Journal of Computers Vol. 28, No. 5, 2017, pp. 1-10
doi:10.3966/199115992017102805001

Improvement in Speech to Text for Bahasa Indonesia Through Homophone Impairment Training

Intan Sari Areni[1#], Indrabayu[2#] and Anugrayani Bustamin[3#*]

[1] Department of Electrical Engineering, Electrical Study Program, Universitas Hasanuddin

[2] Department of Electrical Engineering, Informatics Study Program, Universitas Hasanuddin

[3] Department of Electrical Engineering, Universitas Hasanuddin

[#] Research Group of Artificial Intelligence and Multimedia Processing Universitas Hasanuddin
Makassar, Indonesia
{intan, indrabayu, anugrayani}@unhas.ac.id

Received 17 June 2016; Revised 9 December 2016; Accepted 9 December 2016

Abstract. In this research, an approach for increasing accuracy in speech to text application is done using Mel Frequency Cepstral Coefficient (MFCC) trained by Backpropagation Neural Network (BPNN). A set of Bahasa Indonesia homophones data speech is used for training and validation. The record is taken from 6 native adults comprising 3 males and 3 females. Working in 16 KHz sampling mode, the data is stored in WAV format. A confusion matrix is used to validate the system with and without homophone locking learning. A significant improvement is observed from the experiment. The percentage of accuracy is increased from 53.33 to 93.4 from male samples. From females' records, the increment is even higher. The accuracy percentage has risen from 36.8 to 93.33.

Keywords: BPNN, confusion matrix, homophone, MFCC, speech to text

1 Introduction

Voice has an entity which gives a unique biometric to each individual in the world. Not only distinctive feature from each person, a wrong perception of voice (information) might occur from several factors. Dialect, articulation, noise, and even semantic dispute can contribute to misperception [1]. Lexical semantics (lexicosemantics) is a subfield of linguistic semantics that studies the relation of meaning from words. The relation is categorized into the context of synonym, antonym, homonym, homophone, homograph, polysemy, hypernyms, collocation, denotation and connotation [2].

Lexical semantics is a fascinating subject in speech to text recognition. Due to its differences in spelling, writing and meaning, it can lead to ambiguity in the resulted text. Additionally, each language has its own uniqueness. To avoid ambiguity, homophones as part of lexical semantics will be combined with other words called phrase in this paper. To recognize a word on lexical semantics, feature extraction is needed to mark its characteristic. Classification is conducted for processing the characteristic features which have been obtained. Some methods of classification require a process of learning. One or more test patterns associated with the speech of the same class are used to create a pattern representative of the class characteristics [3].

Related research about lexical semantics of the homonym in Japanese has been proposed by Murakami. The homonym is two or more words that have the same shape both in writing and pronunciation but have different meanings. Data evaluation consists of 11 pairs of homonyms recorded from 9 speakers at a sampling frequency of 16 kHz with a window length of 25ms. The method used is the comparative Mel Frequent Cepstral Coefficient (MFCC) and FBANK for feature extraction and Hidden Markov Model

[*] Corresponding Author

(HMM) for classification. In the case of acoustic parameters, MFCC produces higher average recognition rates than FBANK. However, MFCC with 12 orders has better accuracy compared to FBANK with 24 orders for male speakers, and this also applies for female speakers. The results show that the achieved accuracy level can be up to 89% [4]. Other research that uses speech recognition of MFCC has been done by Mishra et al, but is still limited to the recognition of vowels in Hindi. MFCC Feature Extraction method is compared with the proposed hybrid method MFCC-QCN (quantile-based Dynamic cepstral Normalization). The process of classification is using Hidden Markov Models (HMM) architecture with 3 emitting states and four Gaussian Mixture Components. The results denote QCN-MFCC managed to improve system performance by 13% and 11% for context-dependent and context-independent classification of midvowels [5]. Classification of speech recognition using Neural Network has been accomplished by Vijayendra & Thakar by utilizing the results of MFCC features and RC (Real cepstral Coefficient). The data used is the ratio of speech data source from conventional microphone and in-ear microphone. The classification process finished by a configuration of 2 and 3 layer Neural Network. The results show that the extraction of MFCC feature is 8% higher compared to the RC based on the NN classification with an average accuracy rate of 95% [6].

