Development of Speech Corpus and Speech Recognition System for Indonesian Language

Sakriani Sakti[1], *Paulus Hutagaol*[2], *Arry Akhmad Arman*[3], *Satoshi Nakamura*[1]

[1]ATR Spoken Language Translation Laboratories, Japan
[2]R&D Division, PT Telekomunikasi Indonesia, Indonesia
[3]Electrical Engineering Department, Bandung Institute of Technology, Indonesia
{ssakti,nakamura}@atr.jp, hutagaol@risti.telkom.co.id, aa@lss.ee.itb.ac.id

Abstract

In this paper, we describe our efforts in developing Indonesian speech corpus and speech recognition system. The major challenge to deal with is the shortage of corpus as well as the lack of speech-enabling technology and research on Indonesian language. Difficulties arise in developing speech corpus since Indonesian is actually most people's second language after their own ethnic native language. Collecting all possible languages and dialects that recognized in Indonesia is still a big problem. In developing speech recognition system, segmented utterances according to labels are usually as a starting point for training speech models. It is basically produced by forced alignment given the transcription. In this case, we required an Indonesian speech recognizer which is not available. One way to solve this problem is to segment the utterance uniformly. However, this initialization method would not give sufficient performance. Here, we used a English speech recognizer to set initial segmentation of Indonesian utterance. This method improves the performance significantly up to 40% absolute.

1. Introduction

Indonesia is the fourth most populous nation in the world and it has a population of 210 million people. Considering population, variability, distribution, religious circumstances and linguistic aspects, Indonesian/Malay was ranked ninth to include in the Global-Phone speech database [1]. It also ranks highly in the speech science community [2]. These results are seemingly odds with the fact that Indonesian language suffers from the lack of speech-enabling technology and research.

During the past view years, most speech-related researchers in Indonesia only played an active role in speech synthesizer technologies and natural language processing. There have been no speech recognition research activities which could develop a full-fledged prototype system. One of the main problems is the shortage of an Indonesian speech corpus. Difficulties arise in developing an Indonesian speech corpus since Indonesian is actually most people's second language after their own ethnic native language. Collecting all of the possible languages and dialects of the tribes recognized in Indonesia is still too difficult to be done. Recently, research proposed by another country was to produce Indonesian speech recognition using cross-lingual pronunciation modeling from other resource languages. However, it was observed that this would result in poor performance [3].

In this paper, we present our work that develop an Indonesian speech corpus. The corpus has successfully covered a wide range of ethnic languages for both clean and telephone speech. A word-based Indonesian speech recognition system is also being developed. This initial phase of Indonesian speech corpus and speech recognition system development is part of a project funded by Asia-Pacific Telecommunity (APT). The detailed about project background will be described in Section 1. Characteristic of the Indonesian language is also explained in Section 2. Detailed experiments will be described in the rest of this paper, the speech corpus (Section 3), initial segmentation issues (Section 4), language modeling issues (Section 5), experimental results and discussion (Section 6), and followed by a conclusion (Section 7).

2. Project Background

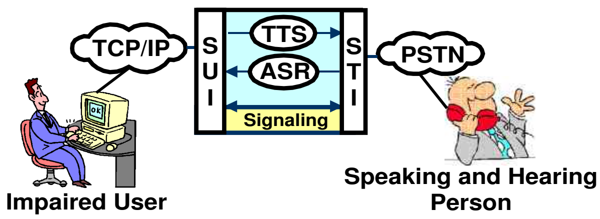

Figure 1: *Overview of system architecture design.*

The project is funded by APT which has been carried out in collaboration between R&D Division PT Telekomunikasi Indonesia (TELKOMRisTI), ATR Spoken Language Translation Laboratories Japan (ATR), Bandung Institute of Technology (ITB), Indonesia University of Education (UPI). The long-term goal is to establish a telecommunication system for hearing and speaking impaired people, in order to give them an opportunity to communicate with others via telephone. The main function of the system is to translate speech messages to the corresponding text and vice versa, using both speech recognition (ASR) and text-to-speech (TTS) technologies respectively. Here, TELKOMRisTI serves as project coordinator. ATR serves as a supervisor, as well as providing an Indonesian speech recognition system. ITB has also joined to provide an Indonesian Text-to-Speech system and speech corpus collection. Analysis of the social aspects of impaired people is being conducted by UPI.

