Target Structured Cross Language Model Refinement

Terrence Martin, Kishan Thambiratnam and Sridha Sridharan

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

Abstract

The task of porting Automatic Speech Recognition (ASR) technology to many languages is hindered by a lack of transcribed acoustic data, which in turn prevents the development of accurate acoustic models necessary for the recognition task. To overcome this problem, recent research has sought to exploit the similarity of sounds across languages, and use this similarity to adapt models from one or more data rich languages for use in recognising data poor *target* languages. Pronunciation variation and cross language context mismatch combine to make the task more difficult then a monolingual ASR application. In this paper, we examine the utility of recent pronunciation modelling approaches and evaluate their performance on the Indonesian and Spanish languages. Finally, we introduce a novel technique which ensures that the state distributions developed using the source language data are more closely aligned with those in the target language, thus improving classification accuracy. This technique achieved an improvement in word recognition accuracy of 19.5% absolute percentage points, when compared to standard knowledge based cross lingual mapping approach.

1. Introduction

Large Vocabulary Continuous Speech Recognition (LVCSR) systems are currently only available for few of the worlds languages. The reason for this is that the resources required to produce an Automatic Speech Recognition (ASR) system for a new language are considerable. Of all the resources required, obtaining sufficient transcribed acoustic data and lexicons represent a significant outlay in terms of both manpower and cost. Accordingly, this research effort extends to the development of generic techniques which can use available resources to produce (ASR) applications in data poor languages.

One method for utilizing resources from a data rich *source* language to produce an ASR capability in a data poor *target* language is to exploit the similarities in the acoustic realisation of sounds across languages. This process involves the creation of an acoustic model set using one or more of these source languages, and then judiciously transforming them to perform target language ASR. Evaluating the performance is problematic. Ideally we desire the performance achievable by ASR systems for data rich languages, however assessing this is impossible given the limited target data.

Currently there is interest in providing ASR applications for the Indonesian language, so a secondary research focus is to use the developed methods for achieving respectable recognition rates for the Indonesian language.

In (Schultz and Waibel 2001) it was shown that multilingual models improve recognition when used for bootstrap training, however whether monolingual or multilingual source language acoustic models are superior when used in cross lingual decoding is largely dependant on the target language, its similarity with available source languages (Kohler 1998) and the degree of phonemic and contextual coverage provided by the pool of languages. In previous research (Martin and Sridharan 2002) we found that the Spanish language, when compared with English and Hindi, provided greater phone set similarity and cross language recognition performance. In contrast to Schultz's findings we also found that Spanish provided superior performance to that achieved by multilingual models. Accordingly, we use the Spanish language as a source language for the cross- lingual experiments outlined in this paper.

The experimental framework used in this paper is based on the availability of a limited amount of data (1-2 hours) from the target language. This training data set size is sufficient to produce a rudimentary set of context independent acoustic models or to use as adaptation data. However the number of gaussian mixture components which can be used to model each state distribution is limited by data sparsity issues. As a result these models may not provide robust performance, particularly in mismatched conditions.

Most reported studies for multilingual and cross-lingual ASR have focused on read speech applications, (Schultz and Waibel 2001), (Nieuwoudt and Botha 2002), or limited vocabulary applications (Kohler 1998). This paper however presents results for word recognition over the telephone. Typically, when studies have been extended to telephone speech, they have largely focused on assessing acoustic model accuracy via phone recognition experiments (Walker, Lackey, Muller, and Schone 2003).

The underlying reason for this is that cross language transfer is a difficult task, and the degree of recognition performance achieved is extremely modest, when compared to the low Word Error Rates (WER) rates achieved in languages such as English. For instance, recent recognition performance as outlined in (Hain, Woodland, Everman, Liu, Povey, Wang, and Gales 2003), report WER less than 20% for Switchboard telephone speech. In comparison, similar performance is reported for the much simpler tasks such as constrained vocabulary or read speech applications.

Transformation based (MLLR) and Bayesian based

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

(MAP) adaptation techniques are commonly used to minimise the train-test mismatch for both channel normalisation and adapting speaker independant models for particular speakers. This approach has also been used for cross lingual transfer, and provides definite improvements. However, as highlighted in (Nieuwoudt and Botha 2002), cross lingual adaptation is a more difficult task and the adaptation process is attempting to cater for different phenomena to those encountered when adapting Speaker Independant (SI) models for Speaker Dependant (SD) applications.

