Language-dependent state clustering for multilingual speech recognition in Afrikaans, South African English, Xhosa and Zulu

Thomas Niesler

Digital Signal Processing Laboratory

University of Stellenbosch
Private Bag X1
Matieland, Stellenbosch 7602

INTRODUCTION

- Multilingual speech recognition particularly relevant in South Africa
  - 11 officially-recognised languages
  - Multilinguality is the norm

- Speech corpora are scarce and expensive to develop

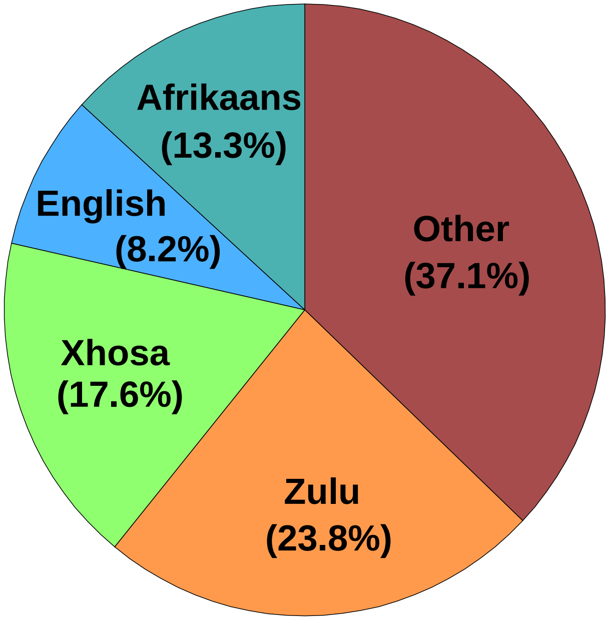
- **<u>Aim</u> :** determine whether data from different languages can be combined to improve the speech recognition performance in any single language

- All spoken in same country ⇒ phonetic and lexical sharing occurs

- Some languages have common origins

LANGUAGES

- We study four widely-spoken languages (first language to 63% of population)



- Afrikaans and English are European Germanic languages
- Xhosa and Zulu are African indigenous Nguni languages
- Phonetically and orthographically annotated data available

SPEECH DATABASES

- Telephone speech data gathered over both mobile and fixed networks
- Speakers were recruited and instructed to read from unique datasheets
  - Phonetically-rich sentences
  - Mix of read and spontaneous items
- Databases have been annotated and validated by human experts
  - Orthographically
  - Phonetically
- Databases gathered in the same manner and datasheets designed in the same way across languages

TRAINING AND TEST SETS

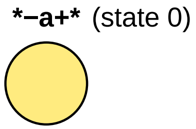
- The acoustic data was divided into testing- and training-sets
  - No speaker overlap
  - Approximate male/female and mobile/landline balance

Database name	Training set				Test set		
	Speech (hours)	No. of speakers	Phone types	Phone tokens	Speech (mins)	No. of speakers	Phone tokens
Afrikaans	6.18	234	84	180 904	24.4	20	11 441
English	6.02	271	73	167 986	24.0	18	10 338
Xhosa	6.98	219	107	177 843	26.8	17	10 925
Zulu	10.87	203	101	285 501	27.1	16	10 008

- Separate development set (not shown) used to optimise recognition parameters

DECISION-TREE STATE CLUSTERING

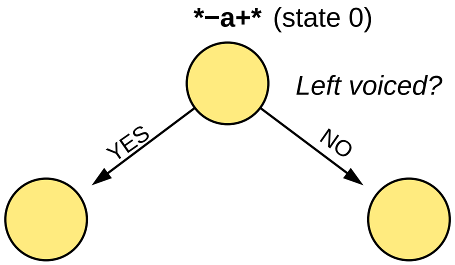
- Begin by pooling all triphones for same basephone in training set
- Create separate pool for each state



