Rapid Development of Initial Indonesian Phoneme-based Speech Recognition Using The Cross-Language Approach

Sakriani Sakti, Konstantin Markov, Satoshi Nakamura

ATR Spoken Language Communication Research Laboratories,
2-2-2 Hikaridai, Seika-cho, Souraku-gun, Kyoto 619-0288, Japan
Phone: +81 774 95 1345, Fax: +81 774 95 1308, URL: www.slt.atr.jp/slc-e/
{sakriani.sakti,konstantin.markov,satoshi.nakamura}@atr.jp

Abstract

This paper presents our systematic development of initial Indonesian phoneme-based speech recognition system using the cross-language approach, where English is the source language and Indonesian is the target language. The available Indonesian speech corpus includes only a small vocabulary covering about 70% of the full Indonesian phoneme-set. To attain a proper Indonesian acoustic model with a full phoneme set, we propose to use the acoustic information from the English language in three ways: (1) Substitute the phoneme label alignments of source language training data with the phoneme labels of the target language and train the model as a seed acoustic model of the target language. (2) Use the seed acoustic model to segment the utterances of target language training data by the Viterbi alignment algorithm, train the new acoustic model of the target language, and fill the missing phoneme model from the seed acoustic model. (3) Adapt the parameters of the seed acoustic model using the target language training data. Each method is explored, and the performance of each resulting acoustic model is also evaluated and compared.

Keywords: *Indonesian phoneme-based ASR, cross-language approach, cross-language substitution, insertion and adaptation.*

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 performed 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].

Indonesia in particular, as the fourth most populous nation in the world - inhabited by more than 200 million people - still lacks speech technology and research, and also suffers from a shortage of Indonesian language data. Difficulties arise in developing an Indonesian large vocabulary speech recognition (LVCSR) system since Indonesian is actually most people's second language after their own ethnic native language. Collecting a speech corpus which can cover all possible languages and dialects of the tribes recognized in Indonesia, therefore, is still the biggest problem. Recently, an Indonesian speech corpus covering several major ethnic dialects spoken in Indonesia was successfully collected, but it includes only a small vocabulary covering about 70% of the full Indonesian phoneme set [5, 6]. In order to apply an Indonesian ASR system in some domain application tasks, a proper acoustic model with a full phoneme set is needed and of course fast development is preferable.

In this study, we consider the rapid development of an initial Indonesian phoneme-based speech recognition system using the cross-language approach, where English is the source language and Indonesian is the target language. One way to achieve this is to substitute the phoneme label alignments of source language training data with the phoneme labels of the target language, train the model as a seed acoustic model of target language, and use it to recognize target language speech, which we called **cross-language substitution**. Another way is to segment the utterances of target language training data using the seed acoustic model based on the Viterbi alignment algorithm, and train a new model of the target language. Since the model does not include a full phoneme set, the missing phoneme models are inserted from the seed model. We refer to this method as **cross-language insertion**. A third way is to adapt the parameters of the seed acoustic model using the target language training data, a method we call **cross-language adaptation** [1, 7]. In this study, we explore each method, and also evaluate and compare the performance of each resulting acoustic model.

In the next section, we briefly describe the framework specification including speech corpora, phoneme set, and the ASR system. The cross-language acoustic modeling approaches with cross-language substitution, insertion, and adaptation, including the issues of English-to-Indonesian phoneme mapping, are described in Section 3, and the results from a comparison of the performances of all cross-language approaches is presented in Section 4. Finally, we draw our conclusions in Section 5.

2. Framework Specification

2.1. Speech Corpora

For English, we use the popular Wall Street Journal (WSJ0 and WSJ1) large-vocabulary speech corpus, which consists of 60 hours of native English speech data spoken by 284 speakers (females and males) [8]. A set of 44 phonemes, which is basically similar to the phoneme set defined by the CMU pronunciation dictionary [9], is used to represent this WSJ data.

The small-vocabulary Indonesian speech corpus used here was collected in a collaborative project between ATR Spoken Language Communication Research Laboratories (Japan), TELKOMRisTI (R&D Center, PT Telekomunikasi Indonesia), and the Bandung Institute of Technology (ITB), which is funded by Asia-Pacific Telecommunity (APT) [5, 6]. It consists of corpus set C1 for digit task and corpus set C2 for simple dialog task, conducted in parallel for clean and telephone conditions. In this study, we use only corpus set C2 (clean speech). It was originally derived from some necessary dialog calls for a telecommunication system of hearing- and speaking-impaired users, such as dialog calls with the 119 emergency department, 108 telephone information department, and a ticket reservation department. One of the dialog scenario examples is shown in Table 1. The speech messages from the 119 emergency department will be covered by ASR while the text messages from impaired user will be covered by TTS. Thus, only the sentences uttered by emergency department staff are collected for the speech corpus.

