Articulatory features for word recognition using dynamic Bayesian networks

Mirjam Wester

Centre for Speech Technology Research, University of Edinburgh

10th April 2007
Motivation
  Why not phones?
  Articulatory features

Articulatory feature recognition
  Data
  Models
  AF Results

Word model
  Pronunciation model
  6-state word models
  Phone-based word models
  Articulatory feature-based word models

Conclusions
  Conclusions
What is wrong with phones?

Spontaneous speech effects

- modelling words as sequences of non-overlapping phone segments ("beads-on-a-string" paradigm) is unrealistic and creates many problems
  - difficult to model the variation present in spontaneous, conversational speech
What is wrong with phones?

Spontaneous speech effects

- modelling words as sequences of non-overlapping phone segments (“beads-on-a-string” paradigm) is unrealistic and creates many problems
  - difficult to model the variation present in spontaneous, conversational speech
- variation arises from the overlapping, asynchronous nature of speech production
  - standard solution: context-dependent phone models, though these can only deal with certain effects, and necessitate parameter tying to alleviate problems of data sparsity
What is wrong with phones?

Language universality

- a universal phone set has to be large (e.g. IPA)
- will contain many rarely-used symbols
- not at all clear that the same IPA symbol is actually pronounced the same in different languages anyway
What is wrong with phones?

Language universality

- a universal phone set has to be large (e.g. IPA)
- will contain many rarely-used symbols
- not at all clear that the same IPA symbol is actually pronounced the same in different languages anyway

A large phone set is problematic for modelling, just like trying to do large-vocab ASR using whole-word models.
What is wrong with phones?

Language universality

- a universal phone set has to be large (e.g. IPA)
- will contain many rarely-used symbols
- not at all clear that the same IPA symbol is actually pronounced the same in different languages anyway

A large phone set is problematic for modelling, just like trying to do large-vocab ASR using whole-word models.

One solution: decompose/factorise phones into a small set of symbols/factors
Articulatory features (AFs) – linguistic motivation

We are building a recognition system in which articulatory features, not phones, mediate between words and acoustic observations.

- AFs are multi-levelled features such as place, manner of articulation, etc
Articulatory features (AFs) – linguistic motivation

We are building a recognition system in which articulatory features, not phones, mediate between words and acoustic observations.

- AFs are multi-levelled features such as place, manner of articulation, etc
- they provide a compact encoding of variation present in natural speech
Articulatory features (AFs) – linguistic motivation

We are building a recognition system in which articulatory features, not phones, mediate between words and acoustic observations.

- AFs are multi-levelled features such as place, manner of articulation, etc
- they provide a compact encoding of variation present in natural speech
- allow simple accounts of spontaneous speech effects
Articulatory features (AFs) – linguistic motivation

We are building a recognition system in which articulatory features, not phones, mediate between words and acoustic observations.

- AFs are multi-levelled features such as place, manner of articulation, etc
- they provide a compact encoding of variation present in natural speech
- allow simple accounts of spontaneous speech effects
- it should be easier to specify a language-universal feature set
Articulatory features (AFs) – linguistic motivation

We are building a recognition system in which articulatory features, not phones, mediate between words and acoustic observations.

- AFs are multi-levelled features such as place, manner of articulation, etc
- they provide a compact encoding of variation present in natural speech
- allow simple accounts of spontaneous speech effects
- it should be easier to specify a language-universal feature set
- this is an articulatory-inspired representation - we are not trying to do articulatory inversion, which aims to recover precise articulator positions.
Articulatory features (AFs) – machine-learning motivation

- AFs are a distributed (factorial) representation
Articulatory features (AFs) – machine-learning motivation

- AFs are a distributed (factorial) representation
- Potential to make better use of limited training data
  - Effectively, train a number of low-cardinality classifiers
  - Fewer classes: less likely to suffer data sparsity
## Feature specification