In a cross lingual setting the adaption is SI to SI and the original acoustic models do not model the expected phonetic contexts very well, in contrast to the SI to SD in a monolingual setting. The acoustic variation across languages is much greater and more complex than same-langauge variation. Additionally, the source language may not have a phonetic counterpart to represent phonemic events in the target language. This becomes more likely when context dependency is required.

In well refined recognition engines, some of the variation which occurs is modelled by introducing additional entries for pronunciation in the lexicon as well as implicitly modelling variation by incorporating context into the acoustic models. It is the areas of pronunciation variation and cross language context mismatch which combine in an integrated manner to hinder the success of cross lingual transfer.

Traditionally, attempts to reduce the impact of pronunciation variation were based on including additional variants in the lexicon (Cremlie and Martens 2000)(Holter and Svendsen 1999), however more recently, work outlined in (Saraclar and Nock 2000), (Fung and Yi 2003), (Yu and Schultz 2003) have deviated from this approach. These studies looked to model the variation implicitly via the acoustic models and minimise the changes required in the lexicon. In this approach, the mixture components are tied based on similarity, regardless of the base phoneme. This approach provided modest but significant improvements for English and Mandarin.

Accordingly, further experiments were conducted to establish the utility of this technique across different languages by evaluation on Indonesian and Spanish. However, the underlying motive for selection of this technique was based on its ability to provide a structured means for accurate synthesis of contexts which occur in the target language, but do not exist in the source language.

To complement this process, a new technique based on the novel work outlined in (Yu and Schultz 2003) is presented in Section 3. This method seeks to structure the mixture component tying of source language models to suit the coarticulatory requirements of the target language, rather than the source language, which has traditionally been the case.

The paper is organised as follows. Section 2 outlines the rationale behind modelling pronunciation variation implicitly via the acoustic models and its benefits to Cross lingual acoustic models. In Section 3 the use of Target Structured Cross Language Model Refinement to capture the coarticulatory effects for the language is explained. Experimental results using both the CallHome Spanish database and OGI Indonesian database for separate recognition tasks in Spanish and Indonesian, as well as a cross lingual recognition of Indonesian using Spanish models are presented in Section 4. Discussion and conclusions drawn from these results are provided in Sections 5 and 6.

2. Pronunciation Modelling

The training of acoustic models in ASR systems is typically conducted using the canonical transcription of the utterances. However in conversational speech, the realised versions of utterances quite often differs from this canonical transcription, and accordingly when decoding is conducted, recognition errors occur because of insertion, substitution and deletion errors. This is quite often the by-product of co-articulation, and certain phenomena such as nasalisation, flapping, voicing, and centralization occur frequently because the vocal tract configuration changes relatively slowly and the realisation of the intended phone is influenced by both the previous and target sounds. This variation can be modelled implicitly via acoustic modelling, or explicitly by including additional entries in the lexicon. The appropriate combination of the two modelling techniques perhaps provides the key to minimising the impact of this effect.

In (Fung and Yi 2003) it was highlighted that more common instances of insertions and deletions should be included in the lexicon. However, the inclusion of substitutions requires more subtlety, as for certain contexts there is little or no difference between the features of *baseform* and *surface* form phonemes. Consider the example below providing *Worldbet* pronunciation annotation for the word *explosion* :

EXPLOSION *I k s p l oU Z & n*
EXPLOSION *I k s b l oU Z & n*

The voicing of the phone $/p/$ is a common phenomenon in this context. Another example is the flapping of $/t/$ in the word *Better* where the $/t/$ is replaced with $/d/$. Given the multitude of possibilities, traditional lexical editing is based on producing a set of rules, such as:

$$rule : s - p + l \Rightarrow \dot{b} \;\; with\, probability\ X \tag{1}$$

Producing these rules however has met with limited success and it was suggested in (Yu and Schultz 2003) that the reason for this is that inappropriate rules result in contamination of the training data for models of both $/p/$ and $/b/$, and this subsequently increases confusion at recognition time. To overcome this phenomenon, (Saraclar and Nock 2000) (Yu and Schultz 2003), and (Fung and Yi 2003) have proposed similar variants of the same basic technique. This *base* technique models context without placing constraints on the base phoneme, allowing tying of mixture components which exhibit similar features. The motivation for this approach is that even the best classifier cannot distinguish between classes if the features are the same, and so it is better to identify those instances and tie the mixture components.

