Language-dependent state clustering for multilingual acoustic modelling ⋆

Thomas Niesler *

Department of Electrical and Electronic Engineering, University of Stellenbosch, Stellenbosch, South Africa

Abstract

The need to compile annotated speech databases remains an impediment to the development of automatic speech recognition (ASR) systems in under-resourced multilingual environments. We investigate whether it is possible to combine speech data from different languages spoken within the same multilingual population to improve the overall performance of a speech recognition system. For our investigation we use recently collected Afrikaans, South African English, Xhosa and Zulu speech databases. Each consists of between 6 and 7 hours of speech that has been annotated at the phonetic and the orthographic level using a common IPA-based phone set. We compare the performance of separate language-specific systems with that of multilingual systems based on straightforward pooling of training data as well as on a data-driven alternative. For the latter, we extend the decision-tree clustering process normally used to construct tied-state hidden Markov models to allow the inclusion of language-specific questions, and compare the performance of systems that allow sharing between languages with those that do not. We find that multilingual acoustic models obtained in this way show a small but consistent improvement over separate-language systems as well as systems based on IPA-based data pooling.

Key words: Multilinguality, multilingual acoustic models, multilingual speech recognition

⋆ Expanded version of a paper presented at the ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (Stellenbosch, South Africa, 2006).

* Corresponding author.
Email address: trn@dsp.sun.ac.za (Thomas Niesler).

1 Introduction

The first step in the development of a speech recognition system for a new language is normally the recording and annotation of a large quantity of spoken audio data. In general, the more data is available, the better the performance of the system. However, data gathering and especially annotation is very expensive in terms of money and time. In a multilingual environment, this problem is compounded by the need to obtain such material in several languages. It is in this light that we would like to determine whether data from different languages that coexist in a multilingual environment can be combined in order to improve speech recognition performance. This process involves determining phonetic similarities between the languages, and exploiting these to obtain more robust and effective acoustic models. The eventual aim of this work is the development of speech recognition systems that are able to deal with multiple languages without the need to explicitly implement a set of parallel language-specific systems.

Multilingual speech recognition is particularly relevant in South Africa, which has 11 official languages and among whose population monolinguality is almost entirely absent. This leads to a large geographic overlap in mother tongue speakers of different languages, as well as habitual and frequent code-mixing and code-switching. We study the four languages Afrikaans, English, Xhosa and Zulu, for which a certain amount of phonetically and orthographically annotated speech data is available. Detailed statistics reflecting the degree of multilinguality in South Africa are unfortunately not available. However, Table 1 shows the percentage of the population found to use each of the four languages under study as a mother tongue, and also as a second language. It has furthermore been estimated that Xhosa and Zulu speakers on average each speak four languages, while Afrikaans and English speakers are usually bilingual (Webb, 2002).

Language	First language speakers	Second language speakers
Afrikaans	13.3%	16.5%
English	8.2%	18.5%
Xhosa	17.6%	18.0%
Zulu	23.8%	24.2%

Table 1
Percentage of the population who use Afrikaans, English, Xhosa and Zulu as first and second languages respectively (Statistics South Africa, 2004).

Multilingual speech recognition has received attention by several authors over the last decade. The seminal work by Schultz and Waibel dealt with 10 languages, each of which is spoken in a different country and forms part of the GlobalPhone speech corpus (Schultz and Waibel, 1998, 2000, 2001). Multilingual recognition systems were developed by applying decision-tree cluster-

ing to tied mixture HMMs. While these techniques proved to be an effective means of constructing recognition systems for an unseen language not part of the training set, the authors were unable to improve on the performance of separate monolingual recognition systems. Similar conclusions were reached by Kohler (Köhler, 2001), who considered agglomerative clustering of monophone models in 6 languages, and by Uebler (Uebler, 2001), who employed IPA-based clustering in 2 and 3 languages. Finally, work by Uebler, Schüßler and Niemann (Uebler et al., 1998), which has focussed on a bilingual speech recognition system, has shown some improvement in recognition accuracy for non mother tongue speech, but also not for first language speakers.

We have considered the application of decision-tree clustering to the development of tied-state multilingual triphone HMMs. Furthermore, we focus on languages spoken within the same country, and hence related at least to some degree by the extensive phonetic and lexical borrowing, sharing, and mixing that takes place in a multilingual society. Finally, strong links exist between certain groups of indigenous languages (such as the Nguni language group which includes Xhosa and Zulu), that may allow more fruitful exploitation of data sharing by speech recognition applications.