<table>
<thead>
<tr>
<th>feature</th>
<th>values</th>
<th>cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>manner</td>
<td>approximant, fricative, nasal, stop, vowel, silence</td>
<td>6</td>
</tr>
<tr>
<td>place</td>
<td>labiodental, dental, alveolar, velar, high, mid, low, silence</td>
<td>8</td>
</tr>
<tr>
<td>voicing</td>
<td>voiced, voiceless, silence</td>
<td>3</td>
</tr>
<tr>
<td>rounding</td>
<td>rounded, unrounded, nil, silence</td>
<td>4</td>
</tr>
<tr>
<td>front-back</td>
<td>front, central, back, nil, silence</td>
<td>5</td>
</tr>
<tr>
<td>static</td>
<td>static, dynamic, silence</td>
<td>3</td>
</tr>
</tbody>
</table>
OGI Numbers

- OGI numbers 30-word subset
OGI Numbers

- OGI numbers 30-word subset
- a little over 6 hours of train and 2 hours test data
OGI Numbers

- OGI numbers 30-word subset
- a little over 6 hours of train and 2 hours test data
- AF labels generated by mapping from time-aligned phone labels, using diacritics where appropriate

<table>
<thead>
<tr>
<th>Worldbet</th>
<th>example</th>
<th>manner</th>
<th>place</th>
<th>voice</th>
<th>front</th>
<th>round</th>
<th>static</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>five</td>
<td>fricative</td>
<td>labdent</td>
<td>-voice</td>
<td>nil</td>
<td>nil</td>
<td>static</td>
</tr>
<tr>
<td>l</td>
<td>six</td>
<td>vowel</td>
<td>high</td>
<td>+voice</td>
<td>front</td>
<td>-round</td>
<td>static</td>
</tr>
</tbody>
</table>

Mirjam Wester
Word recognition using AFs
OGI Numbers

- OGI numbers 30-word subset
- A little over 6 hours of train and 2 hours test data
- AF labels generated by mapping from time-aligned phone labels, using diacritics where appropriate

<table>
<thead>
<tr>
<th>Worldbet</th>
<th>example</th>
<th>manner</th>
<th>place</th>
<th>voice</th>
<th>front</th>
<th>round</th>
<th>static</th>
</tr>
</thead>
<tbody>
<tr>
<td>f</td>
<td>five</td>
<td>fricative</td>
<td>labdent</td>
<td>-voice</td>
<td>nil</td>
<td>nil</td>
<td>static</td>
</tr>
<tr>
<td>l</td>
<td>six</td>
<td>vowel</td>
<td>high</td>
<td>+voice</td>
<td>front</td>
<td>-round</td>
<td>static</td>
</tr>
</tbody>
</table>

- 39-dimensional acoustic observation vector: 12 Mel-frequency cepstral coefficients and energy, plus 1st and 2nd derivatives.
Word segmentations

- Word segmentations are derived from phonetic transcriptions
Word segmentations

- Word segmentations are derived from phonetic transcriptions
- Output from Fiona’s semi-automatic dictionary generating procedure
Word segmentations

- Word segmentations are derived from phonetic transcriptions
- Output from Fiona’s semi-automatic dictionary generating procedure
- Timing information is used to train word models
Evaluating AF recognition performance

No ideal metric with which to evaluate AF recognition

- framewise accuracy: comparison with phone-derived feature labels penalizes asynchrony
Evaluating AF recognition performance

No ideal metric with which to evaluate AF recognition

- framewise accuracy: comparison with phone-derived feature labels penalizes asynchrony
- recognition accuracy:

\[ 100 \times \left( n(\text{correct}) - n(\text{insertions}) \right) / n(\text{total labels}) \]

more useful, though has capacity to penalize events would like to capture, e.g. where assimilation should lead to the deletion of a feature value
Evaluating AF recognition performance

