CROSS-LANGUAGE USE OF ACOUSTIC INFORMATION FOR AUTOMATIC SPEECH RECOGNITION

C. Nieuwoudt and E. C. Botha

Department Electrical and Electronic Engineering, University of Pretoria (South Africa)

botha@ee.up.ac.za

ABSTRACT

Techniques are investigated that use acoustic information from existing source language databases to implement automatic speech recognition (ASR) systems for new target languages for which little data are available. Strategies for cross-language use of acoustic information are proposed and are implemented via maximum *a posteriori* probability (MAP) and transformation-based techniques, as well as via discriminative learning techniques. The discriminative learning technique used is based on a cost-based extension of the minimum classification error (MCE) approach. Experiments are performed using relatively large amounts of English speech data from either a separate database or from the same database as smaller amounts of Afrikaans speech data to improve the performance of an Afrikaans speech recogniser. Results indicate that a significant reduction in word error rate is achievable (between 14% and 48% for experiments), depending on the method used and the amount of target language data available.

1. INTRODUCTION

The development of automatic speech recognition (ASR) systems for the large number of "minor" languages of the world is becoming increasingly relevant as technology becomes more affordable. The standard methods used for constructing speech recognition systems, in particular use of hidden Markov models (HMMs), have been shown to work well for a large number of languages [1], but necessitate a large amount of training data. The collection of large speech databases is an expensive and time consuming process, thereby limiting the development of speech recognition technology for many languages.

Sharing of acoustic information between languages by constructing multilingual phone sets has been researched [2, 3], but results indicate that recognition performance degradation is generally achieved in return for simplified modelling of acoustic parameters and easily integrated multilingual recognition. Only a few studies [3, 4, 5] have considered using cross-language acoustic information for the explicit goal of improving the performance of a speech recogniser in a new target language.

Previously used strategies for the cross-language use of acoustic information include (i) pooling multilingual data to construct explicitly multilingual phoneme models, as used in multilingual systems [2, 3] and (ii) using target language data to adapt models trained on source language(s) [4, 5]. We experiment with two new approaches namely (iii) the training of models on combined source and target language data, followed by subsequent adaptation on target language data only and (iv) the transformation of source language data to augment target language data for model training, followed by target language specific adaptation. Maximum *a posteriori* probability (MAP), transformation-based and minimum classification error (MCE) techniques are used to implement the cross-language adaptation of acoustic models or data. The modelling framework used is that of hidden Markov models in conjunction with Gaussian mixtures for the distribution of acoustic features in each state. Experiments are performed to evaluate the performance of various approaches from the framework, using English speech from the TIMIT database and English and Afrikaans data from the bilingual SUN Speech database [6].

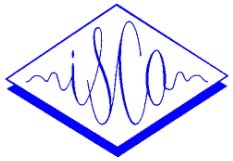
The organisation of this paper is as follows. In Section 2 we discuss different strategies for the cross-language use of acoustic information. In Section 3 issues involving the adaptation of acoustic models across language boundaries are discussed. Section 4 presents experiments to evaluate the various approaches that combine different strategies and algorithms for cross-language use of acoustic information. We conclude in Section 5.

2. CROSS-LANGUAGE USE OF ACOUSTIC INFORMATION

Cross-language use of acoustic information attempts to exploit the acoustic-phonetic similarities between languages. These similarities are evident from the use of international phonetic inventories, such as the International Phonetic Alphabet (IPA), that serve to classify the sounds of many languages. There are still, however, differences with respect to the acoustic properties of sounds from different languages that share the same labels. Also, often a target language may contain sounds that do not occur in languages for which large databases are available. Labelling conventions, recording conditions and the type of speech recorded may also differ between databases, making cross-language and cross-database use of acoustic information a formidable task.

2.1. Language and database issues

When cross-language use of acoustic information is attempted, it is important to use databases of languages that



6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

are as similar as possible to ensure that maximal overlap of phonetic inventory as well as overlap of phonetic context between the languages occur. When more than one database is used, labelling conventions generally differ and a mapping may have to be determined from source language labels to target language labels. For experiments detailed in this paper, a phonetic expert determined a mapping from TIMIT to the SUN Speech database.

2.2. Strategies for cross-language use of acoustic information

The simplest strategy for using acoustic information across language boundaries is to train acoustic models on pooled source and target language data. This may lead to a loss of accuracy compared to using target language data only [7]. However, when only a limited amount of target language data is available and when data from a closely matching language is available, some performance gain has been achieved with multilingual pooling [3].

