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
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:
Separate accent-specific recognisers for each accent

AE recogniser
EE recogniser
Hypothesised transcription
AE speech
EE speech
Hypothesised transcription

Parallel:
Two accent-specific recognisers operating in parallel

Select output with highest likelihood

AE recogniser
EE recogniser
AE & EE speech
Hypothesised transcription

H. Kamper (Stellenbosch University)
Reclassification of SAE accents
PRASA 2011
Oracle and parallel recognition of AE and EE

**Oracle:** Separate accent-specific recognisers for each accent

- **AE** recogniser
  - AE speech
  - Hypothesised transcription

- **EE** recogniser
  - EE speech
  - Hypothesised transcription

Parallel:

- Two accent-specific recognisers operating in parallel
- Select output with highest likelihood
Oracle and parallel recognition of AE and EE

**Oracle:** Separate accent-specific recognisers for each accent

- AE recogniser
- EE recogniser
- Hypothesised transcription

**Parallel:** Two accent-specific recognisers operating in parallel

- AE & EE recogniser
- Select output with highest likelihood
- Hypothesised transcription
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

- AE speech
- EE speech

**Parallel:** Two accent-specific recognisers operating in parallel

- AE & EE speech
- Select output with highest likelihood
- Hypothesised transcription
Oracle and parallel recognition of AE and EE

**Oracle:** Separate accent-specific recognisers for each accent

Oracle:

- AE recogniser
- EE recogniser
- Hypothesised transcription

**Parallel:** Two accent-specific recognisers operating in parallel

Parallel:

- AE recogniser
- EE recogniser
- AE & EE speech
- Hypothesised transcription

Select output with highest likelihood

Small improvements of parallel over oracle for AE+EE

H. Kamper (Stellenbosch University)
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

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

Transcriptions with original accent labels

Train accent-specific HMMs

Last iteration?

Yes

Reclassified accent-specific HMMs

No

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition

Transcriptions with original accent labels

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Last iteration?

Yes

Reclassified accent-specific HMMs

No

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition

Transcriptions with original accent labels

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Last iteration?

No

Create transcriptions with new accent labels

Use HMMs to reclassify training set

Reclassified accent-specific HMMs

Yes

Reclassified accent labels

Multi-accent speech recognition

Reclassified accent labels

Transcriptions with original accent labels

Create transcriptions with new accent labels

Use HMMs to reclassify training set

Last iteration?

Train accent-specific HMMs

Yes
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Last iteration?

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition

Reclassified accent-specific HMMs

Yes

No

H. Kamper (Stellenbosch University)
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Create transcriptions with new accent labels

Reclassified accent-specific HMMs

Multi-accent speech recognition

Use HMMs to reclassify training set

Reclassified accent labels

Last iteration?

Yes

No

H. Kamper (Stellenbosch University)

Reclassification of SAE accents

PRASA 2011
Accent reclassification

- Transcriptions with original accent labels
  - Last iteration?
    - Yes
      - Train accent-specific HMMs
    - No
      - Reclassified accent labels
        - Use HMMs to reclassify training set
          - Multi-accent speech recognition

- Create transcriptions with new accent labels
  - Reclassified accent labels
    - Use HMMs to reclassify training set
      - Last iteration?
        - Yes
          - Reclassified accent-specific HMMs
        - No
          - Transcriptions with original accent labels
            - Create transcriptions with new accent labels
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Last iteration?

Yes

Reclassified accent-specific HMMs

No

Create transcriptions with new accent labels

Use HMMs to reclassify training set

Reclassified accent-specific HMMs

Multi-accent speech recognition
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Last iteration?

Yes

Reclassified accent-specific HMMs

No

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition

Transcriptions with original accent labels

Create transcriptions with new accent labels

Recategorized accent labels

Use HMMs to reclassify training set

Last iteration?

Yes

Reclassified accent-specific HMMs

No

H. Kamper (Stellenbosch University)
Accent reclassification

Transcriptions with original accent labels

Train accent-specific HMMs

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Reclassified accent-specific HMMs

Multi-accent speech recognition

Last iteration?

No

Yes

Reclassified accent-specific HMMs

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition

Last iteration?

Yes

No

Reclassified accent-specific HMMs

Create transcriptions with new accent labels

Reclassified accent labels

Use HMMs to reclassify training set

Multi-accent speech recognition

Last iteration?

