p-ISSN : 2443-2210
e-ISSN : 2443-2229

Jurnal Teknik Informatika dan Sistem Informasi
Volume 6 Nomor 2 Agustus 2020

Building Acoustic and Language Model for Continuous Speech Recognition in Bahasa Indonesia

http://dx.doi.org/10.28932/jutisi.v6i2.2684

Vincent Elbert Budiman[#1], Andreas Widjaja[✉#2]

[#] *Program Studi Magister Ilmu Komputer, Fakultas Teknologi Informasi, Universitas Kristen Maranatha*
Jl. Surya Sumantri No.65, Bandung 40164, Indonesia
[1]1879007@maranatha.edu
[2]andreas.widjaja@it.maranatha.edu

Abstract — **Here a development of an Acoustic and Language Model is presented. Low Word Error Rate is an early good sign of a good Language and Acoustic Model. Although there are still parameters other than Words Error Rate, our work focused on building Bahasa Indonesia with approximately 2000 common words and achieved the minimum threshold of 25% Word Error Rate. There were several experiments consist of different cases, training data, and testing data with Word Error Rate and Testing Ratio as the main comparison. The language and acoustic model were built using Sphinx4 from Carnegie Mellon University using Hidden Markov Model for the acoustic model and ARPA Model for the language model. The models configurations, which are Beam Width and Force Alignment, directly correlates with Word Error Rate. The configurations were set to 1e-80 for Beam Width and 1e-60 for Force Alignment to prevent underfitting or overfitting of the acoustic model. The goals of this research are to build continuous speech recognition in Bahasa Indonesia which has low Word Error Rate and to determine the optimum numbers of training and testing data which minimize the Word Error Rate.**

Keywords— **Acoustic Model; Language Model; Speech Recognition; Word Error Rate**

I. INTRODUCTION

Building a good Continuous Speech Recognition for Bahasa (Indonesian Language) is a challenge. Speech recognition is a wide subject, every language on earth has its pronunciation, sentence form, and words. The Continuous Speech Recognition was built using three components, which are, acoustic model, language model, and a dictionary. The acoustic model is a model that contains training data from speech corpus and audio. The algorithm used to build the acoustic model is Hidden Markov Model. A language model is a model that contains relations between words using n-gram to calculate the probability of an utterance appeared. Dictionary is a file that contains every single word on the speech recognition system within its phoneme.

The motivation of this research is to make a speech recognition for Bahasa Indonesia using only common words. There are two points that can be achieved by doing this research, firstly, making a high accuracy speech recognition by lowering down Word Error Rate. Word Error Rate is a percentage of errors words over all the testing data. Generally, Word Error Rate is the number of errors divided by the total words. Secondly, making the speech recognition continuous. Those two points can be achieved by making or improving the acoustic model and language model.

Lowering down the Word Error Rate makes the speech recognition reliable. The target of this research is to reach a Word Error Rate less than 25%. Hopefully by reaching low Word Error Rate, the method, dataset, and model provided in this research can contribute future Speech Recognition in Bahasa Indonesia.

Other than Word Error Rate, context (Language Model) also matters in speech recognition, as stated by Microsoft Research, "More important than word error rate reduction, the language model for recognition should be trained to match the optimization objective for understanding" [1].

Making Continuous Speech Recognition can also be helpful for further research since there are still a limited amount of research on Continuous Speech Recognition in Bahasa Indonesia [2]. Continuous Speech Recognition can be used in many and flexible ways than a common speech recognition to produce the result. The continuous speech recognition can be implemented to make a streaming voice command or to record a speech and document it in real-time.

Creating a Speech Recognition in Bahasa Indonesia also helps widen the usage of people using it (not only from

Indonesia), therefore, our research was focused on people who learn Bahasa Indonesia.

The problem arose from limited research of Bahasa speech recognition, especially continuous, meaning, datasets, model, and method for Speech Recognition in Bahasa Indonesia that can contribute other researchers.

Another issue arose when real-time usage of speech recognition is needed. Continuous speech has an advantage compared to common speech recognition because of the ability to record the voice and process it in real-time. Continuous speech is different from common speech recognition because it processes the user's speech in real-time, therefore the Speech Recognizer can be in an ambiguous state when the user has not finished the speech. The Speech Recognizer can find a lot of similar utterance to form a sentence.

The challenge of making a better accuracy for speech recognition is to gather enough dataset for the acoustic model [3], also, the threshold of the minimum confidence rate have to be configured, if the minimum confidence rate is too low, the result will get a lot of false positive, while setting it too high will make it harder for speech recognizer to recognize the word.

Making the speech recognition to be continuous is very challenging. A discrete speech recognition processes the voice data after it was spoken from any input device or from a datafile. The result will be provided after the Speech Recognizer completed the process.

