using Matlab. 8 selected hijaiyah letters are س ص ذ ز ق ك ء ع take fathah. The feature extraction process produces several different parameter values, including pre-emphasis, windowing, fast fourier transform, discrete cosine transform, coefficient cepstrum and the duration. The backpopagation experiment using the maximum number of epoch and training functions varies as much as 15 times from each scenario capable of producing training regression 0.91019, test 0.93486, validation 0.99772 and MSE 0.2048. The test of hijaiyah pronunciation using trainlm with the number of hidden layer 10, obtained accuracy of 25%.

Keywords: Signal processing; Mel-Frequency Cepstral Coefficient; Artificial Neural Network (Backpropagation); Simulation;

1 Introduction

Sound is a major part of language. Oral communication will not be carried out if there is no sound spoken and heard. If this sound element is not considered, the spoken language will not be well understood, or it may be understood with a meaning that is far different from the speakers' intention [1]. It's the same as reading or reciting the Qur'an. For the sake of smoothness and goodness in reciting Arabic reading, each letter must be sounded according to its articulation. Errors in articulation can lead to differences in meaning or error in the reading being read [2].

Ifnani Ifka [3], in her research, found 53 words that experienced sound changes in the Saradan Village community. The details are as follows:

Tabel 1. Changes in pronunciation

No.	Number of Letter Changes	That Should be Pronounced	Pronounced Letter

ICONQUHAS 2018, October 02-04, Bandung, Indonesia

DOI 10.4108/eai.2-10-2018.2295283

1	3	ص	س
2	1	ع	ح
3	12	ح	ك
4	2	ح	ه
5	7	خ	/ko/
6	1	ق	ك
7	1	ت	ز
8	2	ذ	ز
9	1	ر	ل
10	2	ء	nga
11	17	ع	nga

From this case, by implementing a signal processing system, the author tries to identify the letters hijaiyah. Haby Bagus Prasetyo, Adiwijaya, and Untari Novia Wisesty [4] revealed that, in order to do voice recognition, the feature extraction and classifier methods are needed. The sound signal that has been extracted character then produces information that can be analyzed for each variation of the existing sound signal.

Therefore, to avoid mistakes that might occur when reciting Arabic reading, reading the Qur'an, praying and communicating, the author tries to identify hijaiyah letter pronunciation using MFCC (Mel frequency Cepstral Coefficient and Artificial Neural Network for classification).

The objectives of this study include:

1. Identifying 8-letter hijaiyah pronunciation using mel-frequency cepstral coefficient as a feature extraction by applying pre-emphasis, frame blocking and windowing stages, fast fourier transform, filter bank, discrete cosine transform, and cepstrum filters and artificial neural networks backpropagation as a classifier.
2. Knowing the effect of learning rate, the number of neurons in the hidden layer, maximum epoch, and training function on backpropagation neural networks against extraction values features mel-frequency cepstral coefficient which uses pre-emphasis, frame blocking and windowing stages, fast fourier transform, filter bank, discrete cosine transform, and cepstrum filter.

2 Study of Literature

The author conducts a literature study that has relevance to the topics discussed, namely the discussion of voice identification and the method of backpropagation and mel-frequency neural network coefficients. The literature study used can be in the form of journals, theses or other publications. The following are some similar studies as ingredients.

Tabel 2. Similar literature studies

Title, Author, Year	Method	Advantages	Deficiency
Algoritma Pengenalan Ucapan Huruf Hijaiyah Bertanda Baca	Linear Predictive Coding (LPC) as feature extraction, and	After several test scenarios obtained the best accuracy for training is	The best accuracy for testing is 58.93%

