

Multilingual Phone Clustering for Recognition of Spontaneous Indonesian Speech Utilising Pronunciation Modelling Techniques

Eddie Wong, Terrence Martin, Torbjørn Svendsen[†], and Sridha Sridharan

Speech and Audio Research Laboratory, Queensland University of Technology,
GPO Box 2434, 2 George St, Brisbane, Australia, QLD 4001.

[†] Dept. of Telecommunications, Norwegian University of Science and Technology
N-7491 Trondheim, Norway

(ee.wong, tl.martin, s.sridharan)@qut.edu.au, torbjorn@tele.ntnu.no

Abstract

In this paper, a multilingual acoustic model set derived from English, Hindi, and Spanish is utilised to recognise speech in Indonesian. In order to achieve this task we incorporate a two tiered approach to perform the cross-lingual porting of the multilingual models to a new language. In the first stage, we use an entropy based decision tree to merge similar phones from different languages into clusters to form a new multilingual model set. In the second stage, we propose the use of a cross-lingual pronunciation modelling technique to perform the mapping from the multilingual models to the Indonesian phone set. A set of mapping rules are derived from this process and are employed to convert the original Indonesian lexicon into a pronunciation lexicon in terms of the multilingual model set. Preliminary experimental results show that, compared to the common knowledge based approach, both of these techniques reduce the word error rate in a spontaneous speech recognition task.

1. Introduction

The resource necessary to produce Automatic Speech Recognition (ASR) systems for a new language are considerable. Of all the resources required, obtaining sufficient transcribed acoustic data and lexicons presents a major problem for many languages. There are still several languages with major population bases which have insufficient resources for the development of speech enabled applications. Our research is focused on producing generic techniques that exploit existing resources from *Source* languages for porting ASR technology to *Target* languages.

Indonesia has a population of 190 million and is the largest Moslem nation on earth. When ranking languages to include in the Global-Phone speech database, population; variability; distribution; religious circumstances and linguistic aspects were the factors considered [1]. Indonesian/Malay was ranked ninth suggesting it also ranks highly in the speech science community. Accordingly, there is particular interest in producing speech enabled applications for the Indonesian languages, and so a secondary focus of our research is to extend the generic methods developed for cross-lingual porting of resources to the Indonesian languages.

The process of utilising resources from a data rich language (referred to as *Source* language in this paper) to a data poor language (referred to as *Target* language), can be broken into

two basic steps. The first step involves the creation of the *Source* language/s model set. Monolingual models can be used for cross-lingual ASR, however in [2] and [3] it was reported that improved recognition results can be achieved using multilingual models. One explanation for this is that multilingual models offer better coverage of the *Target* language because a model set covering multiple languages is more likely to provide a better match to the *Target* language than a monolingual *Source* set.

The second step in the process is to then express the lexicon of the *Target* language (Indonesian) in terms of the *Source* language models (a multilingual model set in this experiment). Past attempts to produce the lexicon have relied on either knowledge based or data driven methods. Knowledge based methods exploit linguistic knowledge to extract mappings between the representational symbols for the sounds of languages. The symbolic representation for the various sounds, and a description of articulatory features is provided by the International Phonetic Alphabet (IPA) [4]. Data driven techniques typically rely on either confusion matrix based approaches or alternatively on distance measures, usually based on relative entropy. Results reported in [5] indicated that data driven methods produce superior results in comparison to knowledge based methods.

In both cases, a single representative from the *Source* language/s is chosen to represent a phonetic event in the *Target* language. However, variation due to context can result in allophonic variants, which are often predictable but not captured. These variants will not be captured if we consider each phonetic event in isolation, such as the one to one mapping provided by the aforementioned methods. To address this issue we implement a technique introduced in a companion paper [6] which utilises pronunciation modelling to capture the systematic variation that occurs in the *Target* language with the models from the *Source* language/s.

The rest of the paper is organised as follows: Section 2 discusses the approach used to create a multilingual phone set from multiple languages. Section 3 then presents the technique used to perform the lexicon mapping from the *Target* to the *Source* language. We then describe the Indonesian speech recognition experiments including results in Section 4. Finally a discussion and conclusions are given in Section 5 and 6 respectively.

2. Multilingual Phone Model Clustering

The multilingual model set in this investigation is derived from the languages of English, Hindi, and Spanish. The process of merging similar phone models is achieved via a relative entropy based decision tree clustering technique, using a set of phonetic related questions. The purpose of employing the decision tree in our experiments differs from that reported in [7] for monolingual context dependent modelling, and that used in [3] for multilingual context dependent modelling. In [3], the underlying motivation was to use the decision tree process to cluster phones with similar contexts in order to overcome the problem of insufficient training data. In our case, the decision tree is used to merge phone models based on acoustic similarity estimated from the training data. The resultant clusters then maximally represent the *Source* languages acoustic evidence with a minimal set of models.