The techniques proposed in these studies are based on manipulating the traditional method for modelling context

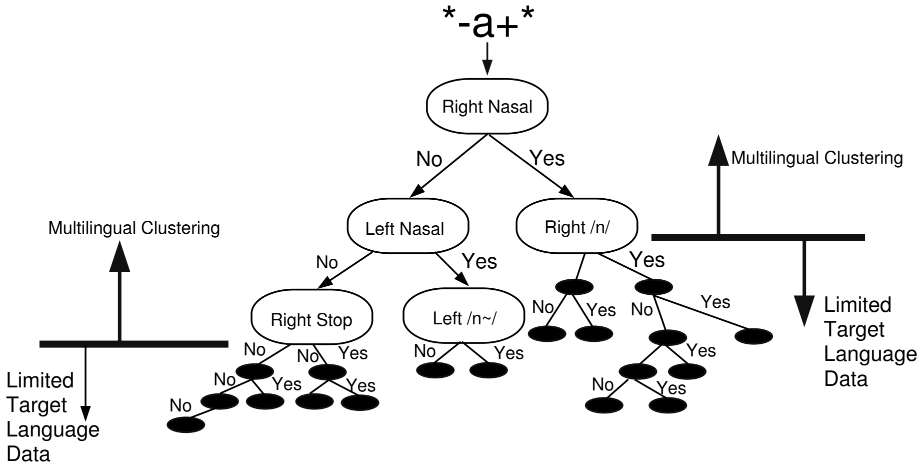

Figure 1: Basic System Overview

outlined by (Odell 1995). This process is a well refined training paradigm where a decision tree is grown for each state of the monophone and is depicted by the upper part of Figure 1 delineated by the bold lines. The lower part of Figure 1 will be further explained in Section 3.

Questions about left and right context are asked, and the selection of which question is the most appropriate is based on a maximum log-likelihood criteria. However, this paradigm prevents the clustering of phones with similar realisations, a phenomenon which we are trying to capture.

In the traditional decision tree clustering approach clustering is performed on a per-state basis, for each phoneme, however we start the clustering process by grouping the phones into either vowels, or consonants and then grow the tree for each state (Models for noise, silence, pauses and other non speech sounds were trained separately). In this way there are 6 trees grown for the vowels and consonants, one for each state in a 3 state HMM topology. This approach also differs from the traditional approach in that questions can be asked about the actual base monophone, and accordingly phones with different base phones, but similar features can be clustered together.

In this paper we have adopted a variant of the technique proposed by Yu in (Yu and Schultz 2003), to circumvent this problem. Performance improvements were obtained using this technique for English in (Yu and Schultz 2003) and (Saraclar and Nock 2000), and a similar variant of this technique for Mandarin in (Fung and Yi 2003). We wanted to confirm the utility of this technique across languages and accordingly an evaluation was conducted using this method for Spanish and Indonesian word recognition and comparing it to the standard context dependent model training method.

This approach has an important application in cross lingual transfer. One of the difficulties involved with cross lingual transfer is the cross language phone mapping process. For example, in Indonesian the /i/ and /I/ difference is an allophonic distinction, whereas in English this is a phonemic distinction. That is, in Indonesian the two ”i” vowels (/i/ and /I/) cannot be used to make a contrastive meaning. The realisation of allophonic variants is typically based on whether the phoneme occurs in an open or closed syllable. Representation of allophones in lexicons do occur but is not common, mostly because allophones are variable. Clearly the most appropriate mapping cannot be achieved using one source language representative.

When a source language does not have a representative for a phone, the mapping conducted is usually based on monophone similarity, via either a knowledge or data driven means. Thus, the mapping is one-to-one, however, this may not be entirely appropriate in certain contexts, as was illustrated with the allophonic variants of the target phoneme $/i/$. By allowing phonemes which exhibit similar features in given contexts to cluster, an alternative selection can be made when no representative for a target language phone exists in the source language. A limiting factor to this is limited target data which will mean that the clusters will contain phones which are grouped according to broad contexts and are not particularly refined. However in Section 3 we outline a novel technique for refining the model accuracy using the source data and allow more accurate synthesis.

3. Capturing Context in Cross Lingual Environment

In (Kohler 1998) and (Schultz and Waibel 2001) it has been shown that context mismatch across languages is a significant impediment to the success of cross language transfer. To overcome this impediment, (Schultz and Waibel 2000) proposed the Polyphone Decision Tree Specialisation technique (PDTS) to reduce the mismatch between represented context in the source model set and observed polyphones in the new language.

