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 ->

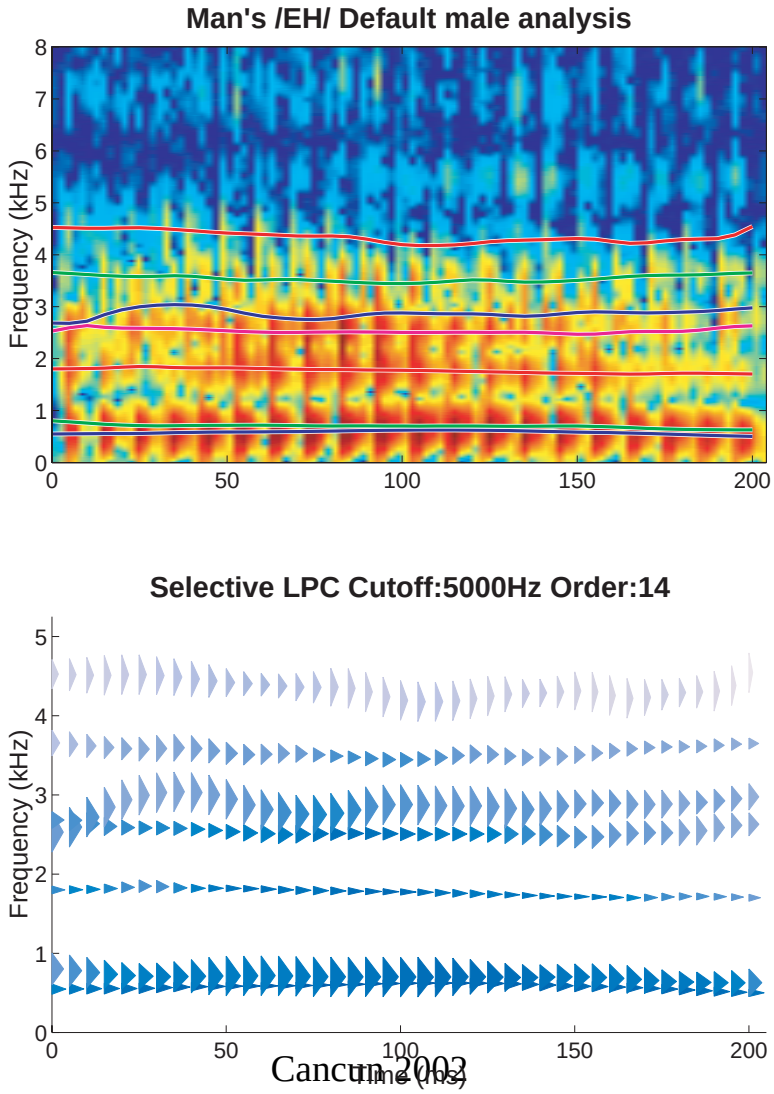


Figure 1b.

(Detail of Figure 1a)

Figure 2. Adult male voice. LPC cutoff 4 kHz, Order 9

Cutoff and order are just right. Easy tracking.