Adaptation of source language acoustic models using limited amounts of target language data has been shown to improve performance in target language experiments [4, 5]. Complex source language models can be estimated by using the large amount of source language data. These models can then be adapted using possibly limited amounts of target language data.

Model training on pooled source and target language data, followed by target language specific adaptation presents a new approach to use available data. Relatively robust, yet imprecise, multilingual models are trained to begin with and these are "fine-tuned" using target language data. The "tuning" attempts to improves the accuracy of the models w.r.t. target language characteristics without sacrificing the robustness of the multilingual models.

A final strategy for cross-language use of acoustic information is to transform source language data to augment target language data for model training. When using multiple databases for cross-language adaptation this may be of specific interest because differences, other than language, may also be removed as part of the process. Due to the fact that the data transformation may be relatively simple and also does not compensate for the differences in variance between the data sets, target language specific adaptation is performed to further improve performance.

3. ADAPTATION TECHNIQUES

Cross-language adaptation of acoustic models is difficult since acoustic variations across languages may be large and a speaker independent (SI) to SI mapping has to be computed. A number of adaptation techniques are discussed next.

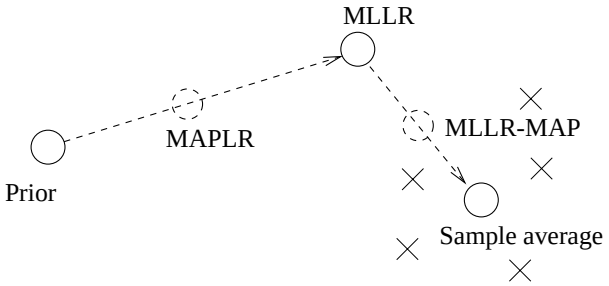


Fig. 1. Graphical comparison of the working of the MAP-MLLR and MLLR-MAP techniques, showing adaptation of the Gaussian mean

3.1. MAP adaptation

Maximum *a posteriori* probability (MAP) adaptation uses prior information about the distribution of parameters to perform improved estimation given a limited amount of observation data. This is particularly applicable in our case, since prior information from existing speech databases can be used to improve model estimates using limited amounts of speech in a new language. In a previous paper [6] simple (single mixture) Gaussian models were estimated. In this paper conditional normal-Wishart prior distributions are used for the mean and variance of Gaussian mixture parameters and Dirichlet priors are used for both the transition probability and mixture weight parameters, as proposed by Gauvain & Lee [8].

3.2. Transformation-based adaptation

The transformation of Gaussian mean components is implemented according to the maximum likelihood linear regression (MLLR) approach [9]. Adaptation of the Gaussian variance components is done separately from the transformation of the Gaussian means. A least squares transformation is calculated in log-space, thereby considering the relative accuracies of the values and also maintaining the constraint that the variance values have to be non-negative. Tieing is used in calculating the transformations by grouping together HMMs according to phonetic categories.

3.3. Combined MAP and transformation-based adaptation

Bayesian and transformation-based adaptation techniques are combined in two different ways in attempts to retain desired properties from both strategies. MLLR-MAP adaptation performs MLLR adaptation in a first step, followed by MAP adaptation. This combines the fast adaptation performance of MLLR with the asymptotic performance of MAP adaptation. The second method defines an *a priori* distribution for the transformation and proceeds to compute the maximum *a posteriori* linear regression (MAPLR) parameters [10]. MAPLR can improve generalisation performance by controlling the amount of adaptation when only a small amount of target language data is available.

3.4. Minimum classification error adaptation

Minimum classification error (MCE) [11] training adjusts model parameters to minimise the classification error rate. We implemented a modified version of MCE by associating a cost with each classification error. The cost (ζ_{kl}) associated with classifying a token from class k as class l is incorporated into the MCE misclassification measure by

$$d_i(X;\Lambda) = -g_i(X;\Lambda) + \log\Big[\sum_{j,j\neq i}^{M} \frac{e^{(\log \zeta_{ij} + g_j(X;\Lambda))\eta}}{M-1}\Big]^{1/\eta}, \quad (1)$$

where $g_i(X;\Lambda)$ is the likelihood function of observation X for class i, M is the number of classes and η is a small integer constant. The advantage of our approach is that the cost (ζ_{kl}) can be determined for target language phonemes and used to improve the generalisation achieved with MCE.