No ideal metric with which to evaluate AF recognition

- framewise accuracy: comparison with phone-derived feature labels penalizes asynchrony
- recognition accuracy:

\[ 100 \times \frac{(n(\text{correct}) - n(\text{insertions}))}{n(\text{total labels})} \]

more useful, though has capacity to penalize events would like to capture, e.g. where assimilation should lead to the deletion of a feature value

- Word models make it possible to compare effect of phones and AFs directly
ANN/HMMs without inter-feature dependencies
GMM/DBNs with inter-feature dependencies

Mirjam Wester

Word recognition using AFs
ANN/DBNs with inter-feature dependencies

=1 =1 =1 =1 =1 =1
=1 =1 =1 =1 =1 =1

Mirjam Wester

Word recognition using AFs
Summary of AF results

<table>
<thead>
<tr>
<th>model</th>
<th>average correct</th>
<th>correct together</th>
<th>accuracy</th>
<th>combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN/HMM</td>
<td>86.7%</td>
<td>71.7%</td>
<td>83.5%</td>
<td>3751</td>
</tr>
<tr>
<td>GMM/DBN</td>
<td>86.2%</td>
<td>79.4%</td>
<td>83.4%</td>
<td>117</td>
</tr>
<tr>
<td>ANN/DBN</td>
<td>89.1%</td>
<td>84.6%</td>
<td>87.8%</td>
<td>54</td>
</tr>
</tbody>
</table>

- Shown that DBNs can match ANN AF recognition accuracy
Summary of AF results

<table>
<thead>
<tr>
<th>model</th>
<th>average correct</th>
<th>correct together</th>
<th>accuracy</th>
<th>combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN/HMM</td>
<td>86.7%</td>
<td>71.7%</td>
<td>83.5%</td>
<td>3751</td>
</tr>
<tr>
<td>GMM/DBN</td>
<td>86.2%</td>
<td>79.4%</td>
<td>83.4%</td>
<td>117</td>
</tr>
<tr>
<td>ANN/DBN</td>
<td>89.1%</td>
<td>84.6%</td>
<td>87.8%</td>
<td>54</td>
</tr>
</tbody>
</table>

- Shown that DBNs can match ANN AF recognition accuracy
- State level coupling of features is indeed beneficial
Summary of AF results

<table>
<thead>
<tr>
<th>model</th>
<th>average correct</th>
<th>correct together</th>
<th>accuracy</th>
<th>combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN/HMM</td>
<td>86.7%</td>
<td>71.7%</td>
<td>83.5%</td>
<td>3751</td>
</tr>
<tr>
<td>GMM/DBN</td>
<td>86.2%</td>
<td>79.4%</td>
<td>83.4%</td>
<td>117</td>
</tr>
<tr>
<td>ANN/DBN</td>
<td>89.1%</td>
<td>84.6%</td>
<td>87.8%</td>
<td>54</td>
</tr>
</tbody>
</table>

- Shown that DBNs can match ANN AF recognition accuracy
- State level coupling of features is indeed beneficial
- Reduced our dependence on phone-derived feature labels and learned set of asynchronous changes
Summary of AF results

<table>
<thead>
<tr>
<th>model</th>
<th>average correct</th>
<th>correct together</th>
<th>accuracy</th>
<th>combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN/HMM</td>
<td>86.7%</td>
<td>71.7%</td>
<td>83.5%</td>
<td>3751</td>
</tr>
<tr>
<td>GMM/DBN</td>
<td>86.2%</td>
<td>79.4%</td>
<td>83.4%</td>
<td>117</td>
</tr>
<tr>
<td>ANN/DBN</td>
<td>89.1%</td>
<td>84.6%</td>
<td>87.8%</td>
<td>54</td>
</tr>
</tbody>
</table>

- Shown that DBNs can match ANN AF recognition accuracy
- State level coupling of features is indeed beneficial
- Reduced our dependence on phone-derived feature labels and learned set of asynchronous changes
- Order of magnitude fewer feature combinations may be a suitable operating point between:
Summary of AF results

