Accent reclassification and speech recognition of Afrikaans, Black and White South African English

Herman Kamper and Thomas Niesler

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

UNIVERSITEIT•STELLENBOSCH•UNIVERSITY
jou kennisvennoot • your knowledge partner

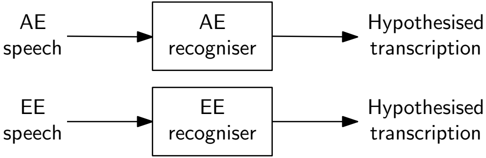
Introduction

- Accented English is highly prevalent in South Africa
- We consider three accents of South African English:
  - Afrikaans English (AE)
  - Black South African English (BE)
  - White South African English (EE)
- For multi-accent speech recognition, **accent labels** must be assigned to training set utterances
- These are assigned by human annotators based on a speaker's mother-tongue or ethnicity and might not necessarily be optimal for modelling purposes
- We consider the unsupervised **reclassification** of training set accent labels

Oracle and parallel recognition of AE and EE

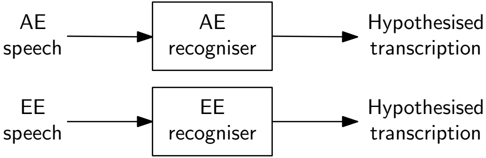
Oracle and parallel recognition of AE and EE

Oracle: Separate accent-specific recognisers for each accent

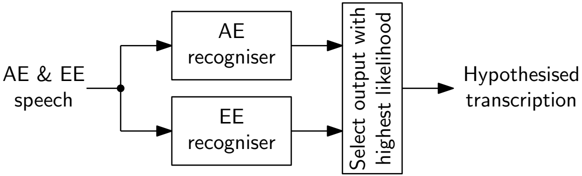


Oracle and parallel recognition of AE and EE

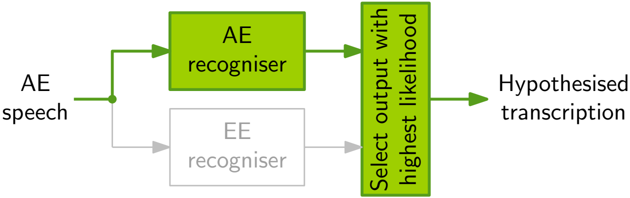
Oracle: Separate accent-specific recognisers for each accent



Parallel: Two accent-specific recognisers operating in parallel

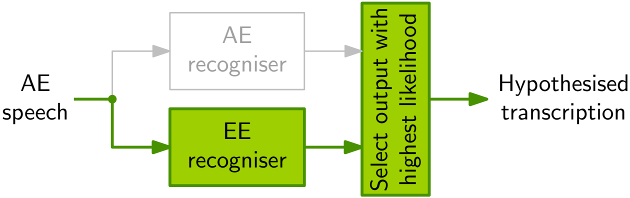


Accent misclassifications



Correctly identified: The matching recogniser is selected

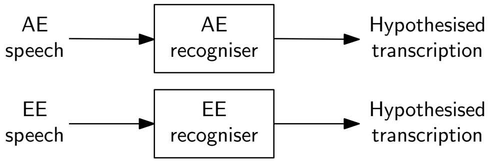
Accent misclassifications



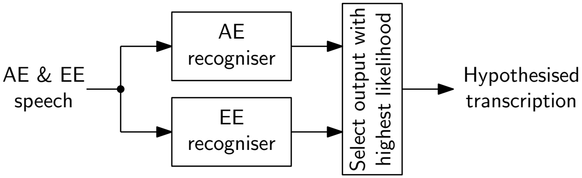
Misclassification: A recogniser from another accent is selected

Oracle and parallel recognition of AE and EE

Oracle: Separate accent-specific recognisers for each accent

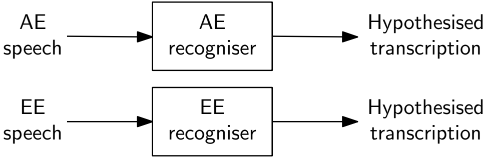


Parallel: Two accent-specific recognisers operating in parallel

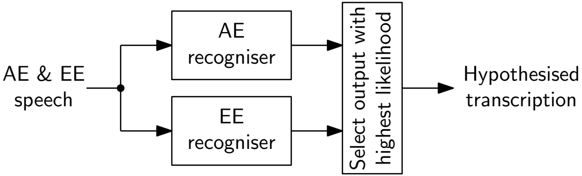


Oracle and parallel recognition of AE and EE

Oracle: Separate accent-specific recognisers for each accent



Parallel: Two accent-specific recognisers operating in parallel



Small improvements of parallel over oracle for AE+EE

Accent reclassification

Conclusions from oracle vs. parallel recognition

- Misclassifications do not always lead to deteriorated accuracies
- The accent labels assigned to training/test utterances might not be the most appropriate

