3pSC12: Evaluation of a strategy for automatic formant tracking

T.M. Nearey (U. Alberta)
P.F. Assmann (U. TX Dallas)
J.M. Hillenbrand (Western MI U.)

Pan-American/Iberian Meeting on Acoustics
Cancun, 4 Dec 2002
Markel and Gray’s (1976) routine “FORMNT”

- Simple formant tracking strategy that works quite well for adult male voices
- Formant candidates from autocorrelation LPC
  - Sample at 10 kHz
  - Use autocorrelation LPC of order ~ 15
    - Allows for max of 7 formants in 5 kHz
M&G’s tracking strategy

- **FORMNT** routine assume exactly 3 formants < 3 kHz for adult male voices
  - M&G observe that in 85-90% of frames have exactly 3 reasonable candidates below 3 kHz
  - If so, assign these to F1, F2 and F3 slots
- If not, align available candidates to minimize discontinuity to previous frame
  - Copy over previous frame value to empty slots
Other voices

• M&G note this strategy is not appropriate for female, child voices

• Why? Consider kid’s voice with formants 30% higher than man's (F3 >> 3 kHz),
  – 15 order LPC in 10 kHz ‘too rich’ for high formant range voices
    • Expect 1 formant/ kHz for male voice (5 formants in 5 kHz)
    • Expect 1 formant/ 1.3 kHz for kid’s voice (<4 formants in 5 kHz)
    • Extra richness produces extra formant candidate that can split formants
Improvements: match settings to each voice

• Can often get good results for ‘high F3 voices’

• Choose appropriately for each voice:
  – Sampling rate
  – LPC order
  – F3 cutoff

• Produces raw candidate sets that are easy to track with simple algorithm
Informal interactive tracking strategy

• Take a few vowels from a given speaker
• Examine spectrograms and raw LPC candidate tracks
• Experiment with different analysis choices
• Determine maximum frequency range (selective LPC) and LPC order that make LPC candidates look easy to track
• Use those settings to analyze rest of vowels
Example analyses with varying cutoffs and LPC order.

Figures 1 to 3 show spectrograms and raw candidate tracks of vowel spoken by a male adult from Hillenbrand et al. (1995)

Figures 4 to 6 show vowel of a female child from Hillenbrand et al. (1995) data.
Figure 1a  Male adult voice, Cutoff 5 kHz Hz, Order 14

Colored lines connect raw candidates in freq. ordered slots; no formant tracking

There are only 5 true formants below 5 kHz. 14th order LPC usually finds 7 complex roots

Extra candidates between F3 and F4 ->

F1 is split ->

Nearey et al.
Figure 1b.
(Detail of Figure 1a)
Figure 2. Adult male voice. LPC cutoff 4 kHz, Order 9
Cutoff and order are just right. Easy tracking.
Figure 3. Adult male voice. LPC cutoff 3.5 kHz, Order 9
F4 is above cutoff. Extra candidates in F1 region.

Selective LPC Cutoff: 3500 Hz Order: 9
Figure 4. Female child’s voice. LPC cutoff 5 k Hz, Order 14
Only 4 formants in range; far too many candidates.
Figure 5. Female child’s voice. LPC cutoff 5.5k Hz, Order 9
Cutoff and order are just right. Easy tracking.
Figure 6. Female child’s voice. LPC cutoff 3.5k Hz, Order 9
Auto tracking strategy: General idea

• Candidates are easy to slot using simple strategy (like M&G’s *FORMNT*) if:
  – selective LPC cutoff is between formants 4 and 5 for a voice
  – order 9 analysis is chosen to allow for exactly 4 formants

• Figure 2 for male and Figure 4 for child show ‘clean’ analyses
Sketch of implementation