As for the introduction of the homophones, Nemoto et al. have studied the Automatic Speech Recognition (ASR) in French. The study aims to improve the accuracy of ASR in the terms of the Acoustic Model and Language Model. There are two problems in this research. Firstly, the recognition of homophones in ASR system relies on language modeling *n*-gram weights. Secondly, the acoustic separability of the two homophones uses appropriate acoustic and prosodic attributes. The automatic classification of the two words using data mining techniques highlights the role duration in voicing and contextual information to distinguish the target words. The classification results are obtained at 78% with an algorithm Logistic Model Trees (LMT) [7]. Chen et al conducted the speech recognition system to text on the Keyword Search (KWS) for Tamil and Vietnam languages. This study highlights the three part-interests in optimizing voice processing submodular acoustic diverse selection of data through gaussian components, as well as the modeling language for modeling sub morpheme keywords and homophones. Speech recognition for homophones in KWS uses two conditions, i.e. FLP (Full Language Pack) and LLP (Limited Language Pack) with two homophones segmentation method SH regardless morpheme and sub homophones (SH_{sub}) that is compared with the following morpheme. The results suggest that the increase in system performance occurs at 49.4% when in segmentation using sub homophones (SH_{sub}) in LLP condition, whereas with the segmentation of homophones (SH), the performance is obtained only at 4.5% in FLP condition [8].

Lee discusses the Language Model on a speech recognition post-processing using Neural Network (NN) with Adaptive Learning approach to handle disambiguation in homophones in Chinese [9]. It controls the distance of preferred and unpreferred pairs and gives improved performance. Contextual language processing has an important role in post-processing speech to text recognition in order to find the candidate of the most commonly used in syllables based on the maximum probability. The performance of the probabilistic model is affected by two major errors, i.e. modeling and estimation errors in training corpus. The adaptive learning algorithm has good performance in this experiment. However, the speed of convergence is the major problem. The results show that the improved accuracy of the sentence reaches 58.13% (from 28.46% to 86.59%) and the accuracy of the character increases by 18.64% (from 79.20% to 97.84%). Neural Network algorithm is used in handling the case of prediction and pattern recognition such as image processing, video processing and speech processing [10]. Speech recognition system for Indonesian has been conducted by Hoesen et al based on sound data which is spontaneous and dictated by a combination of Gaussian Mixture and Hidden Markov Model (HMM). Data recording is obtained from 244 Indonesian speakers during 73 hours for dictated speech and 43.5 hours for spontaneous speech. The success rate of the system is measured by Word Accuracy Rate (WAR). In order to improve the recognition accuracy, the adaptation process is applied to the acoustic model. In this study, acoustic modeling adaptation techniques consist of Maximum A-posteriori Probability (MAP), Maximum Mutual Information (MMI) and Feature-space Maximum Likelihood Linear Regression (fMLLR). The results indicate that the adaptation of MAP improves the accuracy by 2.60% and 1.36% for the spontaneous and dictated speech, respectively. Adaptation of WAR MMI only increases by 1.48%. On the other hand, fMLLR actually reduces system performance for both spontaneuous and dictated speeches [11].

Previous research has described that the success of speech recognition systems is limited to the words and vowels and also the presence of homophones words in several languages. However in this study, the authors explain it in different approaches from other research. The use of homophones in everyday life, especially in Indonesian can provide insight and ambiguity in communication. Therefore, this research will involve the use of the homophone phrases in speech recognition as a solution.

Based on the previous research, the authors propose a speech to text recognition system based on Indonesian homophone phrases by using MFCC and BPNN for feature extraction and classification, respectively. The rest of the paper is organized as follows. Section 2 describes the research method which consists of a dataset, preprocessing and a combination of feature extraction and classification methods proposed. Section 3 presents the results of a research experiment. The conclusions of this paper are pointed out in Section 4.

2 Research Method

2.1 Homophones

There are four terms in lexical semantics that are often used to show the relationship between words in a language, namely homophones, homonyms, homograph, and polysemy. Homophones, homonyms, and homograph can also be classified as homonymy, ie two or more words that have different meanings but have the same shape. If two or more words are spelled differently in writing and have the same phrase, the words are called homophones, for example *bang-bank, massa-masa,* and *sanksi-sangsi* [12].

