Multilingual Training of Acoustic Models in Automatic Speech Recognition

C Nieuwoudt[a] EC Botha[b]

Department of Electrical and Electronic Engineering, University of Pretoria, Pretoria 0002, South Africa
[a]chris@.ee.up.ac.za, [b]botha@.ee.up.ac.za

Abstract

This paper evaluates the performance of a speech recognition system using acoustic models trained on multilingual data. The reason in our case for using data from more than one language is that there may not be enough data available for a new language to train a robust recogniser. Two general strategies are employed: firstly the pooling of data from the different languages for training and secondly the training of models on the data from one language and subsequent adaptation of the models using data from the new target language. For the first approach English data and Afrikaans training data are pooled in order to train hidden Markov models (HMMs) for the target language, Afrikaans. For the second approach the parameters of HMMs trained on English data are adapted using maximum a posteriori probability (MAP) and maximum likelihood linear regression (MLLR) methods on Afrikaans data. Continuous density HMMs are used to model context independent phones found in Afrikaans. Cross-language adaptation performance is evaluated in terms of phone recognition performance as well as for a continuous speech recognition task in Afrikaans. The interesting result is that for continuous recognition the best performance is obtained by simple pooling of the data and this performance far exceeds the performance achievable using only data from the target language. The improvement is due to the fact that in our database there exists no mismatch between the English and Afrikaans data (other than the language difference) and both languages were labelled with a consistent set of labels. Adaptation results indicate that both MAP adaptation and MLLR transformation of English models using Afrikaans adaptation data significantly improves model performance and also achieves better performance than achievable by direct training on the adaptation data.

Keywords: *adaptation, multilingual, speech recognition*
Computing Review Categories: *G.3, I.5*

1 Introduction

Automatic speech recognition is currently one of the most important applications driving the development of the personal computer in terms of computational needs and has the potential to fundamentally change the way in which we interact with computers. Unfortunately current large vocabulary continuous speech recognition (LVCSR) systems are limited to the major languages of the world. The development of an LVCSR system for a new language, or even for a different accent of English, is an expensive undertaking and currently necessitates the collection a large amount of acoustic data and language information in the new language.

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 (say English) to obtain improved models for target languages (e.g. Afrikaans) 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[4, 11], or even to construct phoneme models that are useful across more than one language[1, 14]. In the construction of multilingual phone sets, some recognition performance degradation is usually accepted in exchange for simplified modelling[10]. Use of explicitly multilingual phones may lead to performance degradation because model precision decreases when modelling contexts from a set of languages.

Our research differs from previous multilingual speech recognition approaches in that we also investigate the explicit use of (transformed) source language acoustic parameters for recognition in a new target language. The assumption is that we do not possess enough data for the new language in order to train a speaker independent recognition system on its own and therefore want to make use of acoustic information from another language. 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.

Some major problems with the development of cross-lingual systems are that the applicable languages have different phoneme sets and that the databases for the different languages often differ in terms of labelling conventions, recording characteristics and the type of speech contained. A single consistent database containing both English and Afrikaans speech is used for the experiments detailed in this paper and we are thus allowed to focus specifically on the actual acoustic agreement or difference between the languages.

Because of the consistent phonetic labelling process followed in setting up the database, the pooling of the English data and Afrikaans training data should lead to more robust estimates of model parameters simply because more data is available. It is expected, however, that the models will exhibit bias towards the acoustic properties of English speech. With the adaptation approach the assumption is that the acoustic models should retain their speaker independent characteristics from the source language (English) while being changed to better reflect the distribution of acoustic parameters in the target language (Afrikaans).

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 presents the experiments performed and results obtained. We conclude in Section 4.

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. Adaptation techniques are often based on a Bayesian paradigm in which knowledge about a prior distribution for model parameters is assumed and then model parameters are estimated to maximise the *a posteriori* probability of the parameters given some adaptation data. A second large class of adaptation techniques attempts to estimate a transformation from the parameters of speaker independent models to those of a specific speaker.