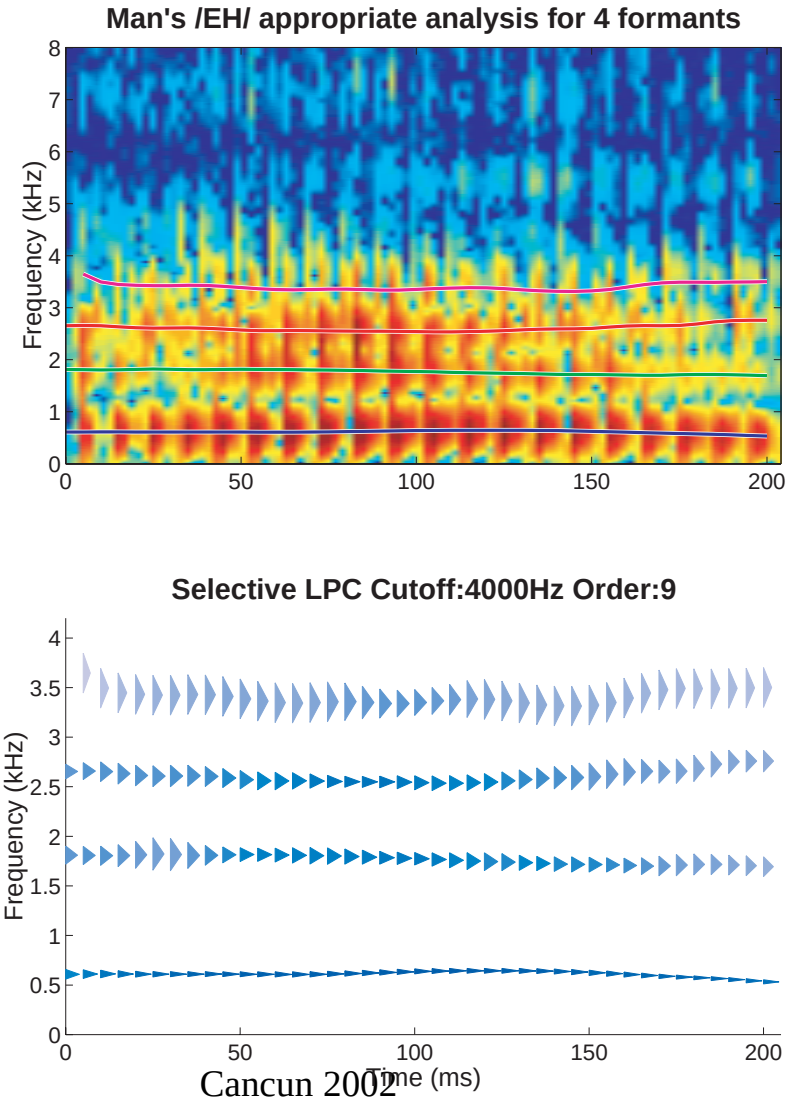


Figure 3. Adult male voice. LPC cutoff 3.5 kHz, Order 9

F4 is above cutoff. Extra candidates in F1 region

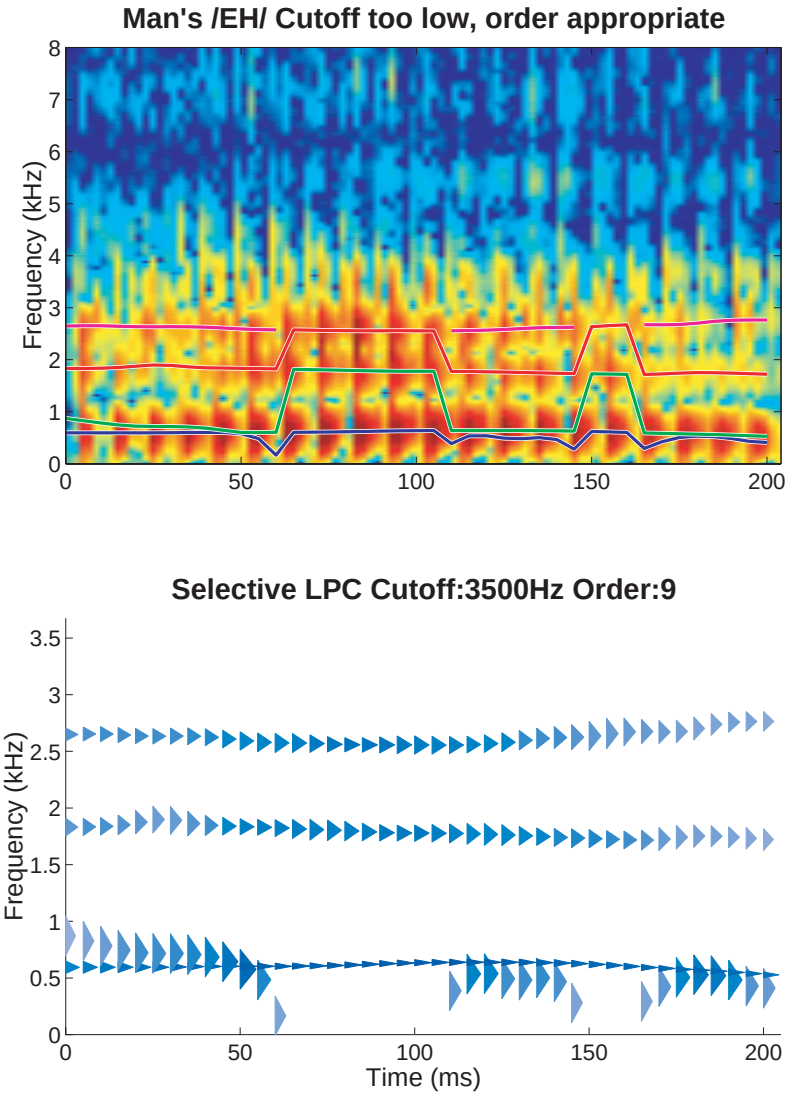


Figure 4. Female child's voice. LPC cutoff 5 k Hz, Order 14

Only 4 formants in range; far too many candidates.