Continuous speech Recognition is processing the audio in real-time, which sometimes makes the result ambiguous caused by similar phonemes between words (homophone), for example "massa" (mass) and "masa" (time) or similar phonemes between several words and unfinished spoken words [4]. For example, "semak" (bush) from "semakin" (more). To improve the continuous speech recognition, language model must be configured and created using a reliable dataset (consist of right usage of sentences and clear silence needed between words) [5].

The purpose of this research is to create a new or improved acoustic model, language model, and dictionary while improving the word accuracy. The purpose can help to contribute further research and development of Speech to text, voice command, and speech documenter.

Benefits of this research for the attempt to make an acoustic model, language model, and dictionary using common words on Bahasa Indonesia are:

1. Provide datasets of common words for Speech Recognition in Bahasa Indonesia.
2. Provides results as an acoustic model, language model, and dictionary that can be used for further research and development.

II. Research Methodology

Previous research [2] shows that 2 hour of data training were able to reach 23% of Word Error Rate. Based on [1], the context of the training and testing data are important to reduce the Word Error Rate, also increasing the quality of speech Recognition. Therefore, on this research used only common words as the context. In this research, all the speakers were told to limit complex and rare words. Common words were provided in a dictionary of around 1000 words for their guidelines. All speakers were free to provide any words or sentences other than the words on the dictionary, but only common words are allowed. (Total of 2792 words at the end of data training, meaning, only 792 new words provided by the speaker).

We also present the ratio of training and testing data followed by the Word Error Rate of each experiment to see the relation between data training and testing.

Our hypotheses are:

1. Using 2 hours of data training (audio speech and transcription) should be sufficient to reach less than 23% of Word Error Rate by giving context to the dictionary.
2. The number of training and testing data affects Word Error Rate

Based on [6], the research design are: (i) Identify the research problem clearly and justify its selection, (ii) review literatures associated with the problem area, (iii) specify the hypotheses clearly and explicitly central to the problem selected, (iv) effectively describe the data necessary for an adequate test of the hypotheses and explain how such data will be obtained, and (v) describe the methods of analysis which will be applied to the data in determining whether the hypotheses are true or false.

TABLE I
RESEARCH METHODOLOGY

No.	Detail of research design
1	We will identify and state all the problem and hypotheses to determine the method of doing the research.
2	We search for any literature about speech recognition and the detailed methodology (including tools, data training, problems, and solutions).
3	The researcher will specify hypotheses based on the knowledge and literature study
4	The data will use: 1. 2 hours of data training containing audio speech and transcription (16khz audio) 2. Total number of words used will be around 2000 words. 3. Sound data taken from 14 person, which contains 8 males and 6 females. For the tools, we will use: 1. Building language model will use n-gram (using Sphinx4 tools) in ARPA Model. 2. Building Acoustic model (using Sphinx4 tools) using HMM algorithm
5	The method of analysis will use in-built speech

No.	Detail of research design
	recognition validator from Sphinx to determine the result and use WER (Word Error Rate).

III. Acoustic and Language Model Creation

Acoustic and Language Model were created using SphinxTrain (The part of Sphinx4) [5] from Carnegie Mellon University. Before models were built, it is required to provide audio and transcript file to train and test the models.

A. Language Model

To create our own Language Model, firstly it is mandatory to know the phoneme of the language used for the Speech Recognition. On Bahasa Indonesia, there are several special phonemes such as "ng, ny, kh" and the total of phonemes on Bahasa Indonesia is 33 [7].

The language model should be created using ARPA Format models [8] which have the utterance of all the text and the probability of each utterance. ARPA Format models used in this study is 3-gram. These are the example of ARPA Format Model (Shown in Figure 1) built from 3 example sentences:

