Abstract
Despite years of improvement in the quality of HMM (Hidden Markov Model) synthesis, this type of synthetic speech still remains significantly less natural than speech output from good concatenative synthesis systems. This is commonly stated as being due to factors such as “over-smoothing” however to the best of our knowledge there have been no formal studies to support this. We aim to investigate the major causes for this loss in quality by observing separate assumptions applied during modelling and effects present in speech following modelling.

Investigation focus areas

1. Parameterisation errors
2. Assumptions applied
3. Poor quality speech

Method
- Natural and vocoded speech included in perceptual listening tests[1, 2, 3]
- Investigated different assumptions applied in HMM synthesis:
  - Source/filter can be separated for modelling[2, 3]
  - Parameter streams can be modelled independently[2]
  - Use of diagonal covariance matrices in modelling spectrum[2]
  - Averaging within context[1, 2]
  - Averaging across different contexts[1]
- Investigated effects present in HMM generated speech:
  - Construct simulation framework of HMM effects on speech[1, 3, 4]
  - Include HMM speech in perceptual listening test[1, 3]
  - Measure perceptual gains from current enhancement techniques (modulation spectrum, GV & formant enhancement)[3]

Conclusions
- Better vocoding required
  - [1, 2, 3] found speech quality was reduced in parametrisation step
- Apply better assumptions in modelling.
  - Source/filter separation for modelling is problematic[2, 3]
  - Source and filter interact and should be reflected in modelling[3]
  - Diagonal covariance modelling of spectrum significantly reduces quality[2]
  - Averaging of spectral parameter within context is somewhat harmful[1, 2]
  - Averaging across different contexts is much more harmful[1]
- Fix effects observed in modelled parameters.
  - Small amount of smoothing has little effect[4] (findings indicate that with correct variance in signal this may improve quality of vocoded speech[1])
  - Correct variance level is important[4], however [1] indicates increased variance level is preferable to decreased variance
- Current HMM enhancement techniques are providing gains however still remain noticeably behind vocoded speech[3]
- HMMs output reduced variance level speech parameters, even with the use of GV[1].

References
Attributing modelling errors in HMM synthesis by stepping gradually from natural to modelled speech

Thomas Merritt, Javier Latorre, Simon King
T.Merritt@ed.ac.uk, javier.latorre@crl.toshiba.co.uk, Simon.King@ed.ac.uk
Centre for Speech Technology Research, University of Edinburgh
Toshiba Research Europe Ltd., Cambridge Research Lab

Abstract
We present a framework for separating each of the effects of modelling in turn to observe their independent effects.

Method
- Simulate effects of modelling on speech parameters
- Listening tests asking “same or different” quality comparisons between simulations
- Multidimensional scaling of results
- Repeat on mceps and mel-LSPs

Framework

Simulated effects
The simulated effects were run over these signal-by-signal (each coefficient separately):
- Temporal smoothing - implemented as a weighted moving average via a sliding Hanning window of variable width.
- Variance adjustment - mean is subtracted from the signal before multiplying by a scalar value and adding the mean back in.
- “Pseudo-HMM” - average over state boundaries from HMM before applying MLPG with deltas and delta-deltas from HMM.

Multidimensional scaling

Conclusions
- Small amounts of smoothing are not harmful.
- Correct variance in speech parameters is preferred, erring on the side of higher variance is better than lower variance.
- Within context averaging is harmful - similar to excessive smoothing.
- Across context averaging is much more harmful.
Abstract

We investigate the perceptual degradations introduced by the independent modelling of vocal source and filter features in statistical parametric speech synthesis. This is achieved using stimuli in which various permutations of natural, vocoded and modelled source and filter are combined, optionally with the addition of filter modifications (e.g. global variance or modulation spectrum scaling). We also examine the assumption between source and filter parameters.

Method

- Simulate effects of modelling on speech parameters
- Add enhancements to HMM spectral parameters
- Listening tests asking “same or different quality” comparisons between simulations
- Multidimensional scaling of results
- Mean Opinion Score listening tests

Framework

<table>
<thead>
<tr>
<th>Speech</th>
<th>Extract source, spectrum, F0, HNR</th>
<th>Area to be simulated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>Learn model</td>
<td></td>
</tr>
<tr>
<td>Text input</td>
<td>Generate from model</td>
<td>Stored model</td>
</tr>
<tr>
<td></td>
<td>Reconstruct</td>
<td>Speech output</td>
</tr>
</tbody>
</table>

Vocoder configurations

<table>
<thead>
<tr>
<th>Condition name</th>
<th>Source</th>
<th>Filter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-natural</td>
<td>natural</td>
<td>natural</td>
</tr>
<tr>
<td>2-nat-voc</td>
<td>natural</td>
<td>vocoded</td>
</tr>
<tr>
<td>3-nat-hmm</td>
<td>natural</td>
<td>HMM</td>
</tr>
<tr>
<td>4-voc-voc</td>
<td>vocoded</td>
<td>vocoded</td>
</tr>
<tr>
<td>5-voc-hmm</td>
<td>vocoded</td>
<td>HMM</td>
</tr>
<tr>
<td>6-hmm-voc</td>
<td>HMM</td>
<td>vocoded</td>
</tr>
<tr>
<td>7-hmm-hmm</td>
<td>HMM</td>
<td>HMM</td>
</tr>
</tbody>
</table>

Modifications

Simulated effects

Simulated effects were run over spectral LSF parameters coefficient-by-coefficient:

- **Temporal smoothing** - implemented as a weighted moving average via a sliding Hanning window of variable width.
- **Variance adjustment** - mean is subtracted from the signal before multiplying by a scalar value and adding the mean back in.
- **Modulation spectrum** - scaled to match alter “smoothness” level of signal between vocoded and modelled.

Enhancement techniques

- **Formant enhancement** - applied to LSFs in the power spectrum domain.

Multidimensional scaling

Conclusions

- Clear differences in quality are noticeable to listeners as source and filter are varied from natural, through vocoded to modelled.
- Current filter enhancement techniques are able to recover some quality loss, yet remains behind vocoded level of quality.
- Final quality more affected by interaction between source and filter than individual quality levels.
Temporal smoothing[1, 2, 3, 4]

(a) Before
(b) After

Variance Adjustment[1, 3, 4]

(c) Before
(d) After

Pseudo-HMM[1]

(e) Set state values to mean value
(f) Apply MLP

Modulation Spectrum[3]

References