Accent reclassification

Conclusions from oracle vs. parallel recognition

- Misclassifications do not always lead to deteriorated accuracies
- The accent labels assigned to training/test utterances might not be the most appropriate

Propose accent reclassification

Use first-pass acoustic models trained on the originally labelled data to reclassify the accent of training set utterances and then retrain the acoustic models

Accent reclassification

Conclusions from oracle vs. parallel recognition

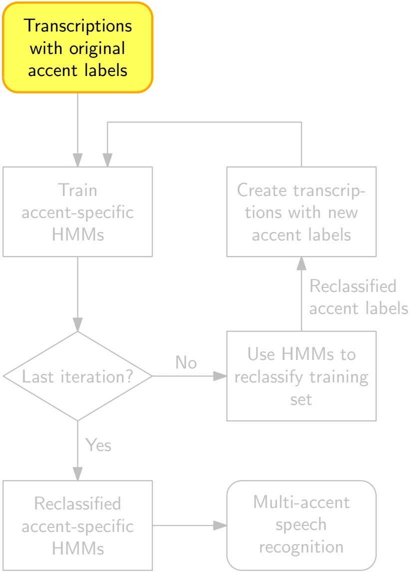
- Misclassifications do not always lead to deteriorated accuracies
- The accent labels assigned to training/test utterances might not be the most appropriate

Propose accent reclassification

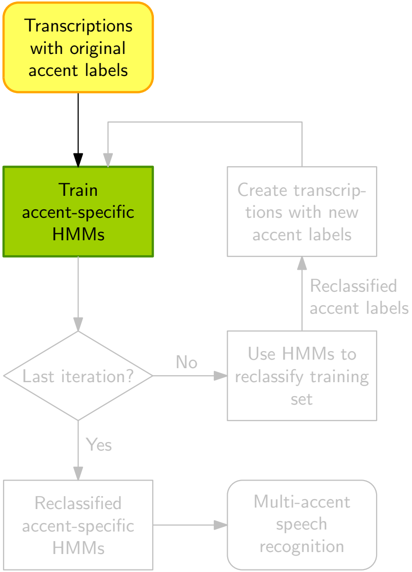
Use first-pass acoustic models trained on the originally labelled data to reclassify the accent of training set utterances and then retrain the acoustic models:

- **AE+EE:** relatively similar accents
- **BE+EE:** relatively dissimilar accents