In PDTS, the initial tree growing process is conducted individually for each state using data from one or more source languages, using the traditional context dependant modelling paradigm. This process is depicted in the top

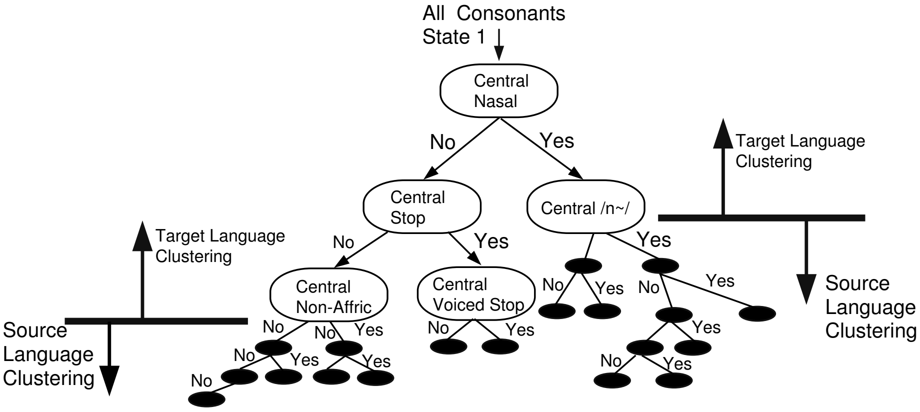

Figure 2: Basic System Overview

half of Figure 1 which is delineated by the bold lines. A limited amount of target language data is then used to further propagate growth of the source language decision tree, so that the dominant (missing) contexts from the target language are also incorporated in the model set, thus reducing the impact of context mismatch. Schultz reported that this technique provided significant gains when applied to cross lingual transfer of multilingual models to the Portuguese language.

In this paper we deviate from this technique in two ways. Firstly we incorporate the technique outlined in Section 2 using vowels and consonants as the starting points for the clustering process. This reduces the negative impact of pronunciation variation and provide scope for synthesis of unseen contexts. Secondly, we use a novel technique for more accurately estimating the state distributions for the target language.

As we have mentioned, there is a large degree of context mismatch between the source and target language. Even for the same monophone, the state distributions will exhibit different model parameters in terms of mean and variance. Importantly the eigenvectors for these distributions are different. That is, if Principal Component Analysis was conducted for these distributions, the eigenvalues and eigenvectors would be be different (Kohler 1996).

The cross lingual differences in state distributions is largely a by-product of the contexts that are dominant in that language. Analysis of the decision tree clustering process conducted using the traditional method for individual monophones revealed that even the first few questions asked exhibit significant differences across language. 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*. Obviously, the child node distributions will exhibit significant differences between the languages in this case. Further propagation down the tree will 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 model parameters the further down the tree you progress. Thus it seems counter-intuitive to use a tree built in a source language, where the most dominant contributor to the state distributions are tailored for representing the source language. More effective use can be made of the limited target language adaptation data by using it to establish the most dominant contexts, even if they are only established at the broad phonetic level. These clusters can then be used as starting nodes for model refinement using the more readily available source data.

Given data restrictions, even if only two child nodes in the target language can be robustly estimated for each monophone, these will be more appropriate for clustering then the first two nodes formed by the source tree. The robustness of the child nodes formed using the target language can be controlled by setting the minimum occupancy threshold to a value which ensures that the clusters formed have sufficient data for model training.

We highlight however, that our starting nodes are actually predetermined not using monophone tree growth, but using the technique proposed by (Fung and Yi 2003) and (Yu and Schultz 2003), and can actually include different phonemes, because we started the tree growth in the target language from the broad classes of vowels and consonants.

Thus in Figure 2 it can be seen that we use the limited target data to establish the most dominant clusters and contexts in the target language. These clusters then act as starting clusters for the source data to continue to propagate the tree. Effectively this acts to constrain the shape of the final distribution of the source data models so that it more closely resembles what we could expect in the target. The source data obviously has much more data, and can then be used to produce more refined model estimates, with a higher number of mixture components allocated to each model then could be trained in the target language.

