



Noise-Robust Fundamental Frequency (F0) Estimation for Speech using Summary Correlograms from Multi-band Comb Filters

Lee Ngee Tan and Abeer Alwan

Speech Processing and Auditory Perception Laboratory

This work is supported in part by DARPA.

Published in *Proc. ICASSP 2011*.

Motivation

- F0 / Pitch information in speech is used in many applications
 - Gender or speaker identification
 - Multi-talker speech separation
 - Query by humming
 - Speech coding

Introduction (1)

F0-containing speech units

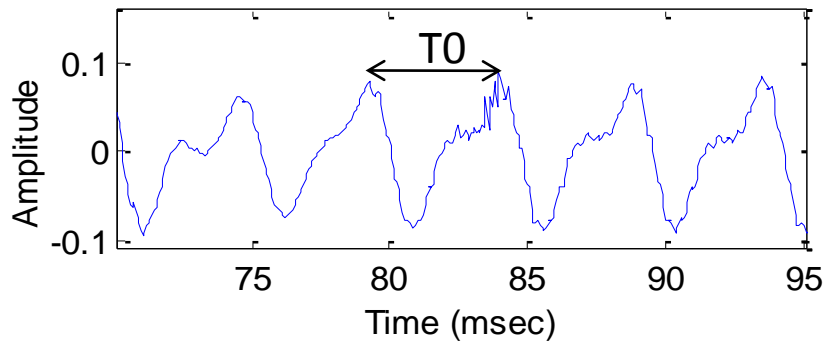
- Unvoiced and voiced phonetic units
 - Non-periodic, unvoiced phonemes
 - E.g. h, p, s, t, etc.
 - Quasi-periodic, voiced phonemes
 - E.g. a, i, e, b, m, z, etc.

Introduction (2)

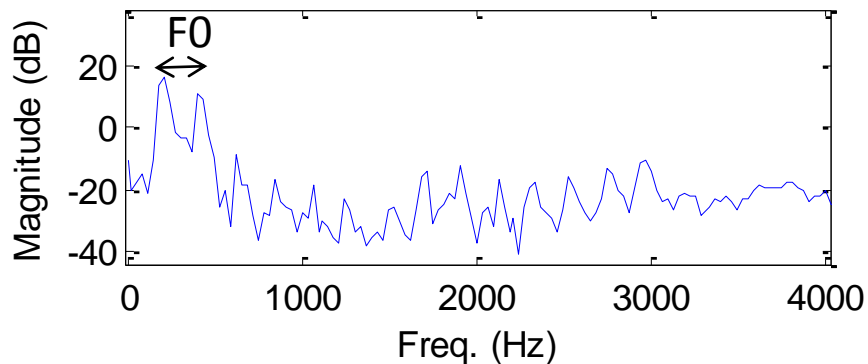
Time and Frequency Domain Representations