- Introduce a set of linguistically-defined questions to split clusters

DECISION-TREE STATE CLUSTERING

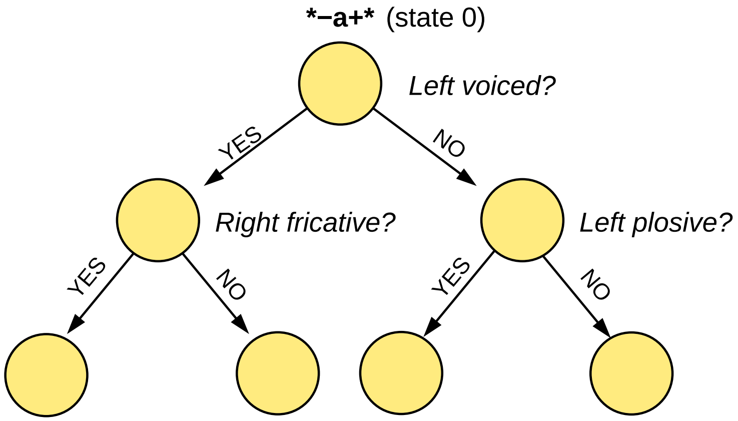
- Begin by pooling all triphones for same basephone in training set
- Create separate pool for each state



- Introduce a set of linguistically-defined questions to split clusters
- Determine question leading to greatest likelihood improvement and split

DECISION-TREE STATE CLUSTERING

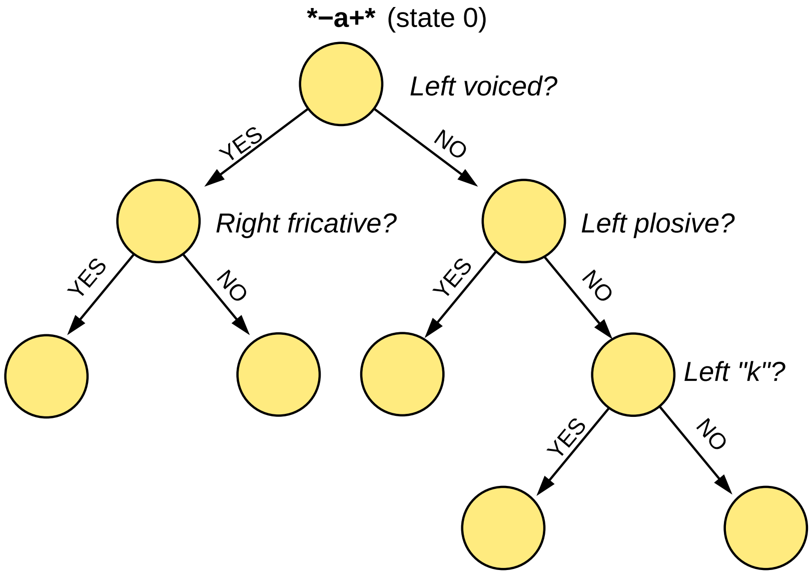
- Begin by pooling all triphones for same basephone in training set
- Create separate pool for each state



- Introduce a set of linguistically-defined questions to split clusters
- Determine question leading to greatest likelihood improvement and split
- Repeat until likelihood improvement too small