This corpus successfully covers about more than 15 major ethnic dialects spoken in Indonesia. It 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), and these utterances are equally split into training and test sets with 10,000 utterances and 100 speakers (50 females, 50 males) in each set. Then, to analyse the performance of the continuous speech recognition system, we removed the single-word utterances from the test set, resulting in about 4,000 utterances. The Indonesian phoneme set is defined based on Indonesian grammar described in [10]. A full phoneme set includes 33 phoneme symbols in total, but the C2 Indonesian corpus only covers 70% of the full set. Since Indonesian is not as popular as English, we will describe the Indonesian phonemes in more detail in the next section.

Table 1: *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?** *(What 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)	

2.2. Indonesian Phoneme Set

The Indonesian phoneme set contains of 10 vowels (including diphthongs), 22 consonants, and 1 silence symbol. The vowel articulation pattern of the Indonesian languages, which indicates the first two resonances of the vocal tract, F1 (height) and F2 (backness), is shown in Fig. 1.

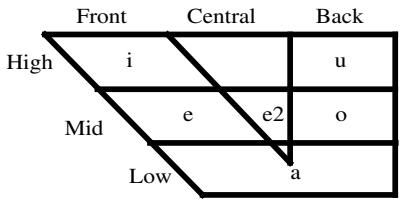

Figure 1: *Articulatory Pattern of Indonesian Vowels.*

It consists of the vowels: /a/ (like "a" in "father"), /i/ (like "ee" in "screen"), /u/ (like "oo" in "soon"), /e/ (like "e" in "bed"), /e2/ (a schwa sound, like "e" in "learn"), /o/ (like "o" in "boss"), and four diphthongs, /ay/, /aw/, /oy/ and /ey/. For the Indonesian consonants, the articulatory pattern can be seen in Table 2.

2.3. ASR System

The experiments were conducted using an ATR speech recognition engine. A sampling frequency of 16 kHz, a frame length of a 20-ms Hamming window, a frame shift

Table 2: *Articulatory Pattern of Indonesian Consonants*

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

of 10 ms, and 25 dimensional feature parameters consisting of 12-order MFCC, Δ MFCC, and Δ log power are used as feature parameters. Three states context-independent HMM acoustic models were used for each phoneme, and two different versions of Gaussian mixture components per state, 5 and 15 were applied. Since only a 70-word dialog vocabulary is used here, unigram language modeling (LM) is applied.

3. Cross-Language Approach

3.1. Cross-Language Substitution

The initial step of cross-language substitution is the phoneme mapping between the English source language into the Indonesian target language. There are many ways to map the phoneme symbols across language, such knowledge-based or data-driven approaches [4, 7]. The most intuitive and straightforward approach to generate a phoneme mapping table is to use knowledge (linguistic) -based phonetic mappings, since they are independent of the bias of recoding properties that may exists between databases [11]. In this case, we use International Phonetic Alphabet (IPA) definition to find evidence of acoustic-phonetic similarities between English and Indonesian. The procedure is performed as follows.

- Convert all English and Indonesian phonemes into IPA symbols.
- For each Indonesian phoneme, find a representative English phoneme which has the same IPA symbol or the closest possible match.
- If necessary, make an approximation of Indonesian phonemes by combining several English phonemes.

Table 3 shows an example of phoneme mapping from the 44-phoneme set of the English source language into the 33-phoneme set of the Indonesian target language. However, this mapping solution might be sub-optimum for the following reasons. First, there are still differences with respect to the acoustic properties of sounds from both languages that share the same labels. For example, the Indonesian /r/ is trill like in Spanish, whereas the English /r/ is liquids. Second, there are also some Indonesian phoneme sounds that do not occur in the English phoneme set inventory. For example, Indonesian has a consonant nasal palatal /ny/ which is similar to “ny” in the English word “canyon”. Since, our English phoneme set does not have a single phoneme symbol for /ny/, we construct it from two English phonemes /n/ and /y/. Another example is that Indonesian has only a single phoneme to represent vowel /i/, while English has more variants for the “i” sounds. In this case, all English variants of “i” are mapped into a single Indonesian phoneme /i/. In the case of “t” sounds, Indonesian has only a single consonant plosive /t/, while English has a consonant plosive /t/ and fricative /th/. Here, we attempted two different mappings. In type 1, we map all English phoneme “t” sounds into Indonesian phoneme /t/ regardless of whether “t” is the plosive /t/ or the fricative /th/. In type 2, only the English consonant plosive /t/ is mapped to the Indonesian consonant plosive /t/. The English consonant fricative /th/ is represented as a combination of two Indonesian /t/ and /h/ phonemes. “D” and “z” sounds are treated similarly to “t” sounds.