Finally this tree can then be used to synthesise the models required for the target application. This provides the advantage that the mapping estimate will be more appropriate, and will have a state distribution that better approximates the target language then that achieved using knowledge based mapping.

We call this approach Target Structured Cross Langauge

Model Refinement(TSCLMR) and the process can be summarised as follows:

1. Use the broad classes of vowel and consonant to build 6 trees using target language data, 1 for each state as outlined in Section 2;
2. Use these *target* derived trees, and associated questions, to subsequently cluster the source data;
3. Use each cluster established in 2 as starting points from which extended growth can be achieved using more plentiful source data,
4. These extended trees can be subsequently used synthesise any unseen phonemes from the target language.

4. Experiments and Results

The experiments we performed were broken into two categories. The first set of experiments was undertaken to quantify the improvements in Indonesian and Spanish ASR using the technique outlined in Section 2 to reduce the impact of pronunciation variation.

Language	Indonesian %A	Spanish % A
Traditional	54.65	31.12
Yu Technique	55.96	31.91

Table 1: *Word Recognition Accuracy for Monolingual ASR task*

The second set of experiments sought to establish the suitability of TSCLMR as a technique for obtaining improved Cross Language transfer. We used Indonesian speech data taken from the Oregon Graduate Institute Multi-language Speech Corpus. 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. For the experiments conducted in this paper the speech data was split into a training set (1.5 hrs) and a test set(40 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 (Martin and Sridharan 2002). 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.

To produce the Spanish models we used the Spanish data used for the 1997 HUB evaluations (NIST 1997), taken from the Callhome Database. This data is transcribed at the utterance level. Some utterances which caused difficulty in training such as non-Spanish speech and excessive background noise were removed from the data to provide a training data base of 10 hours and a test set of 40 minutes. The speech contained in the Spanish data represents an extremely difficult recognition task, with telephone conversations taking place between friends over a 30 minute period.

Using this data, HMM acoustic models were trained for Spanish and Indonesian. We used a 3 state left-to-right model topology for both languages. For the Indonesian speech we experimented with 8, 16 and 32 mixture components to model the state emission densities, but found that best results were obtained with 8 mixtures and accordingly the results presented here are for 8 mixtures. For the Spanish models we used 32 mixture components. 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. For both languages we used the training data to train a bigram language model.

Table 1 provides word recognition accuracy statistics for both languages using the standard approach for context modelling as well as the approach designed to minimise the impact of pronunciation variation.

Technique Description	Word Recognition %A
Dir.Map	23.7
Dir.Map+VIT	32.74
SPANCLUST+VIT	33.91
TGTCLUST+VIT	**50.05**
TGT+SCECLUST+VIT	**51.11**
TSCLMR+VIT	**52.29**

Table 2: *Using Cross Lingual Adapted Spanish Models to Decode Indonesian Speech*

In table 2 we provide results depicting word recognition accuracy using Spanish acoustic models to decode Indonesian speech using the various techniques we have outlined. *Dir.Map* refers to a knowledge based mapping from the source language based on corresponding IPA symbols. As can be seen this provided poor recognition results, made worse by channel mismatch. To overcome this we conducted a single pass of Viterbi based adaptation (Schultz and Waibel 2001) using 20 minutes of Indonesian speech, which provided nearly 10% in absolute improvement. For all other experiments shown in Table 2 we conducted Viterbi adaptation.

SPANCLUST+VIT refers to models which used the pronunciation modelling technique to cluster the Spanish phones followed by adaption. *TGTCLUST+VIT* refers to the use of the target tree to cluster the Spanish data while *TGT+SCECLUST+VIT* builds on this process by then continuing to grow the tree using source language data. Finally *TSCLMR* uses this tree to allow synthesis of triphones that exist in the target language but which have no counterpart mapping in the source language.

5. Discussion

The results presented in Table 1 reinforce that the technique outlined in Section 2 can be used for other languages and can help to reduce the impact of pronunciation modelling. An examination of the decision tree grown using this technique revealed that in the majority of cases the vowels from both languages formed separate clusters. This may be due to the fact that both languages have a relatively small phonemic vowel set and accordingly the features are quite distinct. Analysis of the consonant cluster however revealed expected clustering of central phones such as ($/t/$, $/d/$), ($/p/$, $/b/$), ($/g/$, $/k/$). One interesting clustering was that of ($/k/$, $/?/$). Glottal stops (denoted $/?/$) are a common pronunciation variant in Indonesian for the phoneme $/k/$ when it occurs syllable finally.