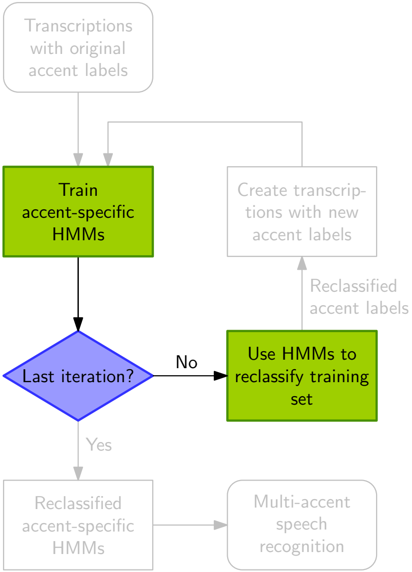
Accent reclassification



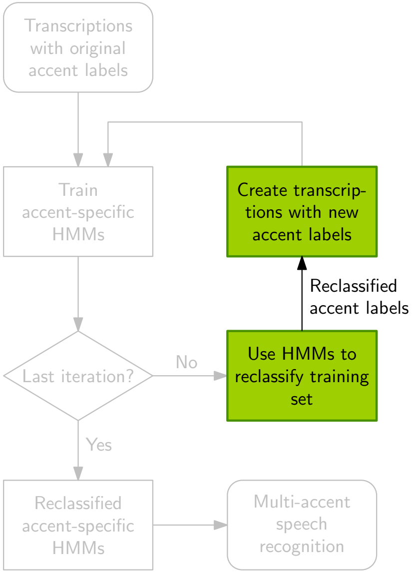
Accent reclassification



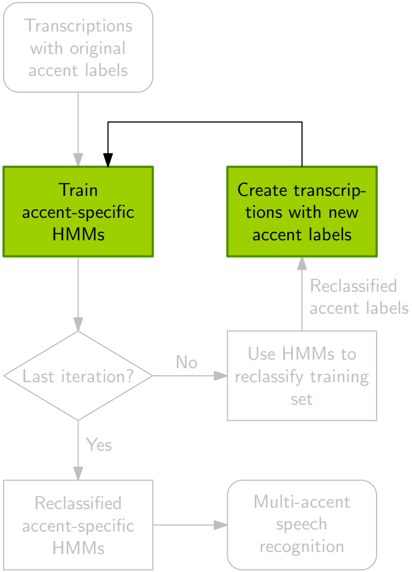
Accent reclassification



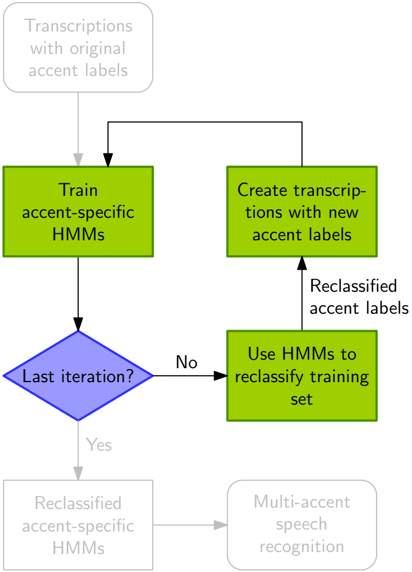
Accent reclassification



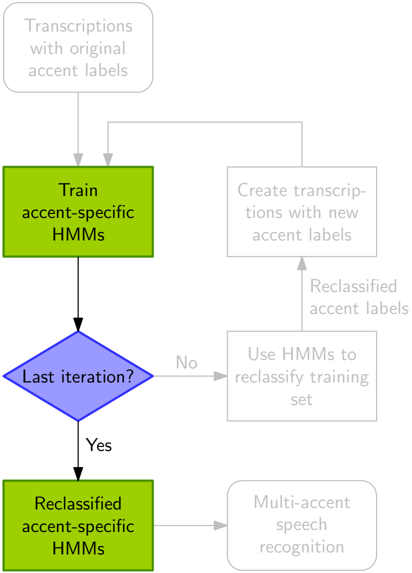
Accent reclassification



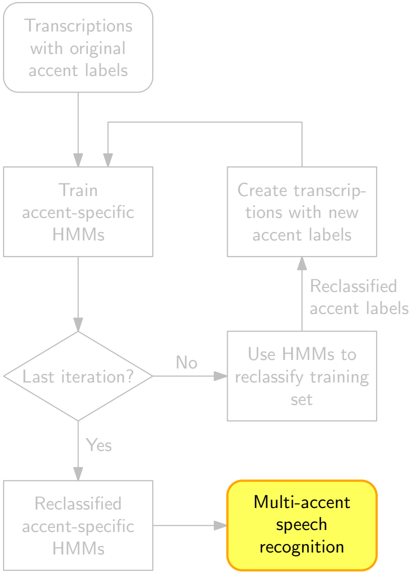
Accent reclassification



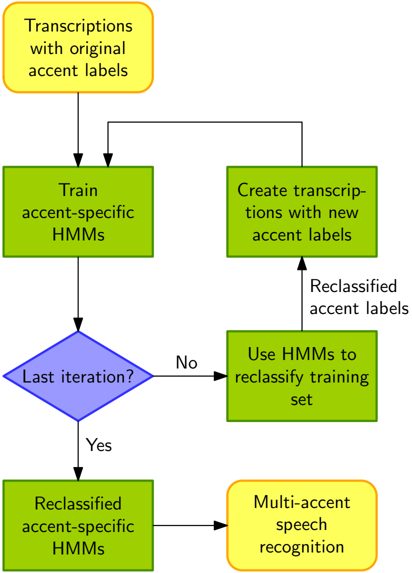
Accent reclassification



Accent reclassification



Accent reclassification



Speech databases

- African Speech Technology (AST) databases:
  - Afrikaans English (AE) database
  - Black South African English (BE) database
  - White South African English (EE) database
- Training set: approximately 6 hours of speech in each accent
- Test set: approximately 24 minutes of speech from 20 speakers in each accent
- Development set: used to optimise recognition parameters

