Building MEDISCO: Indonesian Speech Corpus for Medical Domain

Muhammad Reza Qorib
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
muhammad.reza42@ui.ac.id

Mirna Adriani
Faculty of Computer Science
Universitas Indonesia
Depok, Indonesia
mirna@cs.ui.ac.id

***Abstract*—In this paper we report our work of building MEDISCO: Medical Indonesian Speech Corpus. The medical text corpus is collected from five Indonesian online medical consultation websites. From the text corpus, we created a speech corpus that consists of 360 sentences read by 13 speakers. In total, our speech corpus contains 731 medical terms and consists of 4,680 utterances with total duration 10 hours.**

Keywords-Indonesian Automatic Speech Recognition; Medical Speech Corpus; Text Corpus

I. Introduction

Automatic Speech Recognition (ASR) has been improved greatly in recent years after five decades of developments. Since there are more speech data available accompanied by the emergence of deep learning techniques and advancement of computational power, speech recognition has been applied at industrial level with many kinds of application. Speech recognition has changed the way we interact with computers. Speech recognition has become a trend in human—machine communication (HMC) and aimed to improve human—human communication (HHC) [1]. Speech recognition has a broad application, such as virtual assistant [3, 4, 5], spoken content retrieval [6], spoken translation [8], Java code-writing [9], robot controlling [10], hearing aid [11, 12], and many more.

However, Speech Recognition for Indonesian language has not been developed that well in general, moreover in specific application like medical or health. Compared to English speech recognition that has achieved a human level accuracy [24, 25], Indonesian speech recognition has a long way to go [15, 22, 26]. The lack of resource in speech corpora for Indonesian language is one of the main reasons why ASR for Indonesian is not as developed as English or other languages. There are only a few publicly available Indonesian speech corpora, which are [13] with total duration around 102.9 hours, [14] with total duration around 95 hours, [15] with 73 hours dictated speech and 43.5 minutes spontaneous speech, [16] with total duration 14.5 hours, and [17] with total duration near 10 hours. These are relatively small compared to English resources, such as LibriSpeech with almost 1000 hours of speech [18] and Fisher with 2000 hours of speech [19].

Automatic Speech Recognition has big potential to make life easier, including the field of health domain. If ASR system can detect disease and drug names correctly, we can implement an ASR system for transcribing medical consultation, self-service drug store, and many more. However, to be able to create an ASR system, we need to have the suitable corpus first. Unfortunately, until now, there is no Indonesian speech corpus made for medical domain yet. For other languages, there are some speech corpus created for medical domain, such as English [27, 28, 29] and Spanish [30] with various length of speech.

Medical Speech Recognition, especially for Indonesian language, pose a certain challenge. In Indonesian language some of the medical terms are taken from English and read as it is, some has been assimilated with Indonesian spelling and pronunciation, and some are original local terms. The first type of medical terms poses a challenge because it means the sentence will have a mixed pronunciation. The second type also poses a challenge because even though the spelling also has been assimilated with Indonesian language, usually the spelling only changed a bit from the original Latin word and does not conform with general Indonesian spelling. Seeing the need of medical speech recognition corpus to develop Indonesian speech recognition for medical domain, we decide to develop MEDISCO.

II. Text Corpus

As a mean to get actual sentence in medical consultation that will be used for the speech corpora, we created an Indonesian text corpus by gathering online articles. From the text corpus we select some sentences to be spoken in the speech corpus. The text corpus can also be used to create a language model. Language model has been proved helpful in decoding speech audio to text in many researches that has been done before [20, 21, 22]. We continue the work on [2] by adding more recent articles along with enriching them with general Indonesian sentences from news articles.

A. Text Corpus Collection

We collected text sentences from two types of website, which are online medical consultation website and news website. For the online medical consultation websites, we scraped the patience's question with the doctor's answer from five most popular online medical consultation websites in Indonesia.

Some websites provide a forum-like question-answering which allow the patience to ask a follow up question and allow other users to join in the discussion thread. In that kind of websites, we only use the original question (the first question in that thread) and the first reply from a doctor. In total we get 157.732 medical consultation articles. The detail of the article sources can be seen in Table 1.