<table>
<thead>
<tr>
<th>model</th>
<th>average correct</th>
<th>correct together</th>
<th>accuracy</th>
<th>combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN/HMM</td>
<td>86.7%</td>
<td>71.7%</td>
<td>83.5%</td>
<td>3751</td>
</tr>
<tr>
<td>GMM/DBN</td>
<td>86.2%</td>
<td>79.4%</td>
<td>83.4%</td>
<td>117</td>
</tr>
<tr>
<td>ANN/DBN</td>
<td>89.1%</td>
<td>84.6%</td>
<td>87.8%</td>
<td>54</td>
</tr>
</tbody>
</table>

- Shown that DBNs can match ANN AF recognition accuracy
- State level coupling of features is indeed beneficial
- Reduced our dependence on phone-derived feature labels and learned set of asynchronous changes
- Order of magnitude fewer feature combinations may be a suitable operating point between:
  - All possible feature value combinations (linguistically implausible)
Summary of AF results

<table>
<thead>
<tr>
<th>model</th>
<th>average correct</th>
<th>correct together</th>
<th>accuracy</th>
<th>combinations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANN/HMM</td>
<td>86.7%</td>
<td>71.7%</td>
<td>83.5%</td>
<td>3751</td>
</tr>
<tr>
<td>GMM/DBN</td>
<td>86.2%</td>
<td>79.4%</td>
<td>83.4%</td>
<td>117</td>
</tr>
<tr>
<td>ANN/DBN</td>
<td>89.1%</td>
<td>84.6%</td>
<td>87.8%</td>
<td>54</td>
</tr>
</tbody>
</table>

- Shown that DBNs can match ANN AF recognition accuracy
- State level coupling of features is indeed beneficial
- Reduced our dependence on phone-derived feature labels and learned set of asynchronous changes
- Order of magnitude fewer feature combinations may be a suitable operating point between:
  - All possible feature value combinations (linguistically implausible)
  - Only combinations which correspond to canonical phonemes (back to the “beads-on-a-string” problem).
Towards a word model

- We have the observation process in place: AF recognizer

observation

features

templates

Mirjam Wester

Word recognition using AFs
Towards a word model

- We have the observation process in place: AF recognizer

- Now we simply add on the rest to build a word recognizer.
Incorporating a pronunciation model

- Complete integration of word-feature layer
Incorporating a pronunciation model

- Complete integration of word-feature layer
  - AF recognition component will form observation process
Incorporating a pronunciation model

- Complete integration of word-feature layer
  - AF recognition component will form observation process
  - Generate word by choosing a template for each feature group, where a template gives a sequence of feature values, but not timings.
Incorporating a pronunciation model

- Complete integration of word-feature layer
  - AF recognition component will form observation process
  - Generate word by choosing a template for each feature group, where a template gives a sequence of feature values, but not timings.

```
<table>
<thead>
<tr>
<th>manner</th>
<th>template (i)</th>
<th>p=0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fricative</td>
<td>vowel</td>
</tr>
<tr>
<td></td>
<td>[f ao r]</td>
<td></td>
</tr>
</tbody>
</table>

"four"
```

```
<table>
<thead>
<tr>
<th>manner</th>
<th>template (ii)</th>
<th>p=0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>fricative</td>
<td>vowel</td>
</tr>
<tr>
<td></td>
<td>[f ao]</td>
<td></td>
</tr>
</tbody>
</table>
```
• Unfortunately it’s not straightforward how to add the word recognition to the observation process.
• Unfortunately it’s not straightforward how to add the word recognition to the observation process.

