Atlantis Highlights in Engineering (AHE), volume 2
International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018)

A Comparison of Naïve Bayes and Bayesian Network on the Classification of Hijaiyah Pronunciation with Punctuation Letters

Adiwijaya, Annisa Riyani, Mohamad Syahrul Mubarok
School of Computing – Telkom University, Bandung 40257, Indonesia
[1]adiwijaya@telkomuniversity.ac.id, [2]riyaniannisa@gmail.com, [3]msyahrulmubarok@telkomuniversity.ac.id

***Abstract*— Arabic is a unique language because it really concerns in *makhraj* (the way of sound is made) that differentiate letters and words. The difference in pronouncing letters and words make the meaning of those words different, because pronuncation in Qur'an letters really concern in harakat (the length of words). According to that matter, it is necessary to build a speech recognition for Hijaiyah with punctuation letters in Qur'an. There are many methods that can be used for building that system. One of the best method is Hidden Markov Model (HMM). Main inference in HMM is Bayes' Rule. Bayes' Rule also used in Naïve Bayes, a part of Bayesian Network. This paper focused on Naïve Bayes and Bayesian Network. Before recognizing the data, first the data will be pre-processed using Linear Predictive Coding (LPC) for extracting cepstral coefficient that will be used as input in classifier. This system give a best micro average F1 score result, 76,67%, with Bayesian Network.**

Keywords—Bayesian Network; Hijaiyah Pronunciation; Naïve Bayes

I. INTRODUCTION

Al-Qur'an is written in Arabic and the guidance for all of muslims [1, 7]. As well-known, Islam is one of the religions in Indonesia with the number of followers more than 200 million people [1, 9]. In Indonesia, Al-Qur'an is easily found in every part of the country, but reciting Qur'an sometimes becomes hard to do for Indonesian because Qur'an is written in Arabic. The followings are the steps for overcome the problem in reciting Qur'an. First, a muslim needs to learn how to pronounce the letters of Qur'an. Next, they can recite Qur'an in the right way. For Indonesian, pronouncing the Hijaiyah letters in the right should be improved because there are many common mistakes. By development of technology grew significantly year by year, some mistakes can be prevented without an instructor. Recently, speech recognition is one of the technology trends. Using this technology, a machine can recognize a person's voice and understand regarding to what the person said [2, 5, 10].

Arabic is a unique language because it really concerns in *makhraj* (the way of sound is made) that differentiate letters and words [3, 4]. Arabic, a language which is used in Al-Qur'an, is an important language to be learned. Al-Qur'an is a holy book that is read by 1.5 million of Muslim in the world [1, 8]. There are many ways of pronouncing letter in Arabic, *tartil* (slow), *tadwir* (medium) and *hadr* (fast) [9]. The difference in pronouncing letters and words make the meaning of those words different, because pronunciation in Qur'an letters really concern in *harakat* (the length of words). According to that matter, it is necessary to build a speech recognition for *Hijaiyah* with punctuation letters in Qur'an.

There are many methods that can be used for building that system. One of the best method is Hidden Markov Model (HMM). There are many successful works done by using HMM, i.e. work of the Dragon System at Carnegie Mellon University (Baker 1975), the longstanding effort of IBM on a voice-dictation system (Averbuch et al. 1987; Bahl, Jelinek, and Mercer 1983; Jelinek 1976), etc [11]. Main inference which is used by HMM is Bayes' Rule that is also used in Naïve Bayes [10], a part of Bayesian Network. This system will focus on Naïve Bayes and Bayesian Network. Before recognizing the data, first the data will be pre-processed using Linear Predictive Coding (LPC) for extracting cepstral coefficient and Principal Component Analysis (PCA) that will be used as input in classifier.

The flow works explained above is the first approach in Arabic Speech recognition system. There are many other speech recognition systems, aside from Arabic Speech recognition system, that use both LPC and PCA for feature extraction and Bayesian Networks and Naïve Bayes method for building the system, that gives a good performance. Therefore, this system tries to build a speech recognition for Arabic speech that gives a good performance by using those method.

