Acoustic modelling of English-accented and Afrikaans-accented South African English

H. Kamper, F.J. Muamba Mukanya and T.R. Niesler

Digital Signal Processing Laboratory
Department of Electrical and Electronic Engineering
Stellenbosch University

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

Introduction

South African accents of English

- English is the language of government, commerce and science
- Only 8.2% of the population use English as first language
- Results in various accents (not regionally bound)
- Multi-accent speech recognition particularly relevant in SA

Introduction

South African accents of English

- English is the language of government, commerce and science
- Only 8.2% of the population use English as first language
- Results in various accents (not regionally bound)
- Multi-accent speech recognition particularly relevant in SA

Aim of research

Determine whether data from different South African English accents can be combined to improve speech recognition performance in any one accent

- Afrikaans-accented English (AE)
- South African English (EE)

Speech Databases

- African Speech Technology (AST) databases
  - Afrikaans English (AE) database
  - South African English (EE) database
- Training set: Approximately 6 hours of speech in both accents
- Test set: Approximately 24 minutes of speech from 20 speakers in each accent
- Development set: Used to optimise recognition parameters

Decision-Tree State Clustering

Acoustic modelling of context-dependent phones

- Acoustic modelling of triphones: [j]−[i]+[k]
- Problems:
  - Not all triphones occur in the training data
  - Not enough data for some triphones which do occur
- Want to determine clusters of similar triphones which can then be used to obtain individual models

Decision-Tree State Clustering

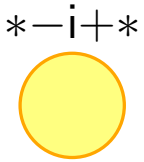
Acoustic modelling of context-dependent phones

- Acoustic modelling of triphones: [j]−[i]+[k]
- Problems:
  - Not all triphones occur in the training data
  - Not enough data for some triphones which do occur
- Want to determine clusters of similar triphones which can then be used to obtain individual models

Solution

Use decision-tree state clustering

Decision-Tree State Clustering

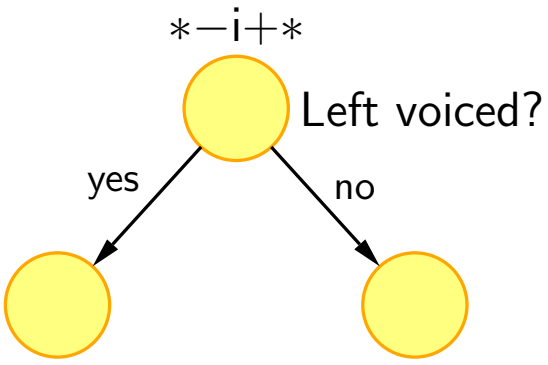


- Begin by pooling all triphones with the same basephone

Decision-Tree State Clustering

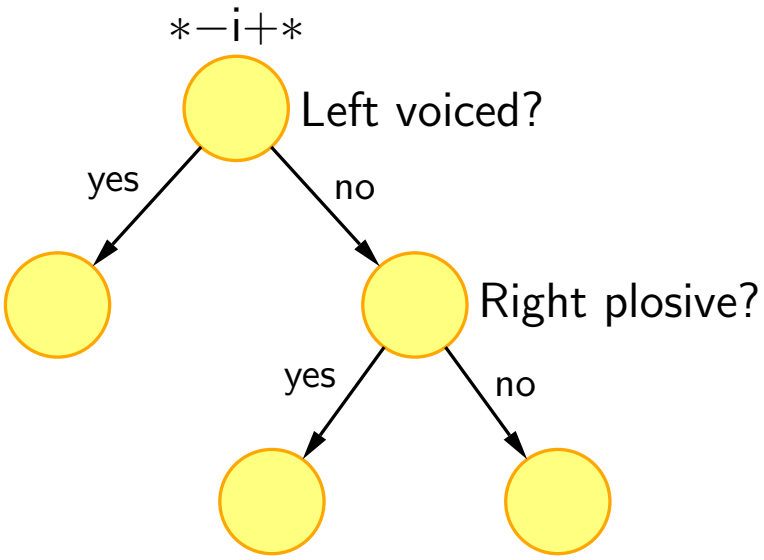
- Begin by pooling all triphones with the same basephone
- Use linguistically-motivated questions to split clusters

