CROSS-LANGUAGE ACOUSTIC MODEL REFINEMENT FOR THE INDONESIAN LANGUAGE

Terrence Martin , Sridha Sridharan

Speech, Audio, Image and Video Technologies
Queensland University of Technology
GPO Box 2434, 2 George St, Brisbane, Australia, QLD 4001.
tl.martin, s.sridharan@qut.edu.au

ABSTRACT

Porting ASR capabilities to many languages is hindered by a lack of transcribed acoustic data. Cross-language adaptation techniques seek to address this problem by substituting models trained in resource-rich source languages to recognise speech in resource-poor target languages. The differences in co-articulatory effects between the source and target languages, together with unwanted pronunciation and channel variation, result in recognition rates that are typically much worse then those achieved by well trained monolingual systems. In this paper, we present a technique which makes more effective use of limited adaptation data by structuring the state distributions to suit the co-articulatory occurrences in the target language. Additionally the proposed technique provides a more suitable method for synthesising unseen contexts. Evaluation of this technique is presented for a word recognition task using English and Spanish source language acoustic models trained using Switchboard and CallHome databases respectively. Using 25 minutes of Indonesian speech for target language adaptation data, this technique achieved an absolute improvement of 3.69% and 6.31% for English and Spanish respectively, when compared to traditional adaptation techniques. Using 90 minutes of adaptation data, an absolute improvement of 3.22% and 3.07% was achieved.

1. INTRODUCTION

Automatic Speech recognition has matured to the point that reasonable accuracy can be achieved in difficult tasks such as conversational telephone speech. However, in order to obtain this type of recognition performance, hundreds of hours of transcribed acoustic data is required. Producing this data requires a large outlay in terms of time, manpower and money, consequently restricting the availability of Automatic Speech Recognition (ASR) technology to only a few of the worlds languages. The potential applications for ASR technology, however, extend beyond these few languages. Accordingly, our primary research focus concentrates on the development of generic techniques which can reduce the target language data requirements and thereby facilitate the production of a Large Vocabulary Conversational Speech Recogniser (LVCSR) for telephone speech.

Indonesia has a population in excess of 200 million, and the technical infrastructure which could benefit from ASR technology. Accordingly, there is considerable interest in developing ASR applications for the Indonesian language. Thus is secondary research focus for our organisation is achieving respectable recognition rates for the Indonesian language.

There are a number of strategies for producing an ASR system for a particular data poor language. All of these approaches seek to exploit the similarities in the acoustic realisation of sounds across languages. Some approaches look to produce a universal set of acoustic models which are capable of recognising multiple languages equally well, however this research is based on adapting acoustic models from one or more languages and improving their discriminatory ability on a specific target language. This approach is commonly referred to as cross-language adaptation.

Cross language transfer is not immune from the problems which plague monolingual ASR; such as channel effects, and speaker and pronunciation variation. However, in addition to this, the differences between languages produce additional problems. As a result most research has sought to isolate these effects by conducting constrained evaluations. Examples of *constrained evaluations* include clean read speech recognition tasks[1], thereby enabling evaluation with reduced channel mismatch and pronunciation variation. Other examples conducted use telephone speech, but are evaluated using isolated word[2] or reduced vocabulary experiments [3].

However the effects which undermine cross language transfer do not act independently, and a more holistic approach is required to address the combined challenges they present. Accordingly in this paper, strategies are presented which seek to reduce the impact of these problems. These techniques are used to adapt English and Spanish acoustic models trained using the *SwitchBoard* and *CallHome* databases respectively, for use in a word recognition task on Indonesian telephone speech.

Section 2 discusses the inter-relationships between factors which can negatively impact on the success of cross language transfer and in Section 3 we propose a technique for reducing their impact. In Section 4 a comparison is conducted between the proposed technique and standard adaptation techniques. Discussion and conclusions are drawn in Sections 5 and 6.