2 Speech databases

We have based our experiments on the African Speech Technology (AST) databases, which consist of recorded and annotated speech collected over both mobile and fixed telephone networks (Roux et al., 2004). For their compilation, speakers were recruited from targeted language groups and given unique datasheets with items designed to elicit a phonetically-diverse mix of read and spontaneous speech. The datasheets included read items such as isolated digits, as well as digit strings, money amounts, dates, times, spellings and also phonetically-rich words and sentences. Spontaneous items included references to gender, age, mother tongue, place of residence and level of education.

The AST databases were collected in five different languages, as well as in a number of non-mother tongue variations. In this work we have made use of the Afrikaans, English, Xhosa and Zulu mother tongue databases.

Afrikaans is a Germanic language with its origins in 17th-century Dutch brought to South Africa by settlers from Holland. It incorporates lexical and syntactic borrowings from Malay, Bantu and Khoisan languages, as well as from Portuguese and other European languages. English was brought to South Africa by British occupying forces at the end of the 18th century, and is today regarded as a specific variety of so-called World Englishes. Xhosa and Zulu both belong to the Nguni group of languages, which also includes Ndebele

and Swazi. The orthography of both languages, as it is commonly accepted today, was standardised as recently as the 1960's. Both are tone languages, and exhibit a range of 'exotic' sounds, such as implosive bilabial stops as well as a range of clicks and associated phonetic accompaniments.

A global set of 135 phones, based on the International Phonetic Alphabet, has been used during phonetic speech transcription of the speech databases (International Phonetic Association, 2001). These phone units were chosen to reflect the variety of distinct sounds that were observed during transcription in each of the four languages. The phone set includes labels for diphthongs and tripthongs that occur across word boundaries in spontaneous speech. Furthermore, in several cases diacritic symbols, in combination with consonants or vowels, have been considered to be unique phones. These diacritics include ejection, aspiration, syllabification, voicing, devoicing and duration. A more detailed description of the phone set and how it differs among the four languages has been presented in (Niesler et al., 2005).

Together with the recorded speech waveforms, both orthographic (word-level) and phonetic (phone-level) transcriptions were available for each utterance. The orthographic transcriptions were produced and validated by human transcribers. Initial phonetic transcriptions were obtained from the orthography using grapheme-to-phoneme rules (Louw et al., 2001; Wissing et al., 2004; Louw, 2005), except for English where a pronunciation dictionary was used instead. These were subsequently corrected and validated manually by human experts.

2.1 Training and test sets

Each database was divided into a training and a test set. The four training sets each contain between six and seven hours of audio data, as indicated in Table 2. Phone types refer to the number of different phones that occur in the data, while phone tokens indicate their total number.

Database name	Speech (hours)	No. of utterances	No. of speakers	Phone types	Phone tokens	Word tokens
Afrikaans	6.18	9 520	234	84	180 904	47 383
English	6.02	9 904	271	73	167 986	47 941
Xhosa	6.98	8 538	219	107	177 843	36 676
Zulu	7.03	8 295	203	101	187 249	35 568

Table 2
Training sets for each database.

Table 3 shows the extent to which the phone set, which corresponds to the set of unique phone types in the training set, of each language covers the phone

types and tokens found in the other language's training set. For example, 75.6% of phone types and 98.4% of phone tokens in the Afrikaans training set are also present among the English phone types. From the table we see that especially Xhosa and Zulu have many phones in common. This is also true for Afrikaans and English, but to a lesser degree.

Phone types of language	Covers % of training set phone types/tokens in:			
	Afrikaans	English	Xhosa	Zulu
Afrikaans	100/100	87.3/99.5	61.9/87.5	60.6/88.1
English	75.6/98.4	100/100	55.2/86.6	51.5/86.5
Xhosa	79.3/97.4	81.7/92.3	100/100	90.9/99.9
Zulu	73.2/92.6	71.8/85.7	85.7/99.8	100/100

Table 3
Degree to which the monophone types of each language cover the training set monophone types and tokens of every other language.

Since our acoustic models will be context dependent, it is useful also to consider the similarity between the triphones of each language. Accordingly, Table 4 shows the extent to which the training set triphone types of each language cover the training set triphone tokens of the other languages. Once again, there is a great deal of commonality between Xhosa and Zulu. English triphones show the greatest overlap with Afrikaans, but this is not true for Afrikaans, which exhibits a slightly better coverage of both Xhosa and Zulu triphones.

Triphone types of language	Covers % of training set triphone tokens in:			
	Afrikaans	English	Xhosa	Zulu
Afrikaans	100	25.6	34.5	32.9
English	29.3	100	18.1	13.1
Xhosa	31.9	26.6	100	88.3
Zulu	34.8	22.9	87.3	100

Table 4
Degree to which the triphone types of each language cover the training set triphone tokens of every other language.