Decision-Tree State Clustering



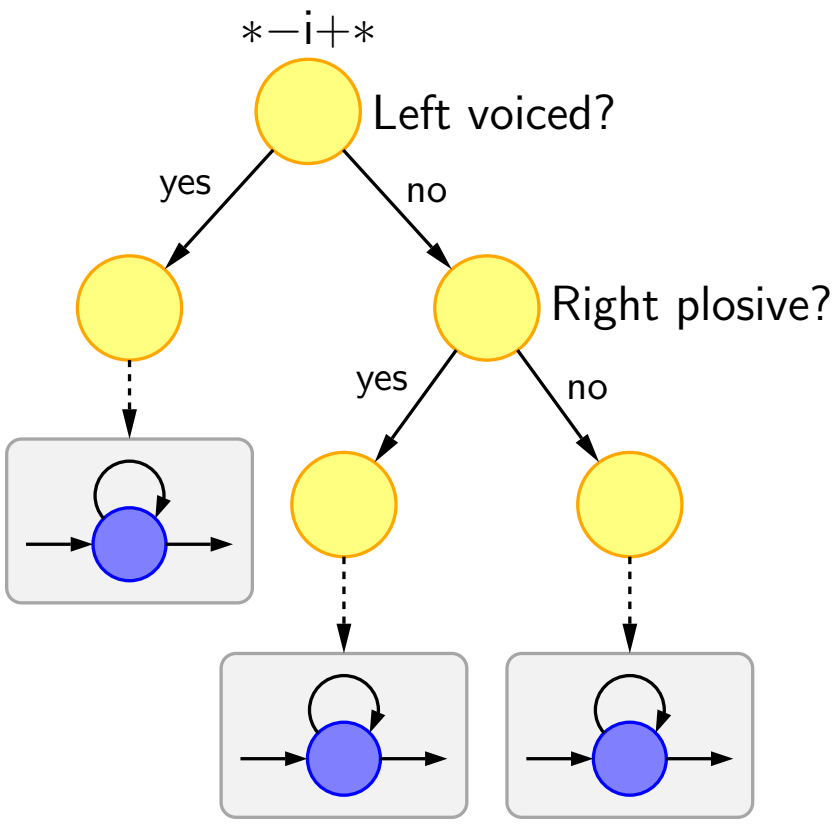
- Begin by pooling all triphones with the same basephone
- Use linguistically-motivated questions to split clusters
- Choose question yielding greatest likelihood improvement and split

Decision-Tree State Clustering



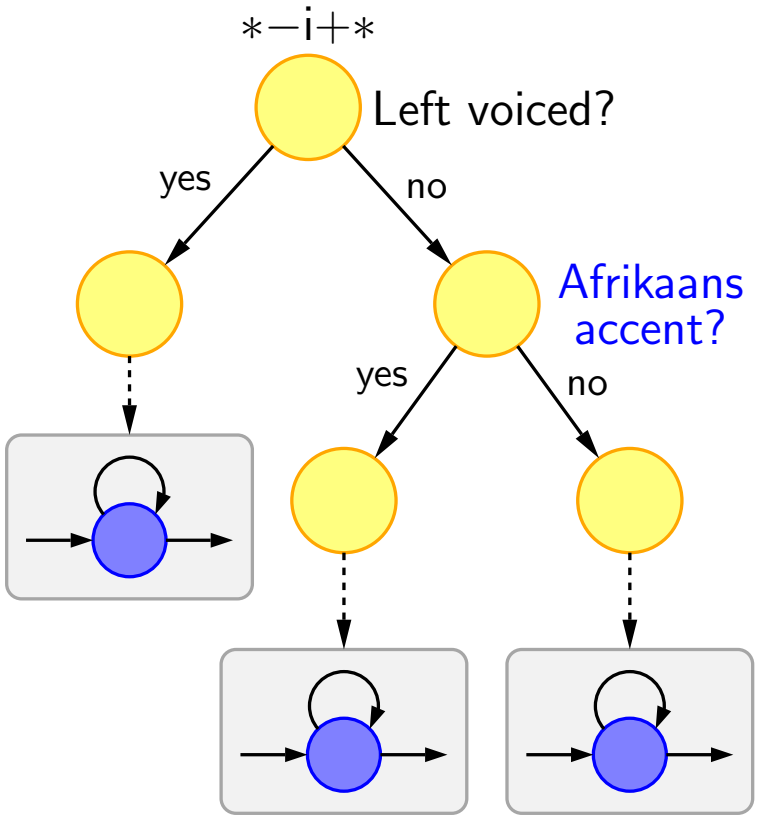
- Begin by pooling all triphones with the same basephone
- Use linguistically-motivated questions to split clusters
- Choose question yielding greatest likelihood improvement and split
- Repeat until likelihood improvement too small

