Optimisation of acoustic models for a target accent using decision-tree state clustering

PRASA 2012

Herman Kamper and Thomas Niesler

Digital Signal Processing Group
Department of Electrical and Electronic Engineering
Stellenbosch University
Introduction

Five major accents of South African English:

- Afrikaans English (AE) 77.8%
- Cape Flats English (CE) 5.7%
- White South African English (EE) 8.8%
- Indian South African English (IE) 3.8%
- Other 2.3%
- Black South African English (BE) 1.6%

H. Kamper & T.R. Niesler (Stellenbosch University)
Modelling accents

- How can we **model** the different accents for speech recognition?

- **AST databases**: approximately 6 hours of speech in each accent

- **Multi-accent acoustic modelling** allows selective sharing across accents

- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements

- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?
Modelling accents

- How can we **model** the different accents for speech recognition?

- **AST databases**: approximately 6 hours of speech in each accent

- **Multi-accent acoustic modelling** allows selective sharing across accents

- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements

- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?
Modelling accents

- How can we **model** the different accents for speech recognition?

- **AST databases**: approximately 6 hours of speech in each accent

- **Multi-accent acoustic modelling** allows selective sharing across accents

- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements

- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?
Modelling accents

- How can we **model** the different accents for speech recognition?

- **AST databases**: approximately 6 hours of speech in each accent

- **Multi-accent acoustic modelling** allows selective sharing across accents

- This approach guarantees **overall** likelihood improvement over all accents, but not **per-accent** improvements

- How do we obtain best acoustic model set for **particular accent**, but still incorporate useful data from other accents?
Modelling accents

- How can we model the different accents for speech recognition?

- **AST databases**: approximately 6 hours of speech in each accent

- **Multi-accent acoustic modelling** allows selective sharing across accents

- This approach guarantees overall likelihood improvement over all accents, but not per-accent improvements

- How do we obtain best acoustic model set for particular accent, but still incorporate useful data from other accents?
Acoustic modelling of context-dependent phones

- Use hidden Markov models (HMMs)
- Acoustic modelling of triphones: [t]−[iy]+[n]
- 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
Acoustic modelling

Acoustic modelling of context-dependent phones

- Use hidden Markov models (HMMs)
- Acoustic modelling of **triphones**: [t]−[iy]+[n]
- 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

Solution

Use **decision-tree state clustering**
Decision-tree state clustering

\[ * - [iy] + * \]

(state i)
Decision-tree state clustering

\[ \text{Left voiced?} \]

\[ \text{Left vowel?} \]

\[ \text{Afrikaans English?} \]

\[ \text{Left silence?} \]

\[ \text{Right vowel?} \]

\[ \text{Left [k]?} \]
Decision-tree state clustering

*—[iy]—*

(state \(i\))

Left voiced?

yes

no

1. Left voiced?
2. Right plosive?
3. Left vowel?

yes

no

1. No
2. Yes

1. Yes
2. No

1. Yes
2. No

Left [k]?

Left silence?

Right vowel?

A

B

C

D

E

−[iy]+*

(state \(i\))

H. Kamper & T.R. Niesler (Stellenbosch University)

Accent-targeted acoustic model optimisation

PRASA 2012 5 / 14
Decision-tree state clustering

\[ * - [iy] + * \]

(state \( i \))

Left voiced?

yes

no

Left vowel?

Right plosive?

yes

no

Afrikaans English?

Left silence?

Right vowel?

A

B

C

D

E

\[ * - [iy] + * \]

(state \( i \))
Decision-tree state clustering

\[
* - [iy] + *
\]

\[(state \ i)\]

Left voiced?

- yes
- no

Left vowel?

- yes
- no

Right plosive?

- yes
- no

H. Kamper & T.R. Niesler (Stellenbosch University)
Decision-tree state clustering

$$- [iy] + *$$

(\text{state } i)

Left voiced?

yes

no

Left vowel?

yes

no

Right plosive?

yes

no

Left [k]?

yes

no

Left voiced?

Afrikaans English?

Left silence?

Right vowel?