The test set for each language contains approximately 25 minutes of speech data, as shown in Table 5. There was no speaker-overlap between the test and training sets, and each contained both male and female speakers.

Database name	Speech (minutes)	No. of utterances	No. of speakers	Phone tokens	Word tokens
Afrikaans	24.4	750	20	11 441	3 051
English	24.0	702	18	10 338	3 652
Xhosa	26.8	609	17	10 925	2 480
Zulu	27.1	583	16	11 008	2 385

Table 5
Test sets for each database.

Finally, a separate development set, consisting of approximately 15 minutes of speech from 10 speakers in each language was also prepared. This data was used only in the optimisation of recognition parameters, before final evaluation on the test-set. There is no overlap between the development set and either the test or training sets.

3 General experimental method

The HTK tools were used to develop and test recognition systems (Young et al., 2002). The speech audio data was parameterised as 39-dimensional feature vectors consisting of 13 Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials, with cepstral mean normalisation (CMN) applied on a per-utterance basis. From this parameterised training set and its phonetic transcription, diagonal-covariance cross-word triphone models with three states per model and eight Gaussian mixtures per state were trained by embedded Baum-Welch re-estimation and decision-tree state clustering (Young et al., 1994).

The decision-tree state clustering process begins by pooling all context-dependent phones found in the training corpus that correspond to the same context-independent phone, termed the basephone hereafter. A set of linguistically-motivated questions is defined with which these clusters can be split. Such questions may, for example, ask whether the left context of a particular context-dependent phone is a vowel, or whether the right context is a silence. The clusters are subdivided repeatedly, at each iteration applying the question that affords the largest improvement in training set likelihood. The process ends either when this likelihood gain falls below a certain threshold, or when the number of occurrences remaining in a cluster becomes too small. Hence the clustering process results in a binary decision tree for each state of each basephone. The leaves of this tree are clusters of context-dependent phones whose training data must subsequently be pooled.

A great advantage of this clustering method is that context-dependent phones not encountered in the training data at all can easily be synthesised by means of the decision trees that have been determined for the corresponding basephone. This is important when using cross-word context dependent models, or when the phone set is large or the training set small and hence sparse.

The related research by Schultz and Waibel has applied decision-tree clustering to multilingual acoustic modelling using tied-mixture systems (Schultz and Waibel, 2001). In these systems the HMMs share a single large set of Gaussian distributions, with state-specific mixture weights. This configuration allows the clustering process to employ an entropy-based distance measure based

on the mixture weights to determine the similarity between states. Our configuration may in contrast be described as a tied-state system. Each set of clustered states shares a much smaller Gaussian mixture distribution, but the distribution is completely separate for each set of clustered states. Since the entropy-based distance measure cannot be used for this configuration, our clustering procedure is based on the reduction in training set likelihood associated with a cluster subdivision.

Since the vocabularies of the AST databases vary widely between languages, comparison of recognition performance will be based on phoneme error rates, as is also proposed in (Schultz and Waibel, 1998; Waibel et al., 2000). All speech recognition experiments are performed using a backoff bigram language model obtained for each language from the training set phoneme transcriptions (Katz, 1987). Absolute discounting was used for the estimation of language model probabilities (Ney et al., 1994).

Database	Bigram types	Perplexity
Afrikaans	1420	11.84
English	1900	14.08
Xhosa	2003	12.72
Zulu	1886	12.57

Table 6
Bigram language model perplexities measured on corresponding test-sets.

Language model perplexities are shown in Table 6. Word-insertion penalties and language model scale factors used during recognition were optimised on the development set. Note that, as indicated in Table 2, the phone sets of each language differ considerably. Of the 135 different phone labels present in the annotations, only 55 are common to all four languages.

4 Language-specific acoustic models

To serve as a baseline, a fully language-specific system was developed, that allows no sharing between languages. Model development begins by pooling all triphones with the same basephone separately for each language. The decision tree clustering process then employs only questions relating to the phonetic character of the left and the right context. The structure of the resulting acoustic models is illustrated in Figure 1 for two languages (Xhosa and Zulu) and a single triphone.

Since no overlap is allowed between the triphones of different languages, this baseline corresponds to a completely separate set of acoustic models for each language.