Decision-Tree State Clustering



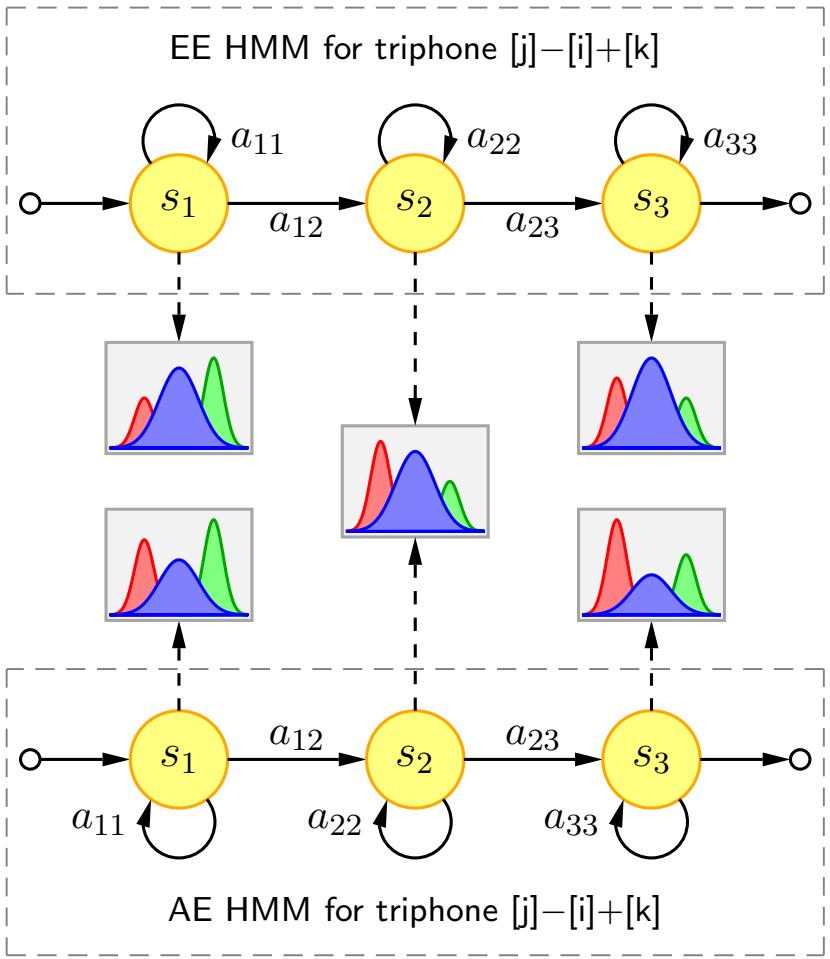
- Begin by pooling all triphones with the same basephone
- Use linguistically-motivated questions to split clusters
- Choose question yielding greatest likelihood improvement and split
- Repeat until likelihood improvement too small
- Each tree leaf corresponds to a cluster of HMM states

Multi-Accent Decision-Tree State Clustering



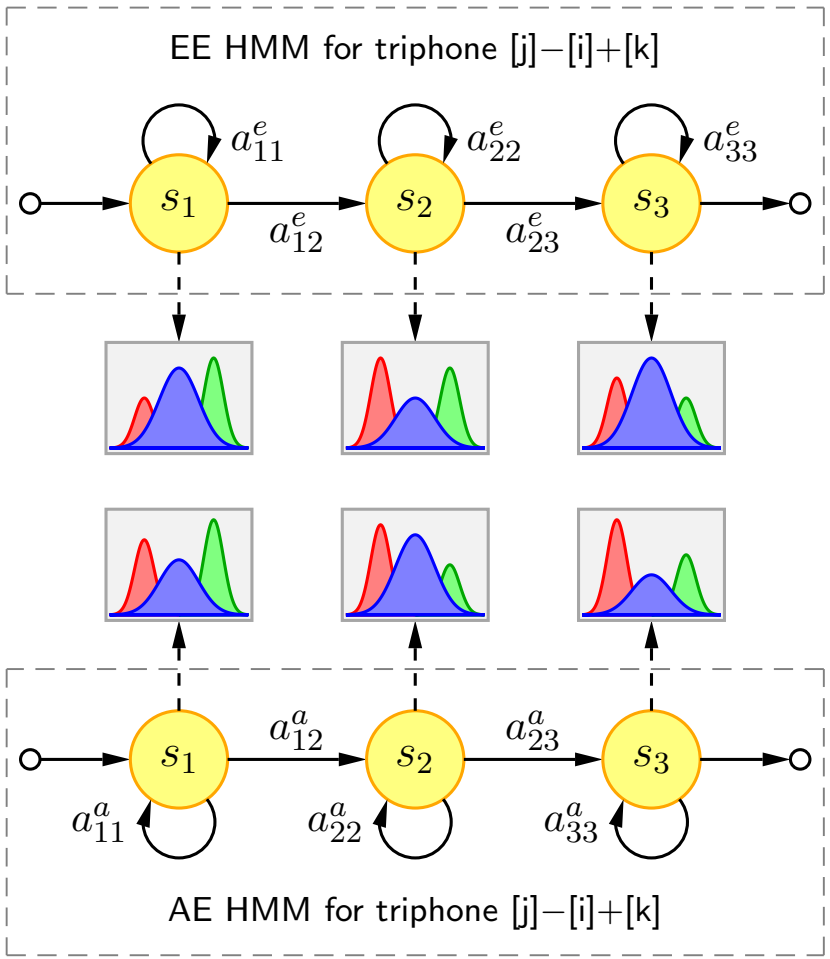
- Tag phones with accent before pooling at root nodes
- Allow decision-tree questions regarding accent as well as phonetic context
- Automatically determine if triphone states from different accents are similar