Voiced

Time waveform of a voiced phoneme

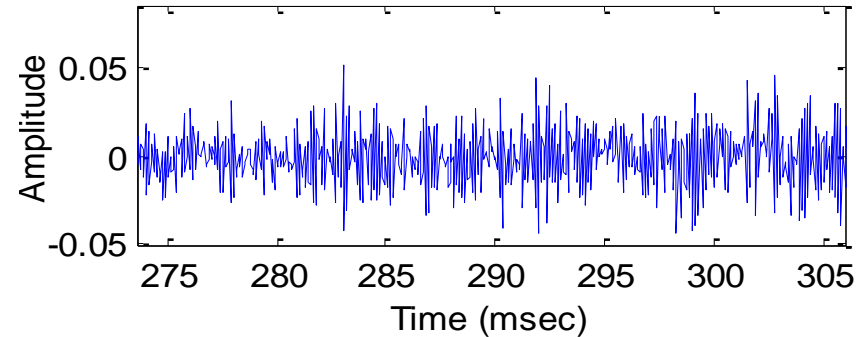


Magnitude spectrum of a voiced phoneme

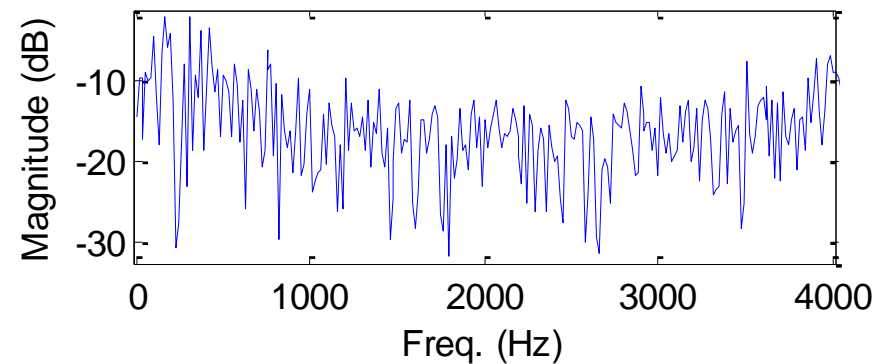


Unvoiced

Time waveform of an unvoiced phoneme

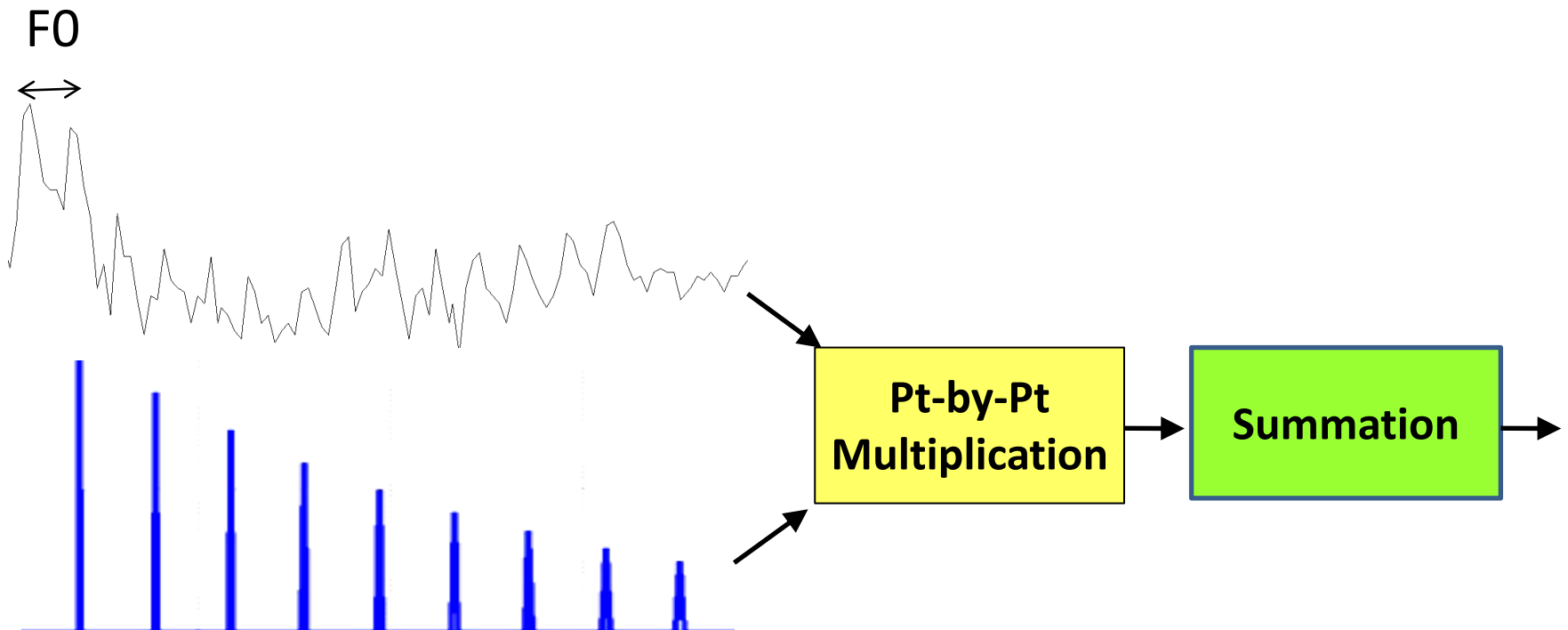


Magnitude spectrum of an unvoiced phoneme



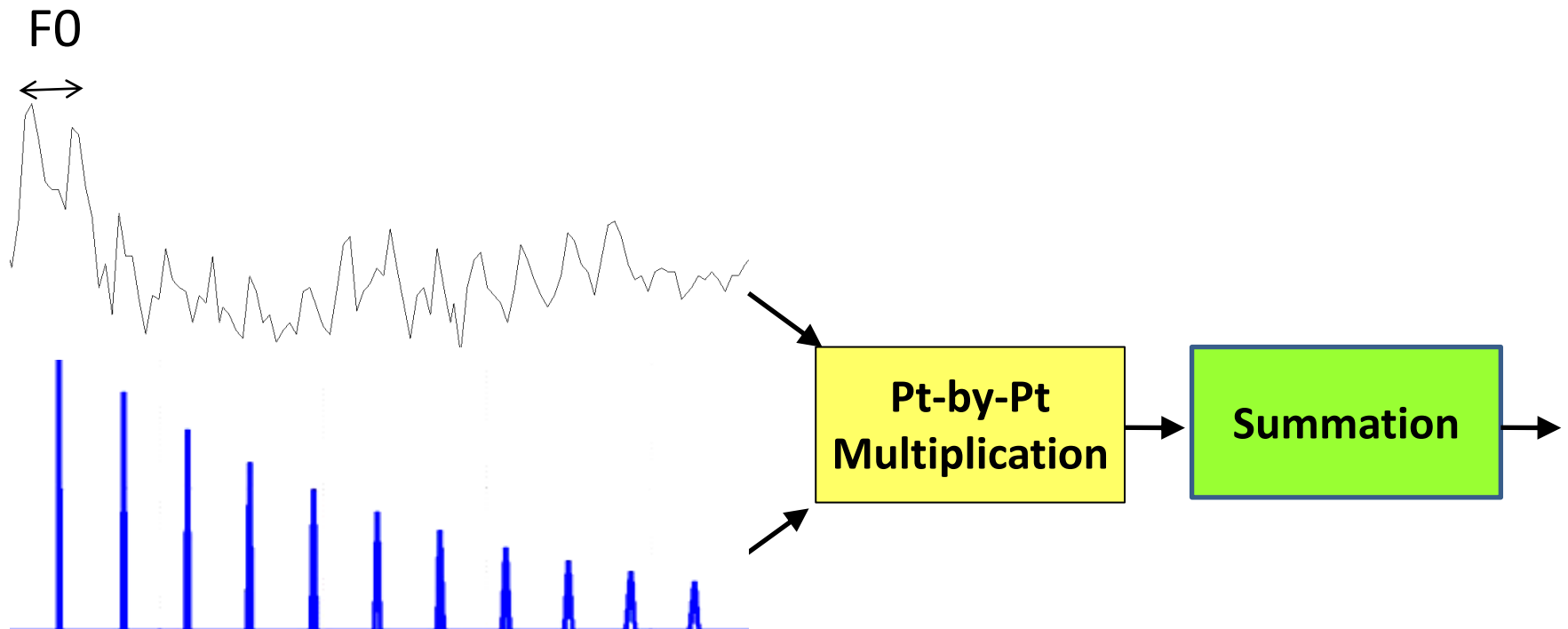
Overview of Freq Domain Comb-Filter-based F0 Est. Techniques (1)

- Subharmonic Summation (SHS) -- Hermes, 1988 [4]



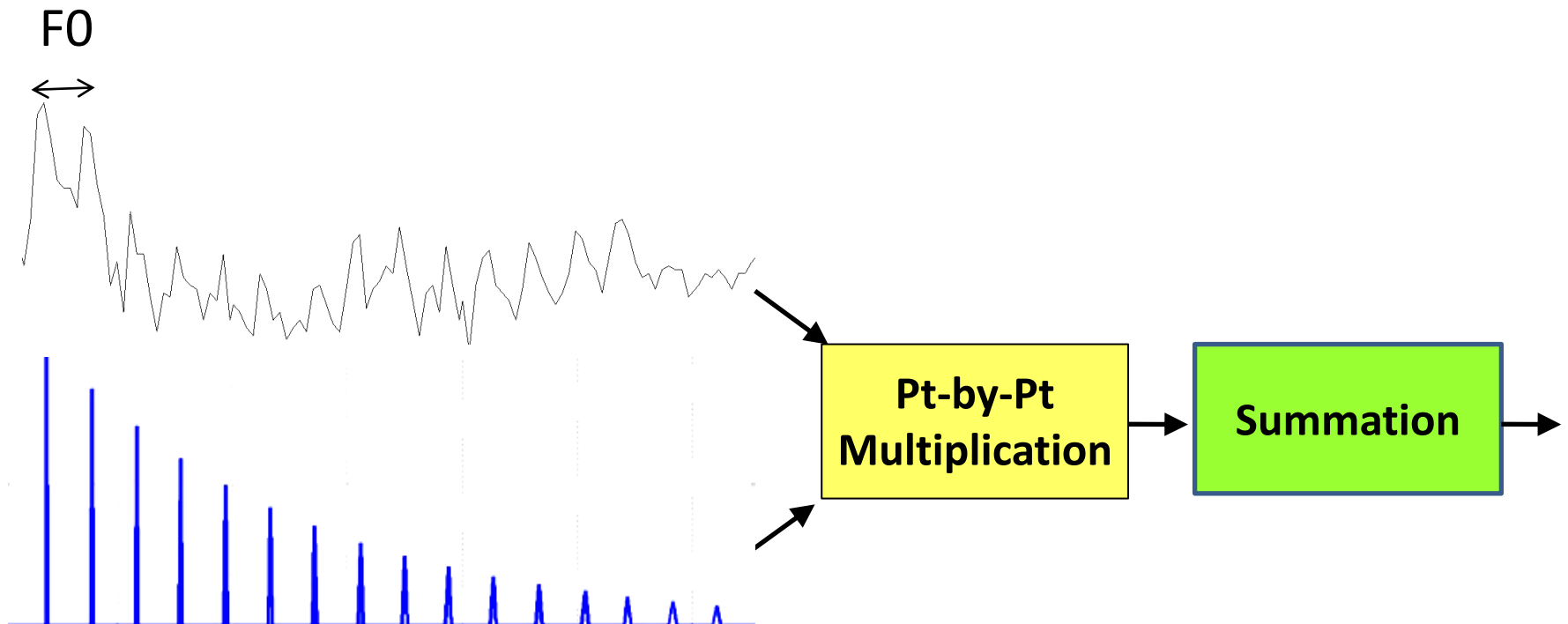
Overview of Freq Domain Comb-Filter-based F0 Est. Techniques (1)

- Subharmonic Summation (SHS) -- Hermes, 1988 [4]



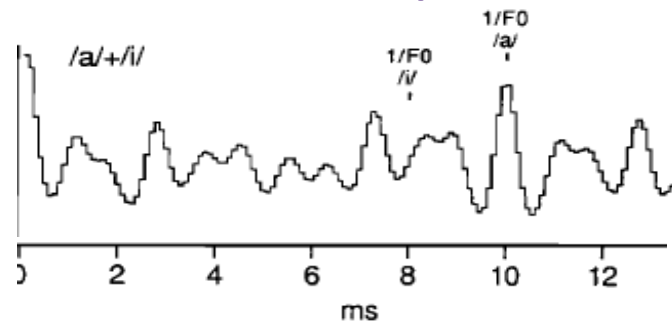
Overview of Freq Domain Comb-Filter-based F0 Est. Techniques (1)