dengan Linear Predictive Coding (LPC) dan Hidden Markov Model (HMM) Haby Bagus Prasety, Adiwijaya, and Untari Novia Wisesty, 2016	Hidden Markov Model (HMM) as a classification.	99.60% with 28 data classes.	
Speech Quality based on Arabic Pronunciation using Mel-Frequency Cepstral Coefficient and LDA N.S.Zahra Zainon, Z.A. Ahmad, M.A. Romli, and S. Yaacob, 2012	Mel-Frequency Cepstral Coefficient (MFCC) as feature extraction and Linear Discriminat Analysis (LDA) as a classification.	Accuracy for ص (Saad), ط (Taa), ظ (Zaa) and ق (Qaaf) is above 80% with different test data and values. The highest level of accuracy achieved is 92.5% for ط (Taa ") when training is 80% and using 35 coefficients.	ض (Daad), even though the training value and coefficient have been manipulated , the accuracy is still around 60%. And the best value is only 71.5% when testing.
Identifikasi dan Aplikasi Pengenalan Spektrum Bunyi Gamelan Menggunakan Jaringan Syaraf Tiruan Pada Matlab Eko Ariyanto and Farid Samsu H, 2014	Artificial Neural Networks as a classification	In testing using training data, the success rate reached 99% in the number of neurons 110 and obtained an MSE value of 0.0001233 on the epoch to 1000	The number of neurons used in hidden layers is not limited to the input provided.
Penggunaan Algoritma Learning Vector Quantization dalam Mengenali Suara	Mel-Frequency Cepstral Coefficient (MFCC) as feature extraction and Learning Vector Quantization	During the testing process with the presence of fan noise with 20 attempts, it produces 15	If there are many noise disorders, from 20 trials, only 2 are recognized

Manusia untuk Kendali Quadrotor. Veronica Indrawati and Yudianto Gunawan, 2014	(LVQ) as a learning algorithm	data that are subjected to good	

Based on the research in the table above, there are several things that are different from previous studies, including:

- Use of other methods in feature extraction process. That is using mel-frequency coefficients and cepstral processes using artificial neural networks.
- Artificial neural network which is backpropagation as an algorithm. Writing of hijaiyah letters that have been pronounced.
- Random changes in the learning rate value, maximum epoch and training functions such as occurring converging faster.
- Using the mel-frequency cepstral coefficients method consists of only a few parameters, namely pre-emphasis, frame-blocking windowing, fast fourier transform, mel frequency wrapping, discrete cosine transform, and cepstrum lifter.
- The preprocessing stage on the cpstral mel-frequency coefficient which in this study only uses a pre-emphasis process without using noise canceling and Voice Activacion Detection (VAD).
- In the extraction feature of the cepstral mel-frequency coefficient does not use the postprocessing stage.
- The sound used as the object of research is the sound or pronunciation of a hijaiyah letter which is only called 8 hijaiyah letters, the س ص ذ ز ق ك ء ع which is a fathah community which is read for one lawless beat without using *mad* (long reading).

3 Basic steps for letters hijaiyah pronunciation

So the writer uses the mel-frequency cepstral coefficient method as a feature extraction method from the 8 letter hijaiyah pronunciation as for the purpose of this stage, which is to answer the question of the problems that have been defined previously. The steps are as follows:

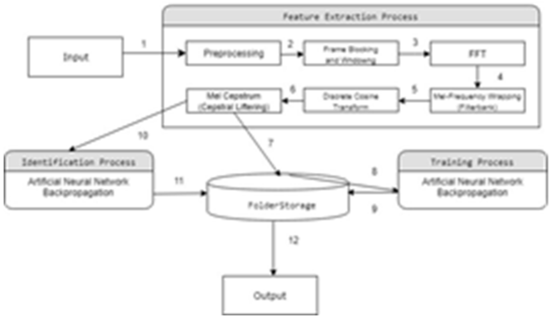

Figure 1. Conceptual flow of the hijaiyah letter pronunciation identification model

The picture above is a conceptual model flow in detail. Which starts from the feature extraction stage, the training process and the identification process.