Most state-of-the-art LVCSR systems use hidden Markov modelling in conjunction with Gaussian mixture distributions to model the parameters of speech. Separate HMMs are estimated for each phoneme to allow the recognition of large vocabularies through use of a pronunciation dictionary. The techniques discussed in this section use trained HMMs as input and adapt their parameters using speech data from a new language. Especially the Gaussian means, but also the Gaussian variance parameters, are important in terms of the performance achieved with adaptation.

2.1 MAP Adaptation

Maximum *a posteriori probability* (MAP)[12, 5], which is an implementation of Bayesian adaptation, uses a limited amount of speaker specific data in combination with prior distributions derived from speaker independent models to estimate the maximum *a posteriori* probability parameters for a new speaker. The MAP parameters are a linear combination of the speaker independent parameters used as prior and the speaker dependent measurements. In the MAP approach suggested by Lee *et al.*[5], only a single Gaussian distribution per HMM state is computed for the transformed parameters. Mean and variance parameters are adapted according to a joint mean and variance prior distribution derived from the set of speaker independent Gaussian mixtures. MAP has been extended to allow for the adaptation of mixture components[3], but presents the problem of choosing suitable parameters for the prior distribution. The implementation of mixture Gaussian MAP adaptation in this paper uses the individual speaker independent Gaussians as priors for the adaptation of the Gaussian means and non-informative priors for the adaptation of the Gaussian variance parameters.

A drawback of MAP adaptation is that it suffers from the problem that only observed mixtures are adapted. Relatively large amounts of adaptation data are therefore needed to adapt distributions consisting of many mixtures. This is the reason why transformation based adaptation in general outperforms Bayesian adaptation when little speaker specific data is available- for example when only a single utterance from a new speaker is available. For our application of language adaptation it is realistic that at least a couple of utterances from a number of different speakers should be available and thus we expect that MAP adaptation should perform reasonably well. A good argument for using MAP adaptation is that it delivers asymptotic performance- if an increasingly large amount of adaptation data is available, the performance of the system converges to the performance expected from a language specific system trained with the same amount of data.

2.2 MLLR Adaptation

The second method commonly used for speaker adaptation is the transformation approach. Advantages of the transformation approach above MAP adaptation are that a suitable prior distribution does not need to be found and that a transformation approach can adapt unobserved distributions by implementing the same transformation for groups of parameters. When very little adaptation data is available this presents an advantage. However, when a large amount of adaptation data is available, a transformation-based adaptation scheme can not guarantee asymptotic behaviour with respect to a language dependent system since the transformation places constraints on the reachable parameter space.

The most commonly used algorithm for transformation based adaptation is the maximum likelihood linear regression[6] algorithm. MLLR estimates a linear transfor-

mation of the Gaussian means that maximises the probability that the transformed models generate the observations. When the dimensionality of the transform (the feature dimension for a full transform) is more than the number of independent components (mean vectors of Gaussians) being transformed, the matrix to be inverted is singular. This is not really a problem, however, and only implies that the approach delivers a zero error transformation and degenerates to maximum likelihood (ML) estimation of the parameters using only the speaker specific data. Care must thus be taken to ensure that the number of independent components is reasonably larger than the feature dimension to ensure robust estimation of the transformation and not simply ML reestimation of the parameters.

A related method for the transformation of both the Gaussian mean and variance parameters was suggested by Digalakis *et al.* [2]. Unfortunately this method is limited to diagonal transformation matrices and has been found[9] not to perform as well as the MLLR approach. In this paper we implemented the standard MLLR approach for the transformation of the Gaussian mean parameters. Our previous experiments with cross-language MAP adaptation [8] have shown the importance of adaptation of the Gaussian variance parameters. The techniques that we experimented with for the adaptation of the Gaussian variance parameters include simple reestimation (on only the Afrikaans data) and use of the mean square error (MSE) criterion to derive a transformation for the Gaussian variance parameters. There are two problems with simply applying the MSE criterion to adaptation of the variance parameters. The first is that the dynamic range of the variance parameters is from 0.0001 to approximately 2.0 and thus the MSE criterion will tend to deliver inaccurate results for variance parameters at the small end of the scale. The second problem is that there are no guarantees that the transformed variance values will be positive. The solution which we propose to this problem is to compute the MSE estimate of the transformation in logarithmic space, i.e. to compute a transform from the log of the source model variance parameters to the log of the target model variance parameters. We find this method to achieve superior performance in continuous word recognition experiments when compared with either plain MSE variance transformation or ML variance reestimation. The experiments we performed are discussed in the next section.