In Indonesia, the sensorineural hearing impairment is a major problem because it affects almost 4.85% of the population or about 10 million cases [4]. Modern styles in big cities have changed the strong communal style life to a relatively individualistic one. Telephone communication has become important. But today, facilities to help people with disabilities are rare in Indonesia. Therefore it is a great start to provide such technologies.

An overview of the system architecture design can be seen in figure 1. It consists of four parts, namely the interface part, the signaling part, the TTS part and the ASR part. The interface part consists of the Speech User Interface (SUI) part and the Speech Telephony Interface (STI) part. These interfaces are dealing with the end user. As stated by its corresponding name, SUI is dealing with user using text messaging client utility while the STI is dealing with the phone user. Any communication between both interface is conveyed by signalling part, TTS part or ASR part depending on the type of communication going on between text messaging client user and phone user:

1. When a text messaging client user is contacting phone user or vice versa, the information of signalling communication is conveyed by the bridge.

2. When a text message is going from text messaging client to the phone user, the TTS part will take over and translating the text to the corresponding voice and send it to the phone user.

3. When a speech going from phone user to the text message client user, the ASR part will take over and translating the speech to the corresponding text message and send it to the text message client.

In one conversation scenario, the speaking and hearing party uses a normal phone set and the impaired user uses a text messaging client terminal. A connection request can be made by either party. If the phone user makes a call request, public switched telephone network (PSTN) will route the call through the Voice IP network to the application system. Then the signaling part will convey the request to the destination party. The text massaging client user can accept or reject the call by pressing a button provided in the application. Once communication channel is established, all speech messages received by STI will be sent to ASR. Then the text message results will be sent to the text-message terminal by SUI through TCP/IP. This is also done in reverse. More detailed information about this system and Indonesian TTS can be found in [5] and [6], respectively.

3. Characteristic of Indonesian Language

The Indonesian language, so-called Bahasa Indonesia, is a unity language formed from hundreds of languages spoken in the Indonesian archipelago. It was coined by Indonesian nationalists in 1928 and became a symbol of national identity during the struggle for independence in 1945.

Compared to other languages, which have a high density of native speakers, Indonesian is spoken as a mother tongue by only 7% of the population, and more than 195 million people speak it as a second language with varying degrees of proficiency. Approximately, there are 300 ethnic groups living in 17,508 islands, speaking 365 native languages or no less than 669 dialects [7]. At home, people speak their own language, such as Javanese, Sundanese or Balinese, though almost everybody has a good understanding of Indonesian as they learn it in school.

Although the Indonesian language is infused with highly distinctive accents from different ethnic languages, there are many similarities in patterns across the archipelago. Modern Indonesian is derived from the literary of the Malay dialect, which was the lingua franca of Southeast Asia. Thus, it is closely related to Malay spoken in Malaysia, Singapore, Brunei, and some other areas. Concerning the number of speakers, today Malay-Indonesian ranks around sixth in size among the world's languages.

The only difference is that Indonesia (which was a Dutch colony) adopted the Van Ophuysen orthography in 1901, while Malaysia (which was a British colony) adopted the Wilkinson orthography in 1904. In 1972, the governments of Indonesia and Malaysia agreed to standardize the ”improved” spelling, which is now in effect on both sides. Even so, modern Indonesian and modern Malaysian are as different from one another as are Flemish and Dutch [7].

The standard Indonesian language is continuously being developed and transformed to make it more suitable to the diverse needs of a modernizing society. Many words in the vocabulary reflect the historical influence of various foreign cultures that have passed through the archipelago. It has borrowed heavily from Indian Sanskrit, Chinese, Arabic, Portuguese, Dutch, and English.

