ISCA Archive
http://www.isca-speech.org/archive

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

ADAPTATION OF ACOUSTIC MODELS FOR MULTILINGUAL RECOGNITION

C. Nieuwoudt and E. C. Botha

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

{chris, botha}@ee.up.ac.za

10.21437/Eurospeech.1999-221

ABSTRACT

This paper evaluates the recognition performance of a system using acoustic models transformed across language boundaries. Parameters of hidden Markov models (HMMs) trained on speaker independent English data are adapted using Afrikaans adaptation data to realise speaker dependent, multispeaker and speaker independent Afrikaans models. Adaptation is performed using maximum *a posteriori* probability (MAP) and maximum likelihood linear regression (MLLR) methods on context independent and context dependent phones. Results show that MLLR transformation of English models using Afrikaans adaptation data significantly improves model performance and for context dependent models achieves better performance on speaker independent tests than achievable by direct training on the adaptation data.

1 INTRODUCTION

For many languages, including ten of the eleven official languages of South Africa (except for English), very little or no labelled speech data are available for training acoustic models. Our research aims at finding techniques that enable the training of robust acoustic models for these languages in the absence of large quantities of speech data. More specifically, we investigate the use of labelled data in a source language to obtain improved models for target languages in which only small amounts of adaptation data are available. Our methods are based on previous research that has shown the applicability of using phoneme data from one or more languages to 'bootstrap' phoneme models for a new language[1, 2], or even to construct phoneme models that are useful across more than one language[3, 4]. In the construction of multilingual phone sets, some recognition performance degradation is usually accepted in exchange for simplified modelling[5]. Use of explicitly multilingual phonemes lead to performance degradation because model precision decreases when modelling contexts from a set of languages. Since we consider only a single target language at a time, the models need not retain the properties of the original language and we use techniques from speaker adaptation to transform the models using adaptation data from a target language. We focus on achieving improved performance for a single target language at a time, but attempt to find methods that are independent of the target language.

Our experiments are performed on the SUN Speech database which was compiled by the University of Stellenbosch to contain phonetically labelled speech in both Afrikaans and English. A unified set of phoneme labels are used which allows us to focus specifically on the differences between the acoustic parameters in the two languages. Experiments are performed which adapt speaker independent English models using limited amounts of Afrikaans adaptation data. We evaluate both MAP adaptation and MLLR transformation in speaker adaptive, multispeaker and speaker independent modes. The results from these adaptation experiments are compared to a set of benchmark experiments which use different amounts of training data from both languages and are tested on the same speaker independent test set. Experiments are performed with both context independent and context dependent phones.

The organisation of this paper is as follows. In Section 2 we discuss the application of speaker adaptation methods for adapting acoustic models across language boundaries. Section 3 describes the bilingual database used in the experiments. Section 4 presents the experiments performed and results obtained and we conclude in Section 5.

2 CROSS-LANGUAGE ADAPTATION OF ACOUSTIC MODELS

When considering the adaptation of acoustic models across languages it is natural to examine the applicability of using speaker adaptation techniques. However, speaker adaptation and language adaptation of acoustic parameters differ in a number of ways. Firstly they differ when speaker adaptation is considered as an adaptation of a prior model, such as MAP adaptation, which uses a speaker independent model as prior distribution. When adapting across language boundaries the assumption that a speaker independent model of the source language is a suitable prior distribution, is probably inaccurate. This argument also implies limited usefulness of speaker clustering techniques for language adaptation as they merely select a subset of an overall distribution. The second important difference between speaker and language adaptation lies in the goal of speaker adaptation, which is to adapt acoustic models to a specific speaker, whereas language adaptation aims to adapt models for use in a multispeaker, or more generally, in a speaker independent environment. This aspect implies limited usefulness of speaker conversion techniques and leads us to consider techniques that start off with speaker independent data or models. Both the aspects of prior distribution consideration and speaker independence lead us also to favour transformation based approaches (say MLLR) over standard Bayesian adaptation approaches (say MAP). Lastly, when one considers the possibly large differences in acoustic modelling between languages, it is unlikely that a general spectral transformation exists that maps the entire acoustic space of one language to that of another. Rather, one would consider an approach that can map regions of the space (e.g. different phone categories) separately, implying the transformation of models rather than raw data.

Two methods commonly used for speaker adaptation are MAP[6, 7, 8] and MLLR[9] . Even though we expect that a transformation-based approach will deliver superior recognition performance, we perform MAP adaptation experiments (based on [7]) for the speaker adaptive case to deliver a reference baseline performance and because MAP results give an indication of how applicable the speaker independent priors are for cross language adaptation. The implementation of MAP adaptation used in this paper was found in [7] and adapts both mean and variance parameters according to a joint mean and variance prior distribution. The implementation of MLLR adaptation[9] that we use performs adaptation of only the Gaussian means. When MLLR adaptation is performed, models are grouped into classes and a separate transformation is calculated for each class. Grouping into classes is done according to broad phonetic groupings, i.e. for two classes vowels/diphthongs are separated from the rest, five classes comprise vowels, diphthongs, fricatives/affricates, stops and nasals/glides/liquids and for an eight class separation all mentioned categories are treated as distinct classes. Grouping transformations into classes has the advantage that each class of similar phones incurs the same transformation, which may be different from that incurred by other classes. The assumption is that the distribution of the acoustic parameters for the new language's speaker(s) exhibit correlation within each class.