978-1-7281-1175-9/18/$31.00 ©2018 IEEE

TABLE I. MEDICAL TEXT CORPUS SOURCES

No	*Website*	*Total Articles*
1	Alodokter[1]	112,713
2	Doktersehat[2]	1,492
3	TanyaDok[3]	16,895
4	KlikDokter[4]	25,839
5	DetikHealth[5]	793
Total		**157,732**

Since we get our text corpus from forum-like websites, then it is natural to get texts with slang languages and typographical errors/mistyping. We clean our text corpus in the following way:

- HTML or XML tags are removed
- Repeating words with “2” writings (exp: the word “pusing-pusing” (headaches), written “pusing2”) are replaced with writing the word repeated two times (“pusing pusing”).
- Words that start with numeral and end with alphabetic characters are split into digit-only word and alphabetic-only word. We do this because usually the tailing characters are a unit measurement for the numeral (exp: “25gr” (25 grams) transformed into “25 gr”, “3bln” (3 months) transformed into “3 bln”)
- Slang and typing errors are replaced by looking up in a mapping dictionary (exp: “bln” transformed into “bulan” (month)).
- Apostrophe symbols (‘) are removed
- Other non-alphanumeric characters are converted into space symbols (exp: “kira-kira” transformed into “kira kira” (approximately)).
- Duplicate whitespaces are replaced with single whitespace.

We also tried to clean the text by changing the numeral representation of number into the form of number words (exp: 37 becomes “thirty-seven”). However, it corrupts the meaning more than making it clearer. It is because there are many usages of numeral “2” for repeating words that are not handled in the previous cleaning step, like writing “kira-kira” (approximately) in the form of “kira 2”, we cannot change them into “kira dua (two)”, since the usage of numeral “2” at that example is to symbolize the repeating of a word not as a numerical expression. Another dominant usage for numerals is for code numbers, like phone numbers, order numbers, etc. It also does not make sense if we convert them into the number word form. Therefore, we decide to keep numeral number as it is, since it gives almost no impact to the model decoding.

Other than medical articles, we also scrape articles from Indonesian news websites to add common Indonesian words to the language model. We scrape the news article from two most popular Indonesian news websites, KOMPAS and DetikNews. From KOMPAS, we gather five categories of article, which are news, economy, life style, science, and education, while for DetikNews we only scrape the news category. Number of news articles that we have gathered can be seen in Table 2

TABLE II. NEWS TEXT CORPUS SOURCES

No	*URL*	*Total Articles*
1	KOMPAS[6]	47,207
3	DetikNews[7]	9,905
Total		**57,112**

Different with medical text corpus, the news articles already use proper Indonesian, so we only need to remove the HTML tags and split the text into sentences. Even though we got fewer news articles than the medical articles, but news articles are mostly longer. Therefore, in terms of number of sentences, the ratio between sentence from news articles and from medical articles are roughly 1:2. The detail information of the text corpora can be seen in Table 3.

TABLE III. TEXT CORPORA DETAILS

Type	*Medical*	*News*
Articles	157,732	57,112
Sentences	2,033,399	1,110,116
Words	36,184,611	18,682,624
Unique words	192,004	161,058

III. DEVELOPMENT OF SPEECH CORPUS

There are two ways in creating a speech corpus. The first one is by transcribing speech recording that was taken from real conversations, usually called spontaneous speech or conversational speech. The second is by asking people to read a script that has been prepared before, usually called dictated speech or read speech [22]. Both have their own advantages, but since we do not have any access to actual medical consultation recordings, we follow the second approach. We create our speech corpus by recording our speakers reading designated sentences which has been selected carefully from the text corpora that we have gathered before.

A. Sentence Selection

From the text corpus, we got 2 million sentences. To help us choosing the sentences for the speech corpus, we sort the sentences based on the phonetical structure and the medical terms contained in the sentence. We want to have a set of sentences with phonetics as balance as possible while having a large variant of medical terms. To get a large variant of medical terms, we calculate the sum of inverse document frequency (IDF) of all the medical terms in that sentence. Inverse document frequency is a way to measure the importance of a word by dividing the total of documents with number of documents which contain that

[1] http://www.alodokter.com

[2] http://doktersehat.com

[3] http://www.tanyadok.com

[4] http://www.klikdokter.com

[5] http://health.detik.com

[6] http://www.kompas.com

[7] http://www.detik.com

word. The formula of IDF can be seen in (1), with N is the total number of documents and n is the number of documents which contain the word in question. In this case we regarded every sentence as document. The reason why we use IDF is because IDF can give more weight to medical terms that are rarely seen (exp: "chlamydia") and less weight to medical terms that exist in many sentences (such as "doctor" or "virus"). Therefore, choosing sentences with greater IDF help us enrich our medical terms collection.

$$idf = \frac{N}{n} \quad (1)$$

There are some variations in defining Indonesian phoneme and diphthong set. We use the Indonesian phoneme set from Dardjowidjojo [23] with slight variation. We agreed with [16] in adding /q/ to the phoneme set because now there are many Indonesian words with 'q' letter, and /q/ has different sound and pronunciation with other letters. The complete Dardjowidjojo's version of Indonesian phoneme set can be seen in Table 4.