Table 3: *English-to-Indonesian Phoneme Mapping*

IND	ENG	IND	ENG	IND	ENG
a	aa	h	hh	oy	oy
ay	ay	i	ih,iy,ix	p	p
aw	aw	j	jh	r	r
b	b	k	k	s	s
c	ch	kh	k+h	sy	sh
d	d,dx,dh	l	l	t	t,th
e	eh,ae	m	m	u	uh,uw
e2	ah,ax	n	n	w	w
ey	ey	ng	ng	y	y
f	f,v	ny	n+y	z	z,zh
g	g	o	ow,ao	sil	sil

After constructing the English-to-Indonesian phoneme mapping table, the next step is to convert all English phoneme labels on WSJ training data, which has been transcribed and segmented previously, into

Indonesian phoneme labels based on that table. Then, train the model as a seed model of the Indonesian target language and use it to recognize Indonesian target language speech. Since the model is built using cross-language substitution, we also call it the CLS model. Figure 2 shows the recognition accuracy rates of the seed CLS model on the Indonesian test set. Both types of mapping described above (type 1 and type 2) are applied here, and will be called ”CLS1” and ”CLS2”, respectively. To find the optimum accuracy, several LM scale parameters are also used. The recognition results show that the performance of CLS2 (where some English fricative is represented as a combination of two Indonesian phonemes) is worse than CLS1 (all variants of English phoneme “t” sounds are mapped into single Indonesian phoneme). The best performance of CLS1 with 5 mixture components is 45.50% word accuracy, and with 15 mixture components it is 49.26% word accuracy, while the best performance of CLS2 with 5 mixture components is only 44.47% word accuracy, and with 15 mixture components it is just 48.60% word accuracy. All the best performances in each type were achieved with LM-scale 1=6 and LM-scale 2=12.

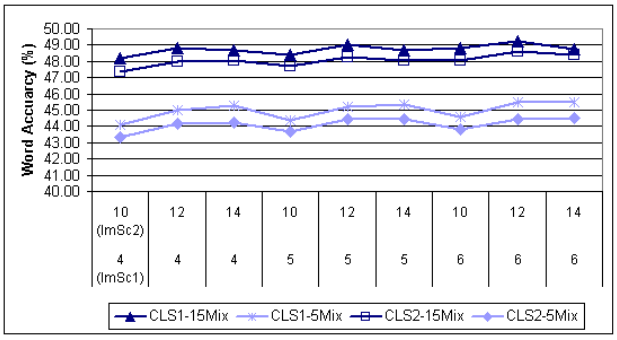

Figure 2: *The recognition accuracy (%) of the CLS model*

3.2. Cross-Language Insertion

In this approach, the initial step is to segment the utterances of the Indonesian C2 training data, based on Viterbi alignment algorithm, using the most optimum seed CLS model resulted in the previous approach. Then, the next step is to train each phoneme HMM using the same procedure and parameters as before. Since the C2 corpus does not include a full phoneme set, the missing Indonesian phoneme HMM models are inserted with the phoneme HMM of the seed CLS model. Finally, all phoneme HMM are combined into one large HMnet, where embedded training is conducted. The final model is referred to the CLI model.

The recognition accuracy rates of the CLI model on the Indonesian test set can be seen in Fig. 3. Several LM scale parameters are also applied here to find the optimum accuracy. For the CLI model with 5 mixture components, the best performance was 87.91% word accuracy, while that for the CLI model with 15 mixture components was 88.97%.

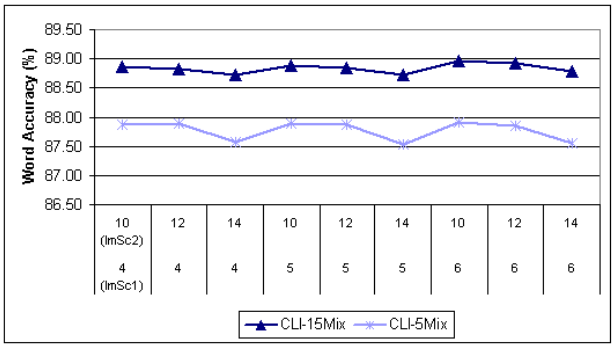

Figure 3: *The recognition accuracy (%) of the CLI model.*

3.3. Cross-Language Adaptation

The method in this approach is to adapt the parameters of the seed CLS model to the Indonesian C2 training data. Here, 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 [12], and also on cross-language adaptation [11, 7].

This scheme principally takes the advantages of prior information about existing models. A Bayesian learning mechanism then adjust the parameters of the seed acoustic model in such a way that the limited Indonesian C2 training data would modify the seed acoustic model parameters guided by the prior knowledge to compensate for the adverse effects of a mismatch [12]. Furthermore, the parameter reestimation is a weighted sum of the prior knowledge and the new estimation of the target language. Note that since the C2 only covers 70% of the total phonemes, only those phoneme model parameters can be adapted. The rest will remain the same.