It has used many writing systems over the years as well [8]. Alphabets from the south of India, very similar to the alphabets used in Thailand or Cambodia today, were used for many centuries to write the native languages of Indonesia. The *huruf jawa* are still used in Central Java for traditional and ceremonial purposes. When Islam arrived in Indonesia, much scholarship was done in the Arabic language, and speakers of Malay and other languages began to write their native words using Arabic script. The alphabet had to be expanded to include sounds such as ”p” which do not occur in classical Arabic, but since such letters had already been invented in Iran and India, this was not a problem. Use of the Arabic alphabet for Malay and Indonesian languages continued well into the colonial period. Arabic script used to write Malay or Indonesian is sometimes called ”Jawi” script. During the period of Dutch colonialism, Indonesian languages began to be written in the Roman alphabet. Until now, modern Indonesian uses Roman script with only 26 letters as in the English/Dutch alphabet [9].

Unlike Chinese language, it is not a tonal language. Compared with European languages, Indonesian has a strikingly small use of grammatically gendered words; the same word is used for he and she or for his and her [10]. Most of the words that refer to people (family terms, professions, etc.) have a form that does not distinguish between the sexes; for example, ”adik” can both refer to a (younger) brother or sister; no distinction is made between girlfriend and boyfriend. In order to specify gender, an adjective has to be added: ”adik laki-laki” corresponds to brother but really means younger male sibling. There is no word like the English man that can refer both to a male person and to a human being in general.

Plurals are expressed by means of reduplication, but only when not implied by the context; thus, ”orang-orang” is people, but one thousand people is ”seribu orang”, as the numeral makes it unnecessary to mark the plural form [10].

The basic word order is S-V-O. Verbs are not inflected for person or number, and there are no tenses; tense is denoted by time adverbs (such as yesterday) or by other tense indicators, such as ”sudah”, meaning already.

It is also a member of the agglutinative language family, meaning that it has a complex range of prefixes and suffixes which are attached to base words [10]. So a word can become very long. For example, there is a base word ”hasil” which means ”result”. But it can be extended as far as ”ketidakber***hasil***annya”, which means his/her failure.

4. Indonesian Speech Database Corpus

4.1. Database Design

The Indonesian speech corpus designed for the project consists of the following three sets:

1. **Digit task (C1)**. This is an adaptation of the official AURORA2 digit task [12], which consists of connected digit tasks among digit words such telephone number, credit card number, etc. Indonesian digit words is shown in Table 1.

Table 1: *Indonesian Digit Words*

Number	English	Indonesian
0	*Oh / Zero*	**Nol / Kosong**
1	*One*	**Satu**
2	*Two*	**Dua**
3	*Three*	**Tiga**
4	*Four*	**Empat**
5	*Five*	**Lima**
6	*Six*	**Enam**
7	*Seven*	**Tujuh**
8	*Eight*	**Delapan**
9	*Nine*	**Sembilan**

2. **Simple dialog task (C2)**. This is based on a word vocabulary which is derived from some necessary dialog calls for impaired users, such as dialog calls with the 119 emergency department, 108 telephone information department, and ticket reservation department. One of the dialog scenario examples is shown in Table 2. The speech message from 119 emergency department will be taken over by ASR while the text message from impaired user will be taken over by TTS. Thus, only the sentences uttered by emergency department staff are collected for speech corpus.

3. **Large vocabulary phonetic-balanced task (C3)**. This consists of phonetically balanced sentences collected from articles in magazines, journals, and daily news.

4.2. Speaker Criteria

The project is initially expected to use at least 200 speakers. Both genders are distributed evenly. The age is limited to middle age (20-40 years), since most people within this age have a strong communal of individual styles life. Regarding the highly distinctive accents described in Section 2, the speakers should present a wide range of spoken dialects from different ethnics groups.

4.3. Recording Set-Up

The recording system is set-up in ITB, Bandung, Java Island. The system configuration is presented in figure 2. It is conducted in parallel for both clean and telephone speech, recorded in 16kHz and 8kHz sampling frequency, respectively. The original 16kHz clean speech is then down-sampled to 8kHz.