Experimental setup

Setup of systems

- Word recognition of continuous telephone speech
- Trained 8-mixture cross-word triphone HMMs
- Parameterisation: MFCCs, 1^{st} and 2^{nd} order derivatives, per-utterance CMN
- Accent-independent language models and pronunciation dictionaries

Experimental setup

Setup of systems

- Word recognition of continuous telephone speech
- Trained 8-mixture cross-word triphone HMMs
- Parameterisation: MFCCs, 1$^{\text{st}}$ and 2$^{\text{nd}}$ order derivatives, per-utterance CMN
- Accent-independent language models and pronunciation dictionaries

Acoustic modelling approaches

Two acoustic modelling approaches for reclassification:

- **Accent-specific models:** trained separately for each accent
- **Multi-accent models:** allows selective cross-accent data sharing

Experimental setup

Setup of systems

- Word recognition of continuous telephone speech
- Trained 8-mixture cross-word triphone HMMs
- Parameterisation: MFCCs, 1st and 2nd order derivatives, per-utterance CMN
- Accent-independent language models and pronunciation dictionaries

Acoustic modelling approaches

Two acoustic modelling approaches for reclassification:

- **Accent-specific models:** trained separately for each accent
- **Multi-accent models:** allows selective cross-accent data sharing

Further baseline: **accent-independent models** trained on pooled data; accent identification and reclassification not possible with these models

Experimental results for AE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	84.01	84.63	84.58
Accent-independent	84.78	84.78	-
Multi-accent	84.78	84.88	84.61

Experimental results for AE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	84.01	84.63	84.58
Accent-independent	**84.78**	**84.78**	-
Multi-accent	84.78	84.88	84.61

- Accent-independent system only as a baseline (no reclassification)

Experimental results for AE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	**84.01**	**84.63**	84.58
Accent-independent	84.78	84.78	-
Multi-accent	**84.78**	**84.88**	84.61

- Accent-independent system only as a baseline (no reclassification)
- Original systems: parallel systems slightly outperform oracle systems

Experimental results for AE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	84.01	**84.63**	**84.58**
Accent-independent	84.78	84.78	-
Multi-accent	84.78	**84.88**	**84.61**

- Accent-independent system only as a baseline (no reclassification)
- Original systems: parallel systems slightly outperform oracle systems
- Original vs. reclassified parallel systems: **original outperform reclassified**

Experimental results for BE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	76.69	76.07	75.86
Accent-independent	75.38	75.38	-
Multi-accent	77.35	76.75	76.60

Experimental results for BE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	76.69	76.07	75.86
Accent-independent	**75.38**	**75.38**	-
Multi-accent	77.35	76.75	76.60

- Accent-independent system only as a baseline (no reclassification)

Experimental results for BE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	**76.69**	**76.07**	75.86
Accent-independent	75.38	75.38	-
Multi-accent	**77.35**	**76.75**	76.60

- Accent-independent system only as a baseline (no reclassification)
- Original systems: oracle outperform parallel (contrast to AE+EE)

Experimental results for BE+EE

Model set	Original HMMs		Reclassified
	Oracle	Parallel	Parallel
Accent-specific	76.69	**76.07**	**75.86**
Accent-independent	75.38	75.38	-
Multi-accent	77.35	**76.75**	**76.60**

- Accent-independent system only as a baseline (no reclassification)
- Original systems: oracle outperform parallel (contrast to AE+EE)
- Original vs. reclassified parallel systems: **original outperform reclassified**

Analysis of training set utterances for AE+EE

Reclassification effect	No. of utterances	Average length (s)
Labels unchanged	19 775	2.28
Relabelled: AE $\rightarrow$ EE	942	1.11
Relabelled: EE $\rightarrow$ AE	505	1.00
Overall	21 222	2.20

Analysis of training set utterances for AE+EE

Reclassification effect	No. of utterances	Average length (s)
Labels unchanged	19 775	2.28
Relabelled: AE $\rightarrow$ EE	942	**1.11**
Relabelled: EE $\rightarrow$ AE	505	**1.00**
Overall	21 222	2.20