3 Experiments and Results

Experiments are performed using the SUN Speech database[13] compiled by the Department of Electrical and Electronic Engineering of the University of Stellenbosch containing phonetically labelled speech in both Afrikaans and English. Details of the database are given in Table 1 with more complete information given in [7]. The database contains read speech from 138 speakers totalling approximately 1500 utterances in English and 500 utterances in Afrikaans. For the purpose of our experiments, the Afrikaans set is divided into training, adaptation and testing sets with disjoint sets of speakers and different sets of sentences for training/adaptation and testing. Experiments reported on in this paper compare performance achieved by models trained directly on the Afrikaans training or adaptation sets with performance achievable by training on pooled English and Afrikaans data, as well as by adapting English models with Afrikaans adaptation data.

Language	Set	Speakers		Sentences	Dur. (s)
		Male	Female		
English		55	21	21-60	7757
Afrikaans	train	21	10	1-10	1239
	adapt	2	6	1-10	316
	test	16	7	11-20	745

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

The software of the system that is used for training and testing of the hidden Markov models was developed at the University of Pretoria. Phones are modelled with context independent continuous density hidden Markov models (HMMs) with 39 mel-scaled cepstral, delta and delta-delta features using a total of 54 phone classes. Pre-processing includes using 16ms frames with a 10ms increment, preemphasis of the input signal and Hamming windowing.

MAP adaptation is used as discussed in Section 2 to adapt both mean and variance parameters. For mixture distribution adaptation non-informative priors are selected for Gaussian variance parameters, amounting to simple reestimation of variance parameters. This was found to outperform the use of prior variance values when many mixtures are used.

When MLLR adaptation is used, the models are grouped into sound classes, with a separate transformation being 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 the 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 speaker exhibit correlation within each class.

The experiments that were performed can be grouped into two main categories namely labelled phone recognition experiments and continuous speech recognition experiments. The two sets of experiments are discussed next.

3.1 Labelled Phone Recognition

The experiments compare the performance of same language acoustic models with the performance of models trained on pooled multilingual data and models adapted

from a different language using only little adaptation data. The first set of experiments perform phone recognition on labelled Afrikaans phone data using a subset of 46 phones which occur in the Afrikaans set. Gaussian mixture HMMs with three states are used to model each of the phones. Baseline recognition performance is determined for speaker independent Afrikaans and English phone recognisers, recognisers trained only on the Afrikaans adaptation set and recognisers trained on the pooled English/Afrikaans training sets and pooled English/Afrikaans adaptation sets. These results are compared to the performance achieved with the MAP and MLLR implementations as detailed in Section 2 when adapting with the Afrikaans adaptation set.

Results are given in Figure 1 and as expected, models trained on the Afrikaans training set deliver the best performance of 61.9% and models trained only on the English set deliver the poorest maximum performance of 46%. Results for pooled data systems fall between these boundaries, indicating that the pooled models are less accurate than models trained only on Afrikaans. Results for models trained only on the Afrikaans adaptation data achieves better performance than the pooled data models, reaching a maximum of 55.1% for an 8 mixtures per state HMM. Both adaptation approaches outperform the adaptation data only system, with MAP adaptation achieving the best performance of 56.6% with 10 mixtures per state and 8 class MLLR adaptation achieving 56.3% with only 6 mixtures per state. Adaptation with fewer classes achieves poorer performance and is not shown. The results indicate that the cross-language adapted models outperform the (unadapted) source language models and can achieve better performance than achievable using only a relatively small amount of adaptation data. The difference in maximum performance between the language adapted (MAP and MLLR) models and the models trained only on the adaptation set is relatively small (between 1.2% and 1.5% absolute) and do not really reflect the usefulness of the adapted models. The poor performance of the pooled data models in terms of phone recognition is also misleading, because these models may actually lead to better generalisation. The next set of experiments attempts to investigate the performance of the models more rigorously by using them in a continuous speech recogniser.

