

Cross-Lingual Pronunciation Modelling for Indonesian Speech Recognition

Terrence Martin, Torbjørn Svendsen*, and Sridha Sridharan

Speech and Audio Research Lab, Queensland University of Technology
GPO BOX 2434, Brisbane, Australia, QLD 4001.

*Dept. of Telecommunications, Norwegian University of Science and Technology
N-7491 Trondheim, Norway

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

Abstract

The resources necessary to produce Automatic Speech Recognition systems for a new language are considerable, and for many languages these resources are not available. This emphasizes the need for the development of generic techniques which overcome this data shortage. Indonesian is one language which suffers from this problem and whose population and importance suggest it could benefit from speech enabled technology. Accordingly, we investigate using English acoustic models to recognize Indonesian speech. The mapping process, where the symbolic representation of the *Source* language acoustic models is equated to the *Target* language phonetic units, has typically been achieved using one to one mapping techniques. This mapping method does not allow for the incorporation of predictable allophonic variation in the lexicon. Accordingly, in this paper we present the use of cross-lingual pronunciation modelling to extract context dependant mapping rules, which are subsequently used to produce a more accurate cross lingual lexicon.

1. Introduction

The resources necessary to produce Automatic Speech Recognition 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; a fact which is seemingly at odds with the lack of speech enabled technology and research on Indonesian languages. Accordingly, a secondary focus of our research is to extend the generic methods developed for cross lingual porting of resources specifically to the Indonesian language. In this paper we outline our preliminary investigations which use available English resources for the production of an Indonesian ASR system.

The challenge in cross-lingual mapping is to determine which variation we should represent in the lexicon, and what should be captured via the acoustic models. Pragmatically we should only attempt to incorporate the predictable variation in the lexicon, leaving the acoustic models to deal with any other variation.

Previous multilingual research efforts have generally focused on exploiting the similarities in the acoustic realization of sounds across languages[2]. Some form of mapping between the symbolic representation for acoustic models of the *Source* language and the phonetic units of the *Target* language is required, so that the *Target* language pronunciation lexicon can be represented in terms of the acoustic models from the *Source* language [3], [4], [2].

Typically the methods used to determine these mappings are either knowledge based or data driven. Knowledge based methods exploit linguistic knowledge to extract mappings between the representational units of two languages. In data driven techniques the aim is to establish the best mapping between two languages using the available training data from both the *Source* and *Target*. Typical data driven methods utilize either confusion matrices or entropy based measures. Results reported in [4] indicate that data driven methods produce superior results in comparison to knowledge driven methods.

However, in employing a one-to-one mapping technique between the phonetic events of different languages, no consideration is given to the acoustic realization of phones in different contexts, and so the predictable variation is not adequately represented in the lexicon. In monolingual ASR, context dependent modelling is an established technique for coping with the allophonic variation which occurs because of context, and provides significant improvement. Context dependent modelling techniques [3] have been incorporated in multilingual ASR, but the original mapping conducted has still been based on one to one mapping techniques, potentially introducing errors from the beginning of the process.

This paper outlines our preliminary investigations directed towards addressing this problem. In an attempt to produce cross language lexicons which capture allophonic variation we introduce *Cross-lingual Pronunciation Modelling*. Pronunciation Modelling has been applied to produce improved monolingual ASR [5][6] and it is our belief that this technique provides an alternative data driven mapping technique with the potential to improve current cross lingual recognition performance.

2. Cross-lingual Pronunciation Modelling

In order to achieve improved mapping performance our aim is to capture the predictable variation which occurs in the *Target* language speech. A basic outline of the procedure used to achieve 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 is only done for those units for which there is 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 any unseen *Target* language lexical entries, including those for which insufficient training examples existed. The result is full coverage of the desired *Target* language vocabulary in terms of the symbols used to represent the *Source* language acoustic models.

Ideally, given sufficient data, it is desirable to produce pronunciation rules based on the widest possible context and in [5] word and even cross word phenomena were examined. However, in a cross-lingual setting, there is insufficient training data available for words to be used to define the boundaries of context width. A unit with smaller context width is required and we chose the syllable; basing this selection on the following considerations:

2.1. Inter-Unit Co-articulation

The word is affected by inter-word co-articulatory phenomena, so any alternative sub-word unit is also likely to be subject to the same fate. However, the unit selected should attempt to minimize this effect. In [7], various properties of the syllable were examined and it was found that the recognized transcription of the syllable *onset* typically maintained its canonical representation, in varying speaking conditions. Accordingly, selection of the syllable at least partially fulfils this criteria.

