LANGUAGE-DEPENDENT STATE CLUSTERING FOR MULTILINGUAL SPEECH RECOGNITION IN AFRIKAANS, SOUTH AFRICAN ENGLISH, XHOSA AND ZULU

Thomas Niesler

Department of Electrical and Electronic Engineering
University of Stellenbosch, South Africa
trn@dsp.sun.ac.za

Abstract

The development of automatic speech recognition systems requires significant quantities of annotated acoustic data. In South Africa, the large number of spoken languages hampers such data collection efforts. Furthermore, code switching and mixing are commonplace since most citizens speak two or more languages fluently. As a result a considerable degree of phonetic cross pollination between languages can be expected. We investigate whether it is possible to combine speech data from different languages in order to improve the performance of a speech recognition system in any one language. For our investigation we use recently collected Afrikaans, South African English, Xhosa and Zulu speech databases. 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 when applied to Afrikaans and English, and to Xhosa and Zulu. The improvement for the latter pair of languages is greater, which is consistent with their larger degree of phonetic similarity.

1. Introduction

Multilingual speech recognition is a particularly relevant challenge in South Africa, which has 11 official languages and among whose population monolinguality is almost entirely absent. We study the four languages Afrikaans, English, Xhosa and Zulu, which are spoken as a mother-tongue by 13.3%, 8.2%, 17.6% and 23.8% of the population respectively [1]. This represents a balance between languages of European descent, and indigenous African languages. Furthermore, a certain amount of phonetically and orthographically annotated speech data is available for these languages.

The first step in the development of speech recognition systems for a new language is normally the recording and annotation of large quantities of spoken audio data. In general, the more data is available, the better the performance of the systems. However, data gathering and especially annotation are very expensive in terms of money and time.

It is in this light that we would like to determine whether data from different languages can be combined in order to improve the performance of a speech recognition system in any single language. This involves determining phonetic similarities between the languages, and exploiting these to determine more robust and effective acoustic models.

In related work, Schultz and Waibel have tried to share the data of 10 languages, each spoken in a different country [2, 3, 4]. Under these conditions, it has not been possible to improve the performance of monolingual speech recognition systems by means of additional data from another language. Kohler [5] and Uebler [6] have come to similar conclusions.

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. Furthermore, strong links exist between certain groups of indigenous languages (such as the Nguni language group), which may allow more fruitful sharing of data that can be exploited in speech recognition applications [7].

2. 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 [8]. For their compilation, speakers were recruited from targeted language groups and given a unique datasheet 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.

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, 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 eleven hours of audio data, as indicated in Table 1. Phone types refer to the number of different phones that occur in the data, while phone tokens indicate their total number. Note that a slightly lower speech rate was observed for Xhosa and Zulu compared with Afrikaans and English.

Each test set contains approximately 25 minutes of speech data, as shown in Table 2. There was no speaker-overlap be-

tween the test and training sets, and each contained both male and female speakers.

Database name	Speech (hours)	No. of speakers	Phone types	Phone tokens
Afrikaans	6.18	234	84	180,904
English	6.02	271	73	167,986
Xhosa	6.98	219	107	177 843
Zulu	10.87	203	101	285,501

Table 1: Training sets for each database.

Database name	Speech (minutes)	No. of speakers	Phone tokens
Afrikaans	24.4	20	11 441
English	24.0	18	10 338
Xhosa	26.8	17	10 925
Zulu	27.1	16	11 008

Table 2: Test sets for each database.

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 [9]. The speech audio data was parameterised as 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 [10]. Since decision-tree state clustering is central to the work we present, it will be described shortly.

The process normally begins by pooling all context-dependent phones found in the training corpus that have the same basephone, effectively resulting in monophone models. 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.

Since the vocabularies in the AST databases vary widely between languages, comparison of recognition performance will be based on phoneme error rates, as is also proposed in [2, 11]. All speech recognition experiments are performed using a backoff bigram language model obtained for each language from the training set phoneme transcriptions [12].

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

Table 3: Bigram language model perplexities measured on corresponding test-sets.

Absolute discounting was used for the estimation of language model probabilities [13]. Language model perplexities are shown in Table 3. Word-insertion penalties and language model scale factors were optimised on the development test-set.

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 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.

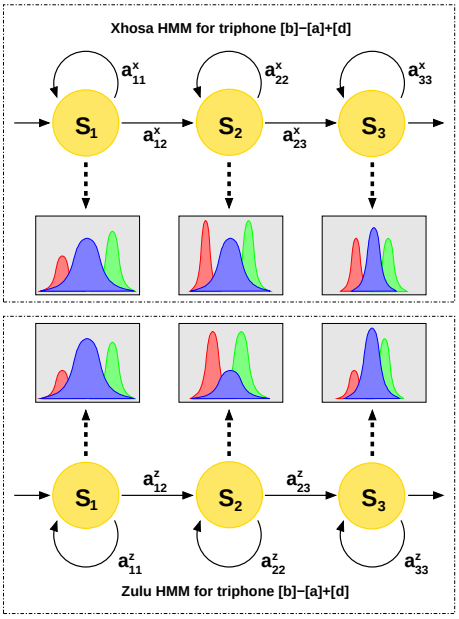

Figure 1: Language-specific acoustic models.