2.1. Clustering Criteria

The phonetic question set is arranged hierarchically from broad questions such as “*Is the phone a Fricative*” to more specific questions such as “*Is the phone the Spanish /b/*”. Effectively the question set places no restriction on the possible number of phone models within each cluster. Language tags are added to each phone so that both language dependent and language independent clusters are possible. No restrictions prevent phone models from the same language being clustered together. In other studies [8], emphasis was placed on ensuring that phones from the same language were not merged. The rationale for this restriction was that if a resultant cluster was too broad, and included phones from the same language, then the possibility for different words having the same phonetic transcription (homophones) increases. However, the intended aim of creating multilingual models for cross-lingual speech recognition is to increase the coverage of the *Target* language’s acoustic/phonetic space. To this end, letting the clustering process be unconstrained, based on the acoustic evidence, is preferable to imposing constraints based on language specific phonemic requirements that may not be applicable to the *Target* language.

2.2. Clustering Process

In the beginning of the clustering process, all monolingual phone models are pooled together at the first node of the decision tree and then split according to the phonetic questions. A threshold is used to stop the splitting process and in turn controls the number of clusters. At the end of the splitting process, phone models that belong to the same cluster are tied together to form a new multilingual model and their corresponding acoustic data are used to re-estimate it’s parameters. Table 1 shows the first twenty questions that were asked during the decision tree clustering process. Questions that belong to broader phonetic events were asked first then progressively followed by more specific questions.

The aim of creating a multilingual phone set in this experiment is to perform Indonesian speech recognition, therefore we would like to choose the threshold to obtain a model set that maximises the acoustic coverage for Indonesian. In addition, the threshold must be set so that the subsequent models are flexible enough to cope with allophonic variation. The last property is important as capturing variation is fundamental to the idea of pronunciation modelling discussed in Section 3. The criteria we adopted for selecting the threshold was based on linguistic consideration, where the cluster size was empirically adjusted

	Phonetic Question	Description
1.	Vowel_en_hi_sp	All vowel phones
2.	Low_en_hi_sp	Low vowel phones
3.	UnStrident_en_hi_sp	Stops, nasals, liquids
4.	Stop_en_hi_sp	Stops
5.	unvoiced_closures_en_hi_sp	Unvoiced closures
6.	Fronting_en_hi_sp	Front Diphthongs
7.	FrontVowel_en_hi_sp	Front vowel phones
8.	unvoiced_closures_hi_sp	Unvoiced closures
9.	Continuent_en_hi_sp	Fricative, liquids, glides
10.	Rounded_en_sp	Rounded vowel phones
11.	voiced_closures_en_hi_sp	Voiced closures
12.	voiced_closures_hi_sp	Voiced closures
13.	Medium_en_hi_sp	Medium vowel phones
14.	UnvoicedStop_en	Unvoiced Stop
15.	Liquids_en	Liquids
16.	High_en_hi_sp	High vowel phones
17.	Glottals_en_hi_sp	Glottals
18.	Short_en_hi_sp	Short vowel phones
19.	EVowel_en_hi_sp	Vowels involved /e/
20.	UnvoicedFricative_en_hi_sp	Unvoiced Fricative

Table 1: *The first twenty questions asked during the decision tree clustering (en - English, hi - Hindi, sp - Spanish).*

until sufficient coverage was obtained for the *Target* language. In our experimental setup, the baseline Indonesian system used 30 phone models. We selected a threshold that yielded 47 multilingual phone models. Table 2 shows the details of the assignment of *Source* monolingual phone models to 47 clusters after the clustering process. The left column indicates the number of phones assigned to a single cluster and the right column illustrates the corresponding clusters which belong to each category. Each string in the right column represents a single cluster, where the ‘-’ character is used to group the monolingual phone models of that cluster and the ‘_’ character is the language tag.

3. Pronunciation Modelling

In a companion paper, [6], we investigated using pronunciation modelling to produce an Indonesian lexicon which captured the pronunciation of Indonesian words, expressed using symbolic representation for the *Source* language acoustic models. The basic procedure for achieving this is as follows:

1. Use data driven pronunciation modelling to obtain a pronunciation lexicon for the *Target* language training data in terms of the symbols used to represent the *Source* language acoustic models. This was only done for those lexicon units for which there was sufficient training examples.
2. Use this pronunciation lexicon to derive a set of rules that map the *Target* language phones to the *Source* language. Prune these rule according to predetermined criteria.
3. Apply the rule set derived to the entries which occur in the *Target* language vocabulary but were not optimised in Step 1. This results in full coverage of the desired *Target* language vocabulary in terms of the symbols used to represent the *Source* language acoustic models.