Table 2: *Dialog scenario example.*

Impaired User (TTS)	Emergency Department (ASR)
Halo ! (Hello !)	**119, Selamat Malam. Ada yang bisa dibantu ?** *(119, Good Evening. May I help you ?)*
Tolong, saya mendapat kecelakaan. Saya terjatuh dari tangga ! *(Help, I've got an accident. I fell down from the stairs !)*	**Dimana alamat anda?** *(Where is your address ?)*
Jalan Gegerkalong 47 *(47 GegerKalong Street)*	**Baik, kami akan kirim bantuan segera** *(OK, We will send you our immediate assistance)*
Terima Kasih (Thank You)	

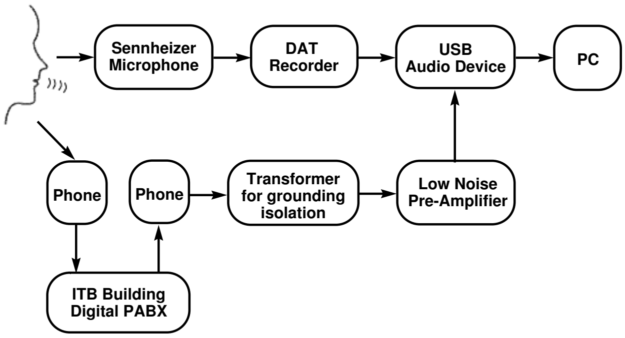

Figure 2: *Recording set-up.*

4.4. Status of Recordings

For this initial phase, we has successfully finished collecting C1 and C2. C3 is not covered yet. As it is close to the official AURORA2 digit task [12], C1 (clean) consists of 8440 training utterances (spoken by 55 Females, 55 Males), and 4004 testing utterances (spoken by 52 Females, 52 Males), which are equally split into four subsets (1001 utterances in each). These training and testing sets consist of about 8 and 4 hours of speech, respectively. C2 (clean) consists of 20,000 utterances (about 18 hours of speech) from the 70-word dialog vocabulary of 100 sentences (including single word sentences) each uttered by 200 speakers (100 Females, 100 Males). These utterances are equally split into training and test sets with 100 speakers (50 Females, 50 Males) in each set. As the recording is conducted in parallel for clean and telephone conditions, both should have the same number of total utterances. However, at the beginning of the recording process, we faced some technical problems to build this parallelism, so that only 70% are successfully recorded for telephone speech.

In order to collect all of the possible languages and dialects of the tribes recognized in Indonesia, the project will require a lot of time, money and resources. In this short phase, we focused only on the ethnic languages in the island for which the population are greatest. Even so, it is still difficult to get a sufficient number of speakers who originally came from non-Java ethnic groups while recording in Bandung (West-Java). Table 3 shows the percentage of population in each island according to a 2000 Census (%A) and the percentage distribution of speakers in the corpus (%B). Modern Indonesian is successfully covered by speakers from Jakarta city. Here, we also include ethnic Tionghoa (Chinese), since there are an estimated 8 million ethnic Tionghoa, including some families who have lived there for centuries.

Table 3: *The percentage of population according to 2000 Census (%A) vs the percentage distribution of speakers in the corpus (%B).*

Island	%A	%B	Native Languages
Java	60%	67%	Sundanese, Javanese, Madurese, Modern Indonesian
Sumatra	21%	21%	Acehnese, Lampung Batak, Minang, Original Malays
Sulawesi	7%	5%	Makassar, Minahasa, Bugis, Gorontalo
Kalimantan	5 %	2%	Banjar
Others	7%	5%	Balinese, Ambonese, Tionghoa

To gain a variety of dialect accents, we asked each speaker to speak naturally without any pronunciation restriction. Consequently, there are some mispronunciations due to their native tongue. For example, ”Nol” is often spoken as ”*e*Nol” by some Javanese speakers, ”D*e*lapan” (with ”e” as in ”Vow*e*l”) is often spoken as ”D*e*lapan” (with ”e” as in ”B*e*d”) by some Batak speakers, ”Tuju*h*” is often spoken as ”Tuju” and ”Saya” is often spoken as ”Saya*h*” by some Sundanese speakers.