2. THE DIFFICULTIES FACING CROSS LANGUAGE TRANSFER

Transformation based (MLLR) and Bayesian based (MAP) adaptation techniques are commonly used to minimise train-test mismatch for both channel normalisation and speaker adaptation in monolingual ASR systems. This approach has also been extended to cross language transfer in [4] and [5] and demonstrated significant improvements. However, as highlighted in [5], cross lingual adaptation attempts to cater for different phenomena to those encountered when adapting Speaker Independent (SI) models for

Speaker Dependant (SD) applications. In a cross lingual setting the adaptation is SI to SI and the original acoustic models do not model the expected phonetic contexts very well which is in contrast to the SI to SD in a monolingual setting. Additionally the acoustic variation across languages is much greater and more complex than same-language variation. As a result, even when MLLR and MAP adaptation techniques have been used, the performance improvement has not been as great as monolingual counterparts.

This performance gap is partially the result of the differences between the phonemic inventories of the source and target languages, as well as the different contexts which can occur. This produces a number of inter-related problems which are outlined below in Sections 2.1, 2.2 and 2.3

2.1. Influence of context on state distributions

To explain this effect, consider the binary decision tree training process. A tree is built for each state of each phone in the phonemic inventory for a particular language. For each state, each parent node is repeatedly partitioned by selecting the question which provides the greatest increase in log-likelihood for the data. The initial questions are typically broad, such as "*is the left context a Fricative*". These questions have the most influence on the overall location and shape of the state distribution and the parameters used to model it. Questions asked further down the tree algorithmically achieve smaller improvements in log-likelihood. These serve to refine the shape of the distribution and its ability to represent the frames associated with each context, but have less significant impact on the principal components, and accordingly the model parameters.

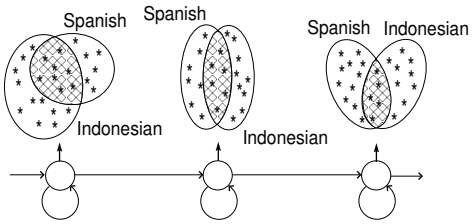

Fig. 1. Impact of Context on State Distributions

Analysis of the decision trees produced using English, Spanish and Indonesian speech revealed that many of the phonemes with the *same IPA symbol*, had significantly different initial questions across languages. For instance in Indonesian, for the phoneme /a/, the first question asked is "*Is the Right Context a Nasal"*. In contrast, for the Spanish language the first question asked is '*Is the left context silence or pause*. The influence of these initial questions will propagate through to the terminal nodes, and subsequently, significant differences in the state distribution for the same context-dependant phone from different languages can exist. A *fabricated* illustration of this effect is depicted in Figure 1. Given that cross language transfer attempts to use source language models to recognise target language speech, some means of reducing this effect is required and a method which seeks to achieve this is outlined in Section 3.

2.2. Allophonic Variants and Missing Phonemes

For each phoneme in the target language, a representative from the source language is typically selected using either knowledge or data-driven techniques. In [5] it was highlighted that when differences between the recording conditions of source and target language exist, knowledge driven techniques provide more robust mappings, and we have a observed similar phenomena in our research. However, knowledge-driven techniques are based on selecting the closest IPA representative, using articulatory descriptions. This effectively enforces a one-to-one mapping,and this hard decision can be inappropriate in certain circumstances. For example, Indonesian uses only one *phoneme* to represent high-front vowels ("**i-sounds**"), but has two distinct context-dependant *allophonic* variants. These variants coincide with the English phonemes /i:/ (eg. **beat**) and /I/ (eg. **bit**). Regardless of which English phoneme is selected to represent the Indonesian sound, it will provide a suboptimal representation for the Indonesian allophone that corresponds to the discarded English mapping candidate. A similar sub-optimal representation will also occur whenever a target language phoneme has no representative in the source language phonemic inventory, necessitating the selection of an alternate.

It is well known that context has a significant influence on the acoustic realisation of each phoneme, especially when the recognition task is conversational speech. Accordingly, the realised versions of some phonemes with the same contexts produce similar features, making discrimination difficult. However, this similarity can be exploited to provide a means for obtaining more appropriate substitutions for missing phonemes, and allow allophonic variants in the target language to be represented using the most appropriate source language data. This is achieved by incorporating more recent ideas to emerge from pronunciation modelling research outlined in [6][7]and [8].