2.2 Dataset

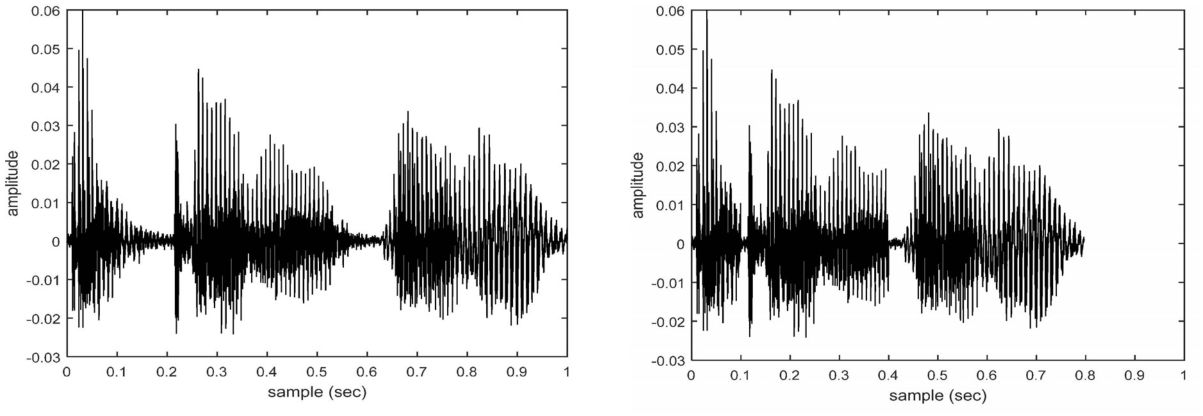
In this research, the data used is in the form of speech data record of Indonesian homophones that consists of two scenarios of data retrieval. Table 1 shows the result of homophones data test for single homophone and homophone phrase. Recording is done in mono channel for 5 seconds by 6 respondents, representing each of the proportional three male and female voices in which the sound recording is performed 5 times for each word. Hence, the amount of speech data is 300, which is training data of 240 and test data of 60. Recorded files are stored in .wav format at a sampling frequency of 16 kHz.

Table 1. Evaluation data (Pairs of homphones and homphones phrase)

Single homophone		Homophone phrase	
Balik	*balig*	*akil balig*	*balik belakang*
Bang	*bank*	*Bang Saleh*	*Bank Mandiri*
dakwa	*dakwah*	*dakwa hakim*	*dakwah ustadz*
dara	*darah*	*dara cantik*	*darah merah*
rock	*rok*	*lagu rock*	*rok mini*

2.3 Preprocessing

Preprocessing stage is a data preparation before entering into the process of feature extraction. In the recording process, respite is sometimes found in order to obtain a state of silence. Therefore, the signal separation must be done to distinguish between silence and deemed valid signal. In this research, the identification of silence is set at 0.03 on a scale of amplitude which is the default threshold as described in [13-14]. Thus, the system will identify the silence by looking at frame with a maximal amplitude of less than 0.03. Removal of silence greatly influences the size of cepstrum coefficient matrix generated from the MFCC. Fig. 1 shows an example of the speech signal before and after removing silence.



(a) before silence removal (b) after silence removal

Fig. 1. Speech signal "akil balig"

2.4 MFCC for Feature Extraction

MFCC maps frequency components using Mel scale and is modeled based on voice perception of the human ear [15]. In this study, the MFCC coefficients method uses a variation of the number and type of window. MFCC stages of the process are shown in Fig. 2.