The methodology used in [6] employed a modification of monolingual techniques outlined in [9] and [10]. We use the

# Phone merged in a cluster	Phone list
> 5	U.en_hi_sp-&.en_hi_sp-Ix_hi_hs_sp Glottal.en_hi_sp-T.en_v.en-z.en-z_hi dr_hi-tr_hi-r({_hi-rr_hi-L.sp-r({_sp-r.sp tS.en_hi.sp-dZ.en-dZ_hi-tSH_hi g.en_hi.sp-kH_hi-k_hi-k.sp
5	d.en-D.en-d{,_hi-t{H_hi-d{,_sp b.en_hi.sp-p_hi-p.sp l=.en-n=.en->i.en-oU.en-x3.sp
4	sil-nos1-nos2-bH_hi w_hi-V.sp-G.sp-D.sp n.en-n_hi-N_hi-n{,_sp j.en_hi.sp-nj.sp
3	h.en-h_hi-dZ.sp f.en_hi.sp vc1.en_hi.sp ph.en-kh.en-x.sp uvcl.en_hi.sp m.en_hi.sp u:hi-o_hi-o.sp S.en_hi.sp I.en_hi.sp
2	s_hi-s.sp w.en-w.sp l_hi-l.sp N.en-N.sp t{,_hi-t{,_sp l.en-u.sp 9r.en-&r.en i_hi-i.sp E_hi-E.sp ei.en-ai_hi ^_en-^_hi A.en-a_hi >.en->_hi aI.en-aI.sp
1	s.en th.en u.en 3r.en i:.en e.sp e_hi E.en aU.en a.sp @_hi @.en

Table 2: A set of 47 multilingual phones obtained from the decision tree clustering process. Each string represents a cluster with the ‘-’ character used to connect each monolingual phone. The ‘_’ character is the language tag, where en - English, hi - Hindi, sp - Spanish.

syllable to constrain the boundary of context width which is then used to derive the rule set. The rationale for the selection of the syllable is outlined in [6]. We use 1.3 hrs of Indonesian training data, which is segmented into isolated syllables. An optimisation technique is then used to select the *Source* language phone-string (*baseforms*) for each syllable, which maximises the likelihood of the training data.

Context dependent pronunciation rules are derived from the optimised baseforms. These rules are then pruned, and subsequently used to produce a lexicon based on the multilingual acoustic data.

4. Experiments and Results

We conducted experiments using spontaneous telephone speech from both the 11 language (for English, Hindi and Spanish data), and the 22 language (Indonesian data) versions of the Oregon Graduate Institute Multi Language Telephone Speech Corpus [11]. Using data recorded in a similar environment provided the opportunity to standardise the training and test environment and hopefully reduce the impact of train/test mismatch and variations in channel effects. No transcriptions for the Indonesian acoustic data existed originally and so two native speakers were employed to transcribe three hours of speech data. This was then verified and corrected for errors. The speech data was split into a training set (1.3hrs), a development test set (54mins) and a test set (25mins). The Indonesian acoustic data transcribed included all utterance categories such as sto-

	Word Recog. No adaptation	Word Recog. With adaptation
	% Accuracy	% Accuracy
Indonesian Baseline	26.92	N/A
Knowledge Driven	13.37	20.67
Pronunciation Modelling	15.54	23.17
Pron. Modelling (English model only)	11.20	20.67

Table 3: Indonesian spontaneous speech recognition results.

ries, age, routes, climates etc. We used a subset of a commercially produced 20 000 word Indonesian lexicon which included syllable demarcation. Further details of the transcription process and lexicon development are outlined in [12]. To avoid out-of-vocabulary errors the subset provided orthographic transcriptions for all the 2519 words that occurred in the train, development test and test data.

A speech recognition engine was developed using the multilingual monophone models derived from English, Hindi, and Spanish. Additionally a baseline Indonesian system was developed which also uses context independent acoustic models. For both recognition systems, the HMM topology was a 3 state left-to-right, with each state emission density comprising 8 Gaussian mixture components. A bi-gram Language Model, trained using the training and development test set, is incorporated. Speech was parameterised using a 12th order MFCC analysis plus normalised energy, 1st and 2nd order derivatives, and a frame size/shift of 25/10ms. Cepstral Mean Subtraction (CMS) was employed. The word recognition performance for this baseline Indonesian system is outlined in the first row of Table 3.

Knowledge based mapping experiments were performed for the purpose of comparison with the pronunciation modelling technique. For the knowledge based approach, the *Target* languages lexicon is mapped in terms of the *Source* language phone set using linguistic knowledge. Experimental results are shown in the second row of Table 3. Compared to the pronunciation modelling technique (the third row), the results suggest that system performance can be improved by incorporating pronunciation modelling. This suggests that it is important to handle the variation in pronunciation in the *Target* language’s lexicon.