The overarching theme espoused by these studies was based on modelling predictable variation implicitly via the acoustic models, rather than explicitly via the lexicon. This was achieved by tying some, or all, of the mixture components from the state distributions, for phonemes with similar contexts which exhibit similar features. These studies originally focused on monolingual applications for English and Mandarin, however in concurrent work [9], we extended these concepts to the Indonesian and Spanish languages, achieving marginal, but significant improvements. More importantly, a more appropriate means for dealing with the problems discussed in this section was discovered.

2.3. Context mismatch

The idea of *context mismatch* was first highlighted by Schultz in [10] and refers to the large number of contexts in the target language which are not represented by context-dependant source language models. To overcome this problem Schultz proposed a data augmentation technique called Polyphone Decision Tree Specialisation (PDTS)[10]. Schultz's work presented a method for producing more accurate models for those dominant contexts in the target language which did not occur in the source language. In PDTS, context-dependant, *source* language acoustic models are built initially using the standard state-tying paradigm outlined in detail in [11]. At the completion of this process, the decision tree was then extended using approximately 25 minutes of target language speech. Schultz reported that this technique provided significant gains when applied to cross lingual transfer of multilingual models to the Portuguese language. However, a number of improvements can be made to this technique. These will be outlined and incorporated in the system discussed in Section 3.

The splitting criteria for binary decision trees is typically based on selecting the question which maximises the average log-likelihood weighted by the state occupancy. Given that a limited amount of target-language data is available, many of the context-dependant phones have few, if any, examples from which to estimate robust model parameters, and are less likely to produce ad-

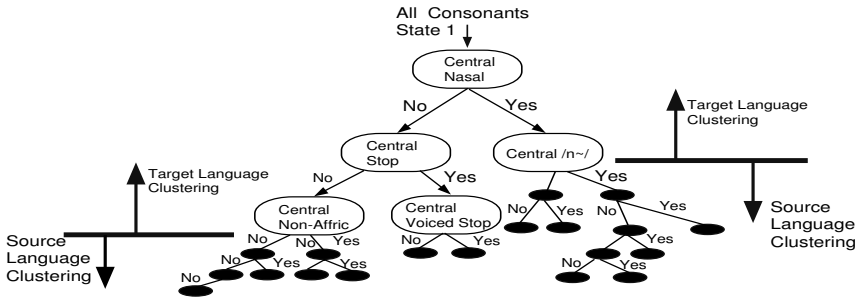

Fig. 2. Example Decision Tree Incorporating Proposed Hybrid Technique

ditional nodes in the original tree. As outlined in Section 2.1, the early questions in the decision tree process have the most influence on the shape and location of the state distributions. Given this, the target language data could be used more productively by influencing the *initial* splits in the tree, rather than extending the tree as in PDTS. The source language models can be used to refine the acoustic models for more specific contexts. This prevents any dominant broad class contexts from the source language subsequently changing the gross target language distribution, but still uses the more plentiful source language data for producing models for more specific contexts, and with higher mixture components.

3. PROPOSED HYBRID TECHNIQUE

The previous sections have highlighted that any cross language technique needs to:

- make more effective use of target language adaptation data,
- constrain the source language so that state distributions reflect co-articulatory occurrences in the target language,
- provide a means of synthesis for unseen contexts and missing phonemes which incorporates both knowledge and data driven information and provides immunity from channel and recording differences,
- reduce the impact of pronunciation variation.

To achieve this a binary decision tree process is again employed, but with several implementational differences. Rather than creating a tree for each individual state of each phoneme, phonemes are initially grouped according to whether they are classed as either a *vowel* or *consonant*, and which state they belong to in a 3 state topology. Accordingly, 6 decision trees are built. Models for noise, silence, pauses etc were trained separately. Questions can also be asked about the actual base monophone, allowing states from different phonemes to cluster, if the realisations exhibit similar features. This section of the technique is similar to that proposed by [8] for reducing the impact of pronunciation variation on *Switchboard* English.