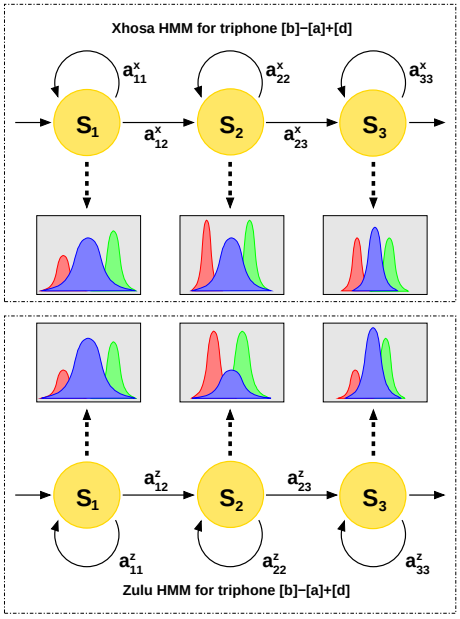

Fig. 1. Language-specific acoustic models.

5 Language-independent acoustic models

In addition to the four language specific acoustic model sets, a single language-independent acoustic model set was trained by pooling the data across all four languages for phones with the same IPA classification. This will no longer allow the models to differentiate between subtly different spectral qualities of the same IPA phone used in different languages. It also no longer allows language-specific co-articulation effects to be taken into account during the decision-tree triphone clustering step of model development. Hence the performance of these models is expected to be a lower bound for the performance of multilingual state clustering that will be introduced in the next section.

Figure 2 illustrates the structure of the language-independent models, again for just two languages and a single triphone. Both languages share the same Gaussian mixture probability distributions, as well as HMM transition probabilities.

6 Multilingual acoustic models

The development of our multilingual model set is similar to that of the language-independent model set. The state tying process begins by pooling all triphones of all four languages corresponding to the same basephone. However in this case the set of decision-tree questions take into account not only the phonetic character of the left or right context, but also the language of the basephone.

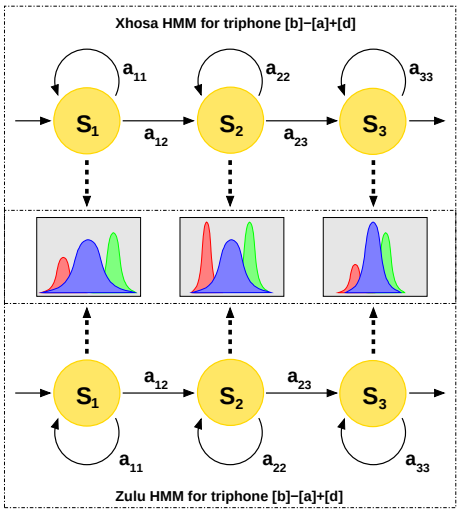

Fig. 2. Language-independent acoustic models.

Two phonemes with the same IPA symbol but from different languages can therefore be kept separate if there is a significant acoustic difference, or can be merged if there is not. For example, a pool of triphones with basephone [a] can be split by a question asking whether the triphone is a Zulu triphone or not. This allows tying across languages when the triphone states are acoustically similar, and separate modelling of the same triphone state for different languages when there are differences.

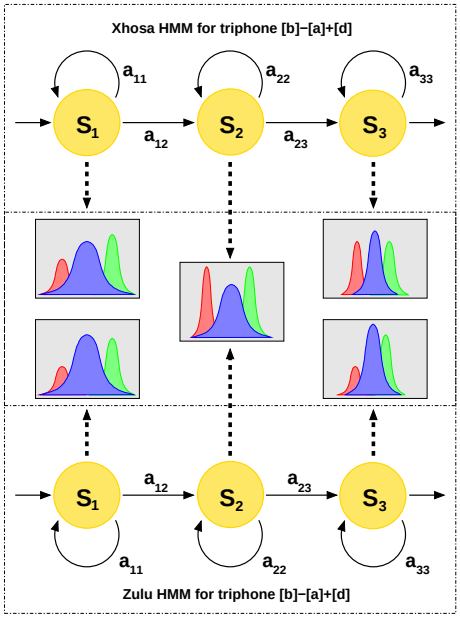

Fig. 3. Multilingual acoustic models.

The structure of such multilingual acoustic model set is shown in Figure 3. Here the centre state of the triphone [b]-[a]+[d] is tied, but the first and last states are modelled separately for each language. In our experiments, the transition probabilities of all triphones with the same basephone were tied,

regardless of language.

7 Experimental results

We have applied the acoustic modelling approaches described in Sections 4, 5 and 6 to the combination of the Afrikaans, English, Xhosa and Zulu training data described in Section 2. Since the optimal number of parameters for the acoustic models was not known, several sets of HMMs were produced by varying the likelihood-improvement threshold used during decision-tree clustering, as described in Section 3. Decision-tree clustering was carried out using HMM sets with single-mixture Gaussian densities per state. Clustering was followed by four iterations of embedded Baum-Welch re-estimation. The number of mixtures per state was then gradually increased to eight, each such increase being followed by a further four iterations of embedded training. The performance, in terms of average phone accuracy, of the final 8-mixture HMM set measured on the evaluation test set is shown in Figure 4. A single curve indicating the average accuracy over all four test-sets is shown, and the number of states in the language-specific systems was taken to be the sum of the number of states in each component language-specific HMM set. The number of states in the multilingual system is the total number of unique states remaining after decision-tree clustering, and hence takes cross-language sharing into account.