Yes

No
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\textsuperscript{st} and 2\textsuperscript{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\textsuperscript{st} and 2\textsuperscript{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, \(1^{\text{st}}\) 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

Further baseline: **accent-independent models** trained on pooled data; accent identification and reclassification not possible with these models
Experimental results for AE+EE

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>84.01</td>
<td>84.63</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>84.78</td>
<td>84.78</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>84.78</td>
<td>84.88</td>
</tr>
</tbody>
</table>

Original systems: parallel systems slightly outperform oracle systems

Original vs. reclassified parallel systems: original outperform reclassified
Experimental results for AE+EE

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>84.01</td>
<td>84.63</td>
</tr>
<tr>
<td>Accent-independent</td>
<td><strong>84.78</strong></td>
<td><strong>84.78</strong></td>
</tr>
<tr>
<td>Multi-accent</td>
<td>84.78</td>
<td>84.88</td>
</tr>
</tbody>
</table>

- Accent-independent system only as a baseline (no reclassification)
Experimental results for AE+EE

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>84.01</td>
<td>84.63</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>84.78</td>
<td>84.78</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>84.78</td>
<td>84.88</td>
</tr>
</tbody>
</table>

- Accent-independent system only as a baseline (no reclassification)
- Original systems: parallel systems slightly outperform oracle systems
## Experimental results for AE+EE

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>84.01</td>
<td>84.63</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>84.78</td>
<td>84.78</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>84.78</td>
<td>84.88</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th></th>
<th>Reclassified</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
<td>Parallel</td>
<td></td>
</tr>
<tr>
<td>Accent-specific</td>
<td>76.69</td>
<td>76.07</td>
<td>75.86</td>
<td></td>
</tr>
<tr>
<td>Accent-independent</td>
<td>75.38</td>
<td>75.38</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Multi-accent</td>
<td>77.35</td>
<td>76.75</td>
<td>76.60</td>
<td></td>
</tr>
</tbody>
</table>

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
## Experimental results for BE+EE

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>76.69</td>
<td>76.07</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>75.38</td>
<td>75.38</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>77.35</td>
<td>76.75</td>
</tr>
</tbody>
</table>

- Accent-independent system only as a baseline (no reclassification)
### Experimental results for BE+EE

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
<th>Reclassified</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
<td>Parallel</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>76.69</td>
<td>76.07</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>75.38</td>
<td>75.38</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>77.35</td>
<td>76.75</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Model set</th>
<th>Original HMMs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Oracle</td>
</tr>
<tr>
<td>Accent-specific</td>
<td>76.69</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>75.38</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>77.35</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Reclassification effect</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels unchanged</td>
<td>19,775</td>
<td>2.28</td>
</tr>
<tr>
<td>Relabelled: AE → EE</td>
<td>942</td>
<td>1.11</td>
</tr>
<tr>
<td>Relabelled: EE → AE</td>
<td>505</td>
<td>1.00</td>
</tr>
<tr>
<td>Overall</td>
<td>21,222</td>
<td>2.20</td>
</tr>
</tbody>
</table>

Relabelled utterances tend to be shorter.
The number of AE → EE training utterances is almost double the number of EE → AE training utterances.
### Analysis of training set utterances for AE+EE

<table>
<thead>
<tr>
<th>Reclassification effect</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labels unchanged</td>
<td>19,775</td>
<td>2.28</td>
</tr>
<tr>
<td>Relabelled: AE → EE</td>
<td>942</td>
<td>1.11</td>
</tr>
<tr>
<td>Relabelled: EE → AE</td>
<td>505</td>
<td>1.00</td>
</tr>
<tr>
<td>Overall</td>
<td>21,222</td>
<td>2.20</td>
</tr>
</tbody>
</table>

- Relabelled utterances tend to be shorter
Reclassification effect | No. of utterances | Average length (s)
--- | --- | ---
Labels unchanged | 19,775 | 2.28
Relabelled: AE → EE | 942 | 1.11
Relabelled: EE → AE | 505 | 1.00
Overall | 21,222 | 2.20

- Relabelled utterances tend to be shorter
- The number of AE → EE training utterances is almost double the number of EE → AE training utterances
## Analysis of test set utterances for AE+EE

<table>
<thead>
<tr>
<th>Recogniser selection</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
<th>Original accuracy</th>
<th>Reclassified accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection unchanged</td>
<td>1241</td>
<td>2.14</td>
<td>85.54</td>
<td>85.08</td>
</tr>
<tr>
<td>Changed: AE → EE</td>
<td>63</td>
<td>1.39</td>
<td>74.21</td>
<td>80.00</td>
</tr>
<tr>
<td>Changed: EE → AE</td>
<td>87</td>
<td>1.63</td>
<td>79.21</td>
<td>78.50</td>
</tr>
<tr>
<td>Overall</td>
<td>1391</td>
<td>2.08</td>
<td>84.88</td>
<td>84.61</td>
</tr>
</tbody>
</table>
Analysis of test set utterances for AE+EE