TABLE IV. INDONESIAN PHONEME SET [23]

No	*Category*	*Phoneme*	*Example*
1	Vowels	/a/	Warna
2		/e/	Merah
3		/ə/	Merona
4		/i/	Itu
5		/o/	Topi
6		/u/	Kamu
7	Diphthong	/ai/	Pakai
8		/au/	Kemilau
9		/oi/	Amboi
10	Semi-vowels	/w/	Menawan
11		/y/	Yang
12	Plosive Consonants	/b/	Biru
13		/p/	Pucat
14		/d/	Dengan
15		/t/	Tali
16		/g/	Garis
17		/k/	Kuning
18		/kh/	Khas
19	Africate Consonants	/j/	Jepang
20		/c/	Cina
21	Fricative Consonants	/v/	Variasi
22		/f/	Figur
23		/z/	Zebra
24		/s/	Sambut
25		/sy/	Syair
26		/h/	Harmoni
27	Liquid Consonants	/r/	Riang
28		/l/	Lantas
29	Nasal Consonants	/m/	Merdu
30		/n/	Nan
31		/ny/	Nyaring
32		/ng/	Ngiang

To get a phonetically balanced set of sentences, we use a greedy approach by checking the least phoneme we currently have in the set of sentences, then try to look for sentences that have the phoneme. If there are more than one candidate for that phoneme, we will pick the sentence with biggest medical term IDF. If we still have more than one candidate, we will choose the sentence with minimum difference of its biggest and smallest phoneme counts (we could say that it is the sentence that has the most balanced phonemes compared to other candidates).

We observe that most of sentences with rich medical terms are very long sentences. We do not want our sentence to be too long, strictly only 10-15 words for each sentence. Thus, the sentences from the text corpus are transformed by separating compound sentences into simple sentences and by adding some linking words to make them more logical and natural.

After that, we reorder the list of sentences again to make the top-n sentences phonetically balanced by the same way as before. The first 360 sentences become training sentence and will be read by training speakers. The test sentences also have 360 sentences with 180 of them are selected sentences from the training set with most medical terms and the other 180 sentences are new sentences that does not appear in training set.

B. Speech Recording

We gather 13 speakers with age in range of 20-25, with 7 of them are male and 6 of them are female. We have a variety of the speaker's race, including Javanese, Arabs, and Chinese. From the sentences we have chosen before, we ask our speakers to read the transcript sentence by sentence. If there are any mistake in reading the sentence, or if there are long pause or external noise, we ask our speaker to repeat the sentence. We record the audio in rooms that are relatively quiet. The recordings are done using noise-cancelling microphone so in the recorded audio there are almost no noise. After recording the audio, we sample the audio to check the quality of the recording. Thus, the recordings are guaranteed to be clean without any need of post-processing.

From 13 speakers, 11 are asked to read the train sentences and the other 2 are asked to read the test sentences. All of the speakers have clear pronunciation of Indonesian without any accent. The recordings are saved in m4a format and then converted into 16-bit wav with 44.1 kHz sampling rate using ffmpeg[8]. In total we get 10 hours of speech audio, the detail can be seen in Table 5.

TABLE V. SPEECH CORPUS PROPERTIES

Type	*Speaker*		*Number of Utterance*	*Hours*
	Male	*Female*		
Train	6	5	3,960	8.52
Test	1	1	720	1.50
Total			**4,680**	**10.02**

C. Medical Terms

In this corpus, most of the sentences contain medical terms and each sentence can have more than one medical terms. As we have said before there are many types of medical terms in Indonesian language. We divided the medical terms into four categories:

a) EN-EN: These are the medical terms that are written in English spelling and read with English pronunciation. The pronunciation in the audio recordings

[8] https://www.ffmpeg.org/

may not be perfect but they are clear as English pronunciation. Exp: *"Filler Injecection"*

b) EN-ID: These are the medical terms that are written in English spelling but read with Indonesian pronunciation. Exp: *"tumor", "syphillis"*.

c) ID-ID: These are the medical terms that written in Indonesian spelling and read with Indonesian pronunciation. This category also includes loan words such as *"tuberkulosis"* (tuberculosis) and original words such as *"cacar"* (smallpox).

d) O-EN: We also prepare a tag for other English words or phrases that are not medical terms such as *"college"* and *"radio frequency"*.