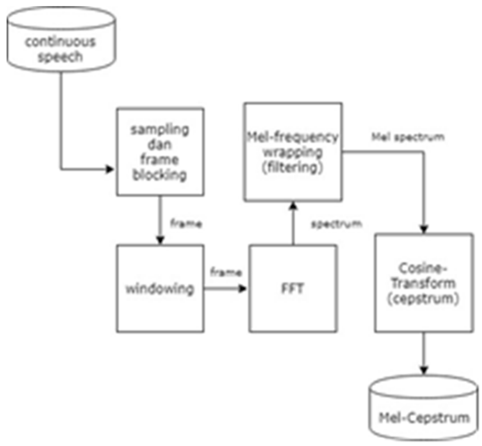

Figure 2. Block diagram for mel-frequency cepstral coefficient

Feature extraction aims to make the signal easily recognizable during the voice recognition process by the system. The main steps of feature extraction include preprocessing, frame blocking and windowing, and feature extraction [5]. The extraction feature here is broken down into several processes including fast fourier transform, mel-frequency wrapping, cosine transform and mel cepstrum.

The following data has been recorded using a smartphone and has been converted in *.wav format.

Tabel 3. Respondent data

No.	Sound type	Sound signal				
	fathah	Training Data			Test Data	
1.	س	sa(1) Wave Sound 162 KB	sa(2) Wave Sound 112 KB	sa(3) Wave Sound 117 KB	sa(4) Wave Sound 112 KB	sa(5) Wave Sound 112 KB
2.	ص	sho(1) Wave Sound 184 KB	sho(2) Wave Sound 90.0 KB	sho(3) Wave Sound 81.0 KB	sho(4) Wave Sound 85.5 KB	sho(5) Wave Sound 94.5 KB
3.	د	dza(1) Wave Sound 135 KB	dza(2) Wave Sound 99.0 KB	dza(3) Wave Sound 99.0 KB	dza(4) Wave Sound 121 KB	dza(5) Wave Sound 90.0 KB
4.	ز	za(1) Wave Sound 184 KB	za(2) Wave Sound 126 KB	za(3) Wave Sound 81.0 KB	za(4) Wave Sound 99.0 KB	za(5) Wave Sound 140 KB
5.	ق	qa(1) Wave Sound 144 KB	qa(2) Wave Sound 85.5 KB	qa(3) Wave Sound 108 KB	qa(4) Wave Sound 121 KB	qa(5) Wave Sound 63.0 KB
6.	ك	ka(1) Wave Sound 153 KB	ka(2) Wave Sound 81.0 KB	ka(3) Wave Sound 117 KB	ka(5) Wave Sound 126 KB	ka(4) Wave Sound 99.0 KB
7.	ء	a(1) Wave Sound 121 KB	a(2) Wave Sound 81.0 KB	a(3) Wave Sound 81.0 KB	a(4) Wave Sound 94.5 KB	a(5) Wave Sound 63.0 KB
8.	ع	'a(1) Wave Sound 144 KB	'a(2) Wave Sound 90.0 KB	'a(3) Wave Sound 94.5 KB	'a(4) Wave Sound 103 KB	'a(5) Wave Sound 81.0 KB

3.1 Pre-processing

In pre-emphasis, which gives emphasis to the sound signal by applying a high pass filter to increase the frequency. In fact, when it spreads via air, the amount of the speech signal decreases when the frequency rises. To compensate for the attenuated speech signal, it is passed through a high-pass filter (limited impulse filter) to recover the signal using a limited im puls filter (1, -0.97) [6] then:

$$Sp(n) = s(n) - 0.97\, s(n-1) \tag{2.1}$$

Information:
Sp(n) = signal of the nth pre-emphasis filter result
s(n) = signal before the pre-emphasis filter
n = signal length