To ensure that the state distributions of the final model set more accurately reflects the target language requirements, the target language is used first to build the formative branches in the tree as outlined in Section 2.1. The source language can then be filtered through this tree with the terminal nodes of this tree become the starting nodes for subsequent source language tree extension. Importantly, the knowledge contained in this tree will be immune to the differences in recording condition and channel effects. The final tree can then be used to synthesise missing context-dependant phones, taking into consideration that more suitable substitutes may come from a different base phoneme with similar contexts.

Figure 2 illustrates this technique for the first state of the consonant tree. For illustration purposes, only a few of the questions are included for the target language. Using 25 minutes of data resulted in basic segregation into monophone classes with some phonemes clustered together in pairings such (t,d), (b,p) (a,&) but only in certain contexts. We observed that using more target data served to refine these groupings, as expected, providing more resolution for the target language requirements.

4. EXPERIMENTS AND RESULTS

The experiments conducted in this paper made use of data from the *Switchboard-1*(SWB-1-ENG) corpus, the 1996 HUB5 evaluation Spanish data (HUB5-SPAN) and Indonesian speech from the Oregon Graduate Institute Multi-language Speech Corpus. Transcriptions for the 3 hours of Indonesian acoustic data were produced in-house and included all utterance categories such as stories, age, routes, climates etc. We used a subset of a commercially produced 20 000 word Indonesian lexicon. 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.

Both the SWB1ENG and HUB5SPAN databases are transcribed at the utterance level. Utterances which caused difficulty in training such as non-Spanish speech in HUB5SPAN and excessive background noise were removed from the data to provide a training database of 10 hours for the Spanish data and 160 hours for the English data. Segment boundaries were modified to split at long pauses and any extended silence.

Context-dependant, HMM acoustic models were trained for all languages using a 3 state left-to-right model topology. Speech was parameterized using 12th order PLP analysis plus normalized energy, 1st and 2nd order derivatives, and a frame size/shift of 25/10ms. Cepstral Mean Subtraction (CMS) was employed to reduce speaker and channel mismatch.

Two hours of the Indonesian speech was used to train a baseline system for comparison purposes. Empirical testing established that 8 mixture component, *context-dependant* models achieved the best word recognition of **58.3**%. 16-mixture context-dependant models were used for the English and Spanish experiments. A bi-gram language model was trained using the Indonesian training data. To prevent problems with Out-of-Vocabulary (OOV) words, those words in the Indonesian test set which did not appear in the training data were assigned a small probability in the language model. Due to limited language model training data, these results will be suboptimal, however the comparison of acoustic model performance is still relevant. Word recognition accuracy results are presented in Table 1. **Know+STD** refers to a knowledge driven mapping technique based on IPA representation, in conjunction with the traditional context-dependant model training paradigm. **NEW-Tech** refers to the proposed technique outlined

		Word Recognition Accuracy	
		No Adaptation Data	With MLLR/MAP
Spanish Models	Know+STD	30.94	37.95/45.33
	NEW-Tech	35.65/42.92	44.06/48.40
English Models	Know+STD	23.09	36.54/45.59
	NEW-Tech	11.21/15.83	40.23/48.81

Table 1. *Cross Language Word Recognition Accuracy Results Using Spanish and English to decode Indonesian*

in Section 3.

Results annotated in column 3 outline those achieved without any adaptation data to enable observation of the impact of mismatched recording and channel conditions. For the **New-Tech** in Column 3, two figures are listed. The first figure reflects using 25 minutes of target data to train the initial part of the tree, and the second reflects the use of 90 minutes. Column 4 provides results using one pass of global adaptation followed by MLLR/MAP adaptation using 25 and 90 minutes of adaptation data respectively.

5. DISCUSSION

It had initially been intended to provide an additional comparison between the proposed technique and the PDTS technique. However, in replicating this technique no additional performance improvement was observed beyond that achieved using the **Know+STD** technique. Further analysis revealed that using PDTS produced very few additional clusters using the target language adaptation data. This may be caused by the Indonesian language, which has a much smaller phoneme set, and importantly less context mismatch, however initial clustering experiments reveal a similar phenomena using English and Spanish source/target combination.