3.2 Continuous Speech Recognition

The continuous speech recognition experiments once again compare the performance of same language acoustic models with the performance of pooled data models and models adapted from a different language using only little adaptation data. The same 46 phone models that were reported on in Section 3.1 are used to construct word models by connecting the phone HMMs according to a phonetic dictionary for all words occurring in utterances 11-20 of the speaker independent test set. The phonetic dictionary is created by analysing the phone labels assigned to the speech of the 8 adaptation speakers for utterances 11-20

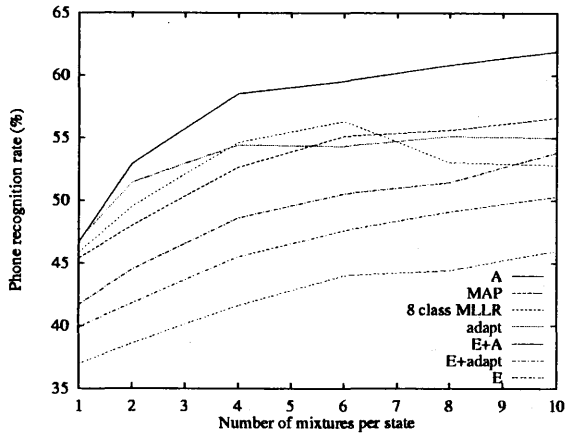

Figure 1: Labelled phone recognition: Recognition performance achieved on a speaker independent Afrikaans test set when training with the speaker independent Afrikaans training set (A), the entire English set (E), the Afrikaans adaptation set (adapt), combinations of the above and when adapting English prior models using the Afrikaans adaptation set with MAP and 8 class MLLR adaptation

and by allowing multiple pronunciations of the same word, as long as at least two or more of the adaptation speakers used the given pronunciation. Using the pronunciation dictionary, a total of 151 models for the 100 distinct words in the test utterances are created. In order to run a continuous speech recognition experiment, a small grammar was devised that allocates each word to one of 5 linguistic categories comprising loosely verbs, nouns, adjectives, pronouns and conjunctives. A total of 18 transitions out of a possible 25 transitions between the 5 categories are allowed, limiting the possible sequences enough to deliver reasonable performance for continuous speech recognition in the absence of statistical language modelling.

Continuous speech recognition results are given in terms of correct word percentage in Figure 2 and in terms of word accuracy in Figure 3. Correct word percentage is computed by aligning the output string from the recogniser with the true transcription and computing the percentage of transcribed words that are present in the output string and are in the correct sequence. Word accuracy is computed by subtracting the insertions and deletions, expressed as a percentage of the number of transcribed words, from the the correct word percentage.

Figure 2 shows that, in contrast with the phone recognition experiments, the best correct word performance of 78.7% is achieved with models trained on the pooled English and Afrikaans training data. This corresponds to a large 23% reduction in error rate over the best performance achieved when models are trained only on the Afrikaans data. Pooling of training data leads to more general models since phones occur in more contexts and may lead to better word recognition performance in the absence of large amounts of language specific data. More general models are also favoured for word recognition because they may compensate for mismatch between the pronunciation dictionary and actual pronunciation. Very good performance

(76.2%) is also achieved when pooling the English set with only the Afrikaans adaptation data. Models trained only on the Afrikaans training set achieve a best correct word rate of 72.3% with 4 mixtures per state.