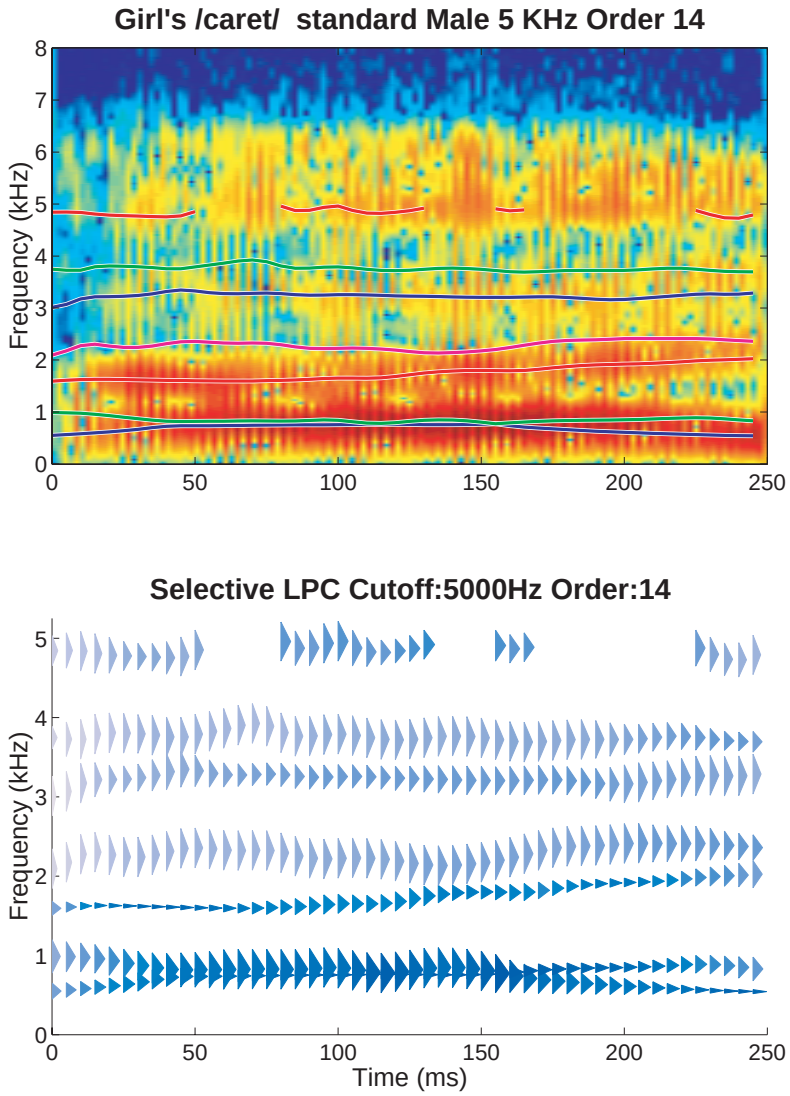


Figure 5. Female child's voice. LPC cutoff 5.5k Hz, Order 9

Cutoff and order are just right. Easy tracking.