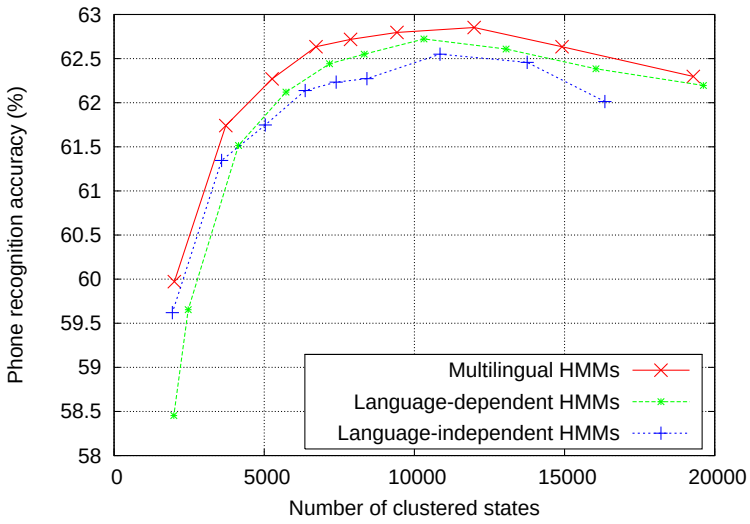

Fig. 4. Average evaluation test set phone accuracies of language-specific, multilingual and language-independent systems as a function of the total number of distinct HMM states.

Figure 4 indicates that, for most systems, pooling the data between languages for phones with the same IPA symbol leads to a deterioration in performance. This agrees with the findings of other studies. However, in contrast to previous findings, the results also indicate that the multilingual systems achieved

modest average performance improvements relative to both the language-dependent and the language-independent systems over all the models in the range considered. This suggests that beneficial cross-lingual sharing is in fact taking place. The improvement achieved by the best multilingual system relative to the best language-dependent system shown in Figure 4 is approximately 0.1% absolute, and has been determined to be statistically significant only at the 80% level using bootstrap confidence interval estimation (Bisani and Ney, 2004). Nevertheless, the improvements are consistently achieved over different model sizes.

Closer inspection of the results shows that the average improvement is usually accompanied by per-language gains. Figure 5 shows the respective recognition performance of the multilingual and the language-dependent systems measured separately on the evaluation test set of each language. The differing phone recognition accuracies are a result of the differing phone sets used by each language, as well as of the different language model perplexities. In particular, Afrikaans, which has the highest phone accuracies, has the lowest language model perplexity. In contrast, Xhosa and Zulu, which have the lowest accuracies, have the largest phone sets. Finally, English displays the second highest phone accuracies, and has the smallest phone set, but has the highest language model perplexity of all.

Figure 5 indicates that in the majority of cases the multilingual system delivers improved performance relative to a corresponding language-dependent system. However, since the clustering process optimises the overall likelihood, improvements for each individual language are not guaranteed. Indeed, Figure 5 demonstrated that unchanged or slightly deteriorated per-language performance does sometimes occur, despite an improvement in the associated overall average.

Inspection of the type of questions most frequently used during clustering reveals that language-based questions are most common at the root of the decision tree, and become increasingly less frequent towards the leaves. Figure 6 analyses the decision tree of the largest multilingual system (with almost 20 000 states), and shows that approximately 45% of all questions at the root nodes are language-based, and that this proportion drops to 22% and 15% for the roots' children and grandchildren respectively.

This behavior is reflected also when considering the contribution to the reduction in log likelihood made by the language-based and phonetically-based questions respectively during the decision tree growing process. In Figure 7 this contribution is shown, again as a function of the depth within the decision tree. It is evident that, at the root node, the greatest log likelihood reduction is afforded by the language-based questions (approximately 74% of the total reduction), while the phonetically-based questions make up the greatest