<s> Saya makan pisang </s>
<s> Ibu makan jagung </s>
<s> Ayah membaca koran </s>

```
Language model created by QuickLM on Wed
Jun 24 14:59:37 EDT 2020
Copyright (c) 1996-2010 Carnegie Mellon
University and Alexander I. Rudnicky

The model is in standard ARPA format,
designed by Doug Paul while he was at
MITRE.

The code that was used to produce this
language model is available in Open
Source.
Please visit
http://www.speech.cs.cmu.edu/tools/ for
more information

This model based on a corpus of 3
sentences and 10 words

\data\
ngram 1=10
ngram 2=14
ngram 3=15

\1-grams:
-0.8451 </s> -0.2341
-0.8451 <s> -0.1963
-1.6232 AYAH -0.2906
-1.6232 IBU -0.2798
-1.6232 JAGUNG -0.2341
-1.6232 KORAN -0.2341
-1.3222 MAKAN -0.2798
-1.6232 MEMBACA -0.2906
-1.6232 PISANG -0.2341
-1.6232 SAYA -0.2798

\2-grams:
-0.6021 </s> </s> -0.3010
-0.6021 <s> <s> -0.1761
-1.0792 <s> AYAH 0.0000
-1.0792 <s> IBU 0.0000
-1.0792 <s> SAYA 0.0000
-0.3010 AYAH MEMBACA 0.0000
-0.3010 IBU MAKAN -0.1761
-0.3010 JAGUNG </s> -0.1761
-0.3010 KORAN </s> -0.1761
-0.6021 MAKAN JAGUNG 0.0000
-0.6021 MAKAN PISANG 0.0000
-0.3010 MEMBACA KORAN 0.0000
-0.3010 PISANG </s> -0.1761
-0.3010 SAYA MAKAN -0.1761

\3-grams:
-0.7782 <s> <s> AYAH
-0.7782 <s> <s> IBU
-0.7782 <s> <s> SAYA
-0.3010 <s> AYAH MEMBACA
-0.3010 <s> IBU MAKAN
-0.3010 <s> SAYA MAKAN
-0.3010 AYAH MEMBACA KORAN
-0.3010 IBU MAKAN JAGUNG
-0.3010 JAGUNG </s> </s>
-0.3010 KORAN </s> </s>
-0.3010 MAKAN JAGUNG </s>
-0.3010 MAKAN PISANG </s>
-0.3010 MEMBACA KORAN </s>
-0.3010 PISANG </s> </s>
-0.3010 SAYA MAKAN PISANG

\end\
```

Figure 1. ARPA Model Example. This are the sample of ARPA trigram Model created from three sentences mentioned.

B. Acoustic Model

To build the acoustic model, it is required to have the Language Model, text corpus, and the audio file for training and testing the data. It is required to configure the SphinxTrain first to get the optimum Word Error Rate. In this study, configurations for the acoustic model were set as follows:

p-ISSN : 2443-2210
e-ISSN : 2443-2229

TABLE II
SPHINXTRAIN CONFIGURATION

PARAMETER	Value	Notes
$CFG_WAVFILE_EXTENSION	wav	The extension of audio files
$CFG_WAVFILE_TYPE	Mswav	The type of audio files
$CFG_WAVFILE_SRATE	16000	The sample rate of audio files
$CFG_MAX_ITERATIONS	10	The number of iterations per data training
$CFG_HMM_TYPE	.cont.	Speech recognition type, .cont. for continuous
$CFG_NPART	10	The number of files splitting for processors to work parallely
$DEC_CFG_BEAMWIDTH	1e-80	The length of beam for each HMM nodes
$DEC_CFG_WORDBEAM	1e-40	The average of sound beam per word
$CFG_FORCE_ALIGN_BEAM	1e-60	The threshold of sound beam while doing alignment, if the difference is more than the value, the file will be ignored.

After all the configuration have finished, as shown on Table I, it is required to install Python, Perl, and Cygwin to compile and run the SphinxTrain. There are several steps of SphinxTrain execution and the result of the process can be seen on the log file. The result of the execution will be the number of Words tested, Words Error, Words Correct, followed by the Word Error Rate.

The training was separated into several modules. The first module (Module_00) is the checker, which will check whether all the file format is correct, such as sample rate and channels for audio, transcript format (including the usage of <s> and </s>), and the words on all the transcript are included in dictionary. This module is necessary to be successfully completed before the training started. After the Module_00 has finished, the second module executed, (Module_20) which is the training module. The configuration on Table II will be reflected on this module, where all the training datasets will be processed to build the acoustic model.

The testing will be on the different module (Decode Module). The Word Error Rate then determined by processing the audio and transcript from testing datasets. While the module tested each sentence on the transcript, each word on each sentence will be differed into 4 types which are Correct, Insertion, Deletion, and Substitution. Correct means the words predicted by the model is correct, Insertion means there are missing word on the sentence, Deletion means there are more words than the expected on the sentence, Substitution means there are incorrect words on the sentence.

IV. SOUND PROFILE

In general, female and male have different average of sound frequency, based on voice spectrum, female sound has higher frequency than male. Different average of frequency had effects on Word Error Rate (Shown on Experiment 1). Each sound profile was drawn using Fast Fourier Transform, which showed the average frequency based on the average of decibels. Figure 2 and 3 show the characteristic of male and female voice taken from the datasets. The horizontal axis is the frequency in Hertz (Hz), and the vertical axis is the sound level in decibels (dB). Before the calculation started, it is important to normalize all the sound file to make all the sounds file equal.