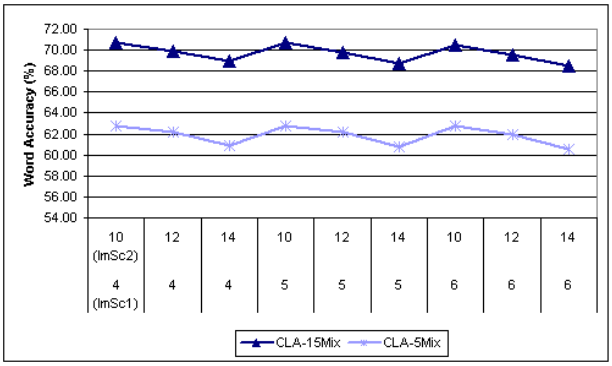

Figure 4: *The recognition accuracy (%) of the CLA model.*

Figure 4 shows the recognition accuracy rates of the CLA model on the Indonesian test set. Several LM scale parameters are also applied here to find the optimum accuracy. For the CLA model with 5 mixture components, the best performance was 62.82% word accuracy, while for the CLA model with 15 mixture components, it was 70.69%.

4. Comparison of Results of Different Approaches

Here, we perform an evaluation comparing the word accuracy from all cross-language approaches, including cross-language substitution, cross-language insertion, and cross-language adaptation. The best performance from CLS, CLI, and CLA models are shown together in Fig. 5.

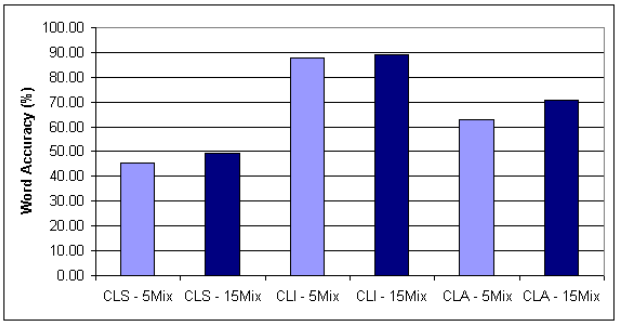

Figure 5: *Overview of the system architecture design.*

The CLS model gives the worst performance, with only a 45.50% word accuracy (using five Gaussian mixture components) and a 49.26% word accuracy (using 15 Gaussian mixture components). By adapting the CLS model to the Indonesian C2 training corpus as the CLA model was, the results show that the MAP-based adaptation can help to improve the accuracy by up to 21.4% absolute. However, this performance by the CLA model is still worst than that of the CLI model, possibly for the following reasons. First, the CLI model is basically a pure monolingual HMM which is trained from the Indonesian C2 corpus, while the CLA model is the adapted CLS model. Second, the major limitation of the MAP-based adaptation approach is that it requires an accurate initial guess for the prior knowledge of the existing CLS model [12], which in this case is difficult to obtain because the CLS model is trained from the English speech data. Moreover, as described in [11], the acoustic variations across languages are much larger and more complex than variations within the same language. Consequently, we need much more Indonesian training data to produce an efficient adaptation. This is why the CLA model's performance is not better than that of the CLI model.

5. Conclusion

We have demonstrated the possibility of rapid development of initial Indonesian phoneme-based speech recognition using the cross-language approach, where English is the source language and Indonesian is the target language. We have attempted the cross-language approach in three ways: (1) cross-language substitution, (2) cross-language insertion, and (3) cross-language adaptation. We have also has shown how the English-to-Indonesian phoneme mapping is generated based on knowledge-driven methods. Evaluation results reveal that the CLI models outperform both the CLS and the CLA models, meaning that in this case the cross-language insertion is the most effective choice for rapid development of the Indonesian ASR.

6. Acknowledgement

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

7. 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] 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.

[6] E. Kelana, R. Harinugroho, and P. Hutagaol, "Development of telcommunication system for dumb and deaf people," in *Proc. AIC*, Kuala Lumpur, Malaysia, 2004.

[7] 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.

[8] D.B. Paul and J.M. Baker, "The design for the Wall Street journal based CSR corpus," in *Proc. DARPA Workshop*, Pacific Groove, California, USA, 1992, pp. 357–361.

[9] "The CMU pronouncing dictionary," http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[10] H. Alwi, S. Dardjowidjojo, H. Lapoliwa, and A.M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia (Indonesian Grammar)*, Balai Pustaka, Jakarta, Indonesia, 2003.

[11] C. Nieuwondt and E.C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, pp. 101–113, 2002.

[12] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall, New Jersay, USA, 2001.