

The Use of Indonesian Speech Corpora for Developing a Filipino Continuous Speech Recognition System

Sakriani Sakti, Ryosuke Isotani, Hisashi Kawai, Satoshi Nakamura
NICT Spoken Language Communication Research Group
3-5 Hikaridai, "Keihanna Science City", Kyoto 619-0289, Japan

{sakriani.sakti, ryosuke.isotani, hisashi.kawai, satoshi.nakamura}@nict.go.jp

Abstract

The development of an automatic speech recognition system for a new language requires collection of a huge amount of speech data, as well as manual annotation and transcription. That is why the feasibility of cross-language transfer of speech technology has become a matter of increasing concern as the demand for recognition systems in multiple languages grows. This paper shows the possibility of developing a Filipino continuous speech recognition system by using Indonesian speech data. It is based on the cross-language approach with following procedure: (1) Normalize the phoneme sets of Indonesian and Filipino, in order to have the same phonetic transcription convention across Indonesian and Filipino language. (2) Train the Indonesian speech corpora with normalized phoneme set and apply it as an initial acoustic model of the Filipino language. (3) Use the initial Filipino acoustic model to segment the limited utterances of Filipino training data by the Viterbi alignment algorithm. (4) Retrain and adapt the parameters of the initial acoustic model using the Filipino training data. Experimental results reveal that even with the initial acoustic model (trained on pure Indonesian speech data), the system could recognize Filipino continuous speech up to 79.50% word accuracy.

1. Introduction

The development of an automatic speech recognition (ASR) system for a new language requires collection of a huge amount of speech data, as well as manual annotation and transcription. However, such a procedure is often difficult, especially because of time and budget constraints. In recent years, the feasibility of cross-language transfer of speech technology has become a matter of increasing concern as the demand for recognition systems in multiple languages grows [1]. The cross-language technique is per-

formed from a source language that has a large amount of data to a target language that has only a few data or even none at all. Many researchers have shown that the cross-language approach is useful for rapid development of a new language ASR system [1, 2, 3, 4].

The Philippines is an archipelago of more than 7,100 islands. With an estimated population of about 92 million people, the Philippines is the world's 12th most populous country. There are more than 100 indigenous language in the Philippines, with Filipino as the national language. Collecting a speech corpus which can cover all possible languages and dialects of the tribes recognized in the Philippines, therefore, is still the biggest problem. Recently, a Filipino speech corpus was successfully developed at the University of the Philippines Diliman [5]. Some researchers in the Philippines has also attempted to build ASR, not only for phoneme recognition [6], but also for continuous word recognition [7]. However, it is reported that the continuous speech recognition could only achieved a 32% accuracy. This maybe because the Filipino speech corpus only includes less than six hours of continuous speech which may not be enough to train a proper acoustic model.

This paper examines the development of a Filipino continuous speech recognition system using other resource languages based on cross-lingual approach. In this study, we use Indonesian as the source language, which currently already has an 80 hours of Indonesian large-vocabulary speech corpora [8]. First, the phoneme sets of Indonesian and Filipino are normalized, in order to have the same phoneme set and the same phonetic transcription convention across Indonesian and Filipino language. Then, we train the Indonesian speech corpora with normalized phoneme set and apply it as an initial acoustic model of the Filipino language. After that, the initial Filipino acoustic model is used to segment the limited utterances of Filipino training data by the Viterbi alignment algorithm. Last, we retrain and adapt the parameters of the initial acoustic model using the Filipino training data.

In the next section, we briefly describe the language

characteristics and phonological system of both Filipino and Indonesian languages. Then, in Section 3, we describe the data resources of Indonesian and Filipino languages that were used in this study. The development of Filipino continuous speech recognition with cross-language approach is described in Section 4. Finally, we draw our conclusions in section 5.

2. Language Characteristics

Filipino is the national language of the Philippines and it is based primarily on Tagalog that belongs to the same family (Malayo-Polynesian branch of the Austronesian language family) as Malay and Indonesian language. So basically, the Filipino language has very strong affinity with Malay and Indonesian languages. But, due to more than 300 years of Spanish colonial rule over the Philippines, the language has incorporated a significant number of Spanish words and expressions. The language also includes words and phrases that are rooted in English and Chinese. However, there still exists many similar words within these languages which are still commonly used in everyday conversation. Table 1 shows some examples of these similar words.