- Fix LPC order at 9 (yields 4 formants max)
- Apply several F4 cutoffs to each token
- For each cutoff:
  - Generate LPC candidate set
  - Use simple tracker, variant of M&G’s FORMNT, with F3 max set to 3/4 of F4 cutoff
  - Slot raw candidates into a trackset
  - Define a ‘goodness score’ (see below) to evaluate each trackset
- Pick cutoff whose trackset has highest score
Goodness score

• Global goodness is product (fuzzy $\text{AND}$) of 8 basic heuristic figures of merit (FOMs)
  – 6 basic FOMs based on information about a single trackset in isolation
    • reflect intuitions about ‘good’ formant tracks i.e., ones deemed likely to be picked by a human judge
  – 2 others involve more complex considerations sketched in notes panels at end
• Each FOM is assigned value between 1 (good) and 0 (terrible)
Simple figures of merit

- **1) Presence** - to what extent are there good candidates available to fill slots?
- **2) BwReason** : Are bandwidths of peaks reasonable?
- **3) AmpReason**: Is Amplitude reasonable?
- **4) ContReason**: Is there reasonable continuity of peaks within each formant?
- **5) DistReason**: Are F2-F1 and F3-F2 distances reasonable?
- **6) RangeReason**: Are formant ranges reasonable for given F4 cutoff?

- Two other FOMs and other details are given in Notes
Evaluation against hand-tracked formants

- Hillenbrand et al. (1995)
  - 12 Vowels spoken by 45 Men, 48 Women, 27 Boys, 19 Girls

  - 12 Vowels by 10 Men, 10 Women, 30 Children
    - 10 kids each, Ages 3, 5 and 7

- Rms errors of predicted and observed
  - Hillenbrand 20 to
## Error report: Hillenbrand data

<table>
<thead>
<tr>
<th>Speaker</th>
<th>N tokens</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1295</td>
<td>27</td>
<td>67</td>
<td>120</td>
<td>100%</td>
<td>98%</td>
<td>94%</td>
</tr>
<tr>
<td>Men</td>
<td>437</td>
<td>20</td>
<td>54</td>
<td>114</td>
<td>100%</td>
<td>99%</td>
<td>95%</td>
</tr>
<tr>
<td>Women</td>
<td>439</td>
<td>25</td>
<td>57</td>
<td>78</td>
<td>100%</td>
<td>99%</td>
<td>96%</td>
</tr>
<tr>
<td>Boys</td>
<td>243</td>
<td>29</td>
<td>81</td>
<td>141</td>
<td>99%</td>
<td>97%</td>
<td>93%</td>
</tr>
<tr>
<td>Girls</td>
<td>176</td>
<td>43</td>
<td>107</td>
<td>210</td>
<td>98%</td>
<td>94%</td>
<td>86%</td>
</tr>
</tbody>
</table>

Note: errors less than 300 Hz are not likely to involve formant skips
## Error report: Assmann Data

<table>
<thead>
<tr>
<th>Speaker</th>
<th>N tokens</th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
<th>F1 %</th>
<th>F2 %</th>
<th>F3 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1295</td>
<td>63</td>
<td>96</td>
<td>172</td>
<td>99%</td>
<td>97%</td>
<td>91%</td>
</tr>
<tr>
<td>Men</td>
<td>437</td>
<td>39</td>
<td>68</td>
<td>143</td>
<td>100%</td>
<td>98%</td>
<td>93%</td>
</tr>
<tr>
<td>Women</td>
<td>439</td>
<td>65</td>
<td>79</td>
<td>130</td>
<td>99%</td>
<td>99%</td>
<td>94%</td>
</tr>
<tr>
<td>7 yr. old</td>
<td>243</td>
<td>76</td>
<td>95</td>
<td>133</td>
<td>97%</td>
<td>97%</td>
<td>93%</td>
</tr>
<tr>
<td>5 yr old</td>
<td>176</td>
<td>94</td>
<td>129</td>
<td>211</td>
<td>96%</td>
<td>94%</td>
<td>87%</td>
</tr>
<tr>
<td>3 yr old</td>
<td>176</td>
<td>115</td>
<td>310</td>
<td>604</td>
<td>95%</td>
<td>78%</td>
<td>64%</td>
</tr>
</tbody>
</table>
Observations