II. RELATED WORKS

Arabic speech recognition development is a multidiscipline work that need an integration between Arabic phonetic, Arabic speech recognition technic and also natural language processing. Many researchers are interested in this field. There are many challenges in building recognition system of Arabic

language because it has a unique characteristic. Because of that, there are many rules that have to be added in ASR decoder. There are many methods that have been proposed by researchers, such as speech corpus transcription for Arabic language in Egyptian Arabic dialect. The are many results from each method that have been proposed, one of them is discarding short vocal letters in the text and merge it with most frequently used words [12].

The first ASR research in Arabic language focus on development using Modern Standard Arabic (MSA). The most difficult part in this research is developing an ASR that has a high accuracy result. The researcher doing an experiment using CMU Sphinx4 system for training and testing data, which is a basis for HMM, speaker-independent, a continue recognition system that can handle words in a large number. The purpose of modeling Arabic language using CMU Sphinx system is that it contains an acoustic model and a language that has been generated and trained with Arabic speech data. The training process works by converting audio data into stream vector feature data, converting text into sequence of triphone linear HMM and searching for the best sequence. The use of this system gave a satisfying result with 90% word detected successfully. For the best result, a large number of corpus in data training is needed (more than 500 different speech) [13].

Besides that, ASR system that uses Hidden Markov Model (HMM) allows a speech recognition works in natural way and gives a high accuracy result in application with large scale. The second method works by using formant readings, a method that converts a shape of vocal into a new better form that can be used for changing a desirable vibration frequency. HMM gives a temporary assumption that speech signal can be well categorized as a random parametric process and parameter form stochastic proses can be predicted accurately in a good shape. Another method is formant that has a goal for evaluating result from first formant (F1) and second formant (F2). This experiment considers every frame that is located in the middle of each vocal for minimizing the effect from co-articulation. The result of experiment shows that formant technic is very effective in classifying vocal with the best accuracy for speech recognition system is 91.6%. Vocal that has a short phase cannot be modeled in recognition system, however there is a big overlap between pair of words in formant plot that also exist in recognition system [14].

Another experiment in Arabic Speech Recognition is building a recognition system in many steps, there are recording a sound, segmentation, feature extraction and recognition with Neural Network (NN) (using Support Vector Machine (SVM)). Because of the goal of this experiment is to give a same result between predicted classifications and desirable classification, the ideal situation is to give a better result. Feature extraction is important because when a large number of data is used and there is a possibility of data redundancy, the feature extraction will transforms the input data into a set of features. Support Vector Machine (SVM) is a set of related supervised learning method that analyze data and recognizing the pattern used for classification and regression analysis. This system uses NN and Collaquial Egyptian dialect with noisy environment. The result reached 94% by using SVM [14].

III. Data and Proposed Scheme

Dataset used in this system is speech of *Hijaiyah* with punctuation letters that pronounced by 6 persons. Every person pronouncing 28 *Hijaiyah* letters with 6 punctuation marks four times with 650 *ms* lengths for each pronunciation. Total dataset obtained is 4032 data (6 persons * 28 *Hijaiyah* letters * 6 punctuation marks * 4 recording). The whole data will be divided into 168 class according to the letter and the punctuation mark. We assume the dataset is all consistently same and accepted, because the data set has been acquired by own authors.