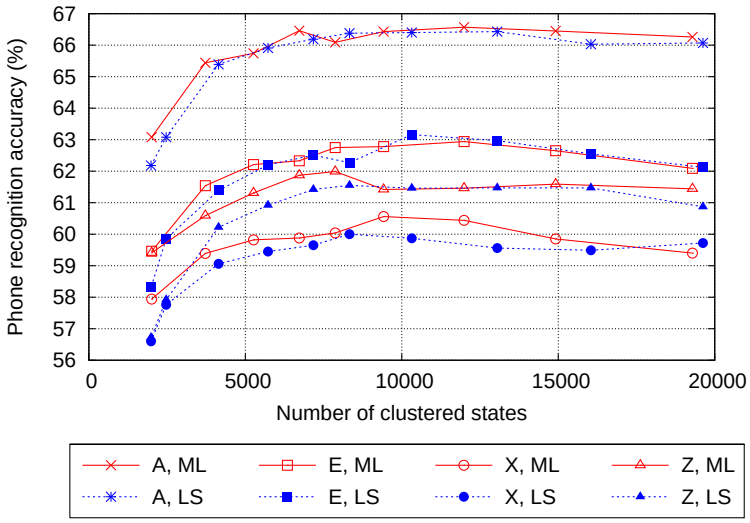

Fig. 5. Phone accuracies of multilingual (ML) and language-specific (LS) systems as a function of the total number of distinct HMM states, measured separately for the Afrikaans (A), English (E), Xhosa (X) and Zulu (Z) evaluation test sets.

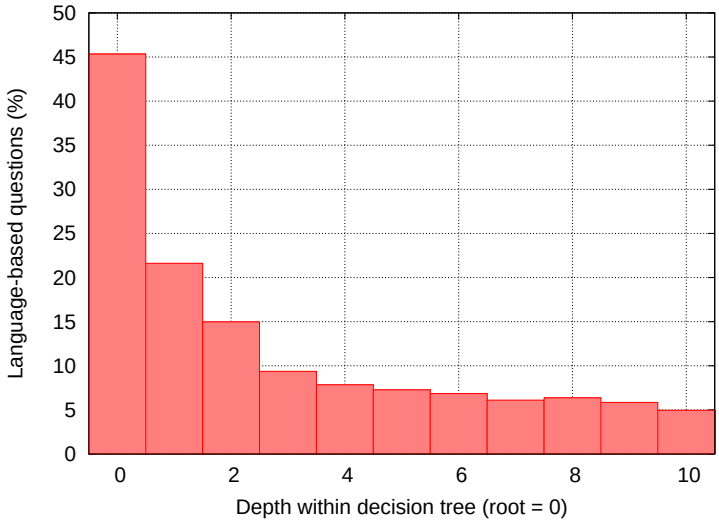

Fig. 6. Analysis showing the percentage of questions that are language-based at various depths within the multilingual decision tree.

contribution thereafter. This indicates that the decision tree often quickly partitions the models into language-based groups, after which further refinements to the tree are based more on phonetic distinctions.

In order to determine to what extent and for which languages data sharing ultimately takes place for the various multilingual systems, we have determined the proportions of decision tree leaf nodes (which correspond to the state clusters) that are populated by states of exactly one, two, three or four languages respectively. A cluster populated by states of a single language corresponds to a unilingual cluster, and indicates that no sharing with other languages has

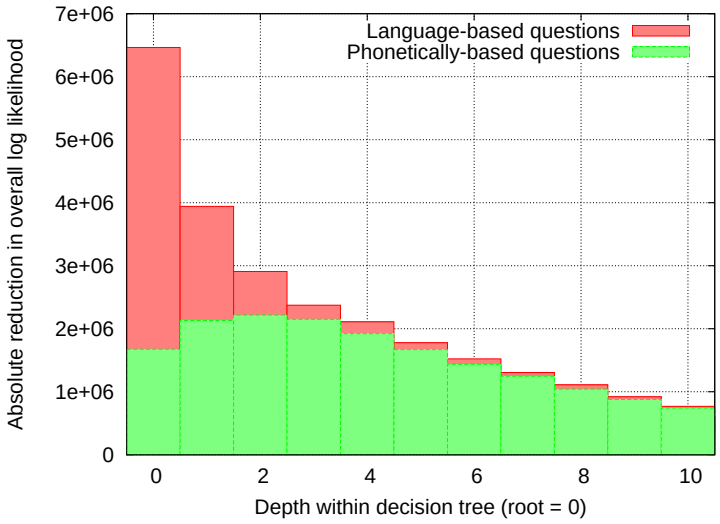

Fig. 7. Analysis showing the contribution made to the reduction in overall log-likelihood by the language-based and phonetically-based questions respectively.

taken place. A cluster populated by states of all four languages, on the other hand, indicates that sharing across all four languages has taken place. Figure 8 illustrates how these proportions change as a function of the total number of clustered states in the system.