• Overall performance on Hillenbrand data is excellent
• Assmann and Katz data shows larger errors, especially for 3 year olds
• Detailed comparison of chosen tracks and hand tracks is underway
  – In some cases, alternate tracksets not chosen by method are considerably better matches to hand tracking
  – Looking for clusters of difficult cases and possible additional heuristics to improve choice
Possible improvements

- More sophisticated formant tracking can be substituted for simple M&G style approach
  - e.g. Dynamic programming (Talkin 1987)
  - Can still take advantage of multiple cutoffs and figures of merit
- Better heuristics, possibly based on statistical distributions, might be substituted for fuzzy FOMs
- Better combinations (e.g. ANN ‘committee of experts’) might be used to combine individual FOMs
References


NOTES ON FIGURES OF MERIT

• Characterization of ‘reasonableness’
  – For each of above figures of merit define a ‘fuzzy membership function’
  – Define a range of ‘clearly good’ anchor values of a trackset property and assign a goodness value near 1.0; Define a range of ‘clearly bad’ trackset anchor values and assign a goodness value near 1.0; Use linear interpolation for values in between
  – Some good and bad anchor point values are adjusted proportionally to F4 cutoff.

• FOM7: Stable analysis \textit{RfStable}: Are formant tracks relatively stable when order is increased by 2?
  – If order 11 and order 9 show essentially same tracks, this suggests no formants have been skipped
  – RfStable is calculated as a correlation coefficient between the order 9 and order 11 tracksets for the cutoff

• FOM8: Analysis-by-synthesis check
  – 8) \textit{Rabs}: Correlation of resynthesized spectrogram with original
    • Synthesize smoothed spectra using method based on Olive(1971)
    • Uses F1, F2, F3 trackset plus a higher pole correction factor based on F4 cutoff
    • Calculate correlation coefficient between synthesized spectrogram and moderately high-order LPC smoothed spectrogram of original signal.
      – Allow optimal global dB/Octave spectral tilt and optimal frame-by-frame gain
    • ABS of entire trackset is a fairly fast, non-iterative process
## Hillenbrand data F1 results

<table>
<thead>
<tr>
<th>Formant 1</th>
<th>N tokens</th>
<th>rms err. (Hz)</th>
<th>Errors &lt; 300 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1295</td>
<td>27</td>
<td>99.6%</td>
</tr>
<tr>
<td>Men</td>
<td>437</td>
<td>20</td>
<td>100.0%</td>
</tr>
<tr>
<td>Women</td>
<td>439</td>
<td>25</td>
<td>100.0%</td>
</tr>
<tr>
<td>Boys</td>
<td>243</td>
<td>29</td>
<td>99.2%</td>
</tr>
<tr>
<td>Girls</td>
<td>176</td>
<td>43</td>
<td>98.3%</td>
</tr>
</tbody>
</table>

Note: Errors less than 300 Hz probably do not involve skipped formants
Hillenbrand data F2 results

<table>
<thead>
<tr>
<th>Formant 2</th>
<th>N tokens</th>
<th>rms err. (Hz)</th>
<th>Errors &lt; 300 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1295</td>
<td>67</td>
<td>97.7%</td>
</tr>
<tr>
<td>Men</td>
<td>437</td>
<td>54</td>
<td>98.6%</td>
</tr>
<tr>
<td>Women</td>
<td>439</td>
<td>57</td>
<td>98.6%</td>
</tr>
<tr>
<td>Boys</td>
<td>243</td>
<td>81</td>
<td>97.1%</td>
</tr>
<tr>
<td>Girls</td>
<td>176</td>
<td>107</td>
<td>93.8%</td>
</tr>
</tbody>
</table>
Hillenbrand data F3 results