5. Initial Segmentation Issues

In speech recognition system, segmented utterances according to labels are usually used as a starting point for training speech models. The automatic segmentation is mostly used since it is efficient and less time consuming. It is basically produced by forced alignment given the transcription. In this case, we required a word-based

Indonesian speech recognizer which is not yet available. One way to solve this problem is to segment the utterance uniformly, the so-called uniform initial models [11]. Here, we assumed that there are silences at the beginning and end of each sentence, but there is no silence that precede or follow any word within the sentence. Based on the above assumptions, the training set is segmented and the waveform duration is divided equally with the number of words (including silences).

Another solution is to do the forced alignment method using an existing speech recognizer from another language, such as an English speech recognizer. Since our available English speech recognizer is phoneme-based, we need to employ a mapping technique between Indonesian words to English phonemes. The pronunciation lexicon used here describes the pronunciation of an Indonesian word in terms of associated English phoneme symbols. Most of the mapping between Indonesian letters to phoneme symbols is basically one-to-one. Then, finding the similar pronunciation between Indonesian and English phoneme, we could also get a simplified one-to-one mapping between Indonesian words to English phoneme symbols.

6. Language Modeling Issues

Another problem is that there is no Indonesian text corpus available to train the language modeling (LM). In this case, we used only no-context LM where all unigram have the same probability, or in other word we can say that there is no LM. Thus, our results strongly depend on the acoustic modeling performance. Note that, for real implementation purpose, finite state automaton (FSA) is used.

7. Experimental Results and Discussion

The experiments were conducted using an ATR speech recognition engine. The setup for both C1 and C2 closely follows the official AURORA2 task evaluation, which based on whole word hidden markov models (HMMs) [12]. The front-end parameters are kept the same with a sampling frequency of 8kHz, a frame length of 25ms, a frame shift of 10ms, and 39 dimensional, included 12-order mel-frequency cepstral coefficients (MFCC) and log power, plus Δ and $\Delta\Delta$ features. 16 states per word with 10 mixture Gaussian per state were used for acoustic model. Any artificial noises are not added here. Of primary interest for us was to gain good results for both clean and telephone speech.

For C1, we began with the uniform segmentation. Clean and telephone speech were trained and tested separately. As described in Section 4.4, the test set utterances are equally split into four subsets. The average results from all test sets are summarized in Table 4. In this simple task, we only gained about 98% performance in average. Some substitution errors happened between the word ”Nol” and ”Enam”, due to strong dialect accents by Javanese speakers, who often said ”Nol” as ”*e*Nol”.

Table 4: *% Word accuracy results of C1 digit task.*

Train Condition	Test Condition	Uniform Segment
Clean	**Clean**	98.83
Telephone	**Telephone**	97.57
Clean+Telephone	**Clean**	98.64
Clean+Telephone	**Telephone**	97.73

Table 5: *% Word accuracy results of C2 dialog task.*

Train Condition	Test Condition	Uniform Segment	English Segment
Clean	**Clean**	52.06	94.74
Telephone	**Telephone**	75.21	96.35
Clean+Telephone	**Clean**	-	92.10
Clean+Telephone	**Telephone**	-	91.36

For C2, we did the same thing as in C1. Unfortunately, the performance with uniform start segmentation is very poor. Especially in the clean condition, we only gained a 52.06% word accuracy (see Table 5). This might be caused by the wider variety of word length in the dialog task (C2). For example, in one sentence there are the word ”ke” (*to*) which only consists of one syllable, and the word ”rencananya” (*his/her plan*) which consists of four syllables. Repeating the process could only rise the performance about 0.3%-0.5% in each iteration. To speed up the process, we need to find another way that could give a good alignment to the acoustic modeling. Therefore, we tried the second method as described in Section 5. We used an English speech recognizer to set initial segmentation of Indonesian utterances.

