Feature Transformation Applied to the Detection of Discontinuities in Concatenated Speech

Barry Kirkpatrick, Darragh O’Brien, Ronán Scaife,
Faculty of Engineering and Computing
Dublin City University
Overview

• Problem definition
• DCU database
• Approach
• Results: PCA
• Results: PCA+ANN
• Results: combining feature sets
• Summary and Conclusion
Concatenative speech synthesis

- Database of recorded speech
- **Chain segments** of recorded speech units
- Natural sounding – State-of-the-art?
- Unit selection
- Inconsistent quality

B. Kirkpatrick *et al.*, – SSW6 Bonn, August 2007
The problem

Several levels:

• how to emulate human judgments of “naturalness” for synthetic audio or video?
• how to optimally match human perception of discontinuity in synthetic speech?
• how to match human perception of spectral (rather than say $f_0$) discontinuity?…
Why we need better models

- If we can accurately model human naturalness judgments, we can:
  - Produce better raw concatenations.
  - Develop spectral interpolation schemes to “repair” bad joins.
  - Optimise size and quality of unit selection database.
Database (I)

- Based on simple perceptual experiment:
- 1 adult male recorded 300 mono-syllabic words from MRT list.
- 1800 CVC words created by PSOLA concatenating left- and right-hand parts with common vowel.
- Task was binary continuous/discontinuous judgment.
Database (II)

- 12 listeners; 3 per subtest (6 words).
- Majority scoring of results.
- Initial use for database was to resolve widely differing reports of “optimal” join cost/distance measure.
- No attempt at spectral interpolation, although results may inform development of such algorithms.
Database (III)

- 4 pitch-period linear fade.
- *Not* yet ready to embed into synthesizer.
Present Study

- Many feature sets (MFCC, LSF, PSD, etc) have been proposed for unit selection join cost calculations.
- Many distance measures have been tried on above feature sets.
- Can more discriminating power be extracted from existing feature sets (a la ASR)?
Approach (I)

• Explore Principal Component Analysis (PCA) and Artificial Neural Networks (ANN) to improve discrimination.

• Which ANN? General Regression Neural Network (GRNN).

• PCA front-end modestly improves discrimination, but mainly allows GRNN with manageable number of nodes.
Approach (II)

• All units represented by time sequence of feature vectors $x$.
• For each candidate join, compute join vector as difference of adjacent frames:

$$x_{\text{join}} = x_{\text{left}} - x_{\text{right}}$$

(rather than scalar distance.)

B. Kirkpatrick et al, – SSW6 Bonn, August 2007
Approach (III)

- PCA resolves raw join vectors into uncorrelated components, ordered as to variance.
- Number of PCA components per frame much less than original feature dimension (39 vs 256 for logPSD.)
- Study restricted to spectral discontinuity (not energy or fundamental frequency.)
PCA performance vs dimension

AUC vs PCA output dimension;
(log PSD feature).

B. Kirkpatrick et al, – SSW6 Bonn, August 2007
PCA alone

- Computed across entire dataset.
- PCA tested with:
  - MFCC
  - LSF
  - LogPSD (from DFT)
- Modest discrimination *increase* for MFCC and logPSD, larger *decrease* for LSF.
- Large reduction in dimensionality.
PCA gains

<table>
<thead>
<tr>
<th>Feature</th>
<th>x</th>
<th>PCA(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.75</td>
<td>0.7696</td>
</tr>
<tr>
<td>LSF</td>
<td>0.7381</td>
<td>0.6966</td>
</tr>
<tr>
<td>PSD (log DFT)</td>
<td>0.7615</td>
<td>0.7841</td>
</tr>
</tbody>
</table>

AUC results for each feature set: without and with PCA.
Combining PCA and ANN

• Can (non-linear) processing improve on PCA-processed features?
• For each of MFCC, LSF and log PSD, computed discrimination when PCA followed by (GRNN) ANN.
• Results assessed by area (AUC) under Receiver Operating Characteristic (ROC).
ANN Details (I)

- Database split equally into training and testing sets.
- Of 1800 concatenated words, 434 perceived discontinuities, equally split between training and testing set.
ANN Details (II)

- GRNN trained with target output (from perceptual test) of 1 for discontinuity, 0 for continuous.
- Results assessed with AUC as for PCA-only case.
## PCA+ANN results

<table>
<thead>
<tr>
<th>Feature</th>
<th>x</th>
<th>PCA+ANN (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>0.7565</td>
<td>0.8413</td>
</tr>
<tr>
<td>LSF</td>
<td>0.7468</td>
<td>0.7955</td>
</tr>
<tr>
<td>PSD (log DFT)</td>
<td>0.7673</td>
<td>0.8744</td>
</tr>
</tbody>
</table>
MFCC+PCA+ANN
LSF+PCA+ANN
logPSD +PCA+ANN
Combining Feature Sets

- Further modest increases in AUC obtained by concatenating feature vectors:

<table>
<thead>
<tr>
<th>Features</th>
<th>x</th>
<th>PCA+ANN (x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC+LSF</td>
<td>0.7468</td>
<td>0.8581</td>
</tr>
<tr>
<td>MFCC+logPSD</td>
<td>0.7673</td>
<td>0.8859</td>
</tr>
<tr>
<td>LSF+logPSD</td>
<td>0.7517</td>
<td>0.8753</td>
</tr>
<tr>
<td>LSF+MFCC +logPSD</td>
<td>0.7517</td>
<td>0.8829</td>
</tr>
</tbody>
</table>
MFCC and logPSD combined
Summary & conclusion

• Application of feature transformation to join cost optimisation.
• PCA to reduce dimensionality.
• ANN learns continuous/discontinuous discrimination function.
• Approach extracts useful extra discrimination.
• Feature sets may be usefully combined.
Feature Transformation Applied to the Detection of Discontinuities in Concatenated Speech

Barry Kirkpatrick, Darragh O’Brien, Ronán Scaife, Faculty of Engineering and Computing
Dublin City University
Unit selection

- Select optimum sequence of units
- Cost criterion
- Target cost – well defined
- Join cost – ill-defined
- Perception of joins
- F0, energy and spectral measures
- Problem – spectral measure
Related work

- Many previous studies addressing this problem:
  - Macon and Wouters (1998)
  - Stylianou and Syrdal (2001)
  - Vepa and King (2004, 2006)
  - Bellegarda (2004, 2006)

- Focused on comparing different feature sets
- Results inconsistent and largely inconclusive
- Sources of inconsistency?
Relating results

- Relating human results to subjective measures
- Receiver Operating Characteristic (ROC) curves
- Performance metric; area under the ROC curve (AUC)