Multi-Accent Acoustic Models



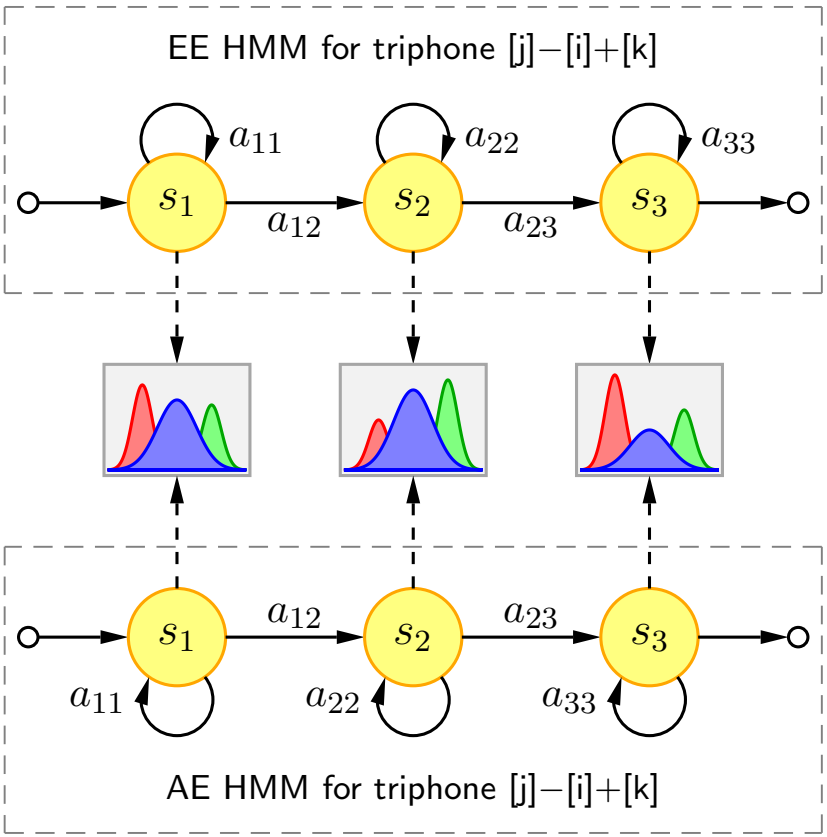
- Allow sharing between accents
- Single set of decision-trees is grown for both accents
- Clustering process employs questions relating to both accent and phonetic context
- States corresponding to the same basephone but different accents may be shared or kept separate

Accent-Specific Acoustic Models



- Allows no sharing between accents
- Separate decision-trees are grown for each accent
- Clustering process employs only questions relating to phonetic context
- Completely separate set of acoustic models for each accent