Our available English speech recognizer was triphone-based, trained using the Wall Street Journal (WSJ) with a sampling frequency of 16 kHz, a frame length of 20ms, and a frame shift of 10ms. 25 dimensional (12-order MFCC, Δ MFCC and log power) was used as feature parameters. Three states were used as the initial model for each phoneme. Then, they were trained using successive state splitting (SSS) algorithm based on minimum description length (MDL) criterion in order to gain the optimal number of states. Details about MDL-SSS can be found in [13]. To minimize the mismatch, we used it to segment the original 16kHz clean speech utterances. Using this time alignment results, we then trained the same way as before. Although not all Indonesian utterances could be successfully transcribed by the English recognizer, the alignment information contained in it is still better than that of the uniform start method. This is proven by its significant improvement up to 40% absolute performance from 52.06% to 94.74% word accuracy.

Most substitutions occurred between similar words. This similar word phenomenon is produced by agglutination rules, for example, in the word ”*bantu*” (help) and ”di*bantu*” (was helped), or word ”*tiket*” (ticket) and ”*tiket*nya” (his/her ticket). There are also some insertions caused by the grammar flexibility of word-order. For example, the sentence ”*Dimana* alamat anda?” (Where is your address?) can also be written as ”Alamat anda *dimana?*”. As a consequence, the recognizer often recognized this as ”*Dimana* alamat anda *dimana?*”.

Here, we also tried a multi condition where both clean and telephone speech segmented data were combined and a single large multi-condition acoustic model was trained. In this case, we were able to gain good results, more than 91% for C2 and 97% for C1 in both clean and telephone conditions.

8. Conclusion

We have presented the development of an Indonesian speech corpus and word-based speech recognition system. The recognition results show that automatic segmentation by an English speech recognizer was able to produce better alignment than just uniform segmentation. Most errors were caused by mispronunciation, agglutination words, and word-order grammar. The speech corpus has covered a wide range of different ethnic dialects, but the percentage of ethnic dialects from East-Indonesia is still minor.

A possible solution for this problem would be to extend the dialect coverage, guide the speakers to correctly pronounce Indonesian words, and advance lexicon and also the LM. These aspects need to be considered for developing C3 speech corpus and Indonesian large vocabulary continuous speech recognition (LVSCR) system in next phase.

9. Acknowledgement

Part of this speech recognition research work was supported by the National Institute of Information and Communication Technology (NICT), Japan.

10. References

[1] Schultz, T, et al., “The Global Phone Project: Multilingual LVCSR with JANUS-3 ”, Proc. 2nd SQEL Workshop, 1997.

[2] Wong, Eddie, et al., “Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilising Pronunciation Modelling Techniques”, EUROSPEECH, Geneva, 2003.

[3] Martin, T., et al., “Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition”, EUROSPEECH, Geneva, 2003.

[4] Sedjawidada, R., et al., “Efforts on Helping the Hearing Impaired in Indonesia: Makassar Case”, HI Newsletter, Series No. 44, P 9-10, 2003.

[5] Kelana, Eka, et al., “Development of Telcommunication System for Dumb and Deaf People”, AIC, Kuala Lumpur, 2004.

[6] Arman, Arry Akhmad., “Indonesian Text to Speech”, http://lss.ee.itb.ac.id/ aa/indotts/.

[7] Johannes Tan, “Bahasa Indonesia: Between FAQs and Facts”, http://www.indotransnet.com/article1.html.

[8] Gimon, Charles A., “Sejarah Indonesia”, http://www.gimonca.com/sejarah/pronounce.html.

[9] Quinn, George, “The Indonesan Language”, http://www.hawaii.edu/sealit/Downloads/.

[10] Wikipedia, “Indonesan Language”, http://en.wikipedia.org/wiki/Bahasa_Indonesia.

[11] Rabiner, Lawrence, at al., “Fundamentals of Speech Recognition”, Prentice Hall, New Jersay, USA, 1993.

[12] Hirsch, H.G., at al., “The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noisy Conditions”, ISCA ITRW ASR2000, Paris, 2000.

[13] Jitsuhiro, T., et al., “Automatic Generation of Non-Uniform CD-HMM Topologies based on the MDL Criterion”, EUROSPEECH, Geneva, 2003.