A

B

C

D

E
Decision-tree state clustering

![Decision-tree diagram]

- *−[iy]++* (state \( i \))
- Left voiced?
  - yes
  - no
- Left vowel?
  - yes
  - no
- Right plosive?
  - yes
  - no
- Left [k]?
  - yes
  - no

H. Kamper & T.R. Niesler (Stellenbosch University)
Decision-tree state clustering

\[ *[iy]+* \]

(state \(i \))

Left voiced?

Yes

No

Right vowel?

Yes

No

Afrikaans English?

Yes

No

Left silence?

Yes

No

H. Kamper & T.R. Niesler (Stellenbosch University)

Accent-targeted acoustic model optimisation
Multi-accent acoustic modelling

AE HMM for triphone [t]−[iy]+[ng]

EE HMM for triphone [t]−[iy]+[ng]
Traditional modelling approaches

Accent-specific models

AE HMM for triphone [t]−[iy]+[ng]

EE HMM for triphone [t]−[iy]+[ng]
Traditional modelling approaches

Accent-specific models

AE HMM for triphone [t]−[iy]+[ng]

AE HMM for triphone [t]−[iy]+[ng]

EE HMM for triphone [t]−[iy]+[ng]

Accent-independent models

AE HMM for triphone [t]−[iy]+[ng]

AE HMM for triphone [t]−[iy]+[ng]

EE HMM for triphone [t]−[iy]+[ng]
## Traditional modelling approaches

### Phone recognition accuracy (%)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AE</th>
<th>BE</th>
<th>CE</th>
<th>EE</th>
<th>IE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent-specific</td>
<td>64.80</td>
<td>56.77</td>
<td>64.59</td>
<td>72.97</td>
<td>64.27</td>
<td>64.68</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>65.97</td>
<td>55.98</td>
<td>66.51</td>
<td>74.45</td>
<td>64.40</td>
<td>65.44</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>66.20</td>
<td>56.56</td>
<td>66.31</td>
<td>73.94</td>
<td>64.60</td>
<td>65.50</td>
</tr>
</tbody>
</table>

H. Kamper & T.R. Niesler (Stellenbosch University)
Problem with multi-accent state clustering

\[ \mu(S), \Sigma(S), L(S) \]

\( S \)
Problem with multi-accent state clustering

\( \mu(S), \Sigma(S), L(S) \)

\( S \) (Accent) question \( q \)?

\( S = S_x \cup S_t \)

Splitting criterion:

\[ \Delta L_q = L(S_1(q)) + L(S_2(q)) - L(S) \]

H. Kamper & T.R. Niesler (Stellenbosch University)
Problem with multi-accent state clustering

\( \mu(S), \Sigma(S), L(S) \)

(Accent) question \( q \)?

yes

\( S_1(q) \)

\( \mu(S_1(q)), \Sigma(S_1(q)), L(S_1(q)) \)

no

\( S_2(q) \)

\( \mu(S_2(q)), \Sigma(S_2(q)), L(S_2(q)) \)
Problem with multi-accent state clustering

\[ \mu(S), \Sigma(S), L(S) \]

(Accent) question \( q \)?

yes \( S_1(q) \)
\[ \mu(S_1(q)), \Sigma(S_1(q)), L(S_1(q)) \]

no \( S_2(q) \)
\[ \mu(S_2(q)), \Sigma(S_2(q)), L(S_2(q)) \]

Splitting criterion: \( \Delta L_q = L(S_1(q)) + L(S_2(q)) - L(S) \)
Problem with multi-accent state clustering

\( \mu(S), \Sigma(S), L(S) \)

(Accent) question \( q \)?

yes

| \( \mu(S_1(q)), \Sigma(S_1(q)), L(S_1(q)) \) |
| \( \mu(S_2(q)), \Sigma(S_2(q)), L(S_2(q)) \) |

no

Splitting criterion: \( \Delta L_q = L(S_1(q)) + L(S_2(q)) - L(S) \)
Problem with multi-accent state clustering

\[ \mu(S), \Sigma(S), L(S) \]

(Accent) question \( q \)?

yes

\[ \mu(S_1(q)), \Sigma(S_1(q)), L(S_1(q)) \]

no

\[ \mu(S_2(q)), \Sigma(S_2(q)), L(S_2(q)) \]