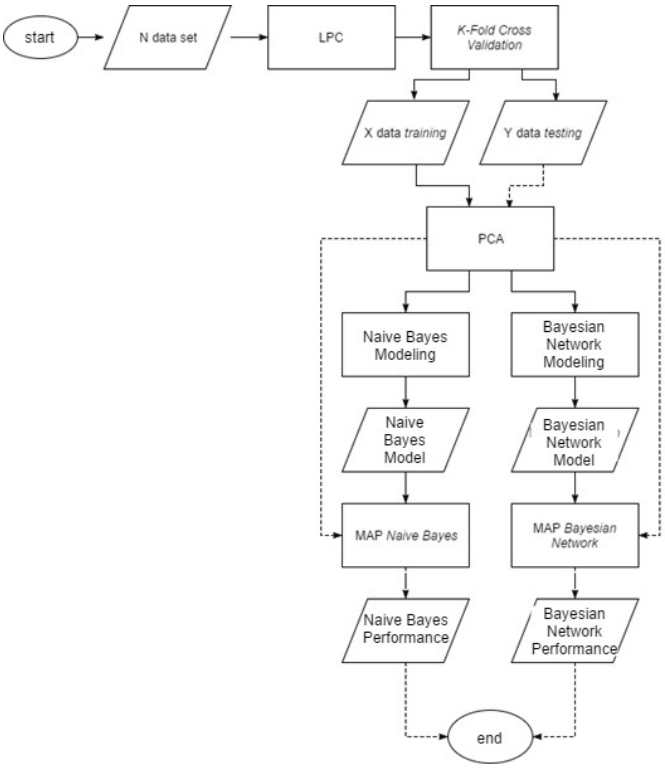

Figure 1. Flowchart of system

Recognition system that will be built has a goal that is classifying data testing. This system has two main steps, there are data pre-processing with Linear Predictive Coding (LPC) and data classification with Bayesian Network and Naïve Bayes. The whole dataset will be first pre-processed using LPC that will convert the data into cepstral coefficient. N dataset will be divided into X data training and Y data testing using K-Fold Cross Validation. The X data that has been converted into cepstral coefficient, will go through PCA step that has a goal to obtain n-principal component that will be used for reconstructing X data training and Y data testing. X

data training will be used for building Naïve Bayes and classifying Y data testing will be done using those models by doing MAP method for each model. The final step is calculating micro average F-1 score for all models. The flowchart of system is shown in Figure 1.

Conditional Probability Table (CPT) that contains prior and likelihood value is needed for building Naïve Bayes model. Evidence value is not needed in this case, because every calculation of posterior value for a class in a data will be done by calculating every feature for that data. Therefore, evidence value for each class will be same in calculating MAP. Prior value obtained by calculating probability of a class appearance in data training. There are two approaches for calculating likelihood value, which are discrete value and continue value.

For calculating likelihood value with discrete approach, first transform the speech data into discrete value by using FCM or K-Means. Every feature value from PCA will be clustered by using FCM and K-Means. Cluster value for each feature is the discrete value of that feature. For every discrete value from FCM and K-Means scenario, there will be two models of Naïve Bayes applied. The first Naïve Bayes model works by grouping (discrete process) every feature of data. The output from that step will be a value for node of that feature. Figure 2 shows the first model of Naïve Bayes. This model will have n + 1 node where n (as much as PC value) node is a node for each feature and 1 node is a node for class.

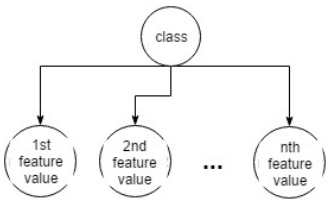

Figure 2. First Naïve Bayes Model

The second Naïve Bayes model, shown in Figure 3, works by grouping (discrete process) data m times with different discrete parameter (cluster for FCM and k for K-Means) for Bayesian Network models. Classification process for each m. This model will have m + 1 node where m node is a random variable with discrete value and 1 node is a node for class.

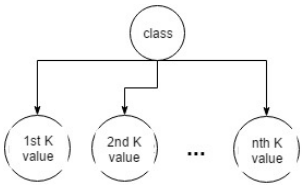

Figure 3. Second Naïve Bayes Model

IV. Results and Analysis

There are five proposed scenarios that will be used for calculating performance of system. The first scenario is changing PC value in PCA. The most optimal PC value is 54 with performance value 74.29%. Table 1 shows the detail result from this scenario. The use of PC value less than 54 will increase the probability of losing an important component so that the data cannot be reconstruct perfectly. The use of PC value more than 54 will not produce a better performance result. This can happen because there will be a probability of an unnecessary additional data in reconstructing process and it will resut in the loss of the important component of the data. Furthermore, the use of big component will increase the performance time.