Accent-Independent Acoustic Models



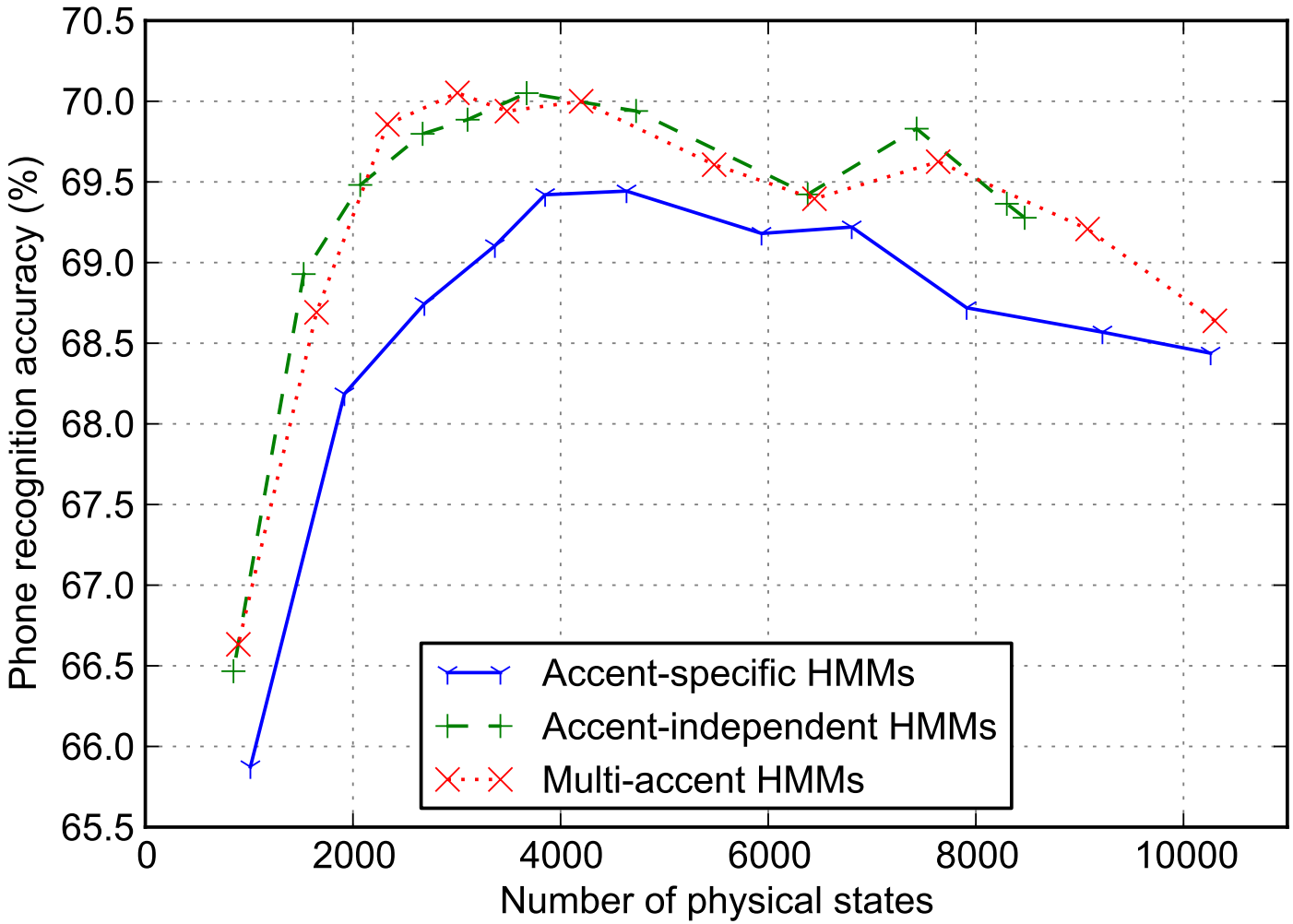
- Data are pooled across both accents
- Single set of decision-trees is grown for both accents
- Clustering process employs only questions relating to phonetic context
- Single set of acoustic models for both accents