4. EXPERIMENTS

The experiments evaluate the performance of the various approaches that combine strategies for cross-language use of acoustic information from Section 2 with adaptation techniques from Section 3. English speech data from the SUN Speech or TIMIT databases is used as source language data in conjunction with either the (full) Afrikaans training set or the Afrikaans training subset as target language data. Word accuracy for continuous speech recognition is measured on the Afrikaans test set.

Table I summarises the methods that were experimented with and their results, which are discussed in detail in this section. Models trained on SUN Speech English data deliver reasonably good performance (57.9% word accuracy) as opposed to using models trained on TIMIT (-5.9% word accuracy). The results indicate that a large bias exists between the databases, probably aggravated by (unavoidable) inaccuracies in mapping between different phoneme sets. Using only the Afrikaans (target) data delivers 67.6% and 45.0% word accuracy when training on the training set and subset respectively. The performance achieved with the training subset (45.0%) is less than that achieved with the SUN Speech English data (57.9%) and can be attributed to its small size. Same-database (i.e. only using SUN Speech) pooling delivers reasonable improvements upon the baseline Afrikaans only results, 68.1% versus 45.0% for the training subset and 73.3% versus 67.6% for the full training set. When cross-database pooling (i.e. using TIMIT and SUN Speech) is attempted, however, there is no improvement and a performance degradation is achieved for pooling with the training set (55.3% versus 67.6%). Overall, the pooling results indicate that same-database (i.e. same recording conditions and labelling) source and target language data availability facilitates easy cross-language use of speech data, while using more than one database may complicate data re-use.

For same-database MAP adaptation poor results are achieved when only the Gaussian means are adapted, but better results are achieved when all parameters (transition probabilities, mixture weights, means and variances) are adapted, achieving 70.2% and 74.9% word accuracies for training subset and training set adaptation respectively. Even better performance is achieved when full MAP adaptation is done from multilingual models (pooling-MAP), achieving 71.4% and 75.3% word accuracy for training subset and training set adaptation respectively. The training subset result (71.4%) represents a relative reduction in word error rate of 48% over baseline Afrikaans training set performance (45.0% word accuracy). For cross-database adaptation, full MAP adaptation delivers better performance than mean-only MAP adaptation and shows a large improvement over using the Afrikaans training subset only (57.0% versus 45.0%), but shows little improvement (0.1%) over using the Afrikaans training set only. Using multilingual priors for MAP adaptation (pooling-MAP) improves performance over using TIMIT priors for Afrikaans training set adaptation (69.0% versus 67.7%), but degrades performance for the Afrikaans training subset (56.0% versus 57.0%).

Transformation-based adaptation in isolation delivers poor performance and does not even achieve the performance of the pooling approach for same-database experiments, or the performance of training only on target data for cross-database experiments. Transformation of mean and variance (using the least squares log space variance transform) consistently outperforms using mean-only MLLR, but still does not deliver useful performance. For same-database adaptation peak performance is achieved with 2 regression classes. For cross-database adaptation, peak performance when adapting on the training subset is achieved with 2 regression classes and with 5 regression classes for adaptation on the training set. For thoroughness, transformation of models trained on pooled source and target data was also experimented with, but also did not deliver useful performance (results not shown).

TABLE I
PEAK WORD ACCURACY ON THE AFRIKAANS TEST SET FOR DIFFERENT APPROACHES TO CROSS-LANGUAGE ADAPTATION ON THE AFRIKAANS TRAINING SET (A) AND AFRIKAANS TRAINING SUBSET (A1)

Source:	SUN Speech		TIMIT	
Target:	A1	A	A1	A
Train source	57.9%	57.9%	-5.9%	-5.9%
Train target	45.0%	67.6%	45.0%	67.6%
Pooling	68.1%	73.3%	45.0%	55.3%
Mean-only MAP	66.1%	71.8%	53.1%	63.6%
Full MAP	70.2%	74.9%	57.0%	67.7%
Pooling-MAP	**71.4**%	75.3%	56.0%	69.0%
Mean-only MLLR	62.7%	67.0%	40.6%	49.6%
Full transform	65.7%	71.8%	43.0%	56.9%
MLLR-MAP	69.9%	74.8%	**64.1**%	**72.0**%
MAPLR	65.9%	73.9%	45.1%	56.1%
Pooling-MCE	71.3%	**76.1**%	51.4%	66.2%
Aug-MAP	-	-	61.8%	71.8 %