As a consequence of limited language model training data the recognition results presented are suboptimal, however the comparison of acoustic model performance is still relevant. The recognition rate for the Spanish data is quite low. However, in the 1997 HUB5 Non-English Evaluation (NIST 1997), the reported recognition rates for Spanish outlined an increase from the previous evaluation from 27% to 34%. Some time has evolved since then, however our Spanish recognition reflect a similar level of performance. The Indonesian results, however, are quite promising given the limited amount of data.

The most impressive result to emerge from the experiments outlined in Table 2 is the improvement of 19.55% absolute percentage points, in comparison to standard mapping techniques. This was achieved using the limited target data to constrain the starting questions for subsequent model refinement using the source data. It should be noted that the most significant increase in absolute performance 17.3% (32.74 to 50.05) was achieved by simply using the target based tree to determine the starting clusters for further source language tree growth. This serves to reinforce our contention that the dominant distributional effects can be captured from the target language by utilising the early broad contextual questions, even with limited data.

Minor improvements (1.06%) were obtained by subsequently refining the acoustic models (TGT+SCECLUST+VIT) and the incorporation of synthesis using the final source refined tree (TSCLMR) also an additional benefit of 1.18% absolute improvement.

6. Conclusion

In this paper we introduced a novel technique for improving the accuracy of acoustic models used for cross language transfer. This technique produced an absolute word recognition improvement of 19.55% in comparison to standard knowledge based mapping. Experiments were also conducted based on previous research to validate the idea of allowing the clustering central contexts in the decision tree process. Results reflected those achieved for other languages. More importantly the use of this technique allowed for a structured means of conducting synthesis of target triphones, when no source data equivalent is available. This also provided additional improvements.

7. Acknowledgements

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

References

Cremlie, N. and J. Martens (2000). In Search of better Pronunciation Models for Speech Recognition. In *Speech Communication*, Volume 29, pp. 115–136.

Fung, P. and L. Yi (2003). Triphone Model Reconstruction for Mandarin Pronunciation Variations . In *Proc. ICASSP*, Volume 1, Hong Kong, pp. 760–763.

Hain, T., P. Woodland, G. Everman, X. Liu, D. Povey, D. Wang, and M. Gales (2003). Automatic transcription of conversational telephone speech - development of the CU-HTK 2002 system. Technical report, Cambridge University.

Holter, T. and T. Svendsen (1999). Maximum likelihood modelling of pronunication variation. In *Speech Communication*, Volume 29, pp. 63–66.

Kohler, J. (1996). Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proc. ICSLP*, Volume 4, Philadelphia,U.S.A, pp. 2195–2198.

Kohler, J. (1998). Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In *Proc. ICASSP*, Volume 1, Washington, U.S.A, pp. 417–420.

Martin, T. and S. Sridharan (2002). Cross Lingual Modelling Experiments for Indonesian. In *Proc of 8th Australian Int. Conf on Speech Science and Technology*, pp. 184–189.

Nieuwoudt, C. and E. Botha (2002). Cross language use of acoustic information for automatic speech recognition. In *Speech Communication*, Volume 38, pp. 101–113.

NIST (1997). The 1997 Hub-5NE Evaluation Plan for Recognition of Conversational Speech over the Telephone in Non-English Language. $http : //www.nist.gov/speech/tests/ctr/hub5ne97 /current - plan.htm$.

Odell, J. J. (1995). *The use of Context in Large Vocabulary Speech Recognition*. Ph. D. thesis, Queens College, Cambridge.

Saraclar, M. and H. Nock (2000). Pronunciation modelling by sharing Gaussian densities across phonetic models. In *Computer Speech and Language*, Volume 14-2, pp. 137–160.

Schultz, T. and A. Waibel (2000). Polyphone decision tree specialization for language adaptation. In *Proc. of ICASSP, Istanbul 2000*.

Schultz, T. and A. Waibel (2001, February). Language independent and language adaptive acoustic modelling. In *Speech Communication*, Volume 35, pp. 31–51.

Walker, B., B. Lackey, J. Muller, and P. Schone (2003). Language Reconfigurable Universal Phone Recognition. In *Proc. Eurospeech*, Geneva, Switzerland, pp. 153–156.

Yu, H. and T. Schultz (2003). Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition. In *Proc. Eurospeech*, Geneva, Switzerland, pp. 1869–1872.