Experiments

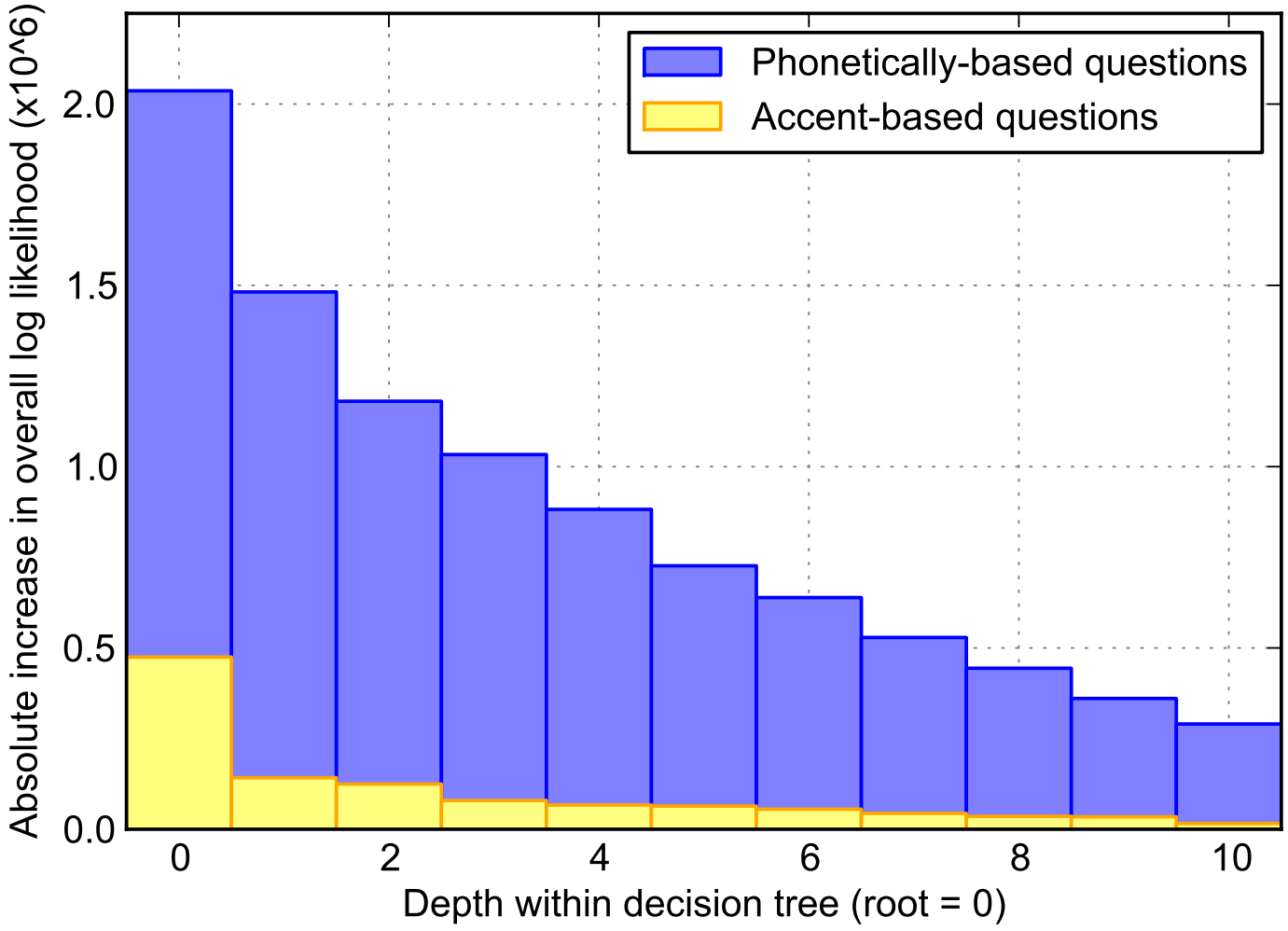
Common setup of systems

- Decision-tree likelihood threshold varied to produce models with different numbers of clustered states
- Used 8-mixture cross-word triphone HMMs
- Speech parameterisation: MFCCs, 1^{st} and 2^{nd} order derivatives, per-utterance CMN