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.

5. Multilingual acoustic models

In order to obtain multilingual models, the state tying process begins by pooling all triphones of all languages corresponding to the same basephone. The set of decision-tree questions now take into account not only the phonetic character of the left or right context, but also the language of the basephone. 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.

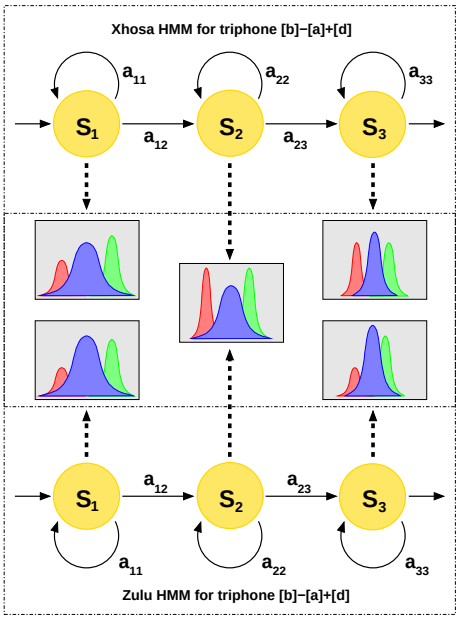

Figure 2: Multilingual acoustic models.

The structure of such multilingual acoustic model set is shown in Figure 2. 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.

6. Results

We have applied the acoustic modelling approaches described in Sections 4 and 5 first to the combination of Afrikaans and English, and then to the combination of Xhosa and Zulu. 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 five iterations of embedded Baum-Welch reestimation. The number of mixtures per state was then gradually increased to eight, each such increase being followed by a further five iterations of embedded training. The performance, in terms of phone accuracy, of the final 8-mixture HMM sets for the Afrikaans/English and the Xhosa/Zulu combinations are shown in Figures 3 and 4. In each case, a single curve indicating the average accuracy over both language's 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.

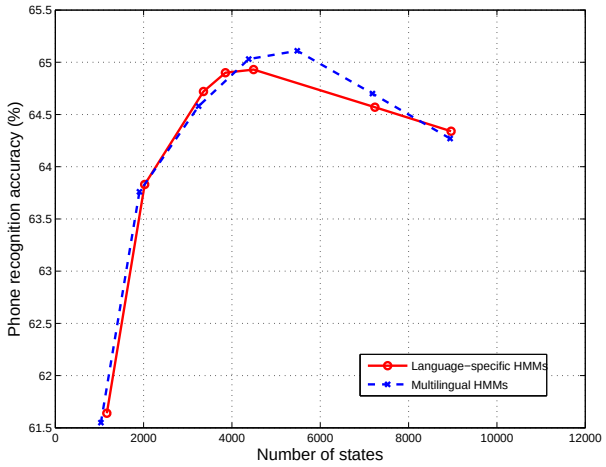

Figure 3: Phone accuracies of language-specific and multilingual systems for Afrikaans and English as a function of the total number of distinct HMM states.

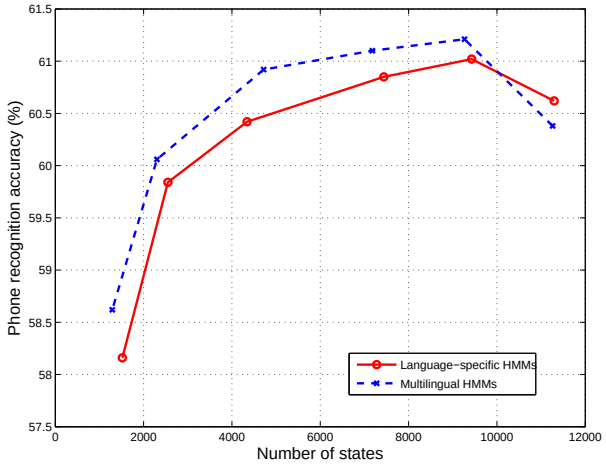

Figure 4: Phone accuracies of language-specific and multilingual systems for Xhosa and Zulu as a function of the total number of distinct HMM states.

7. Discussion and conclusions

It is evident from Figures 3 and 4 that the multilingual systems achieved performance improvements for both the English/Afrikaans and the Xhosa/Zulu combinations. However, the improvement is greater and achieved over a a larger range of HMM set sizes for Xhosa and Zulu. We believe that this may be ascribed to the much greater phonetic similarity between the latter pair of languages, which both belong to the Nguni language

group. Figures 5 and 6 illustrate a measure of this similarity for Afrikaans and English, and for Xhosa and Zulu respectively. The graphs show the extent to which the most frequent training-set triphones cover the test-set triphones of each language.