DECISION-TREE STATE CLUSTERING

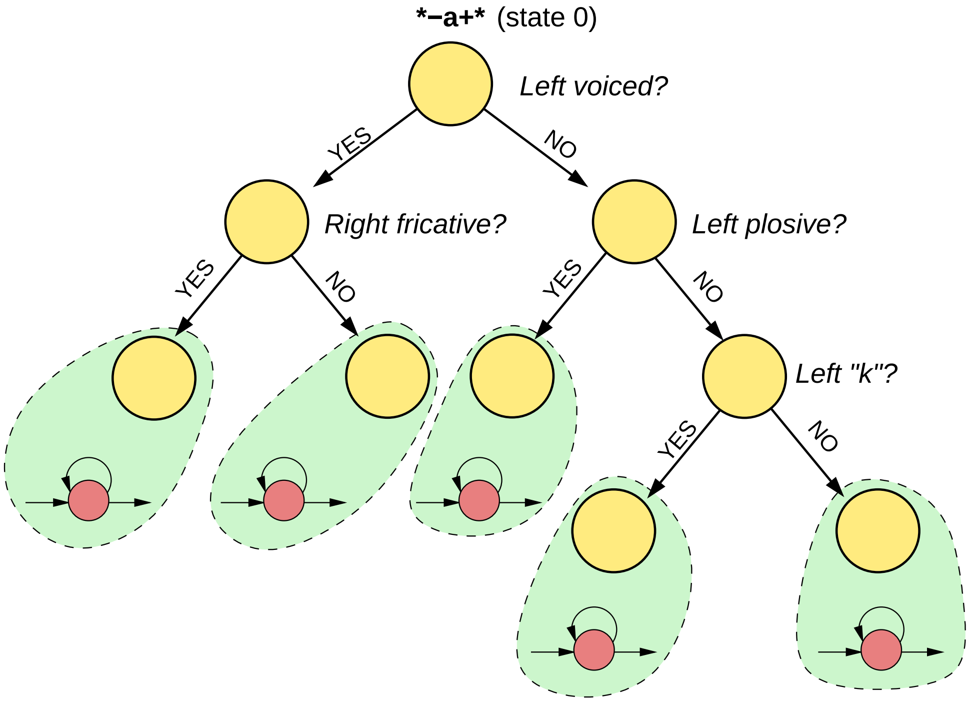
- Begin by pooling all triphones for same basephone in training set
- Create separate pool for each state



- Introduce a set of linguistically-defined questions to split clusters
- Determine question leading to greatest likelihood improvement and split
- Repeat until likelihood improvement too small

DECISION-TREE STATE CLUSTERING

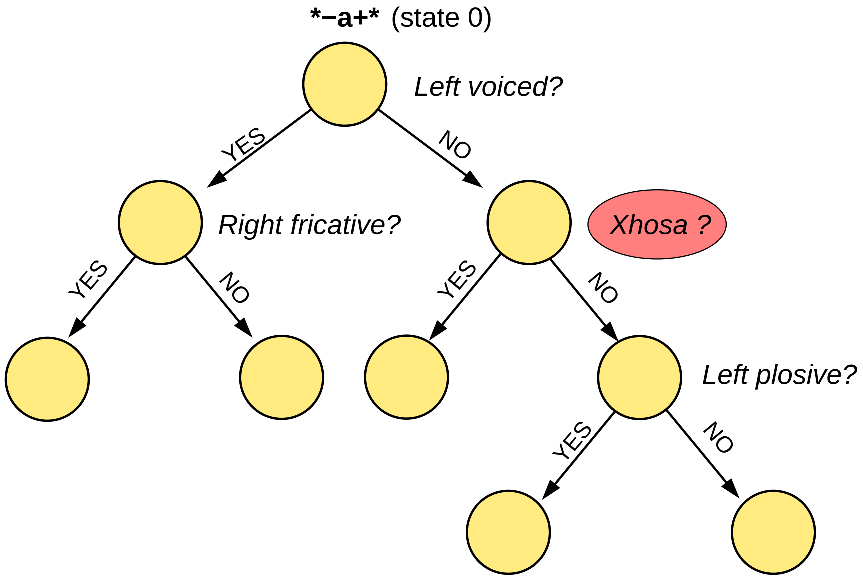
- Finally, each tree leaf corresponds to a cluster of HMM states