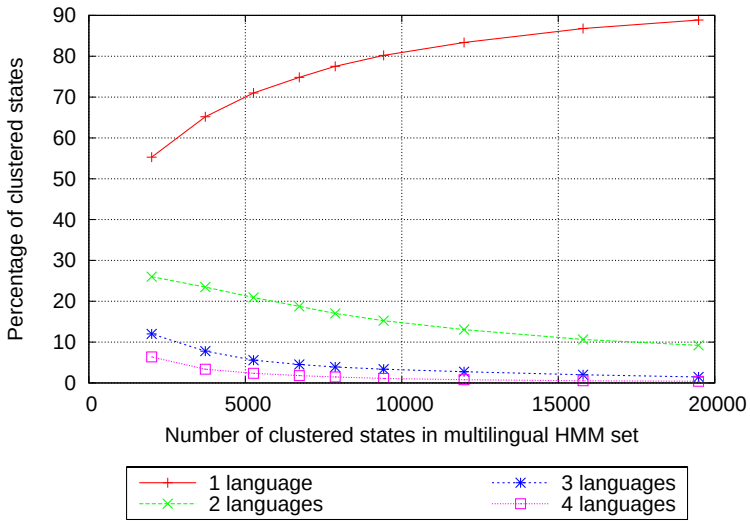

Fig. 8. Proportion of state clusters combining data from one (unilingual), two (bilingual), three (trilingual) and four (quadralingual) languages.

From Figure 8 it is apparent that, as the number of clustered states is increased, the proportion of clusters consisting of a single language also increases. This indicates that the progressive specialisation of the decision tree is tending to produce separate clusters for each language, as one would find in a language-dependent system. The proportion of clusters containing two, three and all four languages shows a commensurate decrease as the number

of clustered states increases. Nevertheless, for an acoustic model with 10 000 states, which according to Figure 4 yields approximately optimal recognition accuracy, around 20% of the state clusters contain a mixture of languages, demonstrating that a significant degree of sharing is taking place.

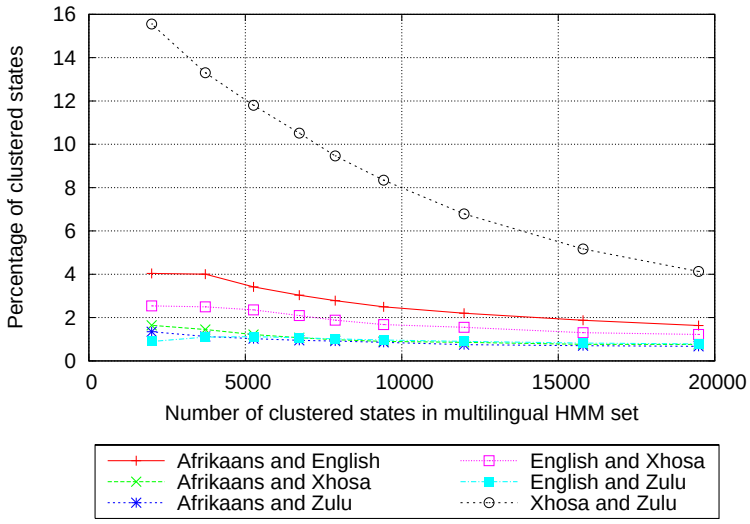

Fig. 9. Proportion of state clusters combining data from various combinations of two languages (bilingual clusters).

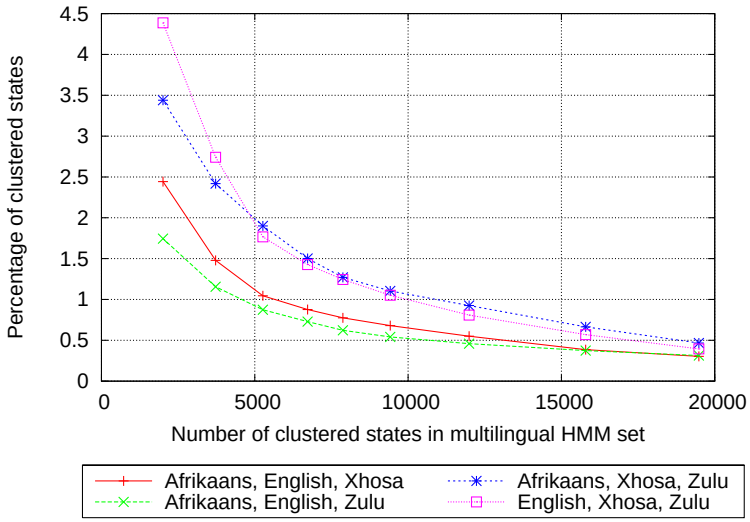

Fig. 10. Proportion of state clusters combining data from various combinations of three languages (trilingual clusters).