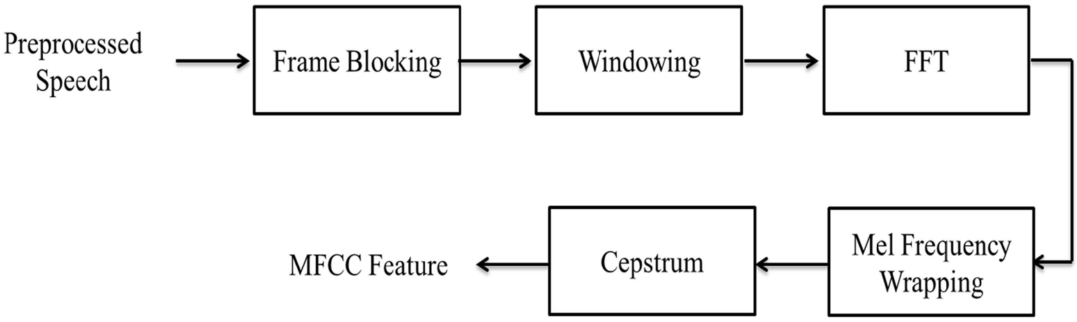


Fig. 2. MFCC Block Diagram

Signals of recorded sound through preprocessing will be divided into several frames to obtain a stable characteristic of the speech signal. In Fig. 3, the speech signals in segmenting into multiple frames overlap so that no signal is deleted when framing is conducted. The length of overlap area that is commonly used is 30% to 50% of the size of the frame. The segmented speech signal has the frame duration of 20 ms [16]. If a sample rate is16 kHz, the sample size in one frame is 320 points (= 16000 Hz x 0.02 sec) with 160 sample points overlapping.

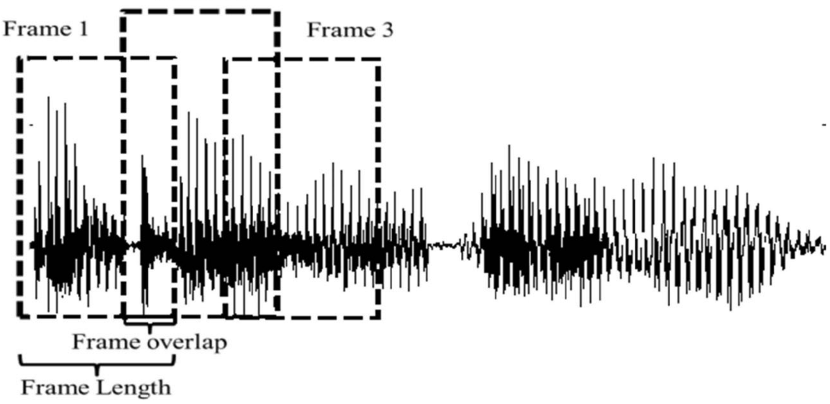


Fig. 3. Frame blocking process for "akil balig"

Windowing will be conducted after frame blocking to minimize aliasing which causes discontinuity in processing the speech signals [14]. Hamming window is used in this paper and described as follows [17].

$$w(n) = 0.54 - 0.46\,x \cos\left(\frac{2n\pi}{N-1}\right), 0 \le n \le N-1 \quad \textbf{(1)}$$

where N is the amount of the speech signal.

The speech signal after windowing is an input of the Fast Fourier Transform (FFT) which converts the signal frame from the time domain to the frequency domain. The next process is Mel Frequency Wrapping for filtering the spectrum of each frame. Each tone with the actual frequency f is measured in Hz and subjective pitch is measured on a scale called "mel" at MFCC. Mel frequency scale is linear frequency under 1000 Hz and logarithmic for frequencies above 1000 Hz. At this stage, the multiplication process will be carried out between Mel-Spaced Filterbank and spectral power of the periodogram, wherein the multiplication results are summed. The final stage of feature extraction is the mel log spectrum transform to obtain the MFCC features.

2.5 Classification with Neural Network

Computational methods of NN are inspired by the workings of the human brain cells. In order to think, human brains get stimulation of neurons found in human senses, then the results of these stimuli are processed so as to produce the information. On the computer, the stimulus given to the input will be multiplied by a value and then treated with a particular function to produce an output. Characteristics of NN are marked on network architecture, the learning method for weighting the connection, and activation of function selection [18].

Backpropagation is a supervised learning training method in NN. This method can handle the case in a complex pattern recognition. Backpropagation Neural Network (BPNN) works by starting with initialization that is weighted and biased. In the BPNN, each unit that is in the input layer is associated with another that is in the hidden layer. MFCC feature extraction results in a speech signal of Homophones as data input on BPNN. Recognized words of 10 homophones (see Fig. 4) are most commonly used in everyday life, such as "balig", "balik", "bang", "bank", "darah", "dara", "dakwah", "dakwa", "rok" dan "rock". Units in the hidden layer are connected to each unit in the output layer. The weight values are possessed by each unit. W_{ij} is the weight of the input layer to the hidden layer and W_{jk} is the weight of the hidden layer to the output layer. This network consists of a multilayer network. When the network is given input pattern as a training pattern, the pattern of the lead units is a hidden layer to be subsequently forwarded to the layer of output units. Then, the output layer units will provide a response as the output of the neural network. When the output is not as expected, the output will be propagated backward in the hidden layer and then from the hidden layer to the input layer as shown in Fig. 4.