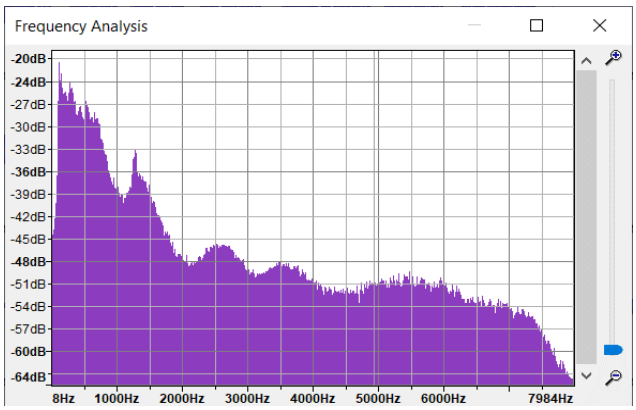

Figure 2. Frequency analysis of male voice. The frequency is more skewed to the lower frequencies. Lower frequencies are used more frequently compared to female voice.

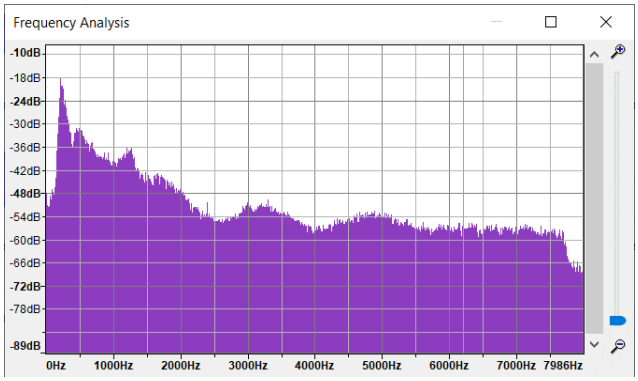

Figure 3. Frequency analysis on female voice. The frequency is more stable through the higher frequency, meaning all frequency are used and higher frequency are used more frequently compared to Male voice.

V. Acoustic And Language Models Experiment

Two categories of experiments are conducted:

1. The testing data used the exact same 50 sentences as training data
2. The testing data used a dictionary that were built before the experiments, containing only common words for Bahasa Indonesia.

Initially a dictionary was built using 1000 common words to be used by the speakers. On later experiments, all speakers had to use only words on dictionary, and they are free to add words to the dictionary. The common words were taken from some of Bahasa Indonesia Learning website such as [9]-[11].

Configuration used on these experiments were $DEC_CFG_BEAMWIDTH (Beam width) set to 1e-80 and $CFG_FORCE_ALIGN_BEAM (Force Align Beam) set to 1e-60.

A. Experiment 1

The first experiment consisted of 8 sub-experiments using different amount of data training and testing, 4 different speakers (6 for training and 2 for testing). The training sentences consisted of 50 pre-stated sentences from each speaker and the rest were wild sentences. The testing sentences consisted of 50 pre-stated sentences transcript from each speaker. Each sentence contained minimal of 3 words. Sub-experiments E and H used same amount of data from the previous experiments, but, on Sub-experiments E, some of the audio data previously were not on the same sample rate, therefore some of the audio data were converted to specific sample rate, unfortunately causing audio loss and corrupted format, therefore, those error files have to be recorded once again to achieve the correct sample rate without having any data loss. On experiments H, we used context-dependant [12] approach (previously used context-independent approach) to reduce the word error rate. Experiment results are shown in Table III.

Table III
Experiment 1 results

No.	Training (Sentence)	Training (Words)	Unique Words	Testing (Words)	Word Error Rate
A	225	1146	307	50	17.8%
C	325	1680	508	50	10.7%
D	525	3018	933	50	7.9%
E	525	3018	933	50	2.4%
F	775	4788	1413	50	1.6%
G	975	6030	1667	100	4.1%
H	975	6030	1667	100	3.6%
I	1225	7416	1790	100	3%