Table 1. Similar words within Indonesian and Filipino languages

Indonesian	Filipino	Meaning
Aku	Ako	<i>I (first person)</i>
Angin	Hangin	<i>wind</i>
Balik	Balik	<i>return</i>
Bangsa	Bansa	<i>country</i>
Itik	Itik	<i>duck</i>
Langit	Langit	<i>sky/heaven</i>
Lelaki/Laki-laki	Lalaki	<i>male</i>
Murah	Mura	<i>cheap</i>
Sakit	Sakit	<i>ill</i>
Saksi	Saksi	<i>witness</i>
Surat	Sulat	<i>letter</i>
Tahun	Taon	<i>year</i>

Although Filipino has been influenced phonologically by many other languages, its grammar construction has remained unchanged and still reflects the Malayan structure, even though it may appear similar to the grammar and syntax of Spanish and English [9]. Similar to Indonesian language, modern Filipino is phonetic based and written in Roman script, which uses 26 letters as the English/Dutch alphabet [10]. But, the seven letters *c, f, j, q, v, x* and *z* are used chiefly in proper names of foreign origin and in certain other borrowings from English or Spanish. Because the original

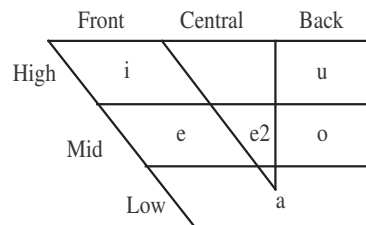
Filipino letters are based on the classic Tagalog alphabet, so called “Abakada” alphabet, which consists of only 20 letters (including digraph /*ng*/) as follows:

a b k d e g h i l m n n g o p r s t u w y

All letters are pronounced much more consistently, and no letters are muted. Some exception exist in Filipino language, such common particle “*nang*” and “*mangah*” are often written with only “*ng*” and “*mga*”, respectively. However, other than these and a few other exceptions, there is fairly good match between spelling and pronunciation.

Following Tagalog structure describe in [11], Filipino language has 16 phoneme consonants, 5 phoneme vowels and 6 diphthongs (/iw/, /ay/, /aw/, /oy/, /ey/, and /uy/). While, the full phoneme set in Indonesian language, as defined in an Indonesian grammar text [12], contains 22 phoneme consonants, 6 phoneme vowels, and 4 diphthongs (/ay/, /aw/, /oy/, and /ey/). Most Filipino phonemes could be covered in Indonesian phonological system. Only glottal stop /ʔ/ and two diphthongs /iw/ and /uy/ which do not explicitly listed in Indonesian phoneme set. The articulatory pattern difference of between Filipino and Indonesian phoneme consonants are given in Table 2, and Fig. 1 illustrates the vowel articulation pattern.

a. Indonesian language



b. Filipino language

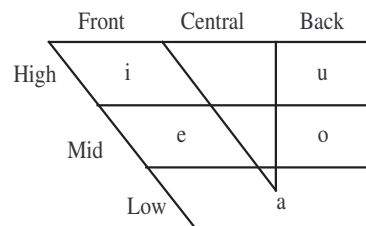


Figure 1. Articulatory pattern of both Filipino and Indonesian vowels.

Table 2. *Articulatory Pattern of Indonesian (left side) and Filipino (right side) Consonants*

	Bilabial	Labiodental	Dental/Alveolar	Palatal	Velar	Glottal
Plosives	b b p p		d d t t		g g k k	?
Affricates				c j		
Fricatives		f	s s z	sy	kh	h h
Nasal	m m		n n	ny	ng ng	
Trill			r r			
Lateral			l l			
Semivowel	w w			y y		

3. Data Resources

3.1. Indonesian Speech Corpora

The Indonesian speech corpora were developed by the R&D Division of PT Telekomunikasi Indonesia (R&D TELKOM) in collaboration with ATR as a continuation of the APT (Asia Pacific Telecommunity) project [8, 13]. These corpora are described below.

1. Daily news task

A raw text source for the daily news task has already been generated by an Indonesian student [14]. The source was compiled from “KOMPAS” and “TEMPO,” which are currently the biggest and most widely read Indonesian newspaper and magazine, respectively. This source consists of more than 3160 articles, with around 600,000 sentences. R&D TELKOM further processed the raw text source to generate a clean text corpus.

From this text data, we then selected phonetically-balanced sentences by using the greedy search algorithm [15]; this produced a total of 3168 sentences. Then, clean and telephone speech were recorded, simultaneously, at sampling frequencies of 16 and 8 kHz, respectively, by R&D TELKOM in Bandung, Java Island, Indonesia. There were a total of 400 speakers (200 males and 200 females). Four main accents were covered: Batak, Java, Sunda, and standard Indonesian (without accent). Each speaker uttered 110 sentences, resulting in a total of 44,000 speech utterances, which amounted to around 43.35 hours of speech.