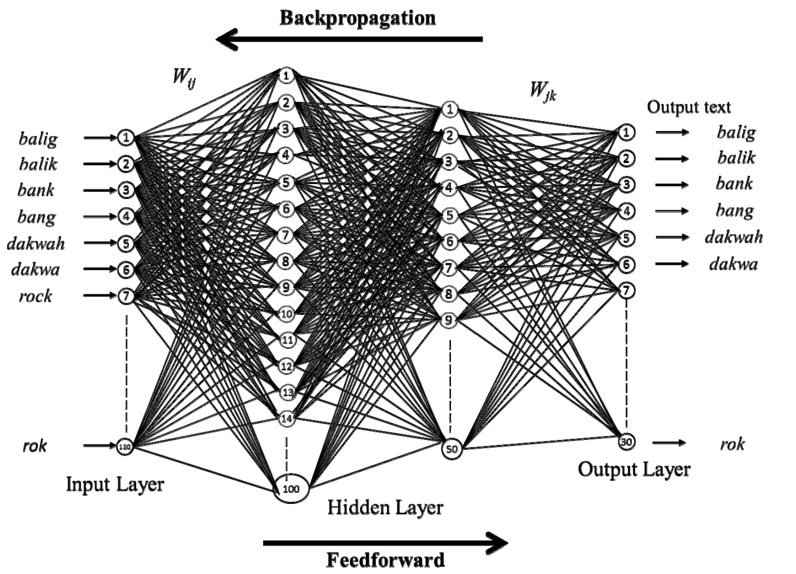


Fig. 4. BPNN architecture

Table 2 shows the structure of the NN with backpropagation training functions that will be applied in this study. Data input used is the end result of the form of a matrix cepstrum of MFCC coefficient with the output neuron representing the target of the input data as a reference to the BPNN training output.

Table 2. BPNN structure

Parameter	Spesification
Architecture	2 hidden layer
Number of input neuron	cepstrum coefficient matrix MFCC
Number of hidden neuron	[100 50]
Number of output neuron	10 (target definition)
Learning function	Backpropagation
Activity function	Log-sigmoid
Maximum epoch	100
Error tolerance	10^{-5}

In Backpropagation, there are many parameters that can be set, but most of these parameters can be used with the value set by default. This is due to variations in the value of these parameters that influence the time required for training [19].

Networks that have trained and achieved the desired results need to be tested to determine the ability when studying the training data given. Testing is conducted to observe the performance of the system that has been created by looking at the value of the minimum error. Training and testing results can be analyzed by observing the accuracy of target network output. After the system is trained, the next step is the validation of the system. In the validation process, the system is tested with other data; it is intended to determine the extent to which the system can alert text output with text input. Output will be compared with the target test data. The training and testing process of speech to text recognition is shown in Fig. 5.

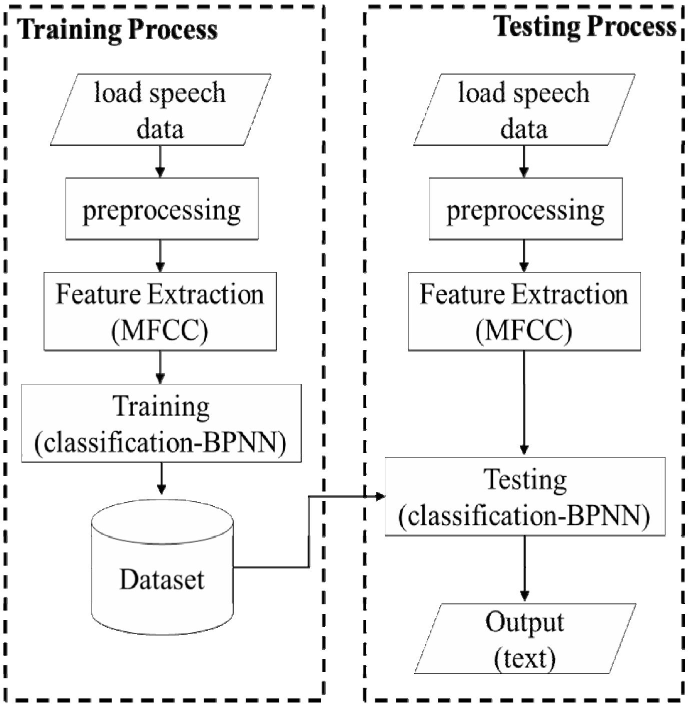

Training Process
load speech data
preprocessing
Feature Extraction (MFCC)
Training (classification-BPNN)
Dataset
Testing Process
load speech data
preprocessing
Feature Extraction (MFCC)
Testing (classification-BPNN)
Output (text)