- Subharmonic Summation (SHS) -- Hermes, 1988 [4]



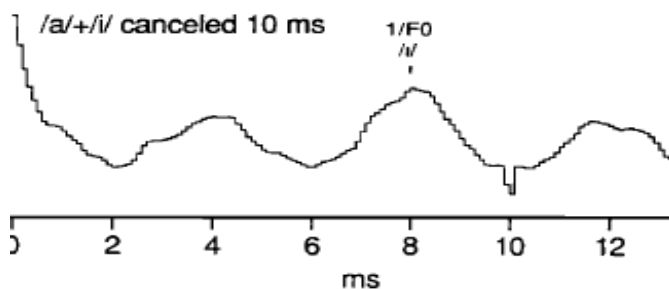
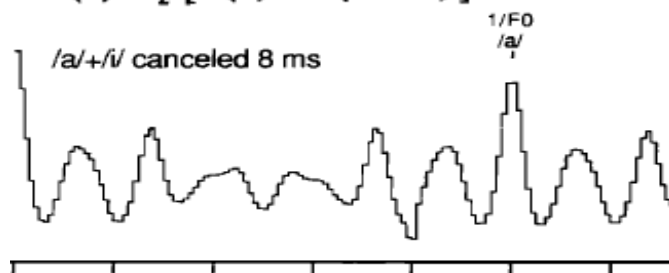
Overview of Freq Domain Comb-Filter-based F0 Est. Techniques (2)

- Concurrent harmonic source separation – Cheveigne, 1993 [5]



Harmonic Cancellation

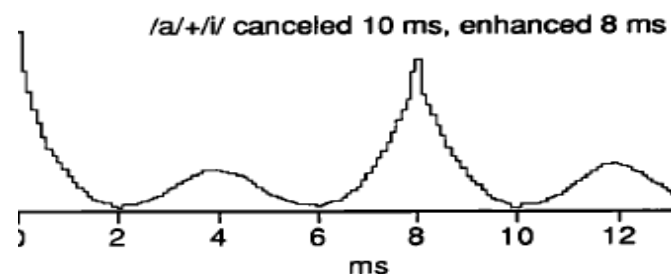
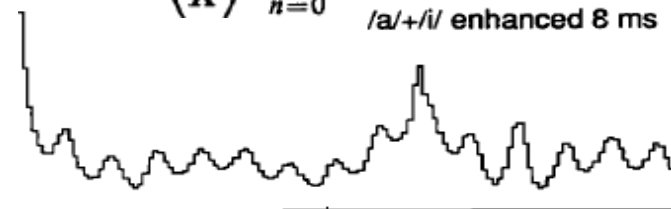
$$h(t) = \frac{1}{2} [\delta(t) - \delta(t - T)]$$



Harmonic Enhancement

$$h(t) = \left(\frac{1}{K}\right) \sum_{n=0}^{K-1} \delta(t - nT)$$

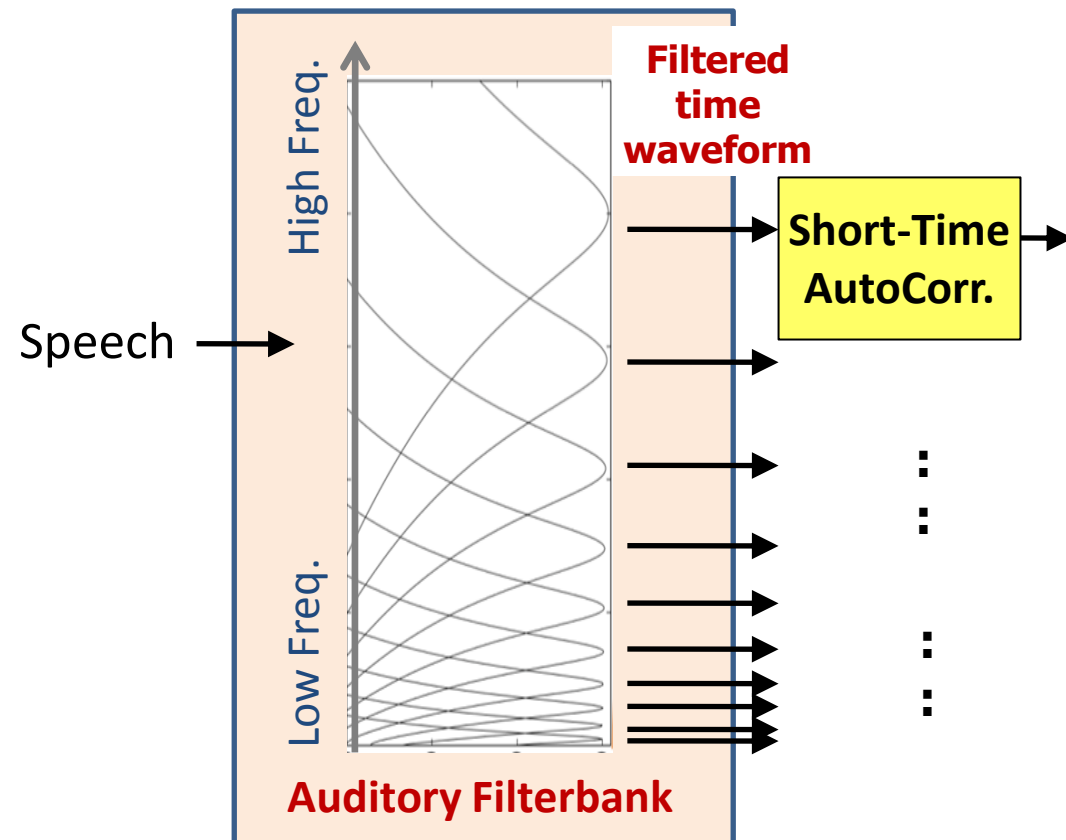
/a+/i/ enhanced 8 ms



Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (1)