In order to determine which languages are being shared most often by the clustering procedure, Figures 9 and 10 analyse the proportion of states that consist of groups of two and three languages respectively. From Figure 9 we see that the largest proportion of two-language clusters are due to a combination of Xhosa and Zulu. This agrees with our earlier observations of the high phonetic similarity of these languages, even at the triphone level. Afrikaans and English are the second most frequent language combination, but are much less common

than the Xhosa and Zulu clusters. Figure 10 shows that, in cases where three languages are clustered together, the combinations of Afrikaans, Xhosa and Zulu, and of English, Xhosa and Zulu are similarly frequent. In contrast, clusters containing Afrikaans, English and one of the Nguni languages are much less common.

8 Summary and conclusions

We have demonstrated that decision tree state clustering can be employed to obtain multilingual acoustic models by allowing sharing between basephones of different languages and introducing decision tree questions that relate to the language of a particular basephone. This mode of clustering was used to combine Afrikaans, English, Xhosa and Zulu acoustic models. Improvements over separate-language as well as language-independent systems were observed. Further analysis showed that language-based decision tree questions play a dominant role at and near the root node, indicating that language specialisation occurs quickly during tree construction. However it is also observed that a significant proportion of state clusters contain more than one language, from which we conclude that language sharing occurs regularly in the multilingual acoustic models. Hence this technique promises to be useful for the development of acoustic models for use within multilingual speech recognition systems.

9 Acknowledgements

The author would like to thank Febe de Wet, Thomas Hain and Tanja Schulz for their very helpful suggestions. This work was supported by the South African National Research Foundation (NRF) under grant number FA2005022300010.

References

Bisani, M., Ney, H., 2004. Bootstrap estimates for confidence intervals in ASR performance evaluation. In: Proc. ICASSP. Montreal, Canada.

International Phonetic Association, 2001. Handbook of the International Phonetic Association. Cambridge University Press.

Katz, S., 1987. Estimation of probabilities from sparse data for the language model component of a speech recogniser. IEEE Transactions on Acoustics Speech and Signal Processing 35 (3), 400–401.

Köhler, J., 2001. Multi-lingual phone models for vocabulary independent speech recognition tasks. Speech Communication 35 (1-2), 21–30.

Louw, P., 2005. A new definition of Xhosa grapheme-to-phoneme rules for automatic transcription. South African Journal of African Languages 25 (2), 71–91.

Louw, P., Roux, J., Botha, E., 2001. African Speech Technology (AST) telephone speech databases: corpus design and contents. In: Proc. Eurospeech. Aalborg, Denmark.

Ney, H., Essen, U., Kneser, R., 1994. On structuring probabilistic dependencies in stochastic language modelling. Computer, Speech and Language 8 (1), 1–38.

Niesler, T., Louw, P., Roux, J., 2005. Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. Southern African Linguistics and Applied Language Studies 23 (4), 459–474.

Roux, J., Louw, P., Niesler, T., 2004. The African Speech Technology project: An assessment. In: Proc. LREC. Lisbon, Portugal.

Schultz, T., Waibel, A., 1998. Language independent and language adaptive large vocabulary speech recognition. In: Proc. ICSLP. Sydney, Australia.

Schultz, T., Waibel, A., 2000. Polyphone decision tree specialisation for language adaptation". In: Proc. ICASSP. Istanbul, Turkey.

Schultz, T., Waibel, A., 2001. Language independent and language adaptive acoustic modelling for speech recognition. Speech Communication 35 (1-2), 31–51.

Statistics South Africa (Ed.), 2004. Census 2001: Primary tables South Africa: Census 1996 and 2001 compared. Statistics South Africa.

Uebler, U., 2001. Multilingual speech recognition in seven languages. Speech Communication 35 (1-2), 53–69.

Uebler, U., Schüßler, M., Niemann, H., 1998. Bilingual and dialectal adaptation and retraining. In: Proc. ICSLP. Sydney, Australia.

Waibel, A., Geutner, P., Mayfield-Tomokiyo, L., Schultz, T., Woszczyna, M., August 2000. Multilinguality in speech and spoken language systems. Proc. IEEE 88 (8), 1297–1313.

Webb, V., 2002. Language policy development in South Africa. In: World Congress on Language Policies. Barcelona, Spain.

Wissing, D., Martens, J.-P., Janke, U., Goedertier, W., 2004. A spoken Afrikaans language resource designed for research on pronunciation variations. In: Proc. LREC. Lisbon, Portugal.

Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2002. The HTK book, version 3.2.1. Cambridge University Engineering Department.

Young, S., Odell, J., Woodland, P., 1994. Tree-based state tying for high accuracy acoustic modelling. In: Proc. Workshop on Human Language Technology. Morgan Kaufmann, pp. 307–312.