Fig. 5. Flowchart stages of identification

2.6 Validation

After finishing the system design, the next process is to evaluate the performance of the system by calculating the degree of accuracy. In this study, the validation method used is confusion matrix which gives a decision based on the results of the training and testing. Confusion matrix provides performance ratings based on the classification of objects correctly or incorrectly. If the dataset consists of only two classes, one class is regarded as positive and the other is negative as shown in Table 3.

Table 3. Confusion matrix model

Correct Classification		Classified as	
		+	-
actual	+	True positives (TP)	False negatives (FN)
	-	False positives (FP)	True negatives (TN)

The result on the diagonal from top left to bottom right is the result of correct classification, and all values outside the diagonal are incorrect, classified as incorrects. This level of accuracy on a model confusion matrix is calculated based on the Equation (2) below:

$$accuiracy\ rate(\%) = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

Where:
True positives (TP): the proportion of positive cases that are correctly classified.
False positives (FP): the proportion of negatives cases that are incorrectly classified as positive.
False negatives (FN): the proportion of positives cases that are incorrectly classified as negative
True negatives (TN): the proportion of negative cases that are classified correctly.

3 Result and Analysis

Accuracy comparisons of speech to text recognition with three spelling of 10 words homophones is tested by confusion matrix as show in Table 4. An example for the spelling word of “balig” has 67% result of accuracy. Only 2 spelling is being recognized by dataset, 1 other word is recognized as “balik”.

Table 4. Classification based on confusion matrix in single word homophones for male data

	Male Data	**Result (%)**									
		balig	*balik*	*bang*	*bank*	*dakwa*	*dakwah*	*dara*	*darah*	*rock*	*rok*
actual class	*balig*	**67**	33	0	0	0	0	0	0	0	0
	balik	33	**33**	0	0	0	0	0	0	0	0
	bang	0	0	**33**	33	0	33	0	0	0	0
	bank	0	0	33	**33**	0	0	0	0	0	0
	dakwa	0	0	0	33	**33**	0	0	0	0	0
	dakwah	0	0	0	0	67	**67**	0	0	0	0
	dara	0	0	0	0	0	0	**67**	0	0	0
	darah	0	0	0	0	0	0	33	**67**	0	0
	rock	0	0	0	0	0	0	0	0	**33**	0
	rok	0	0	0	0	0	0	0	0	67	**100**

The results of Table 4 will be calculated by the equation confusion matrix based Equation (2) below:

$$accuracy\ (male\ data) = \frac{(67+33+33+33+33+67+67+67+67+33+100)}{(100+67+67+67+67+67+67+33+33+33+33+33+33+33+33+33+33+33+33)} \times 100$$

$$= \frac{533}{898} \times 100$$

$$= 59{,}354\%$$

The accuracy rate of single homophone word for female data in Table 5 showed in calculation below:

$$accuracy\ (femals\ data) = \frac{(33+67+67+0+67+0+0+67+67+0)}{(67+67+67+67+67+67+67+67+67+33+33+33+33+33+33)} \times 100$$

$$= \frac{36.8}{801} \times 100$$

$$= 45{,}9426\%$$

Table 4 and Table 5 show that the levels of accuracy obtained for the speech of males and females are 59.35% and 45.94% respectively. These results indicate that the speech to text recognition system of homophones by itself will not be maximized because the noise generated between homophone words is very precise. To increase the accuracy of homophones speech to text recognition, a homophone is combined with other words to form a phrase which makes the meaning of the homophone is increasingly clear. Table 6 and Table 7 describe a comparison of the results of the introduction of homophones with phrases that have been done with the confusion matrix system. The calculation of total accuracy is shown in the formula below:

$$accuracy\ (mals\ data) = \frac{(100+67+100+100+100+100+67+100+100+100)}{(100+100+100+100+100+100+100+100+67+67+33+33)} \times 100$$

$$= \frac{934}{100} \times 100$$

$$= 93{,}4\%$$

Table 5. Classification based on confusion matrix in single word homophones for female data