- Unseen context-dependent phones can be synthesised using the decision tree

MULTILINGUAL DECISION-TREE STATE CLUSTERING

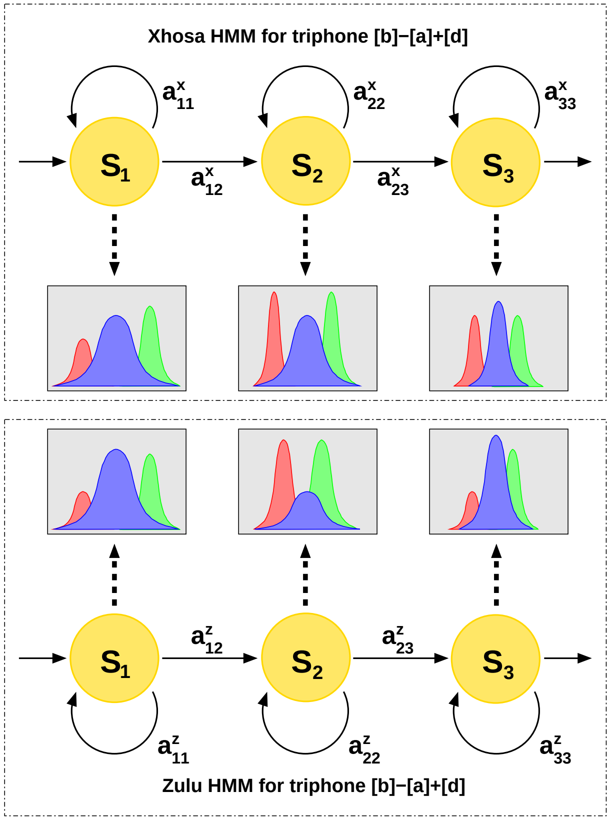
- Allow decision-tree questions to concern language as well as phonetic context



- Tag phones with language before pooling at root nodes

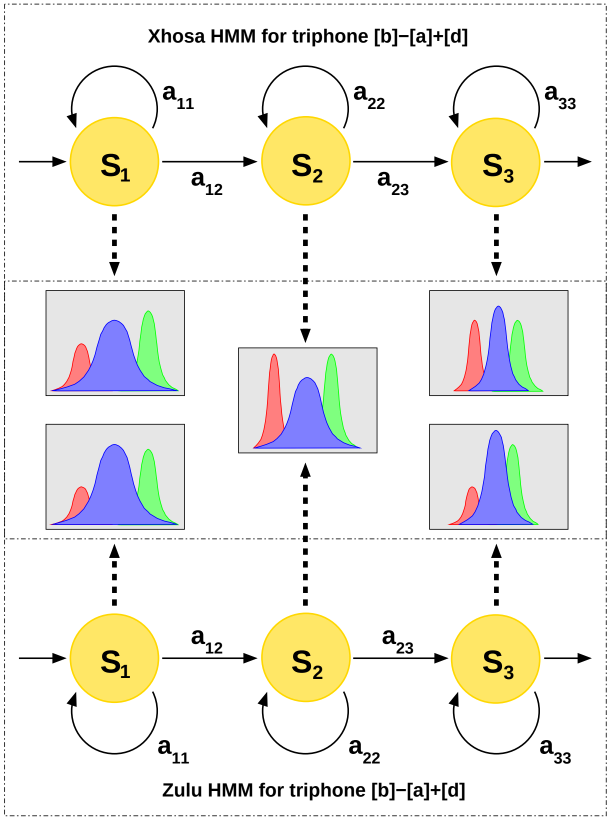
LANGUAGE-SPECIFIC ACOUSTIC MODELS

- Baseline allows no sharing between languages
- Pool triphones with same basephone for each language separately
- Decision-tree clustering questions concern phonetic character only
- Completely separate set if acoustic models for each language