2.2. Coverage

As reported in [8], English has approximately 10000 syllables for the Switch Board Corpus, but only 300 syllables provided 80 % coverage for the acoustic realizations in this corpus. Similarly we examined the coverage in a 20 000 word Indonesian lexicon and found only 1900 syllables. Importantly, only 200 syllables provided 90% coverage for our training/test and development data.

2.3. Segmentation Issues

To capture allophonic variation, context dependent models are normally used when the phone is chosen as the sub-word acoustic model. The short duration of the tri-phone, results in an increased chance of erroneous alignments during segmentation, whether they are obtained manually or automatically. These erroneous alignments result in *leakage* of the frame statistics from surrounding tri-phones, thereby corrupting model accuracy. In contrast to the tri-phone, the syllable is far easier to segment, with fewer errors occurring in both manual and automatic segmentation.

3. Problem Formulation

The *Target* language is defined by a set of acoustic models, $\Lambda = \{\lambda_1^T, \lambda_2^T, \dots, \lambda_p^T\}$ which describes the set of subword units $P_T = \{p_1^T, p_2^T, \dots, p_p^T\}$ and a lexicon, which is a set of entry/baseform pairs, $\mathcal{L}_T = \{(W, B_T(W)); W \in V\}$ where W is the lexical entry (word or syllable) and V denotes the vocabulary. The baseform is a linear string of subword units, $B_T(W) = (u_1^T, u_2^T, \dots, u_{N(W)}^T)$ where $N(W)$ is the length of the baseform and $u_i^T \in P_T$.

For the *Source* language we have available a set of acoustic models, $\Lambda = \{\lambda_1^S, \lambda_2^S, \dots, \lambda_q^S\}$ which describes the *Source* language set of subword units $P_S = \{p_1^S, p_2^S, \dots, p_q^S\}$.

We wish to be able to describe the *Target* language in terms of the *Source* language subword units and acoustic models. I.e., we wish to find a lexicon with baseforms which are sequences of the *Source* language subword units: $\mathcal{L}_S = \{(W, B_S(W)); W \in V\}$ with baseforms $B_S(W) = (u_1^S, u_2^S, \dots, u_{M(W)}^S)$ where $M(W)$ is the length of the baseform and $u_i^S \in P_S$.

Data-driven pronunciation modelling can be applied to find the baseforms for the lexical entries for which there is sufficient training data. In order to create a complete pronunciation lexicon, we must however also be able to predict mappings for unseen lexical entries. To achieve this, we employ mappings that are derived from the pronunciation modelling of the data-rich lexical entries.

4. System Implementation

The process used to generate the lexicon in terms of the *Source* models is achieved using the following steps:

- Production of Source Language Lexicon
- Production of Reference and Alternate Transcriptions
- Reference and Alternate Transcription Alignment
- Rule Generation and Pruning
- Cross-Lingual Lexicon Creation

4.1. Production of Source Language Lexicon

The purpose of the pronunciation modelling is to describe the pronunciation of a syllable in the *Target* language in terms of the symbols associated with the acoustic units of the *Source* language. We employ a data-driven pronunciation modelling technique which selects the baseforms which maximize the likelihood of the training data [6][9].

Given a set of training utterances of a lexical entry, W , $\mathcal{T}_W = \{U_1^W, U_2^W, \dots, U_{K_W}^W\}$ the *Source* language acoustic models, Λ_S , and subword units, P_S , the most likely baseform can be found from the set of all valid baseforms \mathcal{B}_S as

$$\begin{aligned} \hat{B}_S &= \arg \max_{B_S(W) \in \mathcal{B}_S} P(B_S(W) | \mathcal{T}_W, \Lambda_S) \\ &= \arg \max_{B_S(W) \in \mathcal{B}_S} \prod_{n=1}^{K_W} p(U_n^W | B_S(W), \Lambda_S) \quad (1) \end{aligned}$$

where the bottom equality assumes that all baseforms are equiprobable and that the training utterances are conditionally independent for any given baseform.

If only a single training utterance is available, the optimization problem reduces to simple Viterbi decoding. However, in

order to ascertain a representative pronunciation lexicon which is robust to inter- and intra-speaker variation, a larger number of utterances are required, yielding a non-trivial optimization problem. The modified tree-trellis algorithm [9] [6] in principle guarantees finding the optimal baseform for multiple training utterances. Sub-optimal strategies, which are more efficient are also proposed in [6].