<table>
<thead>
<tr>
<th>Formant</th>
<th>3</th>
<th>N tokens</th>
<th>rms err. (Hz)</th>
<th>Errors &lt; 300 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>1295</td>
<td>120</td>
<td>93.9%</td>
<td></td>
</tr>
<tr>
<td>Men</td>
<td>437</td>
<td>114</td>
<td>94.7%</td>
<td></td>
</tr>
<tr>
<td>Women</td>
<td>439</td>
<td>78</td>
<td>96.4%</td>
<td></td>
</tr>
<tr>
<td>Boys</td>
<td>243</td>
<td>141</td>
<td>93.4%</td>
<td></td>
</tr>
<tr>
<td>Girls</td>
<td>176</td>
<td>210</td>
<td>86.4%</td>
<td></td>
</tr>
</tbody>
</table>
### Assmann & Katz data F1 Results

<table>
<thead>
<tr>
<th>Talker group</th>
<th>N Tokens</th>
<th>rms err. (Hz)</th>
<th>Errors &lt; 300 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3434</td>
<td>63</td>
<td>98.5%</td>
</tr>
<tr>
<td>Men</td>
<td>1232</td>
<td>39</td>
<td>99.8%</td>
</tr>
<tr>
<td>Women</td>
<td>1230</td>
<td>65</td>
<td>98.9%</td>
</tr>
<tr>
<td>7 yr. olds.</td>
<td>498</td>
<td>76</td>
<td>97.4%</td>
</tr>
<tr>
<td>5 yr. olds.</td>
<td>251</td>
<td>94</td>
<td>96.0%</td>
</tr>
<tr>
<td>3 yr. olds.</td>
<td>223</td>
<td>115</td>
<td>94.6%</td>
</tr>
</tbody>
</table>
## Assmann & Katz data F2 Results

<table>
<thead>
<tr>
<th>Talker group</th>
<th>N tokens</th>
<th>rms err. (Hz)</th>
<th>Errors &lt; 300 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3434</td>
<td>96</td>
<td>96.5%</td>
</tr>
<tr>
<td>Men</td>
<td>1232</td>
<td>68</td>
<td>98.0%</td>
</tr>
<tr>
<td>Women</td>
<td>1230</td>
<td>79</td>
<td>98.6%</td>
</tr>
<tr>
<td>7 yr. olds</td>
<td>498</td>
<td>95</td>
<td>96.8%</td>
</tr>
<tr>
<td>5 yr. olds</td>
<td>251</td>
<td>129</td>
<td>94.4%</td>
</tr>
<tr>
<td>3 yr. olds</td>
<td>223</td>
<td>310</td>
<td>78.0%</td>
</tr>
</tbody>
</table>
## Assmann & Katz data: F3

<table>
<thead>
<tr>
<th>Talker Group</th>
<th>N tokens</th>
<th>rms err. (Hz)</th>
<th>Errors &lt; 300 Hz</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>3434</td>
<td>172</td>
<td>90.9%</td>
</tr>
<tr>
<td>Men</td>
<td>1232</td>
<td>143</td>
<td>92.5%</td>
</tr>
<tr>
<td>Women</td>
<td>1230</td>
<td>130</td>
<td>94.3%</td>
</tr>
<tr>
<td>7 yr. olds.</td>
<td>498</td>
<td>133</td>
<td>93.0%</td>
</tr>
<tr>
<td>5 yr. olds.</td>
<td>251</td>
<td>211</td>
<td>87.3%</td>
</tr>
<tr>
<td>3 yr. olds.</td>
<td>223</td>
<td>604</td>
<td>63.7%</td>
</tr>
</tbody>
</table>
Practical Applications

• Method is sufficiently promising that it can be used as basis of semi-automatic scheme
• Large number of files is processed in batch mode
  – All candidate tracksets are saved
• User views spectrogram with overlayed ‘best’ automatic formant tracks
  – User can view alternate tracksets and substitute one for ‘best’ choice.