MULTILINGUAL ACOUSTIC MODELS

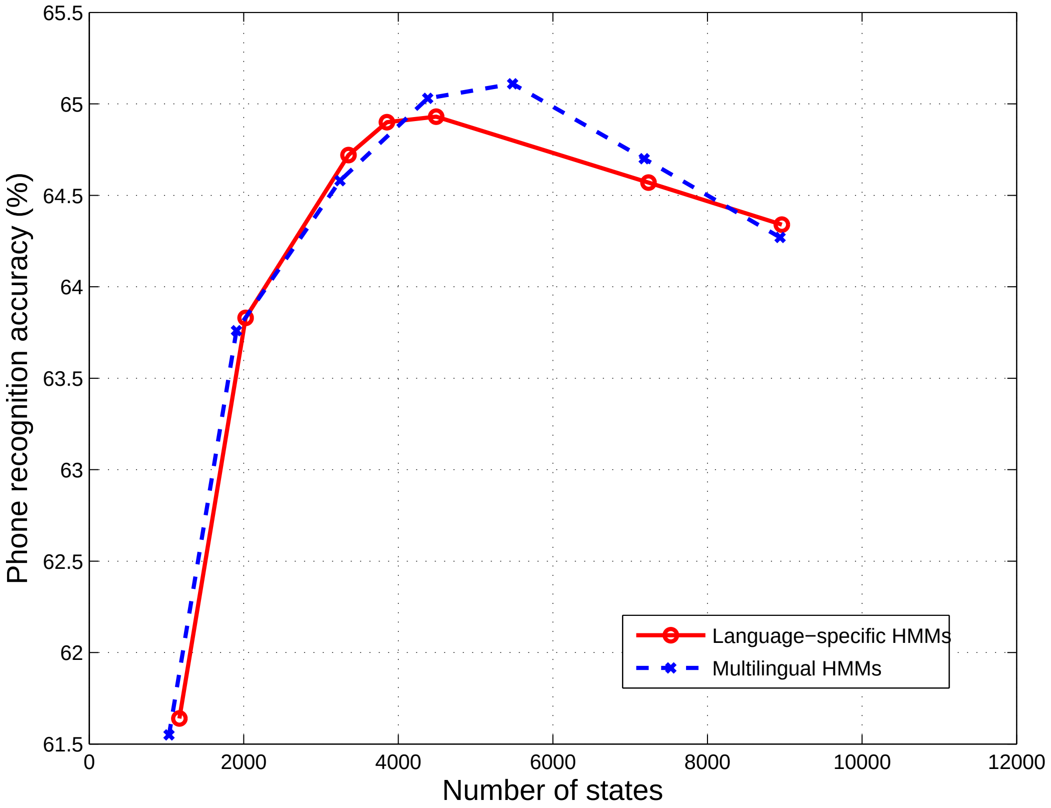
- Allow sharing between languages
- Pool triphones of all languages with same basephone
- Decision-tree clustering questions concern phonetic character of context and language of basephone
- States corresponding to the same basephone but different languages may be shared or kept separate



EXPERIMENTS

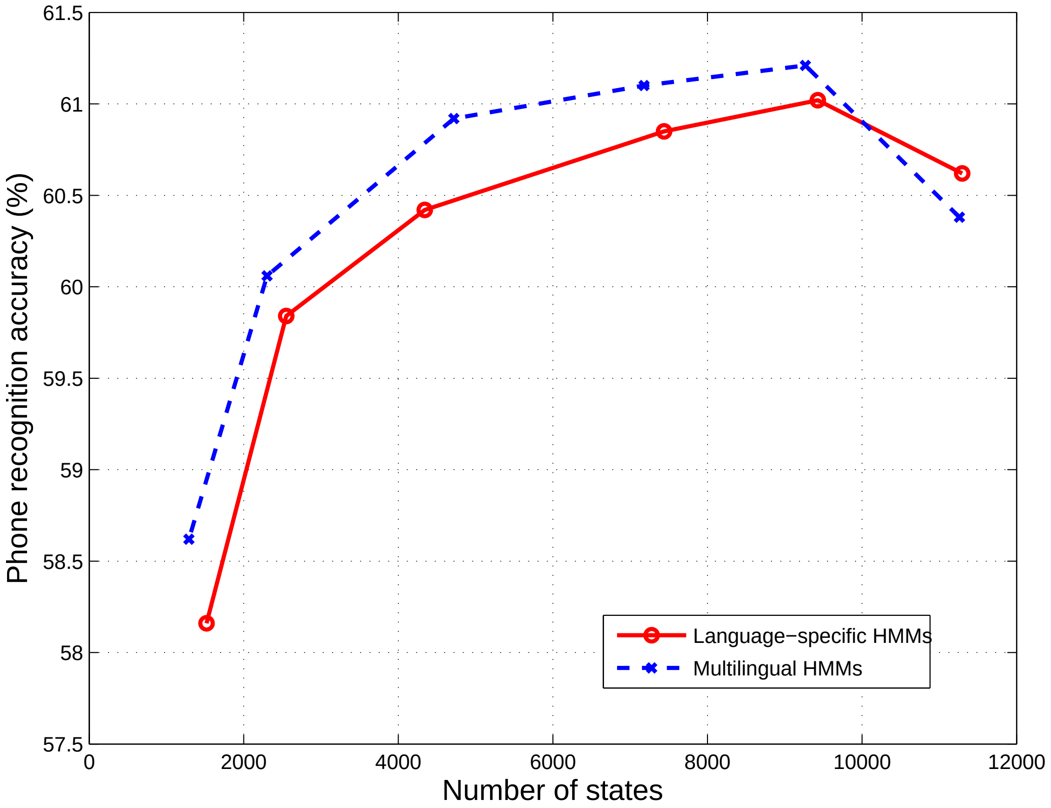
- Combine language pairs:
  - (a) Afrikaans and English
  - (b) Xhosa and Zulu
- Decision-tree likelihood threshold varied to produce models with different numbers of clustered states
- Clustering carried out for single-mixture cross-word triphones
- Number of mixtures increased to 8 after clustering
- Speech parameterisation: MFCCs, 1st & 2nd differentials, per-utterance CMN