- Relabelled utterances tend to be shorter

Analysis of training set utterances for AE+EE

Reclassification effect	No. of utterances	Average length (s)
Labels unchanged	19 775	2.28
Relabelled: AE $\rightarrow$ EE	**942**	1.11
Relabelled: EE $\rightarrow$ AE	**505**	1.00
Overall	21 222	2.20

- Relabelled utterances tend to be shorter
- The number of AE $\rightarrow$ EE training utterances is almost double the number of EE $\rightarrow$ AE training utterances

Analysis of test set utterances for AE+EE

Recogniser selection	No. of utterances	Average length (s)	Original accuracy	Reclassified accuracy
Selection unchanged	1241	2.14	85.54	85.08
Changed: AE $\rightarrow$ EE	63	1.39	74.21	80.00
Changed: EE $\rightarrow$ AE	87	1.63	79.21	78.50
Overall	1391	2.08	84.88	84.61

Analysis of test set utterances for AE+EE

Recogniser selection	No. of utterances	Average length (s)	Original accuracy	Reclassified accuracy
Selection unchanged	1241	2.14	85.54	85.08
Changed: AE $\rightarrow$ EE	63	**1.39**	74.21	80.00
Changed: EE $\rightarrow$ AE	87	**1.63**	79.21	78.50
Overall	1391	2.08	84.88	84.61

- Test set utterances for which classification has changed generally shorter

Analysis of test set utterances for AE+EE

Recogniser selection	No. of utterances	Average length (s)	Original accuracy	Reclassified accuracy
Selection unchanged	1241	2.14	**85.54**	**85.08**
Changed: AE $\rightarrow$ EE	63	1.39	74.21	80.00
Changed: EE $\rightarrow$ AE	87	1.63	79.21	78.50
Overall	1391	2.08	84.88	84.61

- Test set utterances for which classification has changed generally shorter
- Drop in performance due to utterances for which classification was unchanged

Analysis of test set utterances for AE+EE

Recogniser selection	No. of utterances	Average length (s)	Original accuracy	Reclassified accuracy
Selection unchanged	1241	2.14	85.54	85.08
Changed: AE $\rightarrow$ EE	63	1.39	**74.21**	**80.00**
Changed: EE $\rightarrow$ AE	87	1.63	79.21	78.50
Overall	1391	2.08	84.88	84.61

- Test set utterances for which classification has changed generally shorter
- Drop in performance due to utterances for which classification was unchanged
- Improved recognition accuracy for for AE $\rightarrow$ EE utterances

Analysis of test set utterances for AE+EE

Recogniser selection	No. of utterances	Average length (s)	Original accuracy	Reclassified accuracy
Selection unchanged	1241	2.14	85.54	85.08
Changed: AE $\rightarrow$ EE	63	1.39	74.21	80.00
Changed: EE $\rightarrow$ AE	87	1.63	**79.21**	**78.50**
Overall	1391	2.08	84.88	84.61

- Test set utterances for which classification has changed generally shorter
- Drop in performance due to utterances for which classification was unchanged
- Improved recognition accuracy for for AE $\rightarrow$ EE utterances
- Slightly deteriorated recognition accuracy for EE $\rightarrow$ AE utterances

Analysis of test set utterances for AE+EE

Recogniser selection	No. of utterances	Average length (s)	Original accuracy	Reclassified accuracy
Selection unchanged	1241	2.14	85.54	85.08
Changed: AE $\rightarrow$ EE	63	1.39	74.21	80.00
Changed: EE $\rightarrow$ AE	87	1.63	79.21	78.50
Overall	1391	2.08	**84.88**	**84.61**

- Test set utterances for which classification has changed generally shorter
- Drop in performance due to utterances for which classification was unchanged
- Improved recognition accuracy for for AE $\rightarrow$ EE utterances
- Slightly deteriorated recognition accuracy for EE $\rightarrow$ AE utterances

Conclusions

- A single iteration of **reclassification** leads to deteriorated performance
- This deterioration is consistent for:
  - Both accent pairs: AE+EE and BE+EE
  - All acoustic modelling approaches considered
- Analysis indicates:
  - Accent label changes from AE to EE occur more often than vice versa
  - Accent label changes from BE to EE and vice versa more consistent
  - Relabelled and reclassified training and test utterances tend to be shorter
- **Final conclusion:** Best to use the originally labelled data