3.2 Frame Blocking and Windowing

The voice signal is divided into several frames and overlapping each other. The overlapping area tested is 25%, 50%, and 75% so that a certain number of frames are obtained. To calculate the number of frames used is the following formula:

$$((I - N) / M) + 1 \tag{2.2}$$

With,
I = Sample rate
N = Frame size (Sample rate * time framing (s))
M = N/2

The next thing to do is windowing each frame in order to reduce signal discontinuity at both ends of the block. Windowing commonly used is the Hamming Window which is calculated as follows [5]:

$$w(k)=0.54-0.46\cos((2\pi k/(K-1)) \tag{2.3}$$

Information:

w(k) = window function

k = frame length

3.3 Fast Fourier Transform (FFT)

An analysis based on Fourier transform is synonymous with a spectrum analyzer, as Fourier transform to change the digital signal from the time domain to the frequency domain. Fast fourier transform is discrete fourier transform using fast calculation techniques that utilize the periodical properties of fourier transforms. As the following formula:

$$F(k) = \sum_{n-1}^{N} f(n) \cos\left(\frac{2\pi kT}{N}\right) - j \sum_{n-1}^{N} f(n) \sin\left(\frac{2\pi kT}{N}\right) \quad (2.4)$$

Because x(n)=x_r (n)+jx_i (n) can be complex, then:

$$X(k) = X_R(k) + jX_i(k)$$

$$X_R(k) = \sum_{n=0}^{N-1} \left[x_r(n) \cos 2\pi \frac{k}{N} n + x_i(n) \sin 2\pi \frac{k}{N} n\right] \qquad X_I(k) = \sum_{n=0}^{N-1} \left[x_r(n) \cos 2\pi \frac{k}{N} n - x_i(n) \sin 2\pi \frac{k}{N} n\right] \quad (2.5)$$

Information:

N = Number of input sample
F_k = the order of k fast fourier transform component output (x (0), x (1), ….,x (n-1))
K = output index of fast fourier transform in the frequency domain (0,1,….., N-1)
n = index sample of input sample in the time domain (0,1,….., N/2-1)
j = constanta of imaginary numbers (√(-1))
π = (180o) degree

3.4 Mel Frequency Wrapping (Filterbank)

This section is one of the most important parts, which is to get relevant information from the greeting block. Many methods are used at this stage [5]. This stage is also called the triangular filter with the following formula:

$$H_i = \frac{2595 \log(1+\frac{f}{700})}{\frac{S_i}{2}} \quad (2.6)$$

H_i = Filterbank
F = linear frequency
$\mathbf{S_i}$ = Signal from fast fourier transform

3.5 Discrete cosine transform (DCT)

At this stage the spectrum will be converted into the time domain. D iscrete cosine transform is the same as fast fourier transform or inverse of fast fourier transform [5].

$$\tau_n = \sum_{k=1}^{K} (\log S_k) \cos\left[\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right] \quad (2.7)$$

S_k = Result of fiterbank on index k
K = Number of coefficients with moder n

3.6 Cepstral Liftering

To improve the quality of recognition, the result of discrete cosine transform cepstrum must pass through the cepstral liftering processfirst [6] as cepstral liftering formula:

$$w[n] = \{N \frac{L}{2} \sin \frac{n\pi}{L-1}\} \quad (2.9)$$

L = Number of cepstral coefficients
N = Index of cepstral coefficients

3.7 Artificial Neural Networks

Sutojo et al. [7] explained that "artificial neural networks have an extraordinary ability to obtain information from complex or incorrect data, are able to solve unstructured and difficult to define problems, can create a pattern of knowledge through self-regulation or learning ability (self-organizing), able to choose an input data into certain categories that have been defined (classification), able to describe an object as a whole even though only given some data from the data object (association), has the ability to process input data without having to have target, and able to find the best answer so as to minimize the cost function".

3.8 Backpropagation

Backpropagation is a type of nonlinear gradient reduction procedure. This can be used for multi-category classification. The aim is to minimize error criteria [8] Gradient drop method to minimize output error. There are three stages that must be carried out in network training, namely the stage of forward propagation, reverse-propagation stage, and weight and bias stages. This network architecture consists of output layer [7].