In determining the medical terms in a sentence, we use a certain convention on what considered as medical terms. Words or phrases that are tagged as medical terms are those that comply with one or more of the following conditions:

- Name of body organs or body parts (exp: *"heart"*, *"wrists"*)
- Name of symptoms and diseases (exp: *"HIV"*, *"osteoarthritis"*)
- Name and types of drugs (exp: *"imiquimod"*)
- Name of chemical substances that exist in the body or usually consumed (exp: *"calcium"*)
- Name of specific medical tools or procedures (exp: *"filler injection"*, *"fluoroscopy"*)

Before we do the recording process, we have consulted the pronunciation of the medical terms to a medical doctor. All speakers are obliged to read the medical terms precisely. In the audio transcription, we annotate the medical terms with XML tag.
Example:

"<EN-EN>Human immunodeficiency virus</EN-EN> menyerang <ID-ID>sistem imunitas</ID-ID> <ID-ID>tubuh</ID-ID> seseorang dan menyebabkan <ID-ID>sistem imunitas</ID-ID> menjadi lemah"

Literal translation:

Human Immunodeficiency Virus attacks someone's (body's) immune system and make the immune system becomes weak.

In total there are 731 distinct medical terms, 612 medical terms are in training set and 119 medical terms are exclusively only in the test set. The detail can be seen in Table 6.

TABLE VI. MEDICAL TERMS

Type	*Train*	*Test (Exclusive)*
EN-EN	49	9
EN-ID	109	17
ID-ID	451	93
O-EN	3	0
Total	**612**	**119**

IV. Neural Network ASR

We evaluate our dataset by training an automatic speech recognizer using Neural Network. We trained an end-to-end speech recognizer using Bidirectional Long Short-Term Memory (LSTM) circuits. We are able to train the network in end-to-end fashion (character level transcription) because of the Connectionist Temporal Classification (CTC) [31] technique. The model we use is similar to Deep Speech [20, 24, 32]. We implement the model using Keras [34] with Tensorflow [35] backend.

A. Model Architecture

For each audio $x^{(i)}$ from the dataset, the audio is transformed into features for each slice of time t so that $x^{(i)} = \left\langle x_t^{(i)} \middle| t = 1, 2, .., T^{(i)} \right\rangle$ where $T^{(i)}$ is the length of that audio in number of time step/slices. Suppose the label for that audio is $y^{(i)}$ with $y^{(i)} = \left\langle y_t^{(i)} \middle| t = 1, 2, .., T^{(i)} \right\rangle$ and $y_t \epsilon \{a, b, .., z, ', space, \varepsilon\}$, the model objective is to map the features at each time slice $x_t^{(i)}$ to the conditional probability of each character $p_t^{(i)} = \mathrm{P}\left(y_t^{(i)} \middle| x_t^{(i)}\right)$. From the output of the model, we measure the error using CTC loss [33]. The model accepts features of speech audio and outputs the probability of 29 characters (26 alphabets with apostrophe, space symbol, and blank symbol). We use Mel Frequency Cepstral Coefficient (MFCC) to get the features of the audio.

The model uses 3 layers of fully connected layers followed by 1 layer of Bidirectional LSTM and another fully connected layer to produce the probability. For each time step t, the first 3 layers are computed by:

$$h_t^{(l)} = g(W^{(l)} h_t^{(l-1)} + b^{(l)}) \tag{2}$$

where $g(z) = \min\{\max\{0, z\}, 20\}$ is clipped rectified-linear unit (ReLU) and $W^{(l)}$ and $b^{(l)}$ are weight matrix and bias for layer l. Bidirectional LSTM is a Bidirectional Recurrent Neural Network (RNN) with LSTM used for the hidden unit. As illustrated in Figure 2, Bidirectional RNN computes *forward* hidden sequence $\vec{h}$ and *backward* hidden sequence $\overleftarrow{h}$. The hidden units and the output are updated using (3), (4), and (5).