3 THE SUN SPEECH DATABASE

The database used in this paper is the SUN Speech database[10] compiled by the Department of Electrical and Electronic Engineering of the University of Stellenbosch containing phonetically labelled speech in both Afrikaans and English. Some details of the database and its subdivision are given in Table 1. Adaptation is done on a set of eight speakers of which six are female (more females than males spoke all 20 Afrikaans sentences). In the Afrikaans set, sentences 1 to 10 are used for training or adaptation, sentences 11 to 20 from the adaptation speakers are used for testing speaker specific or multispeaker adaptation and sentences 11 to 20 of a disjoint set of speakers are used for speaker independent (SI) testing. When training English models for adaptation to Afrikaans, all English data are combined, totalling more than 86000 phone labels, which is approximately three times the total amount of Afrikaans data.

Language	Set	Speakers		Sen-
		Male	Female	tences
Afrikaans	train	21	10	1-10
	adapt	2	6	1-10
	adapt test	2	6	11-20
	SI test	16	7	11-20
English	train	33	17	21-40
	test	22	4	41-60

Table 1: Subdivision of SUN Speech database into training, testing and adaptation sets for Afrikaans and English

4 EXPERIMENTS AND RESULTS

The goal of the experiments is to compare and evaluate the performance of cross-lingual adaptation versus either same-language training or same-language adaptation. Four sets of experiments are performed. The first three sets of experiments use context independent phones (monophones) and comprise speaker adaptive experiments, multispeaker adaptation experiments and adaptation for speaker independent recognition experiments. The last set of experiments utilise context dependent phones (triphones) and evaluate only speaker independent recognition performance. All experiments perform phone recognition from labelled phone data without using any form of language (phone sequence) modelling. Phones are modelled with context independent and context dependent (triphone) continuous density HMMs with 39 mel-scaled cepstral, delta and delta-delta features using a total of 58 phone classes. The four sets of experiments are discussed next.

4.1 Speaker adaptive experiments

In these experiments the adaptation set is used to adapt speaker independent Afrikaans and English prior models to each of the eight specific adaptation speakers in turn and performance is tested on the respective speaker's test sentences. The experiments evaluate the use of a prior model from a different language than that of the target speaker. For comparison, baseline performance of speaker independent Afrikaans and English phone recognisers, along with the performance of speaker trained (small speaker dependent set) models are also given with the summarised results in Table 2. As expected, adaptation using an Afrikaans prior outperforms using an English prior. The best performance (65.0%) is achieved with MLLR adaptation using 2 phone classes and an Afrikaans prior model. Recognition performance using MLLR and an English prior is also good (61.8% for the two class transformation) and exceeds the best speaker dependent result (58.4%) while using fewer mixtures (4 vs. 10) to achieve the same performance as the comprehensively trained Afrikaans SI models. MLLR also far outperforms MAP, but at the cost of more mixtures.

Description	Configuration *state*×*mix*	Result
Afrikaans SI	3x1	50.6%
	3x10	61.8%
English SI	3x1	40.7%
	3x10	49.4%
Speaker dependent	3x1	58.4%
	2x2	58.1%
Adaptive with Afrikaans Prior		
MAP	3x1	58.6%
MLLR (1 class)	3x10	64.2%
MLLR (2 class)	3x8	65.0%
MLLR (5 class)	3x4	63.9%
MLLR (8 class)	3x4	63.9%
Adaptive with English Prior		
MAP	3x1	56.1%
MLLR (1 class)	3x2	59.4%
MLLR (2 class)	3x4	61.8%
MLLR (5 class)	3x4	61.4%
MLLR (8 class)	3x4	61.4%

Table 2: Speaker adaptive experiments: Recognition performance achieved on the speaker adaptation test set when training with the speaker independent (SI) Afrikaans training set, the SI English set, the (small) speaker dependent adaptation set and when adapting the Afrikaans and English SI prior models

4.2 Multispeaker adaptation experiments

In this set of experiments the data of eight adaptation speakers are pooled for adaptation and performance is measured on pooled test sentences of the same speakers. This experiment evaluates the performance of adaptation when applied to a multispeaker scenario and results are summarised in Table 3. The best performance (64.2%) is achieved with MLLR adaptation and an Afrikaans prior model. Baseline training with only the adaptation data (multispeaker) delivers good performance (63.2%), almost equalling that of the Afrikaans MLLR approach (64.2%), indicating that there is enough adaptation data to reflect the characteristics of the adaptation speakers quite well. Recognition performance using MLLR and an English prior is not as good, achieving a best performance of 58.5%, which is, however, still much better than the 49.4% achieved with (unadapted) English models. Once again, using the Afrikaans prior has delivered better performance than using the English prior. It is interesting to note that the performance increases as more transformation classes are used, indicating that a more complex transform is necessary to accurately convert to the multispeaker case, and also that since more data is available (data is pooled), it allows for a larger number of transformation parameters to be estimated accurately. These experiments, compared to the speaker adaptive experiments, indicate that the MLLR transformation may be better suited to speaker dependent transformation than for the multispeaker case.