Splitting criterion: \( \Delta L_q = L(S_1(q)) + L(S_2(q)) - L(S) \)

The question is: what happens to \( L_{AE}(S) \)?
Problem with multi-accent state clustering

\[ \mu(S), \Sigma(S), L(S) \]

\[ S = S_x \cup S_t \]

(Accent) question \( q? \)

\text{yes} \quad \text{no}

\[ \mu(S_1(q)), \Sigma(S_1(q)), L(S_1(q)) \quad \mu(S_2(q)), \Sigma(S_2(q)), L(S_2(q)) \]

Splitting criterion: \( \Delta L_q = L(S_1(q)) + L(S_2(q)) - L(S) \)

The question is: what happens to \( L_t(S) \)?
Targeted multi-accent acoustic modelling

Proposal: replace $L(S)$ with $L_t(S)$ in the standard clustering procedure.
Targeted multi-accent acoustic modelling

Proposal: replace $L(\mathcal{S})$ with $L_t(\mathcal{S})$ in the standard clustering procedure

But can we calculate $L_t(\mathcal{S})$?
Targeted multi-accent acoustic modelling

Proposal: replace $L(S)$ with $L_t(S)$ in the standard clustering procedure

But can we calculate $L_t(S)$?

$$L_t(S) = \log \prod_{f \in F_t} p(o_f|S)$$

($F_t$ is frames generated by states $S_t$)
Targeted multi-accent acoustic modelling

Proposal: replace $L(S)$ with $L_t(S)$ in the standard clustering procedure

But can we calculate $L_t(S)$?

\[
L_t(S) = \log \prod_{f \in F_t} p(o_f | S)
\]

\[
= \sum_{f \in F_t} \log [\mathcal{N}(o_f | \mu(S), \Sigma(S))]
\]  

($F_t$ is frames generated by states $S_t$)  

(Gaussian observation PDFs)
Targeted multi-accent acoustic modelling

Proposal: replace $L(S)$ with $L_t(S)$ in the standard clustering procedure

But can we calculate $L_t(S)$?

\[
L_t(S) = \log \prod_{f \in F_t} p(o_f | S) \tag{\text{\(F_t\) is frames generated by states \(S_t\)}}
\]

\[
= \sum_{f \in F_t} \log [N(o_f | \mu(S), \Sigma(S))] \tag{\text{Gaussian observation PDFs}}
\]

\[
= -\frac{1}{2} N_t \left\{ \log [(2\pi)^n |\Sigma(S)|] \right\} - \frac{1}{2} n (N_x + N_t)
\]

\[
+ \frac{1}{2} \text{tr} \left\{ \Sigma^{-1}(S) N_x \left[ \Sigma(S_x) + (\mu(S_x) - \mu(S))(\mu(S_x) - \mu(S))^T \right] \right\}
\]
Targeted multi-accent acoustic modelling

Proposal: replace $L(S)$ with $L_t(S)$ in the standard clustering procedure
But can we calculate $L_t(S)$?

\[
L_t(S) = \log \prod_{f \in F_t} p(o_f | S) \\
= \sum_{f \in F_t} \log [\mathcal{N}(o_f | \mu(S), \Sigma(S))] \\
= -\frac{1}{2} N_t \left\{ \log[(2\pi)^n | \Sigma(S)|] \right\} - \frac{1}{2} n(N_x + N_t) \\
+ \frac{1}{2} \text{tr} \left\{ \Sigma^{-1}(S) N_x [\Sigma(S_x) + (\mu(S_x) - \mu(S))(\mu(S_x) - \mu(S))^T] \right\}
\]

Since $\mu(S), \mu(S_x), \Sigma(S)$ and $\Sigma(S_x)$ are calculable from only the the means and covariance matrices of the states in the corresponding clusters, the calculation of $L_t(S)$ for each possible cluster split is computationally tractable.
Targeted multi-accent acoustic modelling

So let us take $L_t(S)$ as *splitting criterion* in our decision-trees.
Targeted multi-accent acoustic modelling

So let us take $L_t(S)$ as strong \textbf{splitting criterion} in our decision-trees . . . problems?
So let us take $L_t(S)$ as **splitting criterion** in our decision-trees . . . problems?