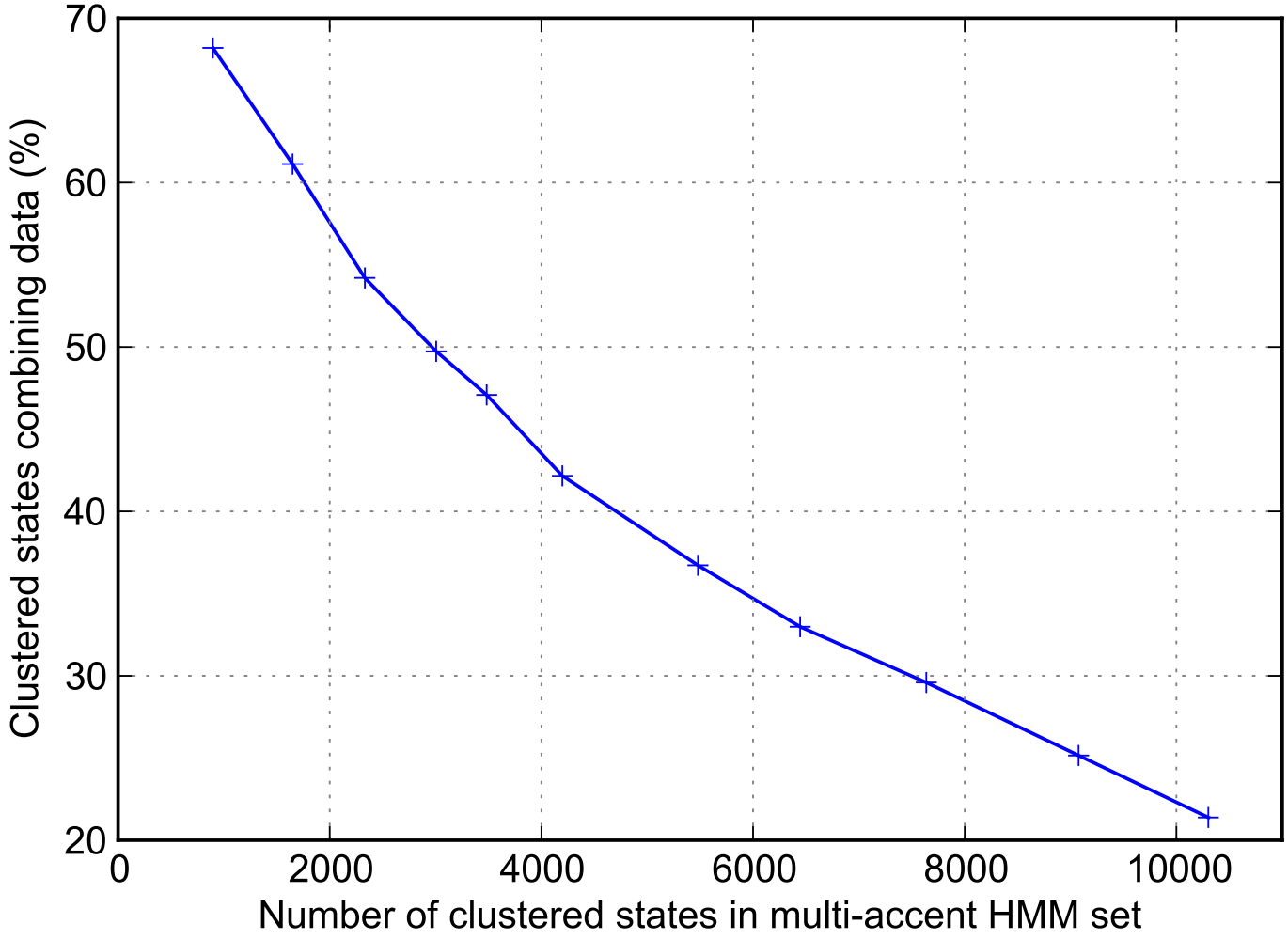
Results: Phone Recognition Performance



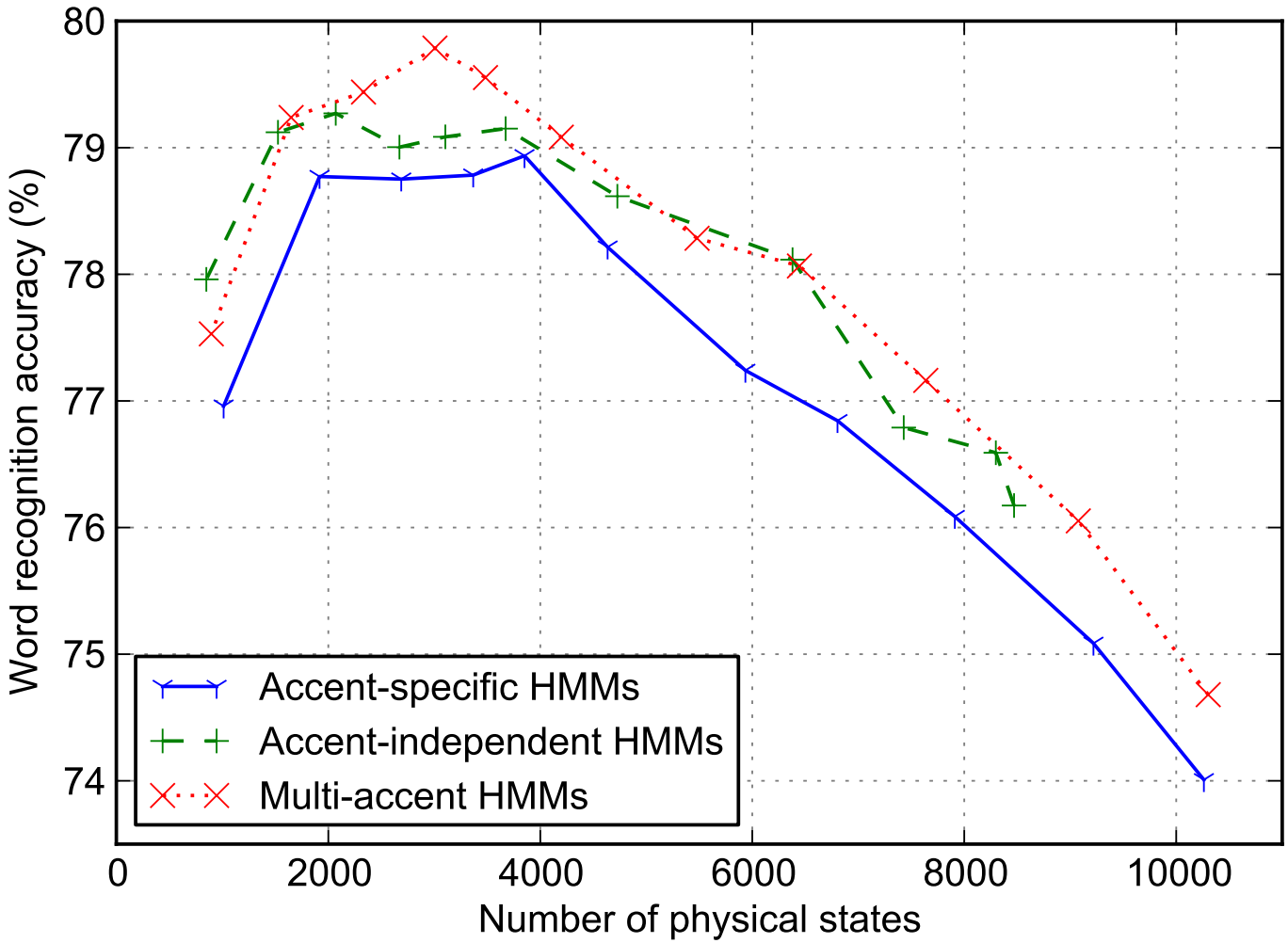
Analysis of Decision Trees



Analysis of Decision Trees



Results: Word Recognition Performance



Conclusions and Future Work

Conclusions

- Accent-specific modelling performs worst
- Accent-independent and multi-accent acoustic modelling yields similar improvements (Afrikaans speaker proficiency)
- Inclusion of accent-based questions (selective sharing) does not impair recognition performance, but does not yield significant gain either
- Supports current practice of simply pooling English accents