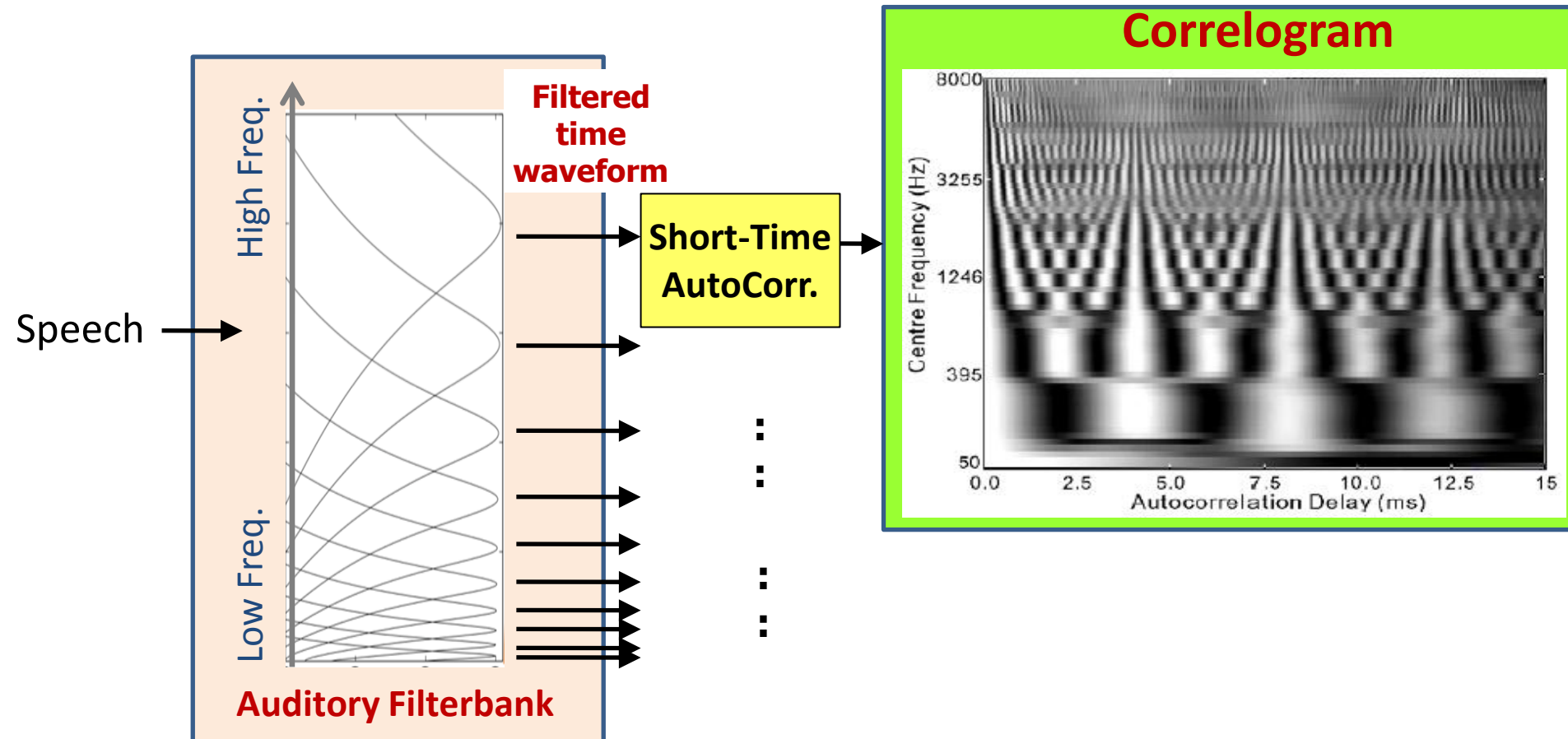
Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (1)

- Licklider, 1951 [1]



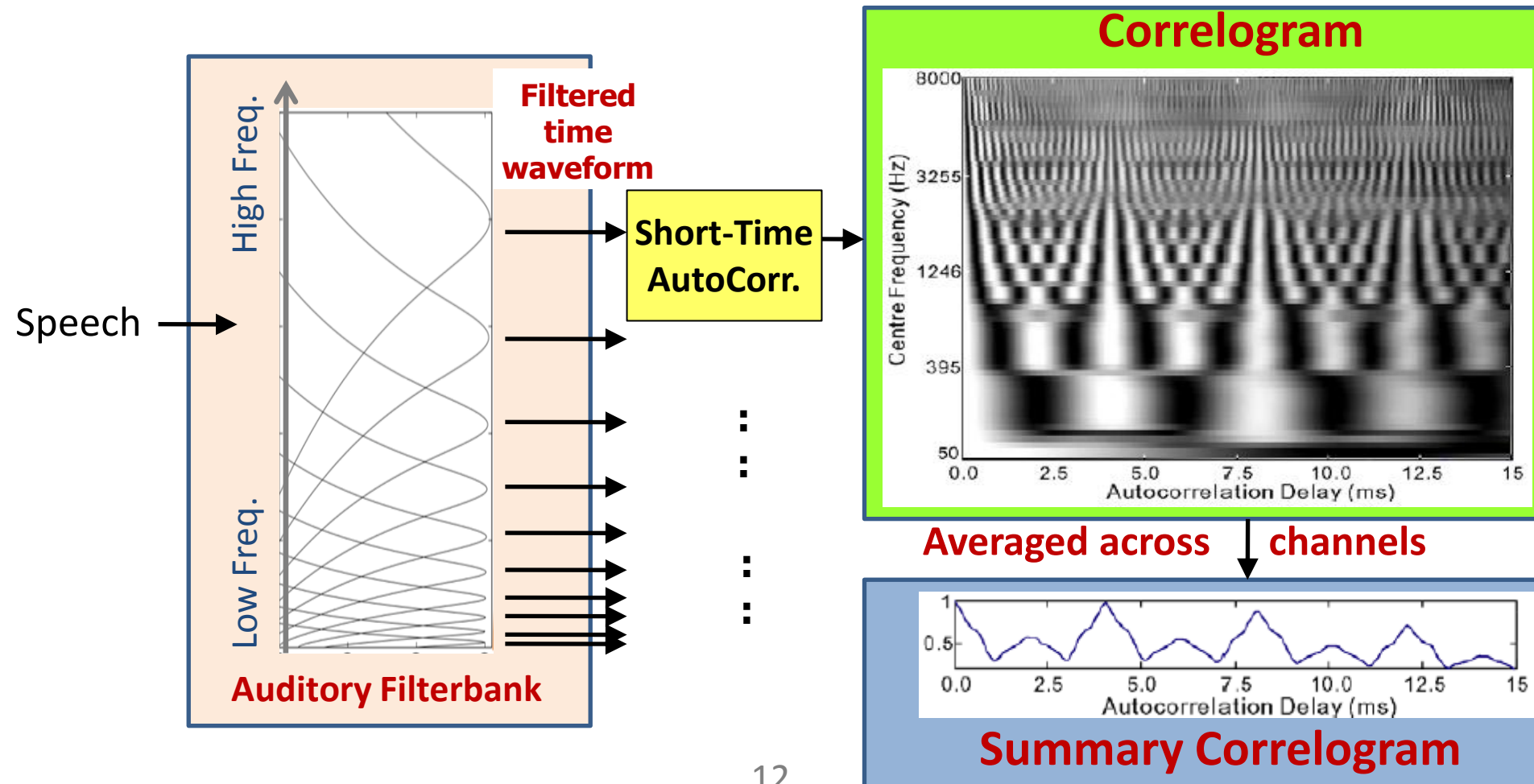
Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (1)

- Licklider, 1951 [1]



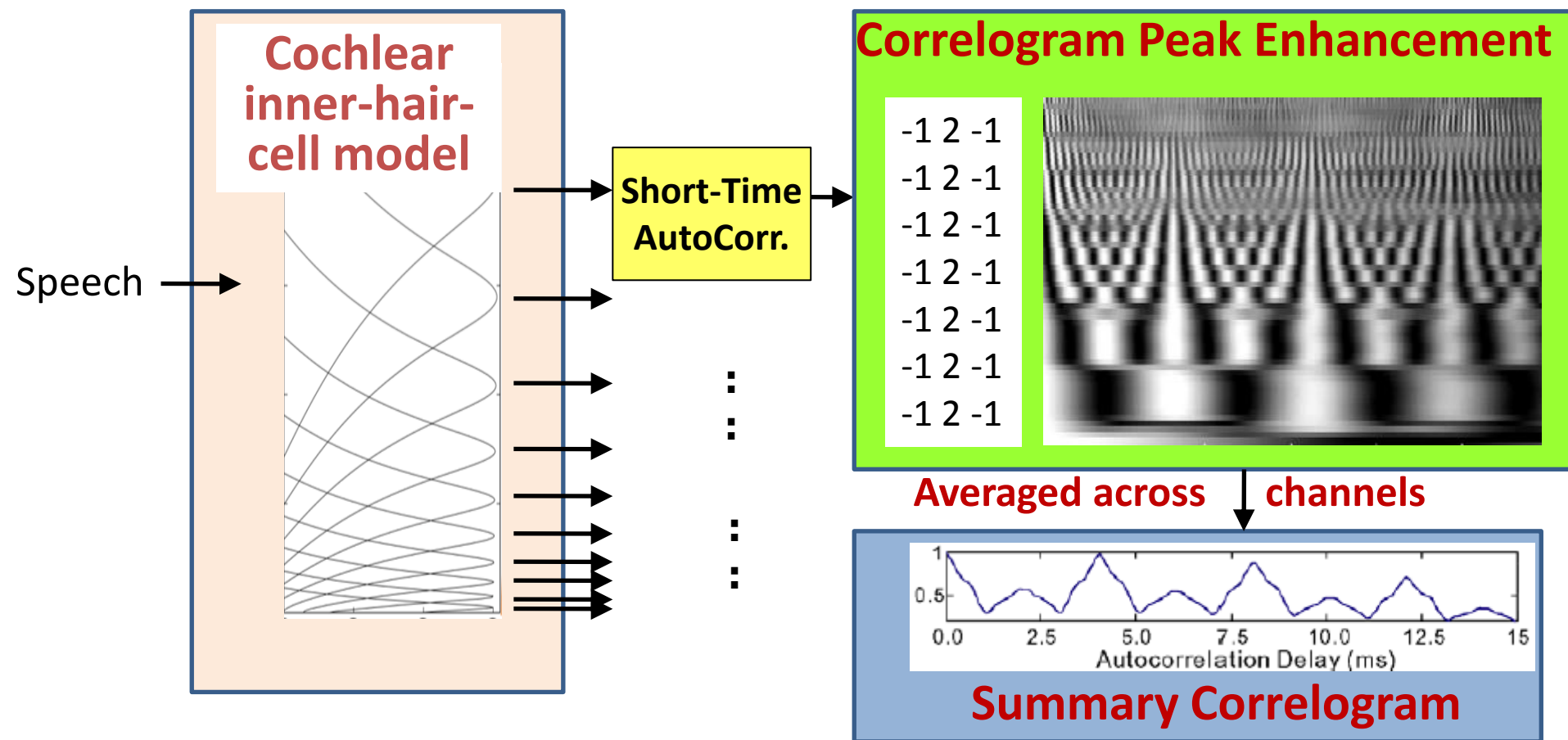
Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (1)