4.2. Production of Reference and Alternate Transcription

The source language lexicon will only contain baseforms for entries where there is sufficient training data in the *Target* language. To produce a lexicon with complete coverage, we need to derive a set of rules that maps the *Target* language phones to the *Source* language. The rules will be inferred from the coupling between the new source language based lexicon and the original *Target* language lexicon. For each syllable, the baseform in the *Target* language lexicon is the *Reference* transcription, while the corresponding *Source* language baseform is the *Alternate* transcription.

4.3. Reference and Alternate Transcription Alignment

There is no constraint on the possible length of the alternate baseforms produced in the optimization process. Accordingly, a comparison between the reference and alternate baseform may reveal the presence of insertions, deletions and substitutions. To ensure that valid rules are derived from these mappings, these insertions and deletions must be symbolically represented in the baseform before further processing is conducted. The sub-syllable phonetic strings are then aligned. We used a lattice based rescoring alignment technique, based on a modified version of the Needleman-Wunsh algorithm [10]. Different penalties were applied for insertions, deletions, and substitutions. However, the original implementation did not discriminate the between substitutions or insertion between dissimilar classes of phonemes (eg. vowels and fricatives) and resulted in some dubious alignments. Several techniques aimed at constraining the possible alignments have been proposed [5][11] however given time constraints we chose to hand correct these alignments.

4.4. Rule Generation and Pruning

Each phoneme (Focus F) that occurs in the *reference* baseform is examined for left (L) and right (R) context. The combination of the Focus and its context is referred to as the condition (LFR). Each condition and alternate transcription constitutes a rule. However some rules are less likely and so application probability statistics are accumulated for each rule and those that are unlikely are culled from the rule set. The mathematical representation for a rule and its application probability are denoted in Equation 2

$$r : LFR \Rightarrow \hat{F} \text{ with probability } \frac{N_2}{N_1} \quad (2)$$

where N_2 represents the number of times that the condition was encountered and was transformed to that output in the alternate baseform transcription and N_1 is the number of times that the condition is encountered.

The rules derived from each condition are organized hierarchically with the most specific rules at the top and least specific, i.e. direct mappings, at the bottom. If insufficient examples of a condition exist, or the application likelihood does not exceed a predetermined threshold, then the rule is culled.

The least specific rules, i.e. those that consider no context, are effective for coping with unseen contexts, or if insufficient instances of more specific contexts resulted in the pruning of the rule for that condition.

4.5. Cross-Lingual Lexicon Creation

The pruned rule set is then applied to each syllable that occurs in the Indonesian lexicon. The result is a lexicon in terms of the source language models.

5. Experimental Procedure and Results

We conducted experiments using Oregon Graduate Institute telephone speech from both the 22 Language and Multi-language Speech Corpus. Using data recorded in a similar environment provided the opportunity to standardize 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 assisted in the transcription of three hours of speech data. This was then verified and corrected for errors. The speech data was split into a training set (1.3 hrs) a development set (54 mins) and a test set (25 mins). The Indonesian acoustic data transcribed included all utterance categories such as stories, 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/test and development data.

A speech recognition engine was developed for both English and Indonesian which employed context independent acoustic models. The model topology was 3 state left-to-right, with each state emission density comprising 8 Gaussian mixture components. Speech was parameterized using a 12th order MFCC analysis plus normalized energy, 1st and 2nd order derivatives, and a frame size/shift of 25/10ms. Cepstral Mean Subtraction (CMS) was employed.

Table 1 outlines the baseline phone recognition for both the English and Indonesian Phone recognition systems. The word recognition results are also shown for the Indonesian baseline system. Word level transcriptions for the OGI English data was not available so word recognition rates were not calculated.

We obtained bi-gram statistics from the training and development data for to give a more indicative result for word recognition. The bi-gram statistics are obviously biased towards the training domain, however the bi-grams were used in all experiments and should not impact on the comparison.

Language	% C	% A
Eng. Phone Rec. (42)	47.28	24.62
Indon. Phone Rec. (30)	54.31	38.18
Indon Word Rec.	28.23	26.92

Table 1: *Recognition System Baseline Performance* %C = % Correct, % A = % Accuracy

Crosslingual recognition results for English models on Indonesian speech with three different mapping approaches were then examined. The approaches considered were knowledge

driven and confusion matrix based methods as well as our cross-lingual pronunciation modelling based method. The results for both phone and word recognition are provided in Table 2.