Female Data		Result (%)									
		balig	*balik*	*bang*	*bank*	*dakwa*	*dakwah*	*dara*	*darah*	*rock*	*rok*
actual class	*balig*	**33**	0	0	33	0	0	0	0	0	0
	balik	67	**67**	0	33	0	0	0	0	0	0
	bang	0	33	**67**	0	0	0	0	0	0	0
	bank	0	0	0	**0**	33	0	67	0	0	0
	dakwa	0	0	0	0	**67**	67	0	0	0	0
	dakwah	0	0	0	0	0	**0**	0	0	0	0
	dara	0	0	0	0	0	0	**0**	0	0	0
	darah	0	0	0	0	0	0	0	**67**	0	0
	rock	0	0	0	0	0	33	0	0	**67**	67
	rok	0	0	0	0	0	0	0	0	0	**0**

Table 6. Classification based on confusion matrix in homophones phrase for male data

Male Data		Result (%)									
		akil balig	*balik belakang*	*Bang Saleh*	*Bank Mandiri*	*dakwa hakim*	*dakwah ustadz*	*dara cantik*	*darah merah*	*lagu rock*	*rok mini*
actual class	*akil balig*	**100**	0	0	0	0	0	0	0	0	0
	balik belakang	0	**67**	0	0	0	0	0	0	0	0
	bang Saleh	0	33	**100**	0	0	0	0	0	0	0
	Bank Mandiri	0	0	0	**100**	0	0	0	0	0	0
	dakwa hakim	0	0	0	0	**100**	0	0	0	0	0
	dakwah ustadz	0	0	0	0	0	**100**	33	0	0	0
	dara cantik	0	0	0	0	0	0	**67**	0	0	0
	darah merah	0	0	0	0	0	0	0	**100**	0	0
	lagu rock	0	0	0	0	0	0	0	0	**100**	0
	rok mini	0	0	0	0	0	0	0	0	0	**100**

Table 7. Classification based on confusion matrix in homophones phrase for female data

	Female Data	**Result (%)**									
		akil balig	*balik belakang*	*Bang Saleh*	*Bank Mandiri*	*dakwa hakim*	*dakwah ustadz*	*dara cantik*	*darah merah*	*lagu rock*	*rok mini*
actual class	*akil balig*	100	0	0	0	0	0	0	0	0	0
	balik belakang	0	**33**	0	0	0	0	0	0	0	0
	bang Saleh	0	67	100	0	0	0	0	0	0	0
	Bank Mandiri	0	0	0	100	0	0	0	0	0	0
	dakwa hakim	0	0	0	0	100	0	0	0	0	0
	dakwah ustadz	0	0	0	0	0	100	0	0	0	0
	dara cantik	0	0	0	0	0	0	100	0	0	0
	darah merah	0	0	0	0	0	0	0	100	0	0
	lagu rock	0	0	0	0	0	0	0	0	100	0
	rok mini	0	0	0	0	0	0	0	0	0	100

Based on the results of the classification in Table 6 and Table 7, from a total of 30 homophone phrases introduced to examine the system, 28 of them are classified correctly. Errors can occur due to recognition of the feature differences which are too large between the voice signal to be identified to the trained voice signals. This problem can be resolved by increasing the variation patterns of words during training so that the network system is more enhanced in knowledge.

The results of the classification of homophone phrases based on Table 6 can improve speech to text recognition accuracy of 34.05% for male respondents data and Table 7 shows the increase of accuracy by 47.39% in female respondent data.

Table 8 shows the average accuracy rate obtained by the speech to text recognition system for homophones in Bahasa Indonesia. The introduction of homophones in a phrase provides the highest accuracy rate of 93.4%.