p-ISSN : 2443-2210
e-ISSN : 2443-2229

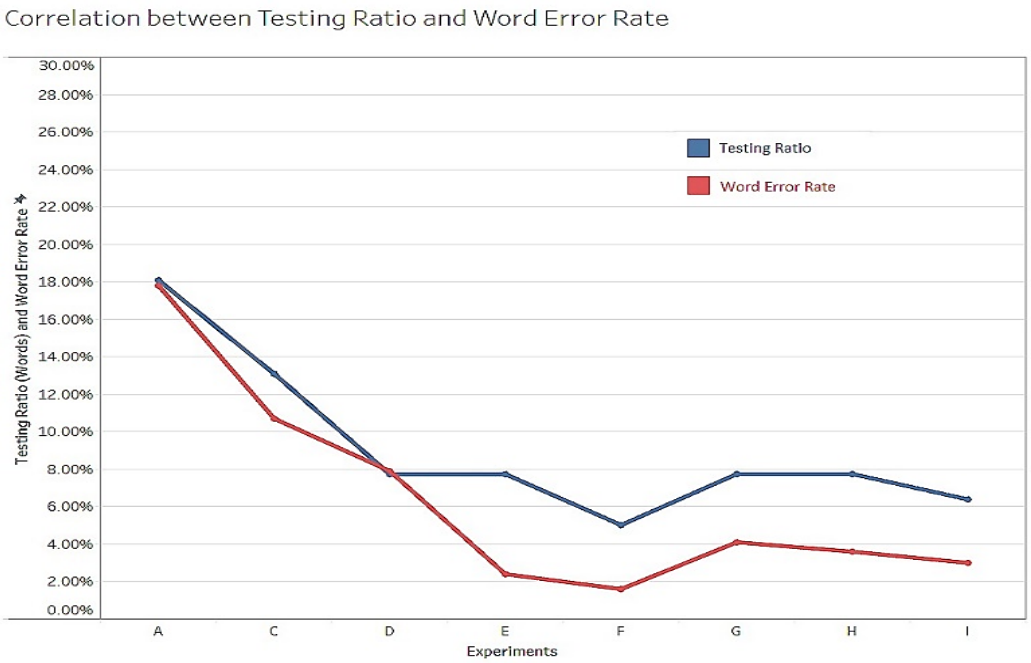

Figure 4. Correlation between Testing Ratio and Word Error Rate. The figure shows that the correlation between Word Error Rate and Testing Ratio are high on Experiment 1, which is 0.95.

On Figure 4, the correlation between Testing Ratio and Word Error Rate is high. Pearson Correlation (r^2) between those two parameters is 0.95, meaning the more training data, the lower the Word Error Rate.

The result above was expected because of the testing and the training dataset used the same 50 pre-stated sentences. On experiments G, higher pitch voices (female) were added as testing data and showing an increase of word error rate from 1.6% to 4.1%, but on H and I, the Word Error Rate were decreasing since we added female dataset for training.

B. Experiment 2

The second experiment consisted of 6 sub-experiments, focused for bigger and more complex sentences (each sentence is consisted of minimal 6 words), different amount of training and testing data were used on each sub-experiment, consisted of 12 different speakers. The dictionary was formed using all the unique words from the training datasets. It is mandatory for all the words on the datasets to be mentioned in the dictionary. Words that are not mentioned in the dictionary will be marked as error. On Figure 5, the correlation between Testing Ratio and Word Error Rate is also high, Pearson Correlation between those two parameters is 0.90, meaning the more training data, the lower percentage of word error rate.

On sub-experiment L, the training data increased by 300 (from 200 to 500) and the sentences contains more words (each sentence contains minimal of 6 words), causing the Word Error Rate increased by 9.1% to 13.3%. Experiment results are shown in Table IV.

C. Experiments Comparison

On this section, we compared the result of experiment 1 and experiment 2. Experiment 1 and 2 concludes that the Testing Ratio is related to the Word Error Rate. The Pearson Correlation value was 0.95 for experiment 1 and 0.90 for experiment 2 (Shown on Figure 6).

TABLE IV
EXPERIMENT 2 RESULTS

No.	Training (Sentence)	Training (Word)	Unique Words	Testing (Word)	Word Error Rate
J	1225	7416	1790	253	4.2%
K	1225	7416	1790	253	4.2%
L	1225	7416	1790	253	13.3%
M	1625	10132	2262	253	11.9%

No.	Training (Sentence)	Training (Word)	Unique Words	Testing (Word)	Word Error Rate
N	1986	12923	2696	253	13.4%
O	2206	14103	2792	506	13.4%