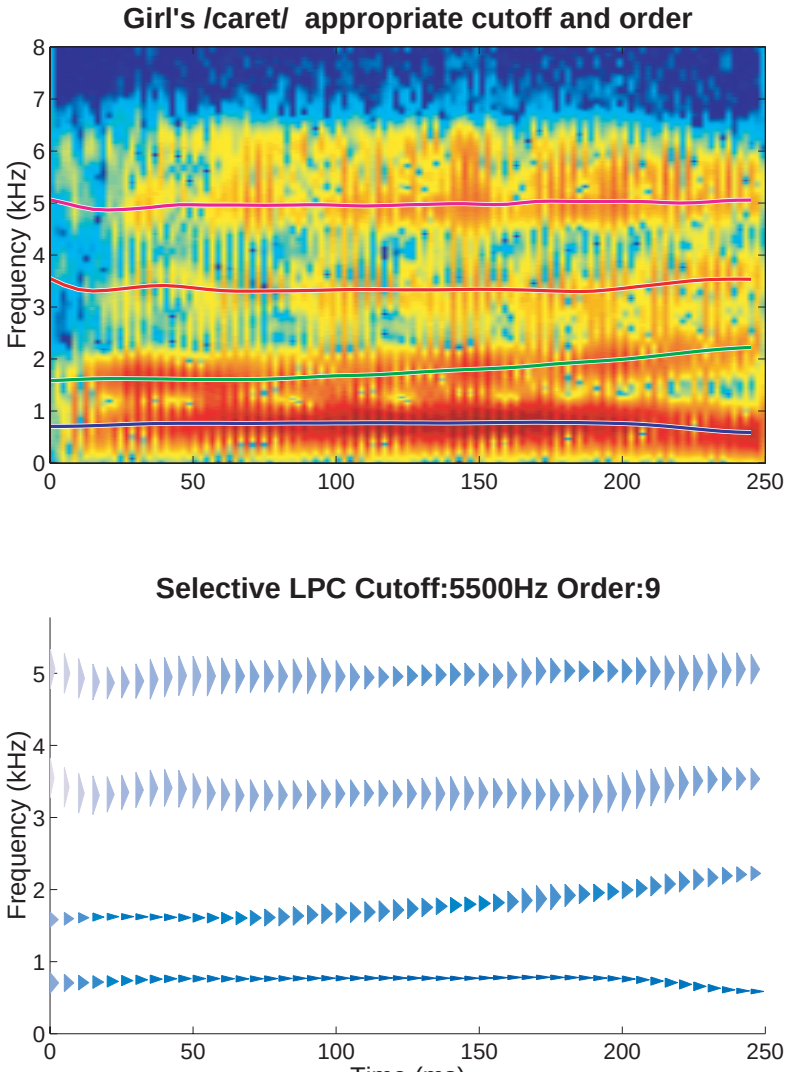
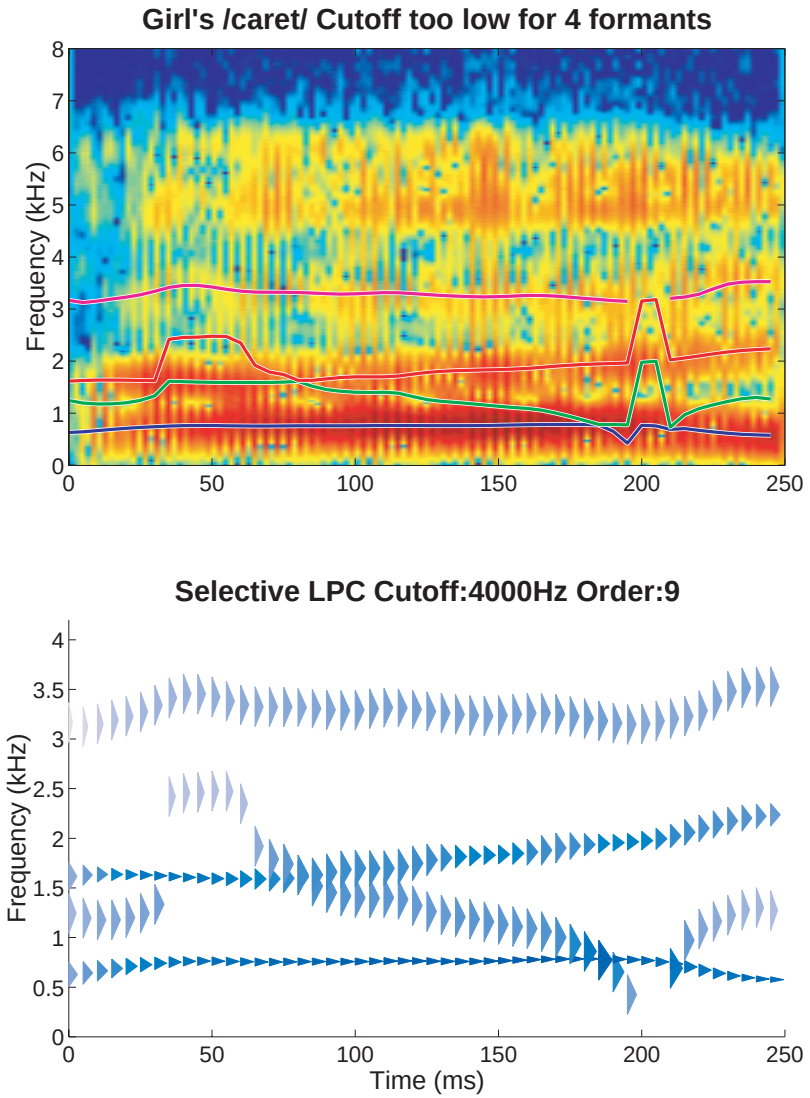


Figure 6. Female child's voice. LPC cutoff 3.5k Hz, Order 9

Cutoff too low for F4. Extra candidate drifts about. Tracking complicated.