<table>
<thead>
<tr>
<th>Recogniser selection</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
<th>Original accuracy</th>
<th>Reclassified accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection unchanged</td>
<td>1241</td>
<td>2.14</td>
<td>85.54</td>
<td>85.08</td>
</tr>
<tr>
<td>Changed: AE → EE</td>
<td>63</td>
<td>1.39</td>
<td>74.21</td>
<td>80.00</td>
</tr>
<tr>
<td>Changed: EE → AE</td>
<td>87</td>
<td>1.63</td>
<td>79.21</td>
<td>78.50</td>
</tr>
<tr>
<td>Overall</td>
<td>1391</td>
<td>2.08</td>
<td>84.88</td>
<td>84.61</td>
</tr>
</tbody>
</table>

- Test set utterances for which classification has changed generally shorter
## Analysis of test set utterances for AE+EE

<table>
<thead>
<tr>
<th>Recogniser selection</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
<th>Original accuracy</th>
<th>Reclassified accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection unchanged</td>
<td>1241</td>
<td>2.14</td>
<td>85.54</td>
<td>85.08</td>
</tr>
<tr>
<td>Changed: AE → EE</td>
<td>63</td>
<td>1.39</td>
<td>74.21</td>
<td>80.00</td>
</tr>
<tr>
<td>Changed: EE → AE</td>
<td>87</td>
<td>1.63</td>
<td>79.21</td>
<td>78.50</td>
</tr>
<tr>
<td>Overall</td>
<td>1391</td>
<td>2.08</td>
<td>84.88</td>
<td>84.61</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Recogniser selection</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
<th>Original accuracy</th>
<th>Reclassified accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection unchanged</td>
<td>1241</td>
<td>2.14</td>
<td>85.54</td>
<td>85.08</td>
</tr>
<tr>
<td>Changed: AE → EE</td>
<td>63</td>
<td>1.39</td>
<td>74.21</td>
<td>80.00</td>
</tr>
<tr>
<td>Changed: EE → AE</td>
<td>87</td>
<td>1.63</td>
<td>79.21</td>
<td>78.50</td>
</tr>
<tr>
<td>Overall</td>
<td>1391</td>
<td>2.08</td>
<td>84.88</td>
<td>84.61</td>
</tr>
</tbody>
</table>

- 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 AE → EE utterances
Analysis of test set utterances for AE+EE

<table>
<thead>
<tr>
<th>Recogniser selection</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
<th>Original accuracy</th>
<th>Reclassified accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection unchanged</td>
<td>1241</td>
<td>2.14</td>
<td>85.54</td>
<td>85.08</td>
</tr>
<tr>
<td>Changed: AE → EE</td>
<td>63</td>
<td>1.39</td>
<td>74.21</td>
<td>80.00</td>
</tr>
<tr>
<td>Changed: EE → AE</td>
<td>87</td>
<td>1.63</td>
<td>79.21</td>
<td>78.50</td>
</tr>
<tr>
<td>Overall</td>
<td>1391</td>
<td>2.08</td>
<td>84.88</td>
<td>84.61</td>
</tr>
</tbody>
</table>

- 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 AE → EE utterances
- Slightly deteriorated recognition accuracy for EE → AE utterances
## Analysis of test set utterances for AE+EE

<table>
<thead>
<tr>
<th>Recogniser selection</th>
<th>No. of utterances</th>
<th>Average length (s)</th>
<th>Original accuracy</th>
<th>Reclassified accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection unchanged</td>
<td>1241</td>
<td>2.14</td>
<td>85.54</td>
<td>85.08</td>
</tr>
<tr>
<td>Changed: AE → EE</td>
<td>63</td>
<td>1.39</td>
<td>74.21</td>
<td>80.00</td>
</tr>
<tr>
<td>Changed: EE → AE</td>
<td>87</td>
<td>1.63</td>
<td>79.21</td>
<td>78.50</td>
</tr>
<tr>
<td>Overall</td>
<td>1391</td>
<td>2.08</td>
<td>84.88</td>
<td>84.61</td>
</tr>
</tbody>
</table>

- 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 → EE utterances
- Slightly deteriorated recognition accuracy for EE → 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