2. Telephone application task

A total of 2500 sentences related to the telephone application domain were generated by R&D TELKOM.

These were derived from the dialog that is commonly employed in the telephone services, including tele-home security, billing information services, reservation services, status tracking of e-Government services and hearing impaired telecommunication services (HITS).

Using the same recording set-up as the one used for the news task corpus, the speech utterances for 2500 sentences pertaining to telephone application tasks were recorded by R&D TELKOM in Bandung, Indonesia; the total number of speakers and their distribution in terms of age and accents were also identical. Each speaker uttered 100 sentences, resulting in a total of 40,000 utterances (36.15 hours of speech).

3.2. Filipino Speech Corpora

The development of Filipino speech corpora is a joint effort between the Linguistics Department and the Electrical and Electronics Engineering Department at the University of the Philippines, Diliman, Philippines [5].

The full read text consists of paragraphs, short sentences, isolated words, syllables and phonemes. These materials was formulated based on the recommendation of the Linguistics Department to elicit the phones and prosodic cues that characterize Filipino speech. However, only paragraphs and short sentences were used in this study, which are described as follows:

1. Paragraph type

Text material for the paragraphs contains five different themes: (1) introducing oneself; (2) relating an emotional event; (3) story-telling; (4) giving directions/advice; and (5) describing a scene. Each paragraph consists of 31 to 109 words.

2. Sentence type

Text materials for the short sentences contains: (1) five questions; (2) three commands; (3) two requests; and (4) two greetings/exclamation. Each sentence consists of 1 to 7 words.

The clean speech were recorded at sampling frequencies of 44.1 kHz, and later on were downsampled to 16 kHz. The data is stored as a mono, 16 bit, wav file. There are 100 speakers (50 male and 50 female) in total with ages 16 and older. The speakers are undergraduate students, lecturers or faculty members at the University of the Philippines, Dili-man. The speakers come from all over the Philippines and may have Tagalog as their first, second or third language.

In this study, all long utterances have been further cut at longer pauses. As a results, we got approximately 6k short utterances (about 6 hours of speech). The last 4 speakers (416 utterances; about 800 vocabulary size) were allocated to the test set, while the remaining data (5320 utterances; about 2k vocabulary size) were used as the training or adaptation set.

4. Development of Filipino ASR

4.1. Parameter Set-up

The experiments were conducted using the following feature extraction parameters: sampling frequency of 16 kHz, frame length of a 20-ms Hamming window, frame shift of 10 ms, and 25 dimensional MFCC features (12-order MFCC, Δ MFCC, and Δ log power).

4.2. Phoneme Normalization

The initial step in this cross-language approach is the phoneme normalization. This is done for both Indonesian and Filipino phoneme sets, in order to have the same phoneme set and the same phonetic transcription convention across Indonesian and Filipino language. There are many ways to normalize the phoneme symbols across language, such knowledge-based or data-driven approaches [4, 16]. The most intuitive and straightforward approach to generate a phoneme symbols across language based on linguistic knowledge, since they are independent of the bias of recoding properties that may exists between databases [17].

Here, we use the articulatory pattern of both Indonesian and Filipino phoneme consonants and vowels in order to find the evidence of acoustic-phonetic similarities between Indonesian and Filipino languages. As described in Section 2, most Filipino phonemes could be covered in Indonesian

phonological system. Only glottal stop /ʔ/ and two diphthongs /iw/ and /uy/ which do not explicitly listed in Indonesian phoneme set. Thus, it is possible to use Indonesian phoneme set across both language, with following consideration:

- Glottal stop /ʔ/ is treated as allophone /k/.
- Diphthong /iw/ is approximated by combination of two Indonesian phonemes (/i+/u/).
- Diphthong /uy/ is also approximated by combination of two Indonesian phonemes (/u+/i/).

The phonetic transcription of Indonesian speech corpora is derived based on the existing Indonesian pronunciation dictionary [8]. The dictionary used here is part of the large-vocabulary dictionary that owned by R&D TELKOM, which was manually developed by Indonesian linguists. On the other hand, there is no available pronunciation dictionary for Filipino language. However, since there is fairly good match between spelling and pronunciation, we simply develop Filipino pronunciation dictionary by one-to-one mapping from graphemes to phonemes. Some exceptions such common particle “*ng*” and “*mga*” are modified manually.