Description	Configuration *state*×*mix*	Result
multispeaker	3x8	63.2%
Adaptive with Afrikaans Prior		
MLLR (1 class)	3x10	62.3%
MLLR (2 class)	3x8	63.3%
MLLR (5 class)	3x10	64.2%
MLLR (8 class)	3x10	64.2%
Adaptive with English Prior		
MLLR (1 class)	3x10	54.7%
MLLR (2 class)	3x4	56.4%
MLLR (5 class)	3x6	58.4%
MLLR (8 class)	3x10	58.5%

Table 3: Multispeaker adaptation experiments: Recognition performance achieved on the speaker adaptation test set when training with the pooled speaker adaptation set and when adapting the Afrikaans and English SI prior models

4.3 Speaker independent adaptation experiments

When the adapted English models of the previous experiment are tested on the speaker independent Afrikaans test set, the best recognition performance achieved decreases to 54.1% (from 58.4% for the multispeaker experiment) for a three state, 6 mixture HMM transformed using 5 transformation classes. The performance (54.1%) is, however, still much better than that achieved with (unadapted) SI English models (49.4%).

The performance of the baseline Afrikaans multispeaker-trained models from the previous set of experiments decreases by a larger margin, dropping to 54.6% (from 63.2% for the multispeaker test) when tested on the speaker independent Afrikaans test set. The models trained on multispeaker Afrikaans data thus perform only slightly better than the MLLR adapted English models (54.6% vs. 54.1%) and if somewhat less adaptation data were available, the MLLR adapted models would probably be superior.

4.4 Context dependent phone modelling

Phone context is modelled in terms of the broad phonetic categories mentioned in Section 2. Since we are interested in only acoustic discrimination, no language, or even context sequence constraints are used and it is therefore not surprising that context modelling by itself does not improve performance by much. An analysis of the triphone distribution in English and Afrikaans in the SUN Speech database reveals that triphone coverage is very sparse and also that overlap of triphone context between the two languages is not good. This is in spite of the overall phone distributions in the two languages being relatively similar. No information about the general triphone overlap between English and Afrikaans is currently available. Baseline recognition results for Afrikaans and (unadapted) English triphone models on the Afrikaans test set show that recognition performance is increased for few mixtures per state, but that peak performance is not significantly increased above that achieved with monophones. Performing adaptation on the English triphone models with the Afrikaans adaptation set delivers a maximum performance of 56.2% for a 15 class transformation, which is better than that achieved with adapted monophones (54.1%), and is better than the best performance using only the adaptation set for training (55.0%).

5 CONCLUSION

The results indicate that speaker adaptation methods can be applied with success in cross-lingual adaptation of acoustic parameters and can deliver reasonably good performance in the absence of large quantities of training data. The results indicate that the adapted models far outperform unadapted models and achieve better performance than when training with the small adaptation data set only. The MLLR method delivers better performance than the MAP method and is a promising technique for cross-lingual adaptation of acoustic parameters. A major obstacle towards achieving good performance with language adapted models is the potentially large mismatch between acoustic contexts of phones in different languages. A potential solution is to use transformations from a large set of languages when training a new language.

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] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *Proc. Eurospeech '97*, (Rhodes, Greece), Sep. 1997, pp. 371–374.

[3] P. Bonaventura, F. Gallocchio, and G. Micca, "Multilingual speech recognition for flexible vocabularies," in *Proc. Eurospeech '97*, (Rhodes, Greece), Sep. 1997, pp. 355–358.

[4] F. Weng, H. Bratt, L. Neumeyer, and A. Stolcke, "A study of multilingual speech recognition," in *Proc. Eurospeech '97*, (Rhodes, Greece), Sep. 1997, pp. 359–362.

[5] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP '98*, Vol. 5, (Sydney, Australia), Nov. 1998, pp. 1819–1822.

[6] R. M. Stern and M. J. Lasry, "Dynamic speaker adaptation for feature-based isolated word recognition," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 35, pp. 751–763, June 1987.

[7] C.-H. Lee, C.-H. Lin, and B.-H. Juang, "A study on speaker adaptation of the parameters of continuous density hidden Markov models," *IEEE Trans. Signal Processing*, vol. 39, pp. 806–841, Apr. 1991.

[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. J. Leggetter and P. C. 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] T. Waardenburg, J. D. Preez, and M. Coetzer, "The automatic recognition of Afrikaans stop consonants in continuous speech," in *Proc. IEEE South African symposium on Communications and Signal Processing*, (Fourways, South Africa), Aug. 1991, pp. 110–115.