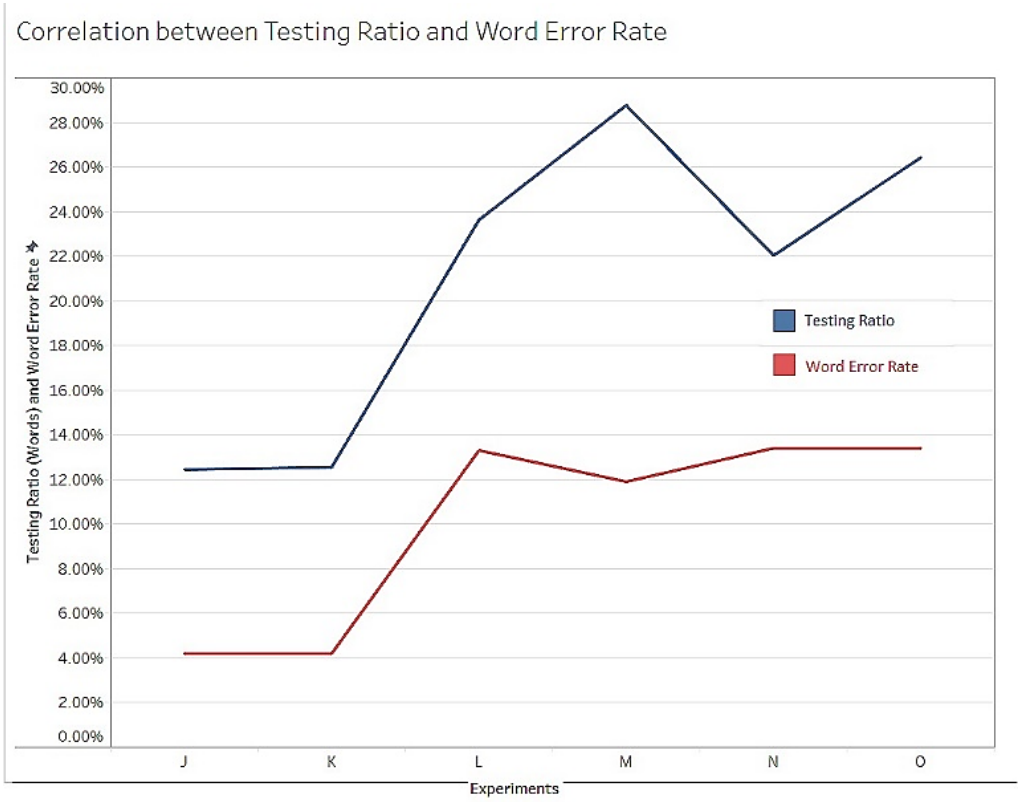

Figure 5. Correlation between Testing Ratio and Word Error Rate. The figure shows that the correlation between Word Error Rate and Testing Ratio are high on Experiment 2, which is 0.90.

Figure 5 shows that the word error rate standard deviation (distribution) was lower than experiment 1. The key difference between experiment 1 and 2 were sentence complexity such as number of words for each sentence, similarity of transcript between training and testing data, and the number of training and testing data. From experiment 2, the Word Error Rate is balanced from experiment L to O by adding more training and testing data while keeping the Testing Percentage Ratio.

Word Error Rate of both of experiments are better than the hypotheses (25%). The log files show that many of the pronunciation errors are very close. Table V shows the example of errors regarding to close pronunciation. As shown on Table IV and from the log file, Phoneme for consonants have more errors than vocals. The error result for this experiment also contains some words that have slightly different spelling but have the same meaning (Shown on Table VI). Error caused by the usage of informal pronunciation of Bahasa Indonesia. Some of the Bahasa Indonesia used on training or testing data were mixed unconsciously by the speakers with regional language such as Sundanese which has similar pronunciation, but have different writing (for example, on Sundanese, letter "P" and "F" are switched, like on the 4th row on Table VI on word "Pernapasan"), it is recommended to clean the data from wrong pronunciation before creating Language and Acoustic Model. It is also important to check if there are mistyped words on the transcript file. Using mistyped words will be trained as it is and cause a new word detection. For example, mistyped word "NEGRI" (Should be "NEGERI") will be trained on the acoustic and language model and can cause future errors (detection on "NEGRI" instead of "NEGERI") and resulting on higher Word Error Rate.