The poorest performance (for more than 2 mixtures) is achieved with models trained only on the adaptation set. This performance is well below that achieved with the English models and is due to the poor generalisation achieved with the small amount of adaptation data. Inspection of individual phone recognition rates for models trained on the adaptation set indicates that only phones with high *a-priori* probabilities achieve good recognition rates, and for many mixtures, phones with low *a-priori* probabilities are over-fitted. For the English models, results in Figure 2 do not vary dramatically as a function of the number of mixtures, reaching a best correct word percentage of 62.4% with 4 mixtures. Figure 2 also shows reasonably good performance achieved with the MAP (67.1%) and MLLR (66.7%) techniques, resulting in respectively in 13% and 11% decreases in word error rate beyond that achieved with only the English set. Best MLLR performance is achieved with a single class transformation, indicating that over-fitting occurs with more regression classes.

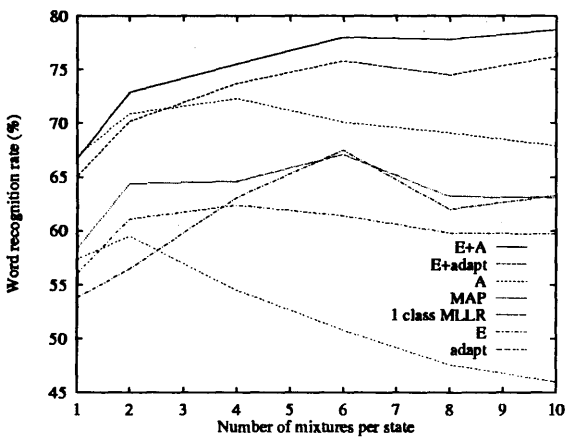

Figure 2: Continuous speech recognition: Correct word percentage achieved on a speaker independent Afrikaans test set when training with the Afrikaans training set (A), the English set (E), the Afrikaans adaptation set (adapt), combinations of the above and when adapting English prior models using the Afrikaans adaptation set with MAP and single class MLLR adaptation

The word accuracy results from Figure 3 show the same general trends as for the correct word percentage, but there is an even larger range in performance between the different methods. Best performance is achieved with pooled models, namely 69.0% for the English/Afrikaans training set and 65.3% for the English/Afrikaans adaptation set. These two figures amount to respectively a 27% reduction in error rate over the Afrikaans trained models (69.0% vs. 57.7% accuracy) and a large 37% reduction in error rate over the English trained models (65.3% vs. 44.9% accuracy) - results are not compared with adaptation set trained models since they perform too poorly. Adaptation improves on baseline English performance with relatively robust performance achieved with the MAP adapted models, but a higher maximum performance achieved with single class MLLR adapted models (52.1% vs. 49.9%). When few mixtures are used , MLLR performance is very close to the performance achieved using only the adaptation set because the transformation degenerates to simple ML estimation.

The continuous speech recognition results are surprising because such a large improvement in performance is achieved by simply pooling the English and Afrikaans data. The results are also in contrast to those achieved with the labelled phone recognition experiments. One should take into account that for phone recognition, more specialised models may deliver better performance, while for continuous speech recognition, more general models deliver better performance. Since the purpose of phone models is to be used as building blocks in a complete speech recognition system, the continuous speech results give a better indication of the usefulness of the models than the phone recognition results.

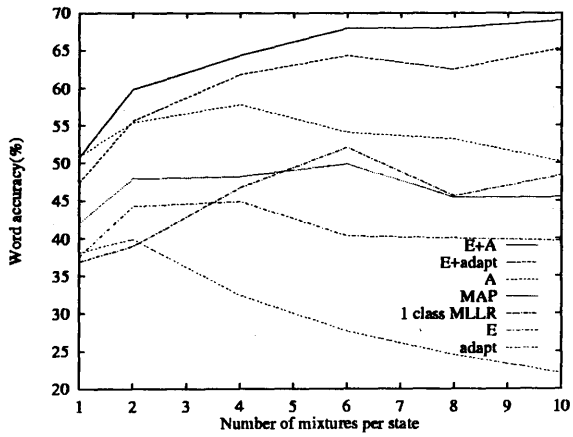

Figure 3: Continuous speech recognition: Word accuracy achieved on a speaker independent Afrikaans test set when training with the speaker independent Afrikaans training set, the entire English set, the Afrikaans adaptation set and when adapting English prior models using the Afrikaans adaptation set

4 Conclusion

In this paper we evaluated different strategies and algorithms for re-using acoustic information across language boundaries. We evaluated the improvement compared to a baseline English system in terms of phone recognition and for continuous speech recognition. We found that phone recognition and word accuracy results do not correlate well, but that emphasis should be placed on word accuracy results, as they most closely represent the use of the models for a practical purpose.