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 *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 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
- Assmann and Katz (2000)
  - 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

		rms(hz)			% < 300 Hz		
Speaker	*N tokens*	*F1*	*F2*	*F3*	*F1*	*F2*	*F3*
All	*1295*	*27*	*67*	*120*	*100%*	*98%*	*94%*
Men	*437*	*20*	*54*	*114*	*100%*	*99%*	*95%*
Women	*439*	*25*	*57*	*78*	*100%*	*99%*	*96%*
Boys	*243*	*29*	*81*	*141*	*99%*	*97%*	*93%*
Girls	*176*	*43*	*107*	*210*	*98%*	*94%*	*86%*

Note: errors less than 300 Hz are not likely to involve formant skips

Error report: Assmann Data

		rms (Hz)			% < 300 Hz		
Speaker	*N tokens*	*F1*	*F2*	*F3*	*F1*	*F2*	*F3*
All	*1295*	*63*	*96*	*172*	*99%*	*97%*	*91%*
Men	*437*	*39*	*68*	*143*	*100%*	*98%*	*93%*
Women	*439*	*65*	*79*	*130*	*99%*	*99%*	*94%*
7 yr. old	*243*	*76*	*95*	*133*	*97%*	*97%*	*93%*
5 yr old	*176*	*94*	*129*	*211*	*96%*	*94%*	*87%*
3 yr old	*176*	*115*	*310*	*604*	*95%*	*78%*	*64%*

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

Assmann, P. F., & Katz, W. F. (2000). Time-varying spectral change in the vowels of children and adults. J. Acoust. Soc. Am., 108(4), 1856-1866.

Markel, J. D., & Gray, A. H. (1976). Linear Prediction of Speech. Berlin: Springer-Verlag.

Olive, J. (1971). acoustic formant tracking in a Newton-Raphson technique. J. Acoust. Soc. Am, 50, 661-670.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. J. Acoust. Soc. Am., 97(5, pt. 1), 3099-3111.

Talkin, D. (1987). Speech formant trajectory estimation using dynamic programming with modulated transition costs. J. Acoust. Soc. Am., 82, Suppl. 1, S55.

NOTES ON FIGURES OF MERIT

- Characterizaton 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 *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) ***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

Formant 1	N tokens	rms err. (Hz)	Errors < 300 Hz
All	1295	27	99.6%
Men	437	20	100.0%
Women	439	25	100.0%
Boys	243	29	99.2%
Girls	176	43	98.3%

Note: Errors less than 300 Hz probably do not involve skipped formants

Hillenbrand data F2 results

Formant 2	N tokens	rms err. (Hz)	Errors < 300 Hz
All	1295	67	97.7%
Men	437	54	98.6%
Women	439	57	98.6%
Boys	243	81	97.1%
Girls	176	107	93.8%

Hillenbrand data F3 results

Formant 3	N tokens	rms err. (Hz)	Errors < 300 Hz
All	1295	120	93.9%
Men	437	114	94.7%
Women	439	78	96.4%
Boys	243	141	93.4%
Girls	176	210	86.4%

Assmann & Katz data F1 Results

Talker group	*N Tokens*	*rms err. (Hz)*	*Errors < 300 Hz*
All	*3434*	*63*	*98.5%*
Men	*1232*	*39*	*99.8%*
Women	*1230*	*65*	*98.9%*
7 yr. olds.	*498*	*76*	*97.4%*
5 yr. olds.	*251*	*94*	*96.0%*
3 yr. olds.	*223*	*115*	*94.6%*

Assmann & Katz data F2 Results

Talker group	*N tokens*	*rms err. (Hz)*	*Errors < 300 Hz*
All	*3434*	*96*	*96.5%*
Men	*1232*	*68*	*98.0%*
Women	*1230*	*79*	*98.6%*
7 yr. olds	*498*	*95*	*96.8%*
5 yr. olds	*251*	*129*	*94.4%*
3 yr. olds	*223*	*310*	*78.0%*

Assmann & Katz data: F3

Talker Group	*N tokens*	*rms err. (Hz)*	*Errors < 300 Hz*
All	*3434*	*172*	*90.9%*
Men	*1232*	*143*	*92.5%*
Women	*1230*	*130*	*94.3%*
7 yr. olds.	*498*	*133*	*93.0%*
5 yr. olds.	*251*	*211*	*87.3%*
3 yr. olds.	*223*	*604*	*63.7%*

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.