$$\overrightarrow{h}_t = \mathcal{H}\left(W_{x\overrightarrow{h}} x_t + W_{\overrightarrow{h}\overrightarrow{h}} \overrightarrow{h}_{t-1} + b_{\overrightarrow{h}}\right) \tag{3}$$

$$\overleftarrow{h}_t = \mathcal{H}\left(W_{x\overleftarrow{h}} x_t + W_{\overleftarrow{h}\overleftarrow{h}} \overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \tag{4}$$

$$y_t = W_{\overrightarrow{h}y} \overrightarrow{h}_t + W_{\overleftarrow{h}y} \overleftarrow{h}_t + b_o \tag{5}$$

where W_{ij} is the weight matrix of unit i and unit j, b_h is the bias of hidden unit h, and $\mathcal{H}$ is the LSTM. Ilustration of LSTM can be seen in Figure 1.

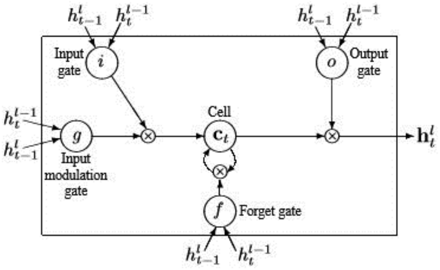

Figure 1. Long Short-Term Memory Cell.

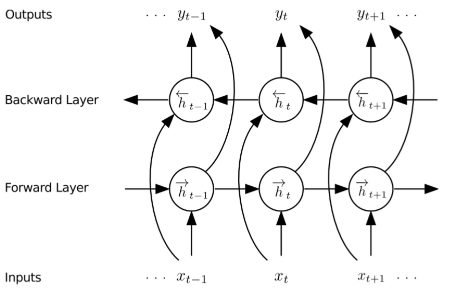

Figure 2. Bidirectional Recurrent Neural Network.

B. Model Inference

As explained before, the model will produce the probability of each character for each time step. There are two ways in inferring the CTC output, the first is the greedy approach by selecting the most probable characters for each time step and the second is by using beam search [31]. In the beam search approach, all possibility of character sequence is calculated with only top-*k* best sequence are saved for the next time step. The coefficient *k* is called the beam size. The greedy method is a specific case of beam search where the beam size is 1.

Beam search can produce a better result, because in CTC, sequences with same end-result will have the probability added up together. Another advantage of beam search is we can combine them with language model to increase the probability of real words. The equation to combine the sequence and language model probability can be seen in (11)

$$Y^* = argmax_Y\ P(Y|X) * P(Y)^\alpha * L(Y)^\beta \tag{11}$$

where $P(Y|X)$ is the CTC conditional probability, $P(Y)$ is the language model probability, and $L(Y)$ is the word count or word bonus. The language model weight *(α)* and word count weight *(β)* are hyperparameters that need to be set. We modify [36] implementation for beam size CTC decoding by adding the language model integration.

V. Evaluation

We use word accuracy as a metrics to evaluate our model. Word accuracy (WA) is defined as the complement of Word Error Rate (WER) (12). Word Error Rate can be seen as the Levenshtein distance between two string at word level instead of character level. The formula of WER can be seen in (13), where *S* is the number of substituted words, *I* is the number of inserted words, *D* is the number of deleted words, *C* is the number of correct words, and *N* is the total words, which is defined by the sum of substituted, deleted, and correct words.

$$Word\ Accuracy = 1 - WER \tag{12}$$

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{13}$$

VI. Experiments

For the experiments, we use our own corpus (MEDISCO) and a general Indonesian speech corpus called TITML-IDN [16]. TITML-IDN has 20 speakers (11 male and 9 female) with total duration 14.6 hours. The testing dataset that is used for all of our experiments is the testing part of MEDISCO. For training, we choose 7 male speaker and 5 female speakers from TITML-IDN for training set, and 2 males and 2 females for validation set. From MEDISCO, we chose 4 female speakers and 4 male speakers as training set and use 1 male speaker and 1 female speaker as validation set. We also create a third dataset which is the union of TITML-IDN and MEDISCO training data. The detail properties of the dataset used for training can be seen in Table 7.