RECOGNITION PERFORMANCE: AFRIKAANS+ENGLISH



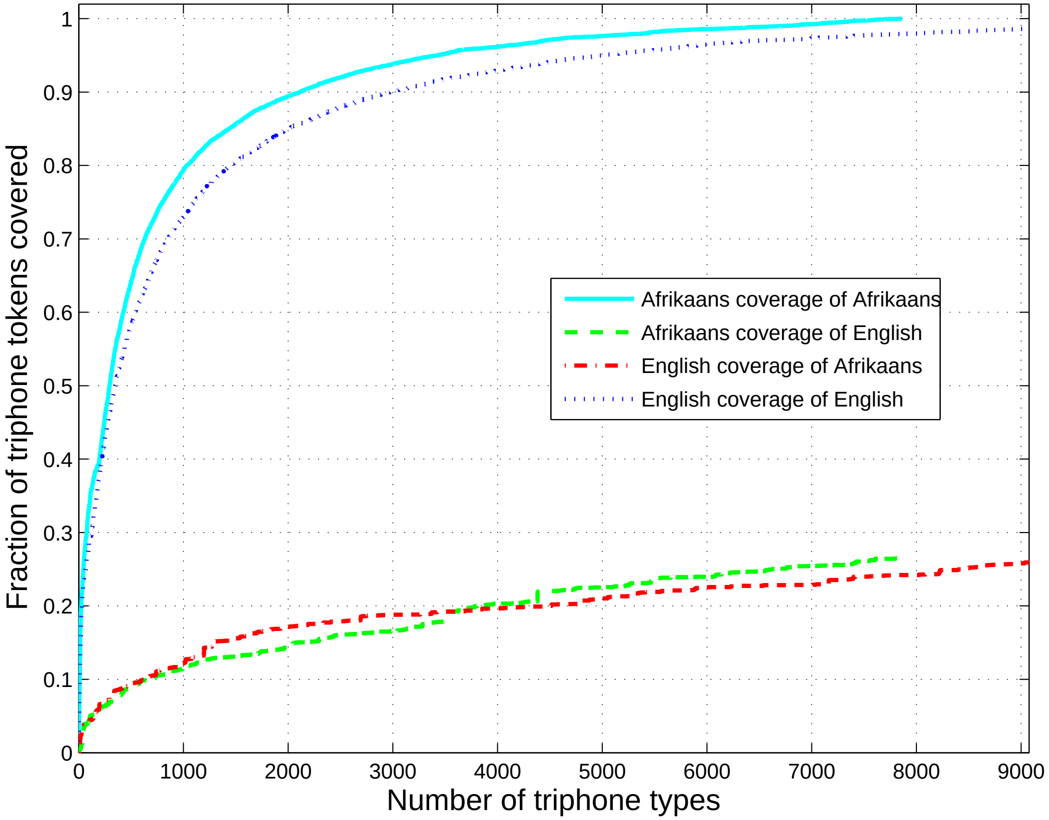
- Small improvement when the number of distinct HMM states is large