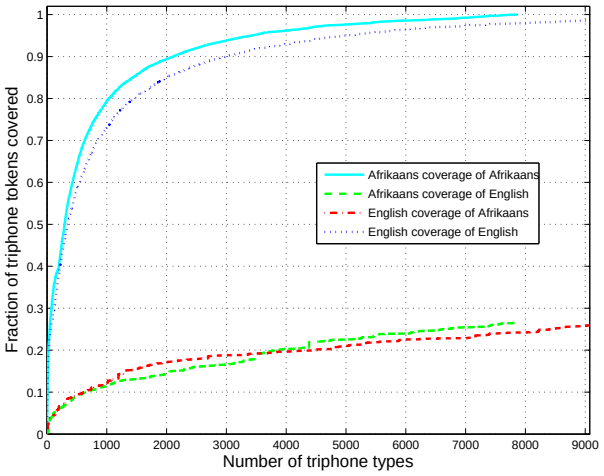

Figure 5: Proportion of triphones in Afrikaans and English test sets covered by the most frequent triphones in the Afrikaans and English training sets.

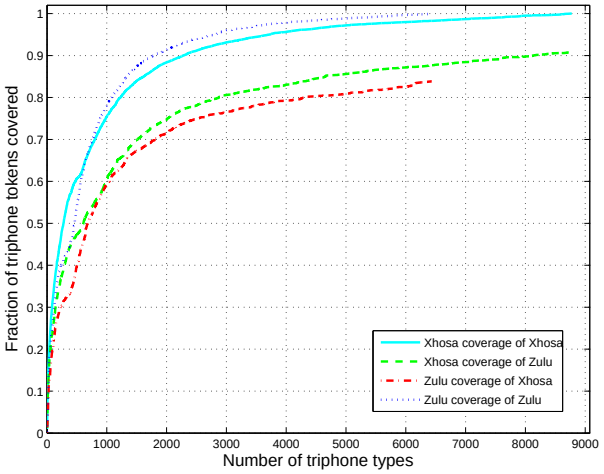

Figure 6: Proportion of triphones in Xhosa and Zulu test sets covered by the most frequent triphones in the Xhosa and Zulu training sets.

The best coverage of test-set triphones is in every case achieved by the corresponding language itself. However from Figure 5 we see that the cross-language triphone coverages for Afrikaans and English are below 30%, while these figure exceed 80% for Xhosa and Zulu. In the light of the small phonetic commonality between Afrikaans and English, we believe even the small improvement in phone recognition accuracy achieved by the Afrikaans/English multilingual model to be promising.

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 and English, as well as Xhosa and Zulu acoustic models. In both cases, improvements over separate-language systems were observed. Furthermore, this improvement was greater for the Xhosa/Zulu combination, which agrees with the empirically-observed greater phonetic similarity between these languages.

9. Acknowledgements

This work was supported by the National Research Foundation (NRF) under grant number FA2005022300010.

10. References

[1] Statistics South Africa, Ed., *Census 2001: Primary tables South Africa: Census 1996 and 2001 compared*. Statistics South Africa, 2004.

[2] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Proc. ICSLP*, Sydney, Australia, 1998.

[3] ——, "Polyphone decision tree specialisation for language adaptation"," in *Proc. ICASSP*, Istanbul, Turkey, 2000.

[4] ——, "Language independent and language adaptive acoustic modelling for speech recognition," *Speech Communication*, vol. 35, pp. 31–51, 2001.

[5] J. Köhler, "Multi-lingual phone models for vocabulary independent speech recognition tasks," *Speech Communication*, vol. 35, pp. 21–30, 2001.

[6] U. Uebler, "Multilingual speech recognition in seven languages," *Speech Communication*, vol. 35, pp. 53–69, 2001.

[7] T. Niesler, P. Louw, and J. Roux, "Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases," *Southern African Linguistics and Applied Language Studies*, vol. 23, no. 4, pp. 459–474, 2005.

[8] J. Roux, P. Louw, and T. Niesler, "The African Speech Technology project: An assessment," in *Proc. LREC*, Lisbon, Portugal, 2004.

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

[10] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using HTK," in *Proc. ICASSP*, Adelaide, 1994, pp. 125–128.

[11] A. Waibel, P. Geutner, L. Mayfield-Tomokiyo, T. Schultz, and M. Woszczyna, "Multilinguality in speech and spoken language systems," *Proc. IEEE*, vol. 88, no. 8, pp. 1297–1313, August 2000.

[12] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recogniser," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 35, no. 3, pp. 400–401, March 1987.

[13] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Computer, Speech and Language*, vol. 8, pp. 1–38, 1994.