TABLE VII. Training Dataset

Dataset	*Speaker*		*Number of Utterance*	*Hours*
	Male	*Female*		
TITML-IDN	7	5	4014	8.63
MEDISCO	4	4	2880	5.86
TITML-IDN + MEDISCO	11	9	6894	14.49

VII. Result

For each dataset, we run the experiment with two inference configurations, combining with language model (LM) and without language model (No LM). For some datasets we also experiment with different beam size (k), with k=100 and k=1000. The result of the experiment can be seen in Table 8.

Our early experiment shows our best model achieve 64.8% word accuracy. The best word accuracy is achieved by combining TITML-IDN and MEDISCO as training data and use beam search with k=1000 combined with language model to decode the model output. The experiment also shows that there is a huge gap of accuracy between training using TITML-IDN only and using MEDISCO only. The experiment also shows that increasing the beam size doesn't have a significant impact if the decoding is not combined with the language model.

TABLE VIII. Experiment Results

Training Data	*Word Accuracy*	
	No LM	*LM*
TITML-IDN (k=100)	2.4%	5.3%
MEDISCO (k=100)	26.4%	48.9%
MEDISCO (k=1000)	26.3%	55.9%
TITML-IDN + MEDISCO (k=100)	32.1%	58.4%
TITML-IDN + MEDISCO (k=1000)	32%	**64.8%**

VIII. Conclusion

In this paper, we report our effort in building Indonesian speech corpus for medical domain. Our corpus consists of 4,680 utterances from 13 speakers with 731 distinct medical terms and total duration 10 hours. Our early evaluation shows 64.8% word accuracy. With this corpus we hope Indonesian Automatic Speech Recognition especially in medical domain can be developed more and would be able to improve health care systems in Indonesia.

ACKNOWLEDGMENT

This work is supported by Information Retrieval Lab of Faculty of Computer Science Universitas Indonesia and Hibah PITTA 2018, funded by DRPM Universitas Indonesia No.1886/UN2.R3.1/HKP.05.00/2018.

REFERENCES

[1] D. Yu and L. Deng, Automatic Speech Recognition: A Deep Learning Approach, 1st ed. London: Springer-Verlag, 2015,pp.1-3.

[2] A. N. Hakim, R. Mahendra, M. Adriani and A. S. Ekakristi, "Corpus development for Indonesian consumer-health question answering system," 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, 2017, pp. 222-227

[3] B. Li, et al. , "Acoustic Modeling for Google Home," INTERSPEECH 2017. Stockholm: Academic, 2017, pp. 271–350.

[4] S. Dasgupta and P. M. D. Raj, "Developing a PDA to control device using proposed algorithm," 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), Kannur, 2017, pp. 1184-1190.

[5] R. Sarikaya, "The Technology Behind Personal Digital Assistants: An overview of the system architecture and key components," in IEEE Signal Processing Magazine, vol. 34, no. 1, pp. 67-81

[6] L. s. Lee, J. Glass, H. y. Lee and C. a. Chan, "Spoken Content Retrieval—Beyond Cascading Speech Recognition with Text Retrieval," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 9, pp. 1389-1420, Sept. 2015.

[7] M. Elaraby, A. Y. Tawfik, M. Khaled, H. Hassan and A. Osama, "Gender aware spoken language translation applied to English-Arabic," 2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP), Algiers, 2018, pp. 1-6

[8] H. Z. Muhammad, M. Nasrun, C. Setianingsih and M. A. Murti, "Speech recognition for English to Indonesian translator using hidden Markov model," 2018 International Conference on Signals and Systems (ICSigSys), Bali, 2018, pp. 255-260.

[9] K. Lunuwilage, S. Abeysekara, L. Witharama, S. Mendis and S. Thelijjagoda, "Web based programming tool with speech recognition for visually impaired users," 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), Malabe, 2017, pp. 1-6..

[10] M. H. Tambunan, Martin, H. Fakhruroja, Riyanto and C. Machbub, "Indonesian speech recognition grammar using Kinect 2.0 for controlling humanoid robot," 2018 International Conference on Signals and Systems (ICSigSys), Bali, 2018, pp. 59-63.

[11] A. Chern, Y. H. Lai, Y. P. Chang, Y. Tsao, R. Y. Chang and H. W. Chang, "A Smartphone-Based Multi-Functional Hearing Assistive System to Facilitate Speech Recognition in the Classroom," in IEEE Access, vol. 5, pp. 10339-10351, 2017.