In this study, the problem that must be solved is how to identify the pronunciation of hijaiyah letters using an envy extraction algorithm, namely mel-frequency cepstral coefficients and backpropagation artificial neural networks as a classification method. By simulating the system that has been made using the parameters of the effect of the classification, learning rate, maximal epoch to the accuracy of the identification of the pronunciation.

4 Result and Discussion

After going through the stages of feature extraction using mel-frequency cepstral coefficient, out the results of the tests using 10 neurons in the hidden layer 1 and 5 neurons in the hidden layer 2 using logsig activation function and learning function trainlm, maximum epoch of5000 and 0.01 as the learning rate shown in fig. 3.

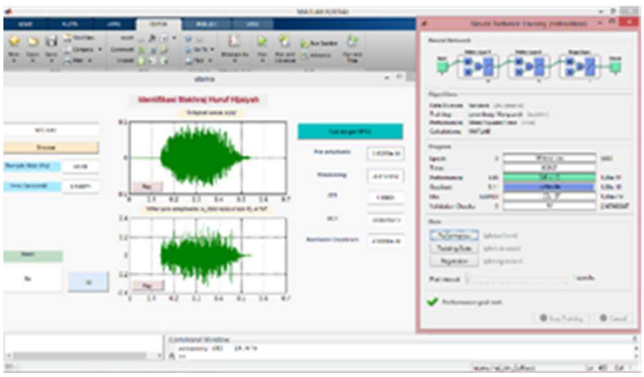

Figure 3. Block diagram for mel-frequency cepstral coefficient test result interface

Figure 2 shows the results of testing the file named "a (1) .wav" which produces the output of the letter "Za". Accuracy obtained is 16.67%, using 14 iterations (epoch), and MSE is 4.4465. Because the results are still not in accordance with the target that should be issued that is worth 7, then it takes repeated experiments so as to produce the same output as the target.

Tabel 4. Result of 19 tests files"a (1) .wav"

Experiment	Target value	Results	Epoch	MSE	Accuracy
1	7	Za	14	4,447	16,67%
2	7	Dza	32	12,95	20,83%
3	7	Za	49	4,67	12,5%
4	7	^A	21	2,797	29,17%
5	7	Dza	15	5,239	16,67%
6	7	Za	35	9,393	12,5%
7	7	Qo	90	4,694	12,5%
8	7	Dza	46	4,005	16,67%
9	7	A	25	0,648	4,17%
10	7	Qo	26	1,786	12,5%
11	7	Sa	16	4,45	12,5%
12	7	A	40	2,494	25%
13	7	Za	19	5,14	12,5%
14	7	Za	153	3,114	58,35%
15	7	Sho	16	7,57	16,67%
16	7	A	29	7.55	54,17%
17	7	Sa	30	15.708	0,0%
18	7	A	51	2,346	16,67%
19	7	A	123	1,05	41,67%

Value The learning rate used in these 12 experimental sections is 0.01, 10 neurons in the hidden layer, using the Logig or binary sigmoid activation function and changing the maximum Epoch value and the type of training function that produces MSE values, regressions from training, validation, and tests. Using maximum epoch values of 3000, 4000, and 5,000 in each experiment using various training functions.

Tabel 5. Results of training using artificial neural networks by different training functions