The unadapted data, as expected, produced results which highlight the negative influence that mismatch in channel and recording conditions can have. The Spanish data appears to have channel conditions which are more closely aligned with Indonesian and this is reflected by the unadapted result of both the **Know+STD** (30.94%) and **NEW-Tech** (42.92%) being much higher than the English counterparts. Additionally, the unadapted English **NEW-Tech** results (15.83%) are worse than the unadapted standard technique (23.09%). However, after adaptation, the **NEW-Tech** outperforms the standard technique for both languages. This suggests that channel effects can mask the improved state distributions of the proposed technique.

Using only 25 minutes of target data for the initial stages of tree growth provides improvement for both languages in the Spanish experiments in comparison to the **Know+STD** technique, both before and after adaptation. The amount of improvement achieved by the English data however is only 3.07% compared to 6.31% for Spanish. Both Indonesian and Spanish vowel sets are much smaller than English and it is our contention that 25 minutes of target data is enough to separate the Spanish data in to clusters which will not adversely affect the Indonesian state distributions. However in the English case, the increased number of phonemes that are allophonic in Indonesian may not be segregated sufficiently using only 25 minutes of data thereby adversely influencing the state distributions of the final models. Increasing the tree-training data to 90 minutes data again results in improvements over the adapted **Know+STD** models from both languages(Spanish 3.07% and English 3.22%), thereby indicating that further information is extracted from the tree based structuring of the state distributions. However, the proposed technique, as with many previously reported methods, still fails to outperform the baseline system, indicating that more research is still required before cross language transfer can be used successfully to produce an LVCSR system on telephone speech.

6. CONCLUSION

In this paper a new technique was proposed for improving the accuracy of acoustic models used for cross language transfer. This technique was evaluated in an Indonesian word recognition task using English and Spanish acoustic models. Evaluations highlighted that the proposed technique makes more productive use of limited target language data by producing state distributions for context-dependant models which more accurately reflect the prominent co-articulatory effects that occur in the target language.

7. ACKNOWLEDGEMENTS

This research was supported by the Office of Naval Research (ONR) under grant N000140310662.

8. REFERENCES

[1] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modelling," in *Speech Communication*, February 2001, vol. 35, pp. 31–51.

[2] J. Kohler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proc. ICASSP*, Washington, U.S.A, 1998, vol. 1, pp. 417–420.

[3] B. Imperl, Z. Kacic, B. Horvat, and A. Zgank, "Clustering of triphones using phoneme similarity estimation for the definition of a multilingual set of triphones," *Speech Comm.*, vol. 39, pp. 101–113, 2003.

[4] P. Fung and M. Chi Yuen, "Map based cross language adaptation augmented by linguistic knowledge: From english to chinese," *in Proc. Eurospeech*, 1999.

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

[6] M. Saraclar and H. Nock, "Pronunciation modelling by sharing Gaussian densities across phonetic models," in *Computer Speech and Language*, 2000, vol. 14-2, pp. 137–160.

[7] P. Fung and L. Yi, "Triphone Model Reconstruction for Mandarin Pronunciation Variations, Hong Kong ," in *Proc.Eurospeech*, 2003, vol. 1, pp. 760–763.

[8] H. Yu and T. Schultz, "Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition," in *Proc. Eurospeech*, Geneva, Switzerland, 2003, pp. 1869–1872.

[9] T. Martin and S. Sridharan, "Target Structured Cross Language Model Refinement," in *Proc of 9th Australian Int. Conf on Speech Science and Technology*, Sydney, 2004.

[10] T. Schultz and A. Waibel, "Polyphone decision tree specialization for language adaptation," in *Proc. of ICASSP, Istanbul 2000*, 2000.

[11] Julian James Odell, *The use of Context in Large Vocabulary Speech Recognition*, Ph.D. thesis, Queens College, Cambridge, 1995.