	Phone Rec		Word Rec	
	%C	%A	%C	%A
Knowledge (28)	36.21	19.98	10.12	8.85
Confusion (24)	39.04	20.60	10.16	9.36
Pron Mod (34)	34.62	17.01	12.43	11.20

Table 2: Comparison of Mapping Techniques

6. Discussion

Examination of Table 2 reveals that the phone recognition rates for both the knowledge and confusion matrix based methods are superior to the pronunciation modelling technique. The one-to-one mapping approach employed by these techniques is designed to achieve better phone recognition rates. In contrast, pronunciation modelling aims to incorporate context information in anticipation of the extension to the task of word recognition. This point is highlighted by the comparatively better word recognition rates. Additionally the pronunciation modelling technique utilized a greater number of phones compared to the other techniques, (phone numbers bracketed). The increased number of phones could potentially add to confusability in a phone recognition experiment. This is highlighted by the fact that the English baseline, with more phones, performed significantly worse than the Indonesian baseline. However, this result may indicate that some knowledge based constraints on the size of the *Source* language model set size before the incorporation of pronunciation modelling may be beneficial.

The results outlined in Table 2 are undoubtedly poor overall, and subsequently no conclusive determination can be made about the suitability of cross lingual pronunciation modelling. However, the amount of training data; its type (continuous telephone speech); and the model topology (8 mixture monophones) are all factors which undermine the recognition process. These disadvantages were accepted given the alternative of having to compare languages obtained from different sources, and the inherent problems associated with channel mismatch. As a result, we examined the qualitative results in the rule set derivation process, in attempt to gain further insight, particularly with regard to the Indonesian language.

One example in which the technique performed well was in distinguishing between released and unreleased unvoiced plosives. Indonesian is characterized by unreleased unvoiced plosives when they occur word finally, a phenomena that this technique captured quite well. Another facet of Indonesian speech which the algorithm captured is glottal stop substitution when the phoneme /k/ occurs word finally.

One area which caused major problems was the effect of glides. The derivation of rules for glides was compromised by two competing factors. The first was that glides are difficult to segment accurately. Examination of the isolated syllable examples revealed that many of the glides were included in the corresponding syllable. This had the effect of including erroneous deletions and insertions in the rule set. This effect highlights that the use of syllable has certain limitations.

The second competing effect is introduced by the occur-

rence of hard and soft glides. The impact of this is best illustrated by an example. Consider the soft version of the phone /l/ when it occurs syllable finally such as in the Indonesian word "BETUL". In this case the /l/ is typically deleted in continuous speech. Compare this to when /l/ is hard, such as in the word "APALAH". In this case /l/ is not deleted. However there was insufficient training examples to cater for every context dependent instance of when /l/ occurred. In this case the less specific context rules are used. The next tier of context dependent rules for /l/ consider either left or right context. Unfortunately, there were more examples where /l/ was deleted, resulting in an inaccurate mapping.

7. Conclusions

No conclusive determination on the merits of cross lingual pronunciation modelling can be extracted from this investigation, due to the low levels of recognition performance. However, the qualitative results suggest that further investigation may provide improved results, if an enhanced testing framework is used and additional data is obtained for experimentation.

8. References

- [1] T. Schultz, M. Westphal, A. Waibel, "The Global Phone Project: Multilingual LVCSR With JANUS-3," *Proc. 2nd SQL Workshop*, 1997.
- [2] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," *Proc. ICSLP*, vol. 4, pp. 2195–2198, 1996.
- [3] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modelling," *Speech Comm.*, vol. 35, no. 1-2, pp. 31–51, February 2001.
- [4] W. Byrne et al, "Towards language independent acoustic modelling," *Proc. ICASSP*, vol. 2, pp. 1029–1032, 2000.
- [5] 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.
- [6] Holter, T., Svendsen, T., "Maximum likelihood modelling of pronunciation variation," *Speech Comm.*, vol. 29, 1999.
- [7] S. Greenberg, "Speaking in Shorthand-A Syllable-Centric Perspective for Understanding Pronunciation," in *Proc. of the ESCA workshop on Modelling Pronunciation Variation for ASR*, 1998.
- [8] A. Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington and J. Picone, "Syllable-based large cocabulary continuous speech recognition," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 4, pp. 358–366, May 2001, 2001.
- [9] T. Svendsen, F.K. Soong, H. Purnhagen, "Optimizing baseforms for HMM-base speech recognition," in *Proc of EUROSPEECH*, September 1995, pp. 783–786.
- [10] S.B Needleman and C.D Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journ. Mol. Biol.*, vol. 48, no. 444, 1970.
- [11] Ingunn Amdal, *Learning Pronunciation Modelling*, Ph.D. thesis, Norwegian University of Science, 2002.
- [12] T. Martin, S. Sridharan, "Cross Lingual Modelling Experiments for Indonesian," in *Proc of 8th Australian Int. Conf on Speech Science and Technology*, Melbourne, 2002.