no	Training Function	lr	Maksimum Epoch	Epoch	Training	Validation	Test	MSE	Inappropriate Output
					R:	R:	R:		
1	Traingdm	0,01	3000	59	0.4109	0.64435	0.9093	0.75923	20
2	Traingdm	0,01	4000	0	0.51275	0.65914	0.87401	2.1594	18
3	Traingdm	0,01	5000	0	0.73542	0.93423	-0.032983	11.3185	14
4	Traingda	0,01	3000	0	0.72112	0.72456	0.85807	1.4598	16
5	Traingda	0,01	4000	1	0.89995	-0.30419	0.99015	0.19848	12
6	Traingda	0,01	5000	0	0.74241	0.99378	0.96088	0.43305	11
7	Traingdx	0,01	3000	0	0.71759	0.8141	0.82872	3.9944	17
8	Traingdx	0,01	4000	3	0.66553	0.98672	0.86361	0.3592	16
9	Traingdx	0,01	5000	0	0.81703	0.97778	0.8235	2.4646	15
10	Trainlm	0,01	3000	0	0.86239	0.88662	0.99772	0.16971	9
11	Trainlm	0,01	4000	0	0.91019	0.99727	0.93486	0.2048	7
12	Trainlm	0,01	5000	0	0.93569	0.98259	0.91181	0.25258	10

The greatest accuracy obtained is 70.83% while the letter sound of the letter "sa" produced by a sound signal called "sa (2) .wav" is most easily recognized even though using a variety of training functions. Of the 12 scenarios, only 2 scenarios cannot recognize.

In the hijaiyah letter pronunciation test using trainlm with the number of hidden layer 10, an accuracy of 25% was obtained which resulted in the same 4 outputs with the target of the 16 data tested. So, to get the MSE value that reaches the minimum level, a random experiment is needed by changing some artificial neural network parameters.

5 Conclusion

This study uses 5 س ص huruf ز ق ك ء ع letter pronunciation sound samples and the fathah community is then identified using Mel-Frequency Cepstral as feature extraction and backpropagation Artificial Neural Network as classifier. Training uses trainlm using the number of hidden layer 10, learning rate 0.01, maximum epoch of 4000 which is done 15 times the experiment produces regression training of 0.91019, test of 0.93486, validation of 0.99772 and MSE value of 0.2048. Produces 7 letters that do not match the target so that the accuracy of the truth obtained is 70.83%. Testing of hijaiyah letters using trainlm with the number of hidden layer 10, obtained an accuracy of 25% which resulted in the same 4 outputs with the target of the 16 data tested.

Therefore, to get the MSE value that reaches the minimum level, a random experiment is needed by changing the number of neurons, maximum epoch, learning rate or training function and increasing training data so that the network can recognize the training process.

Acknowledgements. This paper in conjuction with The 1st International Conference on Islam, Science and Technology (ICONIST 2018) in Malang, East Java.

References

[1] Nasution, A. S. A. Bunyi Bahasa 'Ilm Al-Ashwat AlArabiyyah. Jakarta: Amzah (2010).

[2] Wahyudi, Moh. Ilmu Tajwid Plus. Surabaya: Halim Jaya (2008).

[3] Ifka, I. Kesulitan pelafalan huruf hijaiyyah yang tidak terdapat di huruf indonesia pada masyarakat saradan wonogiri. Semarang: Jurusan Bahasa dan Sastra Asing, Universitas Negeri Semarang (2013).

[4] Bagus Prasetyo, H., & Novia Wisesty, U. Algoritma Pengenalan Ucapan Huruf Hijaiyah Bertanda Baca Dengan Linear Predictive Coding (Lpc) Dan Hidden Markov Model (Hmm). Bandung: Telkom University. Indosc 2016, (Agustus), 149–164 (2016).

[5] Nilsson, M. Speech Recognition using Hidden Markov Model performance evaluation in noisy environment. Sweden: Departmen of Telecommunication and signal processing, Blekinge Institute (2002).

[6] Hassine, M. Hybrid Techniques for Arabic Letter Recognition. International Journal of Intelligent Information Systems, 4(1), 27 (2015).

[7] Sutojo, T, Edy M & Vincent S. Kecerdasan Buatan. Yogyakarta: Andi Publishing (2011).

[8] Cohen, Donna L.H. Neural Networks and Artificial Intelligence for Biomedical Engineering. New York: IEEE PRESS The Institute of Electrical and Electronics Engineers (2000).