Conclusions and Future Work

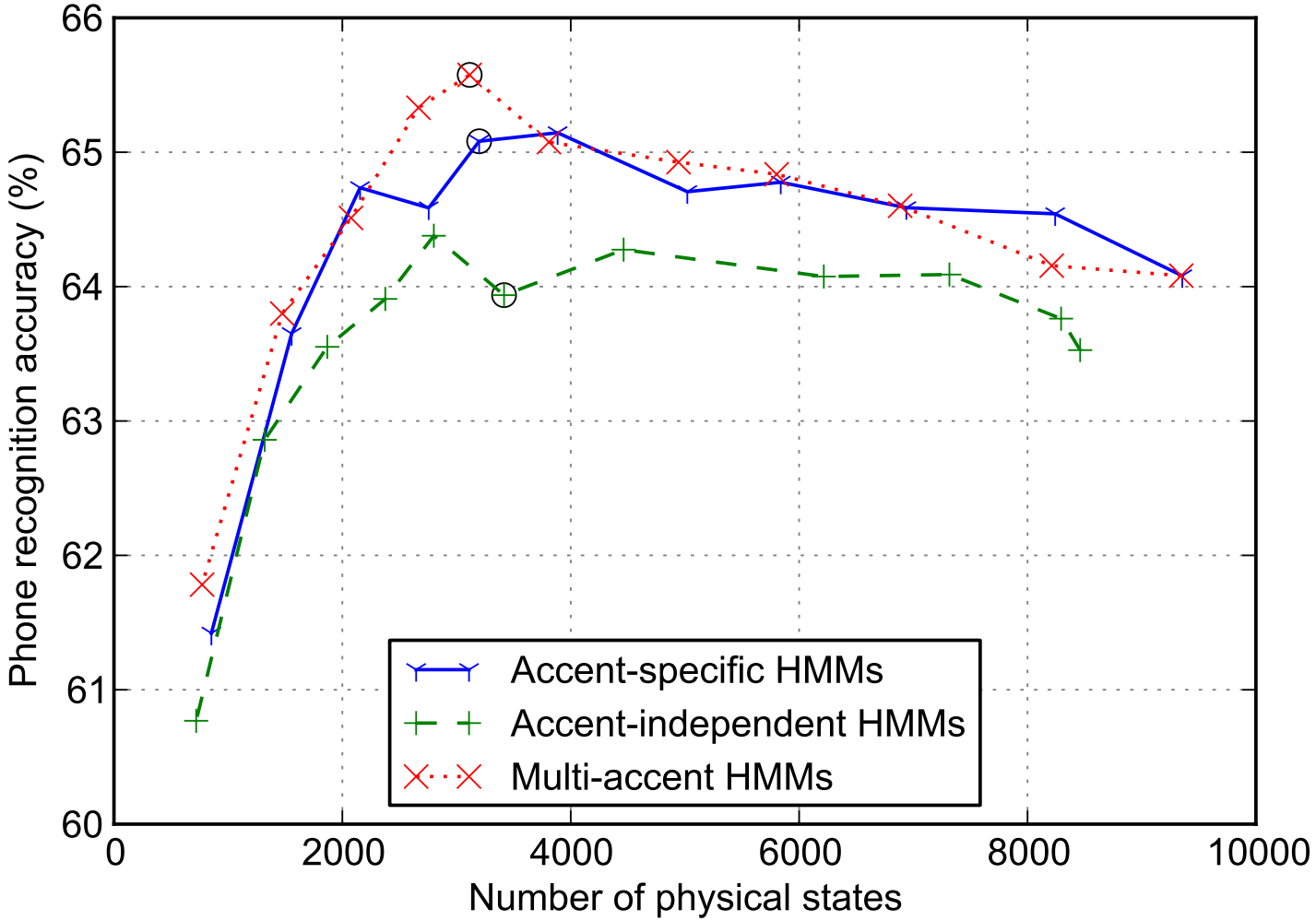
Conclusions

- Accent-specific modelling performs worst
- Accent-independent and multi-accent acoustic modelling yields similar improvements (Afrikaans speaker proficiency)
- Inclusion of accent-based questions (selective sharing) does not impair recognition performance, but does not yield significant gain either
- Supports current practice of simply pooling English accents

Future work

- Less similar accents: Black English and South African English
- Multi-accent acoustic modelling of all five SA English accents

Phone Recognition Performance: BE & EE



Language Modelling: Phone Recognition of AE