In particular, we compared training on pooled (English and Afrikaans) data to a model adaptive approach. We

found that when data from multiple languages are available that have been recorded under the same conditions and have been labelled phonetically with a consistent set of phone labels, a considerable improvement may be obtained by training directly on the pooled multilingual data. Even when a small amount of adaptation data is available, for example when the 80 utterance adaptation set is used, pooling this data with the English data delivers a 37% improvement in accuracy over the baseline English system. The performance (65.3% accuracy) exceeds the best performance achieved with the 310 utterance Afrikaans training set (57.8% accuracy), and also exceeds the performance achieved with the adaptation-based approaches.

The MAP and MLLR adaptation approaches were found to deliver reasonable improvements of 9% and 13% in accuracy over the baseline English system, but did not perform as well as recognisers trained on pooled data. This implies that adaptation does not utilise the available information efficiently when compared to training on the pooled data. This does not mean that the adaptation approach is not useful since if there were significant differences in e.g. recording conditions, or type of speech, we would expect adaptation-based approaches to deliver better performance than achievable by pooling of the data. This is the subject of planned future research.

Other future work should include use of more data sets to ensure that the results are not specific to the database and languages used in this study. We should also investigate the use of mappings from more than one source language.

5 Acknowledgement

The authors would like to thank Darryl Purnell for use of the base HMTSR system for training and testing hidden Markov models.

References

[1] Patrizia Bonaventura, Filippo Gallocchio, and Giorgio Micca. Multilingual speech recognition for flexible vocabularies. In *Proc. Eurospeech '97*, pages 355–358, Rhodes, Greece, September 1997.

[2] V.V. Digalakis, D. Rtischev, and L.G. Neumeyer. Speaker adaptation using constrained estimation of Gaussian mixtures. *IEEE Transactions on Speech and Audio Processing*, 3(5):357–366, September 1995.

[3] 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*, 2(2):291–298, April 1994.

[4] 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*, 17:1–18, August 1995.

[5] 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*, 39(4):806–841, April 1991.

[6] C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, April 1995.

[7] C. Nieuwoudt and E.C. Botha. Multilingual English and Afrikaans phoneme recognition. In *Proceedings of the Ninth Annual South African Workshop on Pattern Recognition*, pages 97–100, Stellenbosch University, Stellenbosch, Nov. 1998.

[8] C. Nieuwoudt and E.C. Botha. Adaptation of acoustic models for multilingual recognition. In *Proc. Eurospeech '99*, pages 907–910, Budapest, Hungary, September 1999.

[9] A. Sankar, L. Neumeyer, and M. Weintraub. An experimental study of acoustic adaptation algorithms. In *Proc. ICASSP '95*, pages 713 – 716, Detroit, MI, May 1995.

[10] T. Schultz and A. Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proc. ICSLP '98*, volume 5, pages 1819–1822, Sydney, Australia, November 1998.

[11] Tanja Schultz and Alex Waibel. Fast bootstrapping of LVCSR systems with multilingual phoneme sets. In *Proc. Eurospeech '97*, pages 371–374, Rhodes, Greece, September 1997.

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

[13] T. Waardenburg, J.A. Du Preez, and M.W. Coetzer. The automatic recognition of Afrikaans stop consonants in continuous speech. In *Proc. IEEE South African symposium on Communications and Signal Processing*, pages 110–115, Fourways, South Africa, Aug. 1991.

[14] Fuliang Weng, Harry Bratt, Leonardo Neumeyer, and Andreas Stolcke. A study of multilingual speech recognition. In *Proc. Eurospeech '97*, pages 359–362, Rhodes, Greece, September 1997.