The second scenario, shown in Table 2, is changing k value in K-Means with first Naïve Bayes model. The experiment shows that the optimal k value is 4. The scenario with k value equal to 4 gives a performance value 44%. The use of k value less than 4 will make the attribute of the data divided into big groups, therefore in one big group there will be a big variance of data. That condition must be avoided because it will increase the probability of different attribute data to be in one group. The use of k value more than 4 also not a good idea. A small number of variance data in one group will be formed because of the big number of data group and it is not an ideal condition. Moreover, the system will have unnecessary group where there is a possibility of one group does not have data inside.

Table 1. Result of PCA scenario

PC Value	*Micro-average* F1-*Score Discrete Naïve Bayes*			*Micro-average* F1-*Score Continue Naïve Bayes*	*Micro-average* F1-*Score Hybrid Naïve Bayes*		Mean
	Model 1		Model 2				
	K-Means = 6	FCM = 6	*K-Means* = 75		*K-Means* = 75 + *continue*	FCM = 6 + *continue*	
50	33,05%	33%	74,37%	44,39%	74,54%	41,36%	50,12%
53	33,18%	32,93%	73,09%	44,84%	73,87%	40,94%	49,81%
54	**33,72%**	**33,1%**	**74,19%**	**45,09%**	**74,29%**	**41,22%**	**50,27%**
55	33%	33,05%	73,85%	45,06%	74,07%	41,24%	50,05%
57	32,83%	32,41%	74,54%	45,76%	74,62%	40,89%	50,18%
60	32,23%	31,64%	74,27%	46,1%	74,42%	40,32%	49,83%
65	31,09%	30,74%	74,14%	45,51%	74,42%	39,45%	49,19%
70	30,27%	73,77%	30,12%	46,05%	74,04%	39,08%	48,89%

Table 2. Result of changing k value in K-Means scenario

K value	Micro-average F1 Score for Discrete	Micro-average F1 Score for Hybrid	Mean
3	29,93%	44,17%	37,05%
4	**33,9%**	**44%**	**38,95%**
5	33,25%	41,96%	37,61%
6	33,1%	41,56%	37,33%
7	32,33%	41,09%	36,71%
8	31,79%	39,7%	35,75%

Table 3. Result of changing random variable in K-Means scenario

Random Variable	Micro-average F1 Score for discrete	Micro-average F1 Score for Hybrid	Mean
65	74,09%	74,12%	74,11%
67	73,8%	73,92%	73,86%
68	**74,49%**	**74,37%**	**74,43%**
69	73,52%	73,75%	73,64%
70	73,42%	73,65%	73,54%
71	73,82%	73,97%	73,90%
74	73,97%	74,17%	74,07%
75	74,19%	74,29%	74,24%
76	73,97%	74,07%	74,02%
80	74,19%	74,19%	74,19%

Table 4. Result of changing cluster value in FCM scenario

Number of clusters	Micro-average F1 Score for discrete	Micro-average F1 Score for Hybrid	Mean
3	29,98%	44,02%	37,00%
4	**33,42%**	**44,09%**	**38,76%**
5	32,63%	41,41%	37,02%
6	33,1%	41,22%	37,16%
7	31,99%	40,37%	36,18%
8	31,66%	39,95%	35,81%

The third scenario is changing the sum of random variable in K-Means with second Naïve Bayes model. The most optimal sum of random variable, shown in Table 3, is 68 with performance of 74.49%. The analysis of this scenario is the same with the second scenario because the second and the third scenario have the goal to determine the effect of Naïve Bayes model used in system. From those two experiments, it show that the second model gives the better performance than the first model.

The fourth scenario is changing cluster value in FCM with first Naïve Bayes model. The most optimal cluster value from this experiment is 4 with performance value of 44.09%, shown in Table 4. The analysis of this scenario is equal with the second and third scenario. The fourth scenario is done to determine which clustering algorithm works better in this system. After comparing the second, third, and fourth experiment, it shows that FCM gives better performance. FCM allows a data to become a member of some group. In this experiment, a data can become a member of up to 2 groups. Therefore in this case, a data with a possibility to become a member of many groups is better than a data that belongs to one group.