Table 8. Average accuracy rate

Data	**Average Accuracy Rate (%)**	
	Male	Female
Homphones	59,35	45,94
Homphone Phrases	93.4	93.33

4 Conclusion

Speech recognition for lexical semantics of the homophones with MFCC-BPNN method has been conducted in this study. The speech data in .wav format consists of 120 training data and 30 test data for each male and female respondent. System testing is done by using a confusion matrix. Speech to text recognition with homophone phrases improves recognition accuracy by 40,07% (from 53.33% to 93.4%) for the data of male respondents and increases the accuracy by 56.53% (from 36.8% to 93.33%) in female respondents. This indicates that the speech to text recognition system for homophones is more effectively if it is in the form of a phrase. The time required for the multilayer on BPNN in completing the training process with multiple iterations is not very practical. In the future, this research can be developed by utilizing the modeling language which forms the corpus to enrich the homophones dataset. Also, feature extraction and classification methods still need to be involved to improve the performance of the system.

References

[1] D.P. Lestari, K. Iwano, S. Furui, A large vocabulary continuous speech recognition system for Indonesian language, in: Proc. 15th Indonesian Scientific Conference in Japan, 2014.

[2] A. Chaer, General linguistics (Indonesia), Rineka Cipta, Jakarta, 1994.

[3] A. Bustamin, I. Areni, I. Sari, Review of speech recognition technology: speech to text method, in: Proc. National Conference of Information Technology (SNATIKA), 2015.

[4] J. Murakami, H. Hotta, Japanese speaker-Independent homonyms speech recognition, Procedia- Social and Behavioral Sciences 27(2011) 306-313.

[5] S. Mishra, A. Bhowmick, M.C. Shrotriya, Hindi vowels classificiation using QCN-MFCC features, Perspectives in Science 8(2016) 28-31.

[6] D. Vijayendra, V. Thakar, Neural network based gujarati speech recognition for dataset collected by in-ear microphone, in: Proc. 6th International Conference on Advances in Computing & Communication, 2016.

[7] R. Nemoto, I. Vasilescu, M. Adda-Decker, Speech errors on frequently observed homophones in French: perceptual evaluation vs. automatic classification, in: Proc. The International Conference on Language Resources and Evaluation, 2013.

[8] F.N. Chen, C. Ni, I-F. Chen, S. Sivadas, V.T. Pham, H. Xu, X. Xiao, T.S. Lau, S.J. Leow, B.P. Lim, C.-C. Leung, L. Wang, C.-H. Lee, A. Goh, E.S. Chng, B. Ma, H. Li, Low resource keyword search strategies for tamil, in: Proc. Acoustic, Speech and Signal Processing, 2015.

[9] Y.-S. Lee, Neural network approach to adaptive learning: with an application to Chinese homophone disambiguation, in: Proc. International Joint Conference Neural Network, 2001.

[10] H.N. Indrabayu, M.S. Pallu, A. Achmad, Statistic approach versus artificial intelligence for rainfall prediction based on data series, International Journal of Engineering and Technology 5(2)(2013) 1962-1969.

[11] D. Hoesen, C.H. Satriawan, D.P. Lestari, M.L. Khodra, Towards robust Indonesian speech recognition with spontaneous-speech adapted acoustic models, in: Proc. 5th Workshop on Spoken Language Technology for Under-resourced Languages, 2016.

[12] B.Y. Cahyono, Kristal-Kristal Ilmu Bahasa, 1st ed., Airlangga University Press, Surabaya, 1995.

[13] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound, in: Proc. the Institute of Phonetics Sciences, 1993.

[14] E. Molina, L.J. Tardon, I. Barbancho, A.M. Barbancho, The importance of F0 tracking in query-by-singing-humming, in: Proc. 15th International Society for Music Information Retrieval Conference (ISMIR), 2014.

[15] R. Sandanalakshmi, V.M. Martina, G. Nandhini, A novel speech to text converter system for mobile applications, International Journal of Computer Applications 73(19)(2013) 7-13.

[16] Md.A. Ali, M. Hossain, M.N. Bhuiyan, Automatic apeech recognition technique or Bangla words, International Journal of Advanced Science and Technology 50(2013) 51-59.

[17] L. Pan, Research and simulation on speech recognition by Matlab, [thesis] Swedia: University of Gavle, 2013.

[18] L.V. Fausett, Fundamentals of neural networks: architectures, algorithms, and applications, Prentice-Hall, Englewood Cliffs, NJ, 1994.

[19] M. Riedmiller, H. Braun, A direct adaptive method for faster backpropagation learning: the RPROP algorithm, in: Proc. the IEEE International Conference on Neural Networks (ICNN), 1993.