- Licklider, 1951 [1]



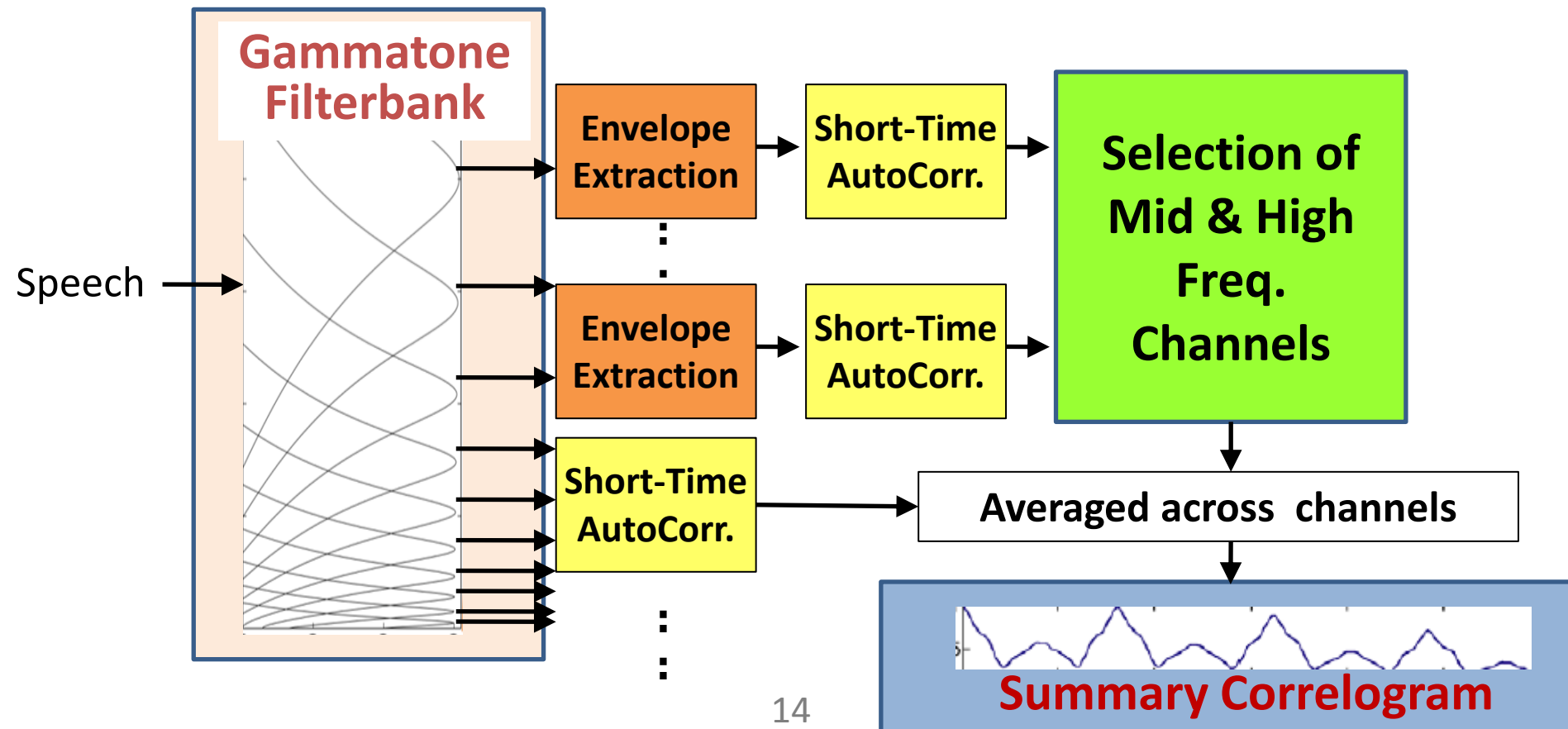
Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (2)

- Slaney and Lyon, 1990 [2]