4.3. Initial Filipino Acoustic Model with Indonesian Speech

Using the normalized phonetic transcription, we trained the initial Filipino acoustic model on 80 hours of Indonesian continuous speech. A hidden Markov model (HMM) is typically employed to represent the acoustic model. Three states were used as the initial HMM for each phoneme. A shared state HMnet topology was then obtained using a successive state splitting (SSS) training algorithm based on the minimum description length (MDL) optimization criterion [18]. The resulting context-dependent triphone had 1,256 states in total with 5, 10, 15 and 20 Gaussian mixture components per state.

Word bigram and trigram language models of Filipino language were trained by using the Filipino training set, yielding bigram and trigram perplexity of 15.5 and 6.1, respectively, with out-of-vocabulary (OOV) rate less than 1% on the Filipino test set. There are some overlapping texts between training and test set which might contribute the low perplexity of bigram and trigram language model. The recognition accuracy rates of this initial model on the Filipino test set are described in Figure 2. As can be seen, without any use of Filipino speech data, the initial ASR is still able to recognize Filipino continuous speech with more than 75% word accuracy. The best performance was 79.50% word accuracy given by the Filipino initial model with 20 Gaussian mixture components. The major errors

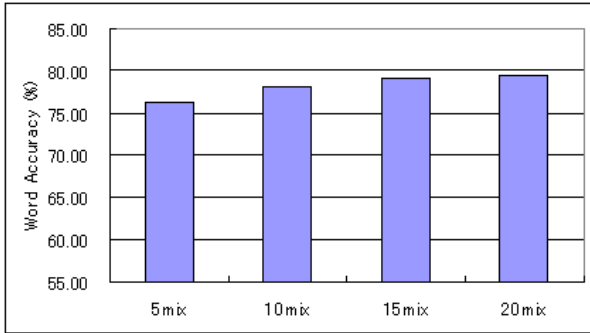


Figure 2. The recognition accuracy (%) of the initial model.

mainly caused by foreign words of English or Spanish. The reason maybe because the Filipino speakers pronounce those foreign words differently than the Indonesian speakers, and thus there exists mismatch pronunciation in the lexicon dictionary.

4.4. Filipino Speech Alignment

The next step is to segment the utterances of the Filipino training data using the initial Filipino acoustic model described in previous section. This is done automatically based on Viterbi alignment algorithm. It is produced by forced alignment given the transcription of Filipino training materials.

4.5. Filipino Acoustic Model Refinement

The last step is to retrain and adapt the parameters of the initial model to the Filipino training data which already segmented from the previous step. As for the adaptation, we use the maximum a posteriori (MAP)-based adaptation scheme, which is commonly used to compensate for either speaker or environmental variations in monolingual ASR systems [19], and also on cross-language adaptation [17, 16].

This scheme principally takes the advantages of prior information about existing models. A Bayesian learning mechanism then adjust the parameters of the initial acoustic model in such a way that the limited Filipino training data would modify the initial acoustic model parameters guided by the prior knowledge to compensate for the adverse effects of a mismatch [19]. Furthermore, the parameter re-estimation is a weighted sum of the prior knowledge and the new estimation of the target language.

The recognition accuracy rates of these adapted and retrained models on the Filipino test set are described in Fig-

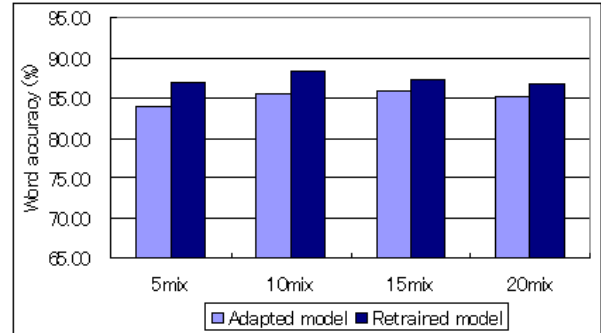


Figure 3. The recognition accuracy (%) of the retrained and adapted model.

ure 3. The results reveal that by adapting and retraining the initial model to the Filipino training corpora, the ASR performance could be further improved. Both the adapted and retrained Filipino acoustic models outperformed the initial models that was trained with only Indonesian speech data. The retrained model performed better than the adapted model, this may indicate that minimizing the speech characteristic difference between Indonesian and Filipino could not be enough by only re-estimating the mean vector of Gaussian mixtures. The best performance of the adapted and retrained Filipino acoustic models were 85.91% and 88.43% word accuracy, respectively. The best adapted model is performed with 15 Gaussian mixture components, while the best retrained model is performed with 10 Gaussian mixture components.