p-ISSN : 2443-2210
e-ISSN : 2443-2229

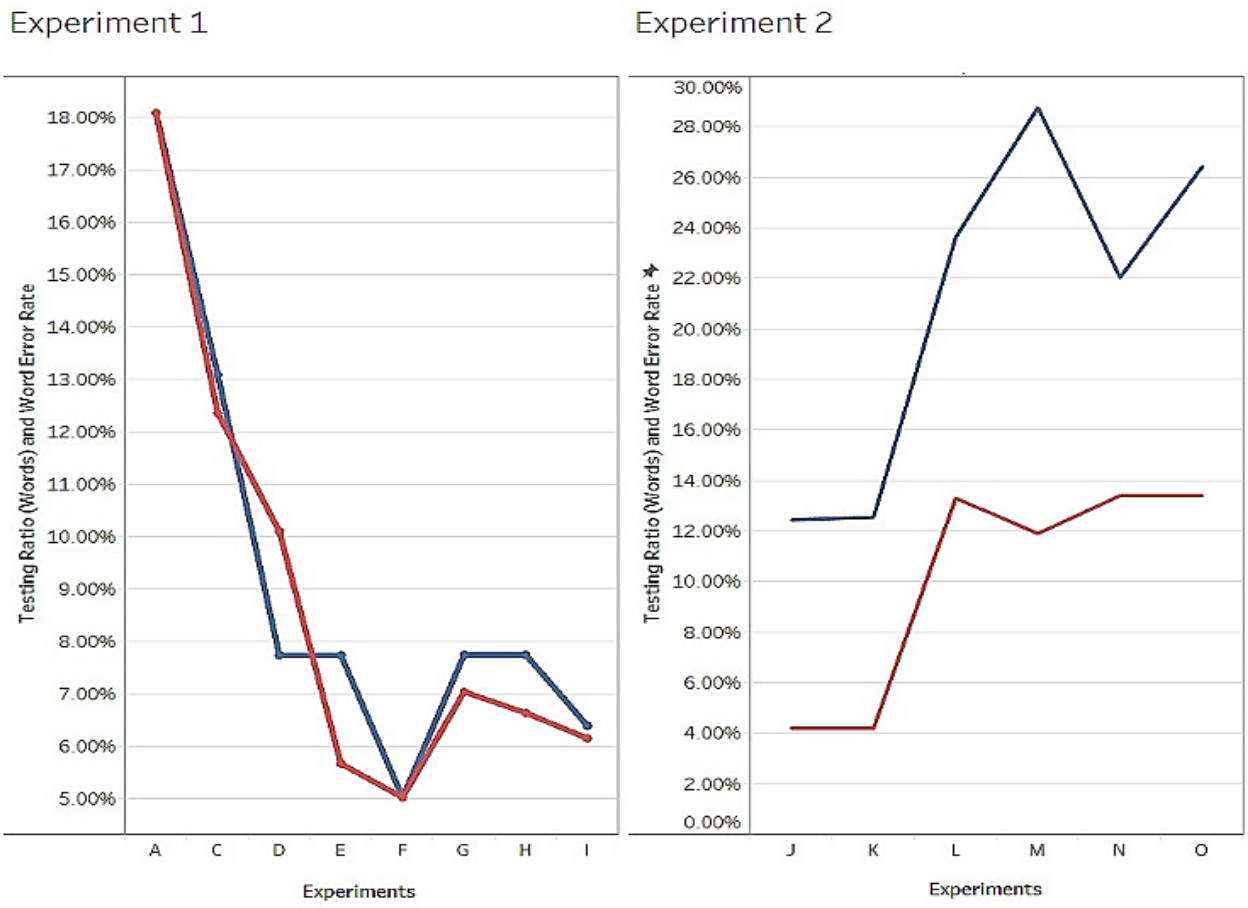

Figure 6. Side by side comparison between experiment 1 and 2. The figure show the comparison between experiment 1 and 2, both having linearity of pearson correlation (r^2) above 0.90.

TABLE V
ERRORS ON EXPERIMENT THAT HAVE CLOSE PHONEME

Expectation	Result
ibu memakaikan jas hujan KEPADA adik	ibu memakaikan jas hujan KU PADA adik
auman singa terdengar HINGGA keluar kebun binatang	auman singa terdengar TIGA keluar kebun binatang
ibu selalu takut untuk NAIK taksi	ibu selalu takut untuk BAIK taksi
aku MERASA MALU karena jatuh di depan teman teman	aku METER SAMA LALU karena jatuh di depan teman teman

TABLE VI
ERRORS ON EXPERIMENT CAUSED BY INFORMAL PRONUNCIATION

Expectation	Result
aku pergi ke luar NEGERI saat liburan	aku pergi ke luar NEGRI saat liburan
bisakah kamu menebak DI MANA aku menyimpan pakaian	bisakah kamu menebak DIMANA aku menyimpan pakaian
pakaian berwarna COKELAT itu terlihat dewasa	pakaian berwarna COKLAT itu terlihat dewasa
ruangan yang kotor dapat menyebabkan masalah PERNAPASAN	ruangan yang kotor dapat menyebabkan masalah PERNAFASAN

VI. Discussion

The context of the speech recognition improves the accuracy of acoustic and language model, increasing the accuracy by 9.6% resulting 13.4% Word Error Rate from compared work by using context on the data gathering. Related works from [2] have 23% Word Error Rate using similar training duration which is approximately 2 hours. The disadvantage of giving context to the Speech Recognizer is the vocabulary recognized is not as wide as the previous works. The advantage is the Word Error Rate is lower, meaning that it is more likely the speech recognizer to give true positive results.

Related works from [13] show that it is possible to create Bahasa Indonesia Speech Recognition using English Acoustic Model. The research resulting in a higher Word Error Rate which is 32% but has the advantage of not requiring to gather datasets, configure, and create own acoustic model. Language Model still also need to be built in [13].

Based on the log file, most errors occurred on consonants. Training and testing data also correlate each other, consequently they affect the Word Error Rate.

Female and male voices also affect the Word Error Rate, meaning average frequency used on training and testing data affect the result of speech recognition. As shown on Sub-experiment G, adding female voice to the testing data (initially all male voices), increase the Word Error Rate.

Configuration used on the SphinxTrain is also important. Force Align beam must be configured depends on the training datasets (May require some trial and errors). Setting up Force Align Beam (threshold) too high may cause wrong pronounced words to be forcedly aligned, while setting up too low may cause a slight mistake on the pronunciation to be not aligned. Aligned wrong pronounced words makes the acoustic model less accurate (underfitted model) while not aligning slight mistakes make the acoustic model overfitted.