RECOGNITION PERFORMANCE: XHOSA+ZULU



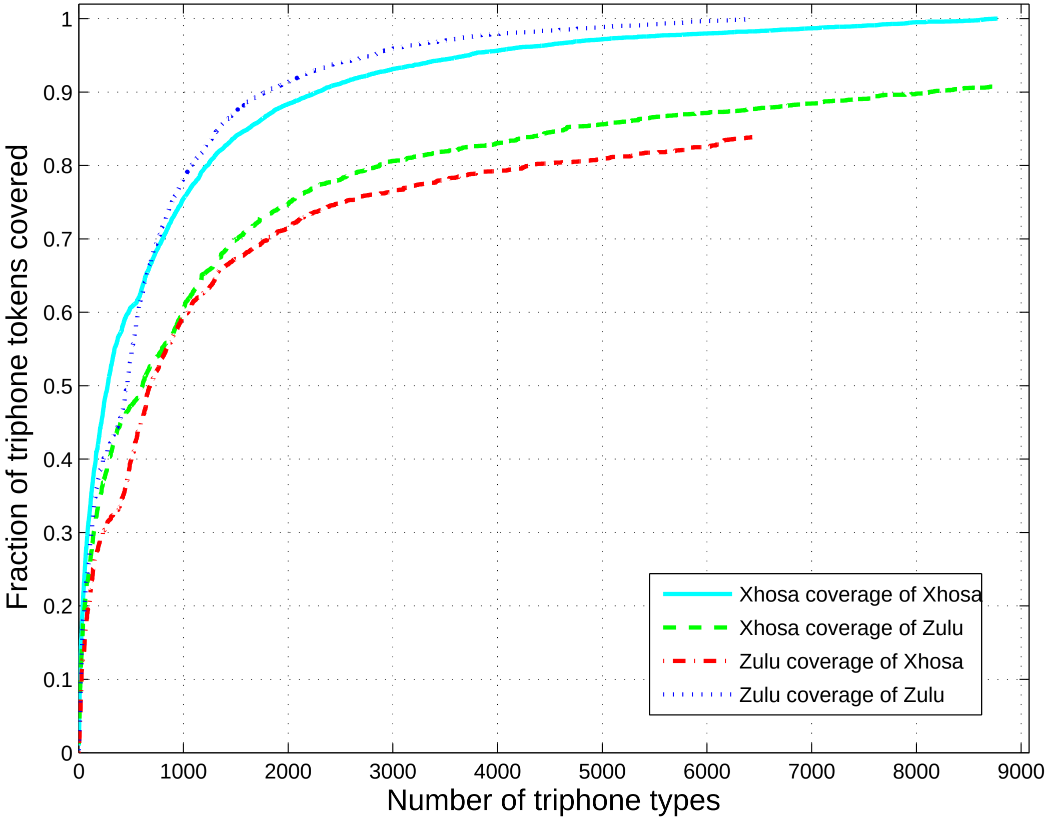
- Improved performance over wider range of HMM complexities

TRIPHONE COVERAGE: AFRIKAANS vs ENGLISH



- Cross-language triphone coverage between Afrikaans and English does not exceed 30%

TRIPHONE COVERAGE: XHOSA vs ZULU



- Cross-language triphone coverage between Xhosa and Zulu exceeds 80%

CONCLUSIONS

- Decision-tree state clustering can be employed to obtain multilingual acoustic models
- Allow sharing between corresponding basephones of different languages
- Small performance gains are seen when combining Afrikaans and English in this way
- Improvements larger for Xhosa and Zulu, which are phonetically more similar
- Future work
  - Apply to more languages
  - Apply to South African English accents