5. Conclusion

We have demonstrated the possibility of development of a Filipino continuous speech recognition based on the cross-language approach, where Indonesian is the source language and Filipino is the target language. We have attempted the cross-language approach with following procedure: First, the phoneme sets of Indonesian and Filipino are normalized, in order to have the same phoneme set and the same phonetic transcription convention across Indonesian and Filipino language. Then, we train the Indonesian speech corpora with normalized phoneme set and apply it as an initial acoustic model of the Filipino language. After that, the initial Filipino acoustic model is used to segment the utterances of Filipino language training data by the Viterbi alignment algorithm. Last, we retrain and adapt the parameters of the initial acoustic model using the Filipino language training data. Experimental results reveal that even with the initial acoustic model (trained on pure Indonesian speech data), the system could recognize Filipino

continuous speech up to 79.50% word accuracy. By adapting and retraining the initial model to the Filipino training corpora, the ASR performance could be further improved. Both the adapted and retrained Filipino acoustic models outperformed the initial models that was trained with only Indonesian speech data. The best performance of the adapted and retrained Filipino acoustic models were 85.91% and 88.43% word accuracy, respectively. The best adapted model is performed with 15 Gaussian mixture components, while the best retrained model is performed with 10 Gaussian mixture components.

6 Acknowledgements

The authors would like to thank the University of the Philippines (UP), Diliman Office of the Vice-Chancellor for Research and Development (OVCRD) and UP Digital Signal Processing Laboratory for providing us the Filipino Speech Corpus (FSC).

References

- [1] B. Wheatly, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *Proc. ICASSP*, Adelaide, Australia, 1994, pp. 237–240.
- [2] V. Bac Le and L. Besacier, "First steps in fast acoustic modeling for a new language: Application to vietnamese," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 821–824.
- [3] T. Martin and S. Sridharan, "Cross-language acoustic model refinement for the Indonesian language," in *Proc. ICASSP*, Philadelphia, USA, 2005, pp. 865–868.
- [4] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," in *Proc. EUROSPEECH*, Aalborg, Denmark, 2001, pp. 2721–2724.
- [5] R. Guevara, M. Co, E. Espina, I. Gracia, E. Tan, R. Ensomio, and R. Sagum, "Development of a Filipino speech corpus," in *Proc. National ECE Conference*, Philippines, 2002.
- [6] R. Guevara, "Filipino phoneme recognition trained on the Filipino speech corpus and TIMIT," in *Proc. National ECE Conference*, Philippines, 2003.
- [7] G. delaRoca, H. Go, and R. Ordonez, "Speaker-independent continuous speech recognition of the filipino speech corpus," in *Proc. National ECE Conference*, Philippines, 2003.
- [8] H. Riza S. Sakai K. Markov S. Nakamura S. Sakti, E. Kelana, "Recent progress in developing indonesian large-vocabulary corpora and LVCSR system," in *Proc. MALINDO*, Cyberjaya-Selangor, Malaysia, 2008, pp. 40–45.
- [9] S. P. Aspillera, *Basic Tagalog*, University of the Philippines, Philippines, 1956.
- [10] B. Comrie, *The World's Major Languages*, chapter Tagalog Phonology and Orthography, Oxford University Press, 1990.
- [11] T. V. Ramos, *Tagalog Structures*, University of Hawaii Press, USA, 1971.
- [12] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)*, Balai Pustaka, Jakarta, Indonesia, 2003.
- [13] S. Sakti, P. Hutagaol, A.A. Arman, and S. Nakamura, "Indonesian speech recognition for hearing- and speaking-impaired people," in *Proc. ICSLP*, Jeju Island, Korea, 2004, pp. 1037–1040.
- [14] F. Tala, *A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia*, Ph.D. thesis, The Information and Language System (ILPS) Group, Informatics Institute, University of Amsterdam, Amsterdam, Netherland, 2003.
- [15] J. Zhang and S. Nakamura, "An efficient algorithm to search for a minimum sentence set for collecting speech database," in *Proc. ICPHS*, Barcelona, Spain, 2003, pp. 3145–3148.
- [16] P. Fung and M. Chi Yuen, "MAP-based cross-language adaptation augmented by linguistic knowledge: From English to Chinese," in *Proc. EUROSPEECH*, Budapest, Hungary, 1999, pp. 871–874.
- [17] C. Nieuwondt and E.C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, pp. 101–113, 2002.
- [18] T. Jitsuhiro, T. Matsui, and S. Nakamura, "Automatic generation of non-uniform HMM topologies based on the MDL criterion," *IEICE Trans. Inf. & Syst.*, vol. E87-D, no. 8, pp. 2121–2129, 2004.
- [19] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall, New Jersey, USA, 2001.