• So back to basics...
Word model

Pronunciation model
- 6-state word models
- Phone-based word models
- AF word models

Talk outline
- Motivation
- AF recognition
- Word model
- Conclusions

Mirjam Wester
Word recognition using AFs
Word model

- Word counter
- Word
- Word position
- Phone
- Acoustic observation

Word transition
Phone transition
Word model

- Motivation
- AF recognition
- Word model
- Conclusions

Pronunciation model
- 6-state word models
- Phone-based word models
- AF word models

Word recognition using AFs
Word model

6-state word models
Phone-based word models
AF word models
Word model

Pronunciation model
- 6-state word models
- Phone-based word models
- AF word models

AF recognition

Word model
- phone
- word
- position
- word counter

Conclusions

Word recognition using AFs
6-state word models

- 6 states per word
- 31 words (30 words + silence)
- No pronunciation model
- 13 iterations of splitting and vanishing scheme
6-state word models

- 6 states per word
- 31 words (30 words + silence)
- No pronunciation model
- 13 iterations of splitting and vanishing scheme
- 7.1% WER
Phone-based word model

- 3 states per phone
- 31 words (30 words + silence)
- No explicit pronunciation variation model
- Top 1 variant in training data for each word
- 13 iterations of splitting and vanishing scheme
Phone-based word model

- 3 states per phone
- 31 words (30 words + silence)
- No explicit pronunciation variation model
- Top 1 variant in training data for each word
- 13 iterations of splitting and vanishing scheme
- 6.9% WER
Articulatory feature-based word model

<table>
<thead>
<tr>
<th>feature</th>
<th># templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>manner</td>
<td>232</td>
</tr>
<tr>
<td>place</td>
<td>312</td>
</tr>
<tr>
<td>voicing</td>
<td>48</td>
</tr>
<tr>
<td>rounding</td>
<td>137</td>
</tr>
<tr>
<td>front-back</td>
<td>223</td>
</tr>
<tr>
<td>static</td>
<td>62</td>
</tr>
</tbody>
</table>

- CPT for $p(\text{lex\_var}|\text{word})$ with AFs observed
Articulatory feature-based word model

<table>
<thead>
<tr>
<th>feature</th>
<th># templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>manner</td>
<td>232</td>
</tr>
<tr>
<td>place</td>
<td>312</td>
</tr>
<tr>
<td>voicing</td>
<td>48</td>
</tr>
<tr>
<td>rounding</td>
<td>137</td>
</tr>
<tr>
<td>front-back</td>
<td>223</td>
</tr>
<tr>
<td>static</td>
<td>62</td>
</tr>
</tbody>
</table>

- CPT for $p(\text{lex_var}|\text{word})$ with AFs observed
- However, too many zero prob utterances and memory allocation problems
Articulatory feature-based word model

<table>
<thead>
<tr>
<th>feature</th>
<th># templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>manner</td>
<td>232</td>
</tr>
<tr>
<td>place</td>
<td>312</td>
</tr>
<tr>
<td>voicing</td>
<td>48</td>
</tr>
<tr>
<td>rounding</td>
<td>137</td>
</tr>
<tr>
<td>front-back</td>
<td>223</td>
</tr>
<tr>
<td>static</td>
<td>62</td>
</tr>
</tbody>
</table>

- CPT for $p(\text{lex_var}|\text{word})$ with AFs observed
- However, too many zero prob utterances and memory allocation problems
- 1 variant per word - add in pronunciation variation later
Articulatory feature-based word model

<table>
<thead>
<tr>
<th>feature</th>
<th># templates</th>
</tr>
</thead>
<tbody>
<tr>
<td>manner</td>
<td>232</td>
</tr>
<tr>
<td>place</td>
<td>312</td>
</tr>
<tr>
<td>voicing</td>
<td>48</td>
</tr>
<tr>
<td>rounding</td>
<td>137</td>
</tr>
<tr>
<td>front-back</td>
<td>223</td>
</tr>
<tr>
<td>static</td>
<td>62</td>
</tr>
</tbody>
</table>

- CPT for $p(\text{lex.var} | \text{word})$ with AFs observed
- However, too many zero prob utterances and memory allocation problems
- 1 variant per word - add in pronunciation variation later
- still working on this...
Conclusions

- WERs for state-based word models and phone-based word models look good.
- Watch this space for AF results