Correlation between Testing Ratio and Word Error Rate were taken from each experiment and calculated using the Pearson Correlation method (r^2). The method shows that the linearity between Testing Ratio and Word Error Rate is high (Shown on Figure 6) (0.9 for Experiment 1 and 0.95 on Experiment 2). High correlation meaning that it is required to train the speech recognition with large clean datasets to get the least possible Word Error Rate. In the future usage, which is also mentioned in [3], training data should be consisted of 20 hours for command control and at least 30 hours for small dictation applications that consist of around 5000 words.

Cleaning the transcript and audio data is also important, mentioned in experiment 1 and 2, mistyped words will make the words to be detected as new. It is mandatory to check the sample rate and channel of each audio file. Audio files that do not have correct format will be ignored on the training process, making the training process incomplete (Some words may not appear on the final model).

For further research, wider words and longer duration of speech can be added to improve number of words detected by the system.

VII. Conclusion

Language and Acoustic models were able to be built using SphinxTrain by giving enough dataset as training and testing data. It is recommended to configure the SphinxTrain based on the language and sound profiles.

Configuration were also required during the acoustic model creation since it affects the training process and can cause underfit or overfit the model. The configuration used for beam width is 1e-80 and force align beam is 1e-60. The result may vary depends on the datasets.

Our experiment produced very good results, since we have 9.6% of accuracy increase compared to our hypotheses, by giving context to the data, meaning the datasets can be used for Speech Recognition in Bahasa Indonesia with a decent quality (86.6% accuracy). Word Error Rate of 13.4% was achieved by using 2206 sentences containing 2206 audio files and transcript rows for training, and 800 for testing data. The total training data are approximately 2.5 hours, within 12 total speakers (mixed of male and female). The dictionary consisted of totally 2792 words.

If a testing word does not exist in the dictionary, then it will be marked as an error, hence it is important to make sure all the words used on the testing data exist in the dictionary. It is also important to provide correct audio files (format and content). Wrong format of audio input can cause errors on the training or force aligned phoneme, therefore affecting the Word Error Rate.

Based on Experiments 1 and 2, Testing Ratio (Testing data / Total data) in words are very correlated with the Word Error Rate, resulting Pearson Correlation of 0.95 for experiment 1 and 0.90 for experiment 2. Therefore, there is a linear relationship between Testing Ratio and Word Error Rate on both experiments.

References

[1] Y.Y. Wang, A. Acero and C. Chelba, "Is Word Error Rate a Good Indicator For Spoken Language," *IEEE Workshop on Automatic Speech Recognition and Understanding*,2003, pp. 577-582.

[2] Muljono, A. Q. Syadida, D. R. I. M. Setiadi and A. Setyono, "Sphinx4 for Indonesian Continuous Speech Recognition System," in *International Seminar on Application for Technology of Information and Communication*, 2017, pp. 264-267.

[3] C. M. University. (2017) Training an acoustic model for CMUSphinx, [Online]. Available: https://cmusphinx.github.io/wiki/tutorialam/

[4] D. B. Roe and J. G, *Voice Communication Between Humans and Machines*, Washington DC: National Academy of Sciences, 1994.

[5] Carnegie Mellon University. (2017) CMUSphinx Documentation, [Online]. Available: https://cmusphinx.github.io/wiki/.

[6] Kirshenblatt-Gimblett. (2016) What is Research Design -Performance Studies Method Syllabus [Online]. Available :

http://www.nyu.edu/classes/bkg/methods/005847ch1.pdf

[7] L. Li, Y. Zhao, D. Jiang and Y. Zhang, "Hybrid Deep Neural Network--Hidden Markov Model (DNN-HMM) Based Speech Emotion Recognition," *IEEE Humaine Association Conference on Affective Computing and Intelligent Interaction*, no. 978-0-7695-5048-0, 2013, pp. 312-317.

[8] Carnegie Mellon University. (2017) ARPA Language Model [Online]. Available: https://cmusphinx.github.io/wiki/arpaformat/

[9] IALF Team. (2018) Learning Bahasa Indonesia, Living In Indonesia - A site for expatries. [Online]. Available: https://www.expat.or.id/info/learnbahasaindonesia.html.

[10] L. Aidart. (2018) The most important vocabulary to know during your trip in Indonesia. [Online]. Available: https://www.evaneos.co.uk/indonesia/holidays/survival-vocabulary/.

[11] Afifahamani36. (2017) 250 Basic Words in Indonesian For Beginner - Mastering Bahasa. [Online]. Available: https://masteringbahasa.com/basic-words-in-indonesian.

[12] A.M. Derouault, *Context-Dependent Phone Markov Models for Speech Recognition*,Berlin: Springer, 1988.

[13] V. Ferdiansyah and A. Purwarianti, "Indonesian automatic speech recognition system using English-based acoustic model," *Proceedings of the 2011 International Conference on Electrical Engineering and Informatics,2011*, vol. 10.1109/ICEEI.2011.6021583.,pp. 1-4.