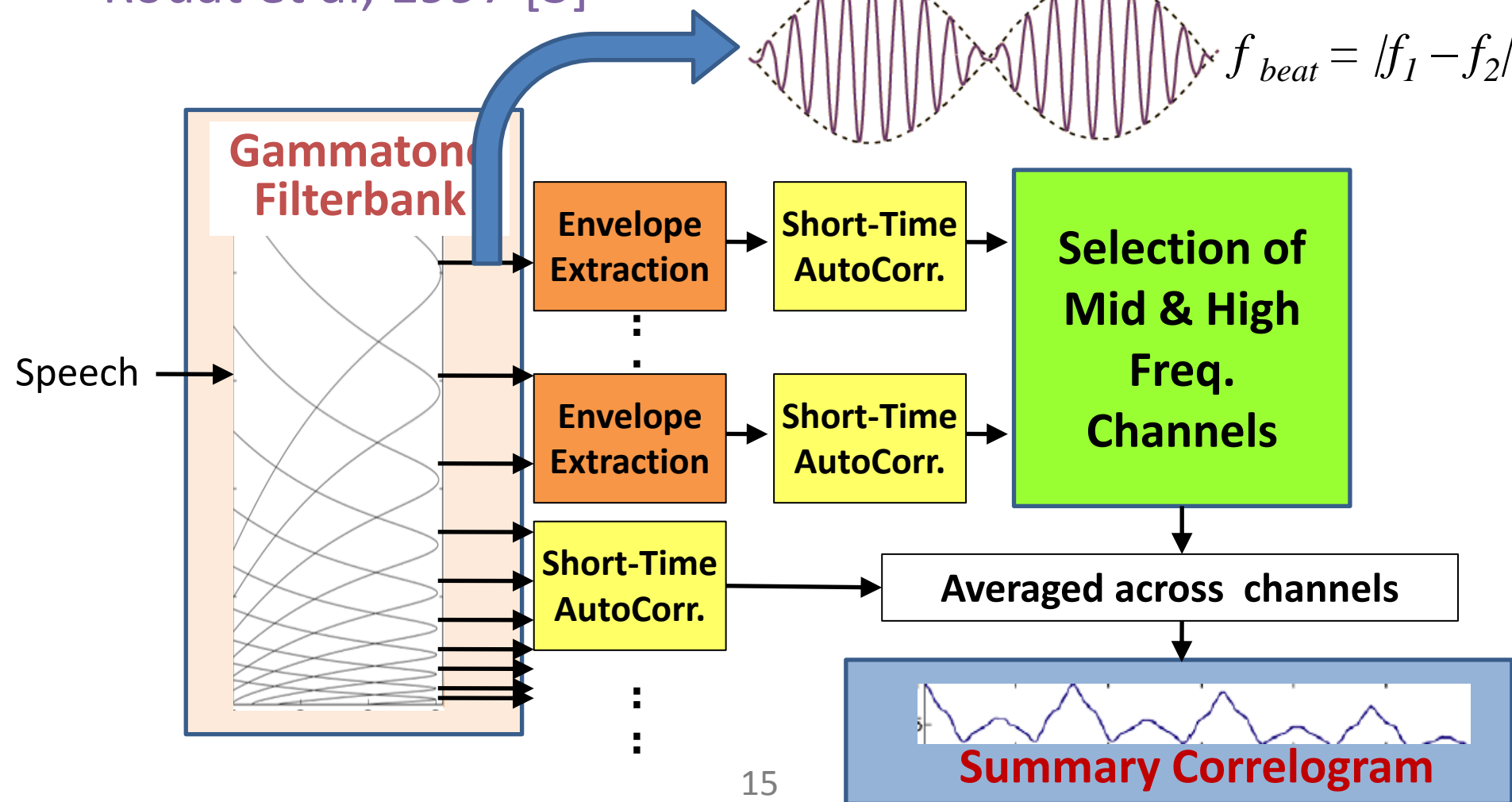
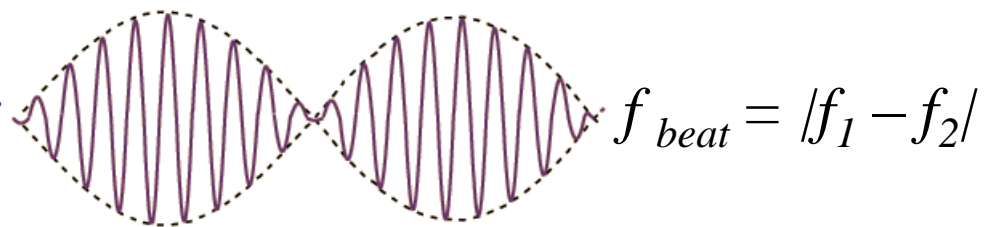
Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (3)

- Rouat et al, 1997 [3]



Overview of Time-Freq Domain Correlogram-based F0 Est. Techniques (3)

- Rouat et al, 1997 [3]

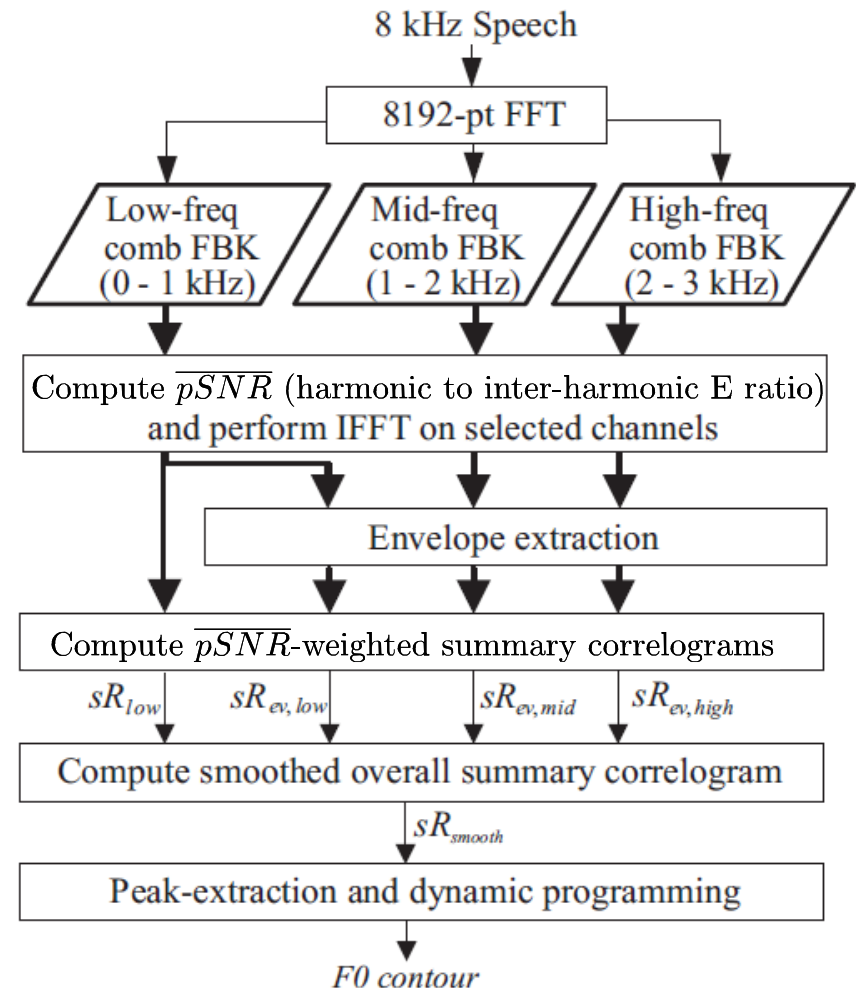


Proposed F0 Estimator

- Motivated by:
 - **Information richness** of the **correlogram** representation
 - **harmonic enhancement** and **suppression** capabilities of **comb filters**
- Proposed:
 - **Correlogram**-based F0 estimation algorithm using **multi-band comb filters**
 - Time-frequency domain technique
 - Deterministic (No statistical models trained)

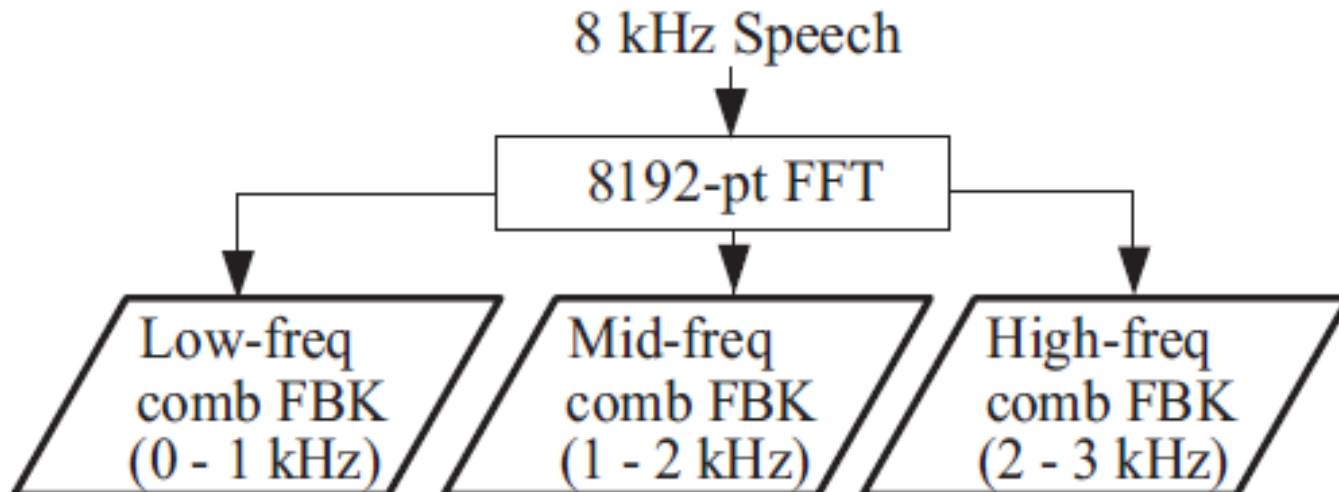
Novelties of Proposed Algorithm

- Multi-band comb-filterbanks (FBKs) for freq. decomposition
- Pseudo-SNR ($pSNR$)-based channel selection
- Inclusion of low-freq. envelope information
- $pSNR$ -weighting for computing each subband's summary correlogram