[12] I. Dabran, T. Avny, E. Singher and H. Ben Danan, "Augmented reality speech recognition for the hearing impaired," 2017 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems, Tel-Aviv, 2017, pp. 1-4.

[13] S. Sakti, et al, "Development of Indonesian large vocabulary continuous speech recognition system within A-STAR project, " in Workshop on Technologies and Corpora for Asia-Pacific Speech Translation (TCAST), Hyderabad, 2008.

[14] G. Gunarso, H. Riza, "An Overview of BPPT's Indonesian Language Resources", 12th Workshop on Asian Language Resources (ALR12), pp.73-77, 2016.

[15] D. Hoesen, C. H. Satriawan, D. P. Lestari, M. L. Khodra, "Towards Robust Indonesian Speech Recognition with Spontaneous-Speech Adapted Acoustic Models", in Procedia Computer Science, vol. 81, pp 167-173, 2016.

[16] D. P. Lestari and S. Furui, "A large vocabulary continuous speech recognition system for Indonesian language, " in 15 Indonesian Scientific Conference in Japan Proceeding, p. 17-22

[17] M. R. A. R. Maulana and M. I. Fanany, "Indonesian audio-visual speech corpus for multimodal automatic speech recognition," 2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Bali, 2017, pp. 381-386.

[18] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. "Librispeech: an asr corpus based on public domain audio books," In ICASSP, 2015

[19] C. Cieri, D. Miller, and K. Walker. "The Fisher corpus: a resource for the next generations of speech-to-text," In LREC, volume 4, pages 69–71, 2004

[20] A. Hannun, C. Case, J. Casper, B. Catanzaro, G. Diamos, E. Elsen, R. Prenger, S. Satheesh, S. Sengupta, A. Coates, and A. Y. Ng. "Deep speech: Scaling up end-to-end speech recognition". Arxiv:1412.5567, 2014. http://arxiv.org/abs/1412.5567

[21] A. Y. Hannun, A. L. Maas, D. Jurafsky, and A. Y. Ng. "First-pass large vocabulary continuous speech recognition using bi-directional recurrent DNNs." abs/1408.2873, 2014.

[22] D. P. Lestari and A. Irfani, "Acoustic and language models adaptation for Indonesian spontaneous speech recognition," 2015 2nd International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA), Chonburi, 2015, pp. 1-5.

[23] Darjowidjodjo, S. "Indonesian Syntax", Ph.D dissertation, Georgetown University, Washington. 19

[24] D. Amodei, et al., "Deep Speech 2: End-to-End Speech Recognition in English and Mandarin". arXiv:1512.02595, 2015.

[25] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, A. Stolcke, "The Microsoft 2017 Conversational Speech Recognition System". arXiv:1708.06073, 2017. https://arxiv.org/abs/1708.06073

[26] 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, Bandung, 2011, pp. 1-4.

[27] C. Chiu, et al., "Speech recognition for medical conversations". arXiv:1711.07274, 2017. https://arxiv.org/abs/1711.07274

[28] W. Salloum, E. Edwards, S. Ghaffarzadegan, D. Suendermann-Oeft, M. Miller, " Crowdsourced Continuous Improvement of Medical Speech Recognition," The AAAI-17 Workshop on Crowdsourcing, Deep Learning, and Artificial Intelligence Agents, San Fransisco, 2017, pp. 432-435

[29] S. Pakhomov, M. Schonwetter, J. Bachenko, "Generating Training Data for Medical Dictations," in Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics on Language Technologies, Stroudsburg, 2001, pp. 1-8

[30] J. R. Orozco-Arroyave, J. D. Arias-Londoño, J. F. Vargas-Bonilla, M. C. Gonzalez-Rátiva, E. Nöth, " New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, 2014.

[31] A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In (ICML), Beijing, 2014

[32] Rob, "Keras Deep Speech". GitHub Repository, 2018. https://github.com/robmsmt/KerasDeepSpeech

[33] A. Graves, S. Fernandez, F. Gomez, and J. Schmidhuber. "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks". In (ICML), pages 369–376. ACM, 2006

[34] F. Chollet, et al., "Keras", 2015, https://keras.io

[35] M. Abadi, et al., "Tensorflow: Large-Scale Machine Learning on Heterogeneous Systems", 2015, tensorflow.org

[36] A. Hanun, "Sequence Modeling with CTC", Distill, 2017.