The combination of Bayesian and transformation-based techniques deliver interesting results, especially for cross-database adaptation. MLLR-MAP adaptation delivers the best results achieved with using the TIMIT database, achieving 64.1% and 72.0% word accuracy for training subset and training set adaptation, which represents 35% and 14% relative reductions in word error rate respectively over the baseline Afrikaans results (45.0% and 67.6% word accuracy). These results are achieved for relatively simple transformations in the first stage, using a single regression class for training subset and 2 regression classes for training set adaptation. Use of a relatively simple transformation (few regression classes) removes bias between the data sets, producing improved priors for subsequent (full) MAP adaptation. Experimentation with block-diagonal and diagonal transformations in the first step did not deliver further improvements in performance and are therefore not reported in detail. Same-database MLLR-MAP adaptation delivers poorer performance than using only MAP adaptation, which is expected since there should not be significant bias between source and target data from the same database. Use of MAPLR mean and variance transformation generally shows improvement over using MLLR, but does not provide consistently useful performance.

MCE adaptation of models trained on pooled data delivers good performance for same-database experiments, achieving best overall performance of 76.1% for Afrikaans training set adaptation. This represents a relative reduction in word error rate of 26% over baseline Afrikaans training set performance (67.6% word accuracy). MCE performance for cross-database adaptation is poor and is attributed to the fact the MCE is an optimisation approach that may converge to local minima when initial seeding is not good.

The last approach for cross-language use of acoustic information that is investigated is the data augmentation approach. Models are trained on target data augmented with transformed source data, providing good prior models for MAP adaptation. Performance of 61.8% word accuracy for Afrikaans training subset adaptation and 71.8% word accuracy for Afrikaans training set adaptation are achieved, which is better than that achieved with either cross-language MAP or pooling-MAP approaches when the TIMIT database is used as source. Only the MLLR-MAP approach delivers better performance.

5. CONCLUSION

The results in this paper show significant improvements in performance for continuous speech recognition in a target language by use of speech data from a source language. Between 26% and 48% reduction in error rate for same-database and between 14% and 35% reduction in error rate for cross-database use of English speech in improving Afrikaans recognition is achieved.

All the general strategies for cross-language use of acoustic information were shown to deliver useful results. The *multilingual pooling-adaptation* strategy, in conjunction with full MAP or MCE adaptation, delivered the best results when source and target data are closely matched i.t.o. recording conditions and labelling conventions. When more than one database is used, cross-language model adaptation utilising MLLR-MAP adaptation delivered the best performance. For cross-database experiments the *cross-language augmentation-adaptation* strategy was also shown to deliver useful performance.

6. ACKNOWLEDGEMENTS

The authors would like to thank Darryl Purnell for use of the base Hidden Markov Toolkit for Speech Recognition (HMTSR) system for training and testing hidden Markov models. The financial support of the Mellon Foundation is also gratefully acknowledged.

7. REFERENCES

1. J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT Voyager system," *Speech Communication*, vol. 17, pp. 1–18, Aug. 1995.
2. F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 359–362, Sep. 1997.
3. P. Bonaventura, F. Gallocchio, and G. Micca, "Multilingual speech recognition for flexible vocabularies," in *Proc. Eurospeech '97*, (Rhodes, Greece), pp. 355–358, Sep. 1997.
4. J. Köhler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *Proc. ICASSP '98*, (Seattle, USA), pp. 417 – 420, May 1998.
5. U. Bub, J. Köhler, and B. Imperl, "In-service adaptation of multilingual hidden Markov models," in *Proc. ICASSP '97*, (Munich, Germany), pp. 1451 – 1454, Apr. 1997.
6. C. Nieuwoudt and E. Botha, "Adaptation of acoustic models for multilingual recognition," in *Proc. Eurospeech '99*, (Budapest, Hungary), pp. 907–910, Sep. 1999.
7. T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP '98*, Vol. 5, (Sydney, Australia), pp. 1819–1822, Nov. 1998.
8. J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, Apr. 1994.
9. C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171–185, Apr. 1995.
10. W. Chou, "Maximum a posterior linear regression with elliptically symmetric matrix variate priors," in *Proc. Eurospeech '99*, (Budapest, Hungary), pp. 1–4, Sep. 1999.
11. B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 257–265, May 1997.