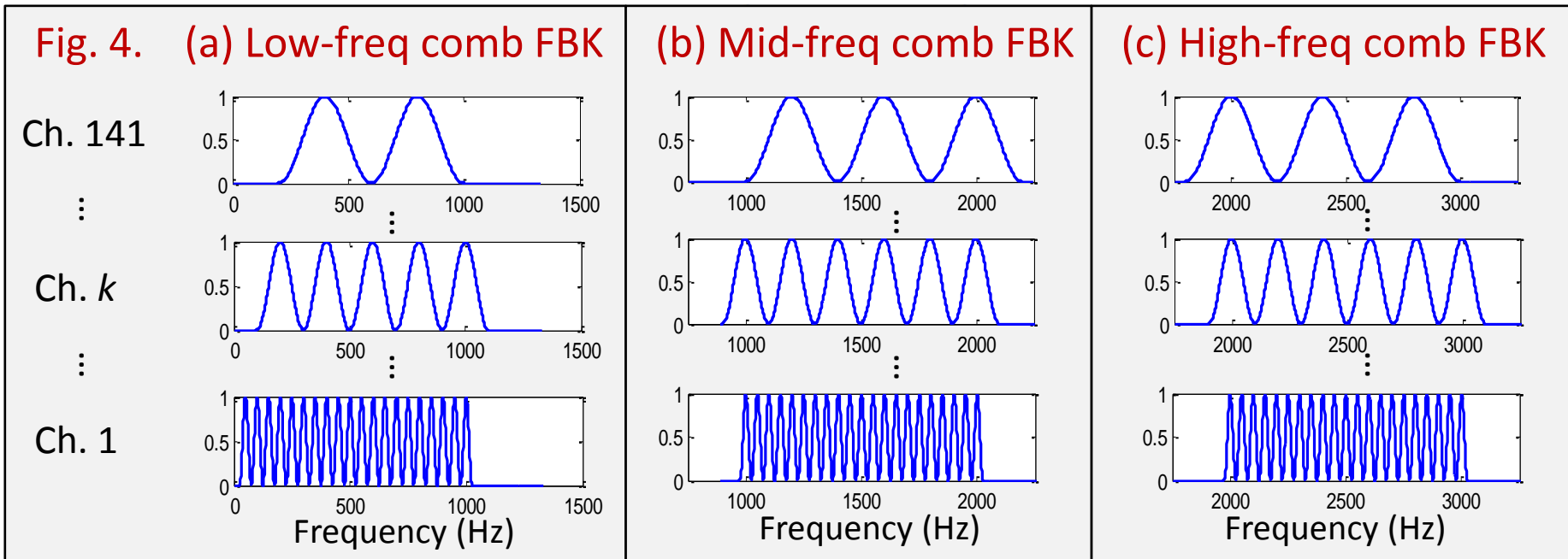
Multi-band Comb Filtering Stage (1)

- Comb filters in each FBK



Multi-band Comb Filtering Stage (1)

- Comb filters in each FBK



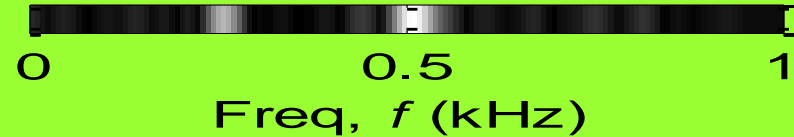
$$comb_k(f) = 0.5 + 0.5 \cos(2\pi f \tau_k)$$

$$\text{where } \tau_k = \{\tau_{max} = 160, \tau_{max} - 1, \dots, \tau_{min} = 20\}$$

$$\tau_{max} = f_s / (\min F0 = 50), \tau_{min} = f_s / (\max F0 = 400), f_s = 8000$$

Multi-band Comb Filtering Stage (2)

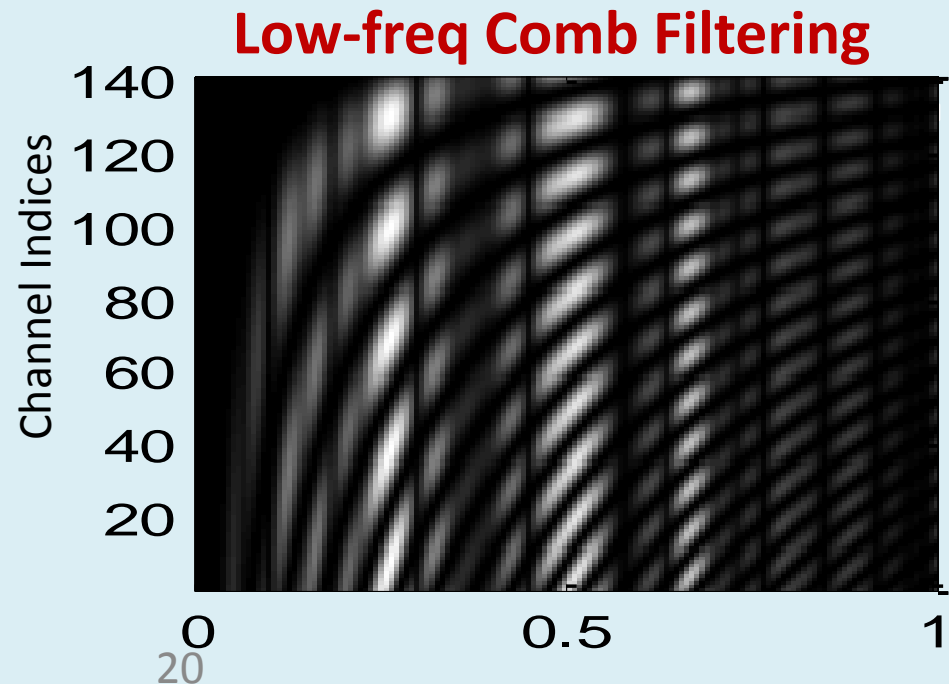
Magnitude spectrum of clean speech



Magnitude spectrum of 0dB white noise-corrupted speech **before** comb filtering

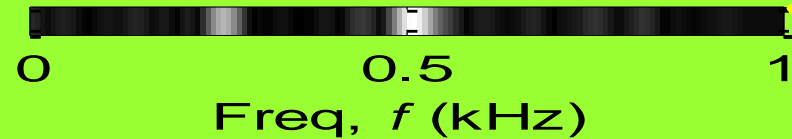


Multi-channel magnitude spectra of noisy speech **after** comb filtering



Multi-band Comb Filtering Stage (2)