The same pronunciation modelling experiment was performed using only English as the *Source* language. The corresponding testing results are shown at the last row of Table 3. This reinforces results previously reported in [3] where the multilingual models provided better recognition results and highlights that this phenomena is applicable for the Indonesian language. The more diverse coverage of acoustic space provided by the multilingual models increase the chances of better modelling the *Target* language and therefore would yield higher accuracy.

All experiments were repeated using the models after adaptation, where the adaptation used a development data set. The adaptation procedural sequence is as follows: conduct a global mean only MLLR; mean only MLLR for each model; mean and variance MLLR for each model; 5 iterations of mean and variance MAP adaptation. The experimental results are depicted in the second column of Table 3. Consistent outcomes were

obtained after adaptation supporting that both the multilingual phone set approach and the pronunciation modelling technique can improve the accuracy of a cross-lingual speech recognition system.

5. Discussion

The limited amount of training data meant that the acoustic models could only reasonably be based on monophones. This subsequently constrained the possible recognition performance to a lower level. Using continuous telephone speech data also served to restrict the recognition accuracy. However, the preliminary indications are that the combined use of multilingual model set and pronunciation modelling are an effective means for obtaining improved cross lingual recognition performance.

In our experiment concerning the creation of the multilingual phone set, we wanted to examine the impact of not restricting the merge of phone models from the same language during the clustering process. As outlined in Section 2.1 the disadvantage of this technique is that it could result in a cluster which is phonetically too broad, with the potential loss in discrimination between words.

Interestingly, in an Indonesian lexicon, defined using English models and pronunciation modelling, two homophones occurred. In comparison, the combination of multilingual phone set with pronunciation modelling introduced only an additional two homophones. Although the lexicon is comparatively small, (2.3k words), this indicates that the impact of our clustering technique does not significantly increase the possibility for homophones, yet allows for a decrease in confusability by removing redundant clusters.

The pronunciation modelling technique we employed constrained the number of variants included in the *Target* language lexicon to one. If the clustering process (without language restriction) produces homophones, then the pronunciation modelling technique can produce an N-best list of pronunciations which can be traversed and substituted into the lexicon.

6. Conclusions

In this paper we presented a two tiered approach to cross-lingual porting of multilingual models. We investigated the use of an unconstrained decision tree based approach, which relies on a set of phonetic related questions, to merge acoustically similar phone models from multiple languages. The utilisation of the multilingual phone set in this task reinforced previous reported results and showed that this technique is applicable for the Indonesian language.

The addition of pronunciation modelling to the multilingual clustering produced improved recognition results. Additionally it has the potential in future extensions to minimise the homophones which can occur because of the unconstrained clustering process by substituting alternate pronunciation variants.

7. References

- [1] T. Schultz, M. Westphal, A. Waibel, "The Global Phone Project: Multilingual LVCSR With JANUS-3," *2nd SQEL Workshop*, 1997.
- [2] J.Kohler, "Multi-lingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds," *ICSLP*, vol. 4, pp. 2195–2198, 1996.
- [3] T. Schultz, A. Waibel, "Language Independent and Language Adaptive Acoustic Modelling," *Speech Comm.*, vol. 35, no. 1-2, pp. 31–51, February 2001.
- [4] IPA, *Handbook of the International Phonetic Association : a guide to the use of the International Phonetic Alphabet*, Cambridge University Press, 1999.
- [5] W. Byrne et al, "Towards language independent acoustic modelling," *ICASSP*, vol. 2, pp. 1029–1032, 2000.
- [6] T.Martin, T.Svendsen, S.Sridharan, "Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition," in *EuroSpeech*, 2003.
- [7] J. Odell, *The use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Queens College, Cambridge, 1995.
- [8] A. Zgank, B. Imperl, F.T. Johansen., "Crosslingual Adaptation of Multilingual Triphone Acoustic Models," in *MSLP*, 2001.
- [9] N. Cremelie, J.P. Martens, "In Search of better Pronunciation Models for Speech Recognition," *Speech Comm.*, vol. 29, no. (2-4), pp. 115–136, 2000.
- [10] T.Svendsen, F.K.Soong, H.Purnhagen, "Optimizing baseforms for HMM-base speech recognition," in *EuroSpeech*, September 1995, pp. 783–786.
- [11] T. Lander and R. Cole and B. Oshika and M. Noel, "The OGI 22 language telephone speech corpus," *The European Conference on Speech Communication and Technology*, 1995.
- [12] T. Martin, S.Sridharan, "Cross Lingual Modelling Experiments for Indonesian," in *8th Australian Int. Conf on Speech Science and Technology*, Melbourne, 2002.