The last experiment is to determine which is better between Naïve Bayes and Bayesian Network. From the four experiments before, it shows that Naïve Bayes gives the best performance of 74.49%. By using heuristic approach in Bayesian Network and changing the dependency between each node to get the best performance, the system gives the best output of 76.67%. Naïve Bayes assumes that each random variable is a conditionally independent given class. In this case, the assumption of Naïve Bayes is not correct. It was proven by Bayesian Network's algorithm that dependency for each random variable is influencing the performance of the system. Therefore, deciding the right model or graph structure will give the best performance.

V. CONCLUSION

Based on five experiments, it shows that clustering algorithm that gives the best performance is FCM. Although the second model of Naïve Bayes is better than the first one, Bayesian Network still gives the best performance with the number of 76.67%. To give a better performance in the next experiment, the data should be enriched with more different speaker to give different characteristic in the data. The last and important thing is to determine the right structure for the system, whether it is Naïve Bayes and/or Bayesian Network.

References

[1] Adiwijaya, Aulia, M.N., Mubarok, M.S., Wisesty, U.N., and Nhita, F., 2017. A Comparative Study of MFCC-KNN and LPC-KNN for Hijaiyyah Letters Pronounciation Classification System. In Information and Communication Technology (ICoICT), 2017 5th International Conference on. IEEE.

[2] Wisesty, U.N. Adiwijaya and Astuti, W., 2015. Feature extraction analysis on Indonesian speech recognition system. In Information and Communication Technology (ICoICT), 2015 3rd International Conference on (pp. 54-58). IEEE.

[3] Wisesty, U. N. , Mubarok, M. S., and Adiwijaya, 2017, A classification of marked hijaiyah letters' pronunciation using hidden Markov model, AIP Conference Proceedings 1867, 020036 (2017)

[4] Wisesty, U.N. Liong, T.H. and Adiwijaya, 2012. Indonesian speech recognition system using Discriminant Feature Extraction—Neural Predictive Coding (DFE-NPC) and Probabilistic Neural Network. In Computational Intelligence and Cybernetics (CyberneticsCom), 2012 IEEE International Conference on(pp. 158-162). IEEE.

[5] Yulita, I. N., Houw Liong The, and Adiwijaya. 2012. Fuzzy Hidden Markov Models for Indonesian Speech Classification. JACIII, 16(3), 381–387.

[7] Dukes, K. and Buckwalter, T., 2010, March. A dependency treebank of the Quran using traditional Arabic grammar. In Informatics and Systems (INFOS), 2010 The 7th International Conference on (pp. 1-7). IEEE.

[8] Newman, D. and Verhoeven, J., 2002. Frequency analysis of Arabic vowels in connected speech. Antwerp papers in linguistics., 100, pp.77-86.

[9] Kirchhoff, K., Vergyri, D., Bilmes, J., Duh, K. and Stolcke, A., 2006. Morphology-based language modeling for conversational Arabic speech recognition. Computer Speech & Language, 20(4), pp.589-608.

[10] AbuZeina, D. and Elshafei, M., 2012. Arabic Speech Recognition Systems. In Cross-Word Modeling for Arabic Speech Recognition (pp. 17-23). Springer US.

[11] Juang, B. H., & Rabiner, L. R. (1991). Hidden Markov models for speech recognition. *Technometrics*, *33*(3), 251-272.

[12] Satori, H., Hiyassat, H., Harti, M. and Chenfour, N., 2009. Investigation arabic speech recognition using CMU sphinx system. Int. Arab J. Inf. Technol., 6(2), pp.186-190.

[13] Alotaibi, Y.A. and Hussain, A., 2010. Comparative analysis of arabic vowels using formants and an automatic speech recognition system.

[14] Shady, Y. and Zayed, S.H.H., 2009. Speaker independent Arabic speech recognition using support