![Diagram](https://example.com/diagram.png)
## Targeted decision-tree state clustering

### Phone recognition accuracy (%)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AE</th>
<th>BE</th>
<th>CE</th>
<th>EE</th>
<th>IE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent-specific</td>
<td>64.80</td>
<td>56.77</td>
<td>64.59</td>
<td>72.97</td>
<td>64.27</td>
<td>64.68</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>65.97</td>
<td>55.98</td>
<td>66.51</td>
<td>74.45</td>
<td>64.40</td>
<td>65.44</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>66.20</td>
<td>56.56</td>
<td>66.31</td>
<td>73.94</td>
<td>64.60</td>
<td>65.50</td>
</tr>
<tr>
<td>Targeted multi-accent</td>
<td>64.60</td>
<td>55.17</td>
<td>64.11</td>
<td>72.65</td>
<td>64.44</td>
<td><strong>64.21</strong></td>
</tr>
</tbody>
</table>
Weighted targeted decision-tree state clustering

Let us weigh the likelihoods:

\[ L_w(S) = w_t L_t(S) + w_x L_x(S) \]
Weighted targeted decision-tree state clustering

Let us weigh the likelihoods: \( L_w(S) = w_t L_t(S) + w_x L_x(S) \)

Phone recognition accuracy (%)

<table>
<thead>
<tr>
<th>Approach</th>
<th>AE</th>
<th>BE</th>
<th>CE</th>
<th>EE</th>
<th>IE</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accent-specific</td>
<td>64.80</td>
<td>56.77</td>
<td>64.59</td>
<td>72.97</td>
<td>64.27</td>
<td>64.68</td>
</tr>
<tr>
<td>Accent-independent</td>
<td>65.97</td>
<td>55.98</td>
<td>66.51</td>
<td>74.45</td>
<td>64.40</td>
<td>65.44</td>
</tr>
<tr>
<td>Multi-accent</td>
<td>66.20</td>
<td>56.56</td>
<td>66.31</td>
<td>73.94</td>
<td>64.60</td>
<td>65.50</td>
</tr>
<tr>
<td>Targeted multi-accent</td>
<td>64.60</td>
<td>55.17</td>
<td>64.11</td>
<td>72.65</td>
<td>64.44</td>
<td>64.21</td>
</tr>
<tr>
<td>Weighted targeted</td>
<td>66.74</td>
<td>56.56</td>
<td>66.13</td>
<td>73.94</td>
<td>64.96</td>
<td>65.65</td>
</tr>
<tr>
<td>Weight ( w_t ) used above</td>
<td>0.51</td>
<td>0.5</td>
<td>0.53</td>
<td>0.5</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>
Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**

- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)

- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained

- **Criticism**: clustering early on in model training process, no guarantees

- **Future**: compare/incorporate to/in classic **adaptation** approaches
Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit optimisation on a target accent

- Showed that when likelihood is calculated only on target accent, performance deteriorates (possibly due to high separation of target)

- Showed that when some weight is also assigned to non-target accents (giving control over separation) very small improvements can be obtained

- Criticism: clustering early on in model training process, no guarantees

- Future: compare/incorporate to/in classic adaptation approaches
Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated **only** on **target accent**, **performance deteriorates** (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained

**Criticism**: clustering early on in model training process, no guarantees

**Future**: compare/incorporate to/in classic **adaptation** approaches
Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**

- Showed that when likelihood is calculated **only** on **target accent**, performance **deteriorates** (possibly due to high separation of target)

- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained

- **Criticism**: clustering early on in model training process, no guarantees

- **Future**: compare/incorporate to/in classic **adaptation** approaches
Summary and conclusions

- Extended the standard decision-tree state clustering algorithm to allow explicit **optimisation** on a **target accent**
- Showed that when likelihood is calculated only on **target accent**, performance deteriorates (possibly due to high separation of target)
- Showed that when some **weight** is also assigned to **non-target accents** (giving control over separation) very **small improvements** can be obtained
- **Criticism**: clustering early on in model training process, no guarantees
- **Future**: compare/incorporate to/in classic **adaptation** approaches