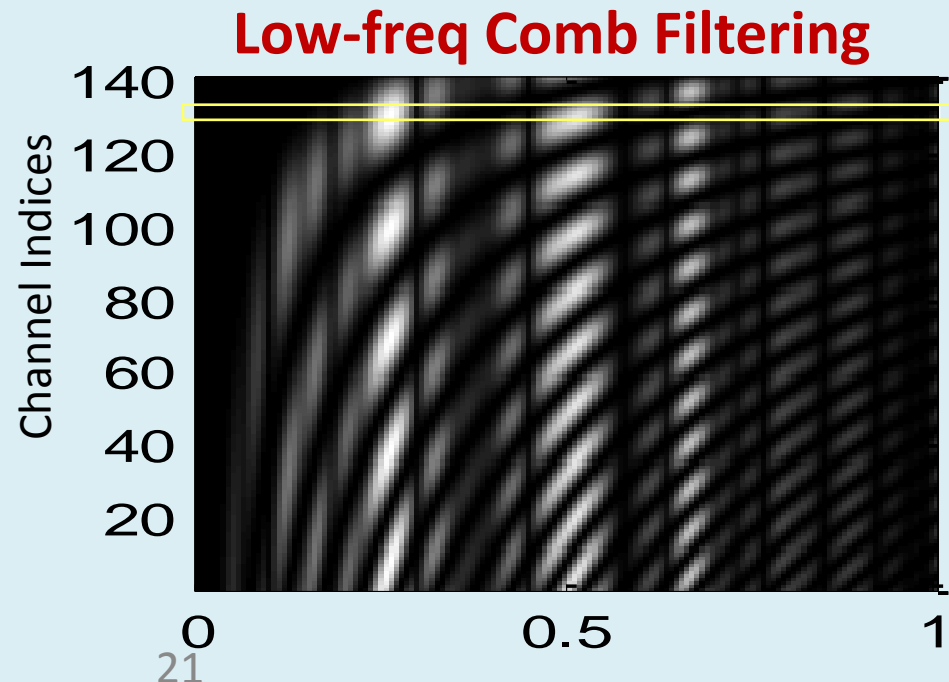
Magnitude spectrum of clean speech



Magnitude spectrum of 0dB white noise-corrupted speech **before** comb filtering



Multi-channel magnitude spectra of noisy speech **after** comb filtering



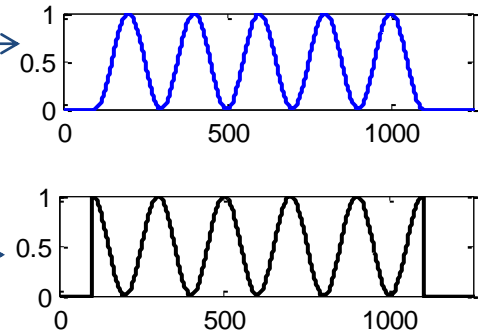
$pSNR$ -based Channel Selection

- Calculating mean pseudo-SNR of k th channel, $\overline{pSNR}(k)$

1. $pSNR$ of k th channel in each FBK:

e.g.

$$pSNR_{low}(k) = \frac{\sum_f |X(f) \overbrace{comb_{low,k}(f)}|^2}{\sum_f |X(f) \underbrace{noise_comb_{low,k}(f)}|^2}$$



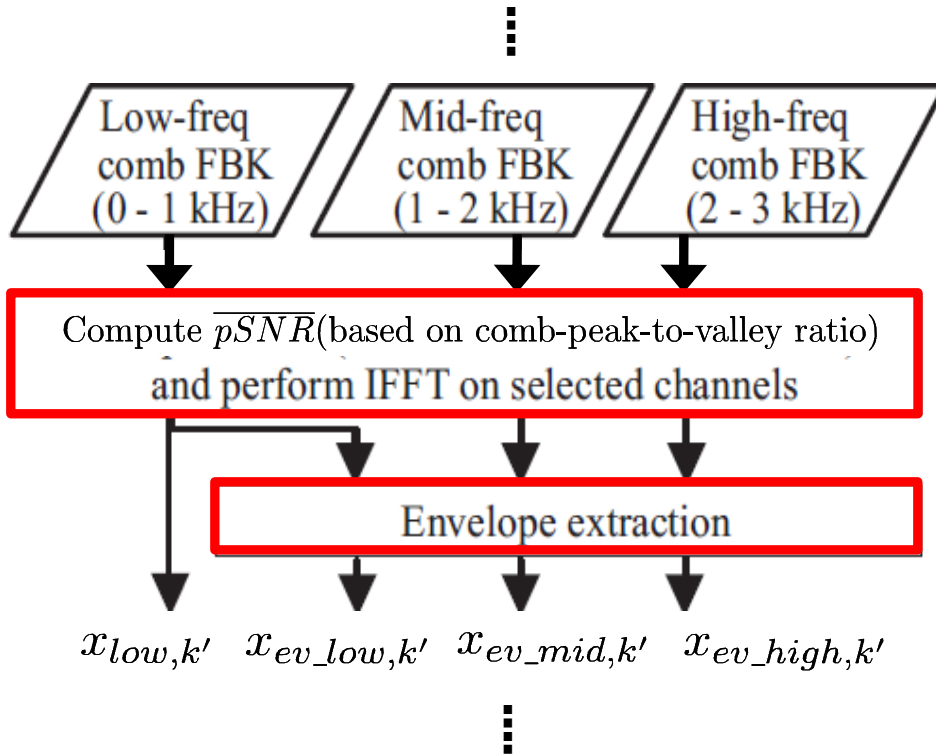
2. Average $pSNR$ of k th channel over the 3 FBKs:

$$\overline{pSNR}(k) = [pSNR_{low}(k) + pSNR_{mid}(k) + pSNR_{high}(k)] / 3$$

- Channel Selection

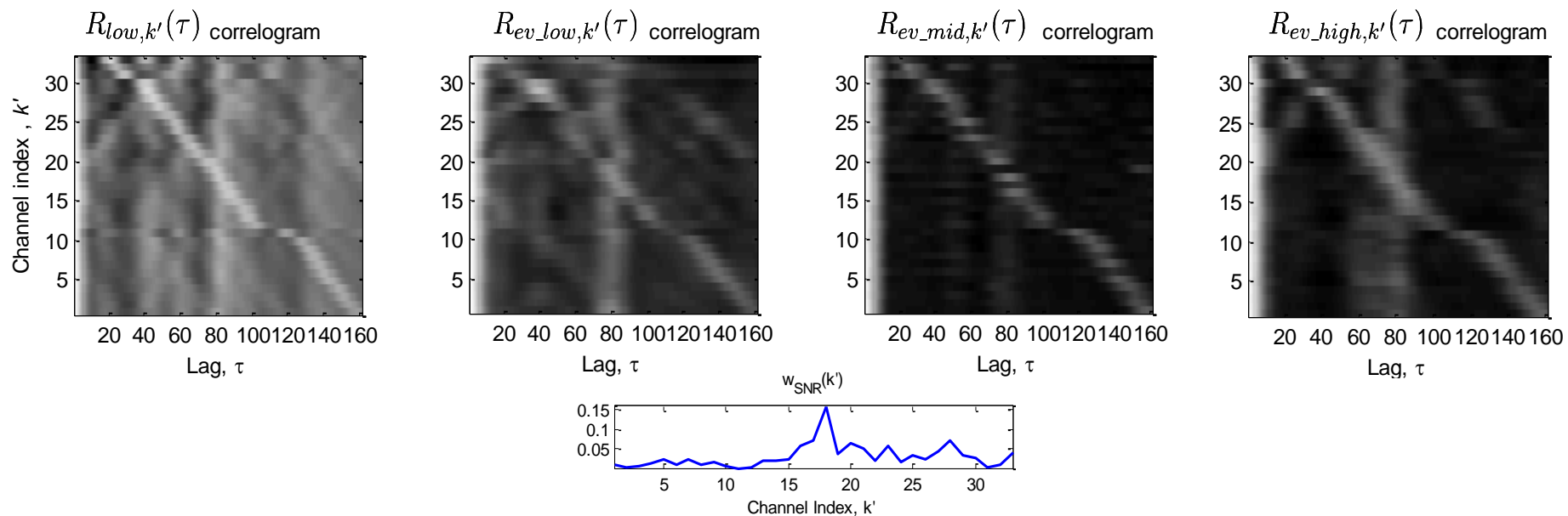
– Peaks in $\overline{pSNR}(k)$ with magnitude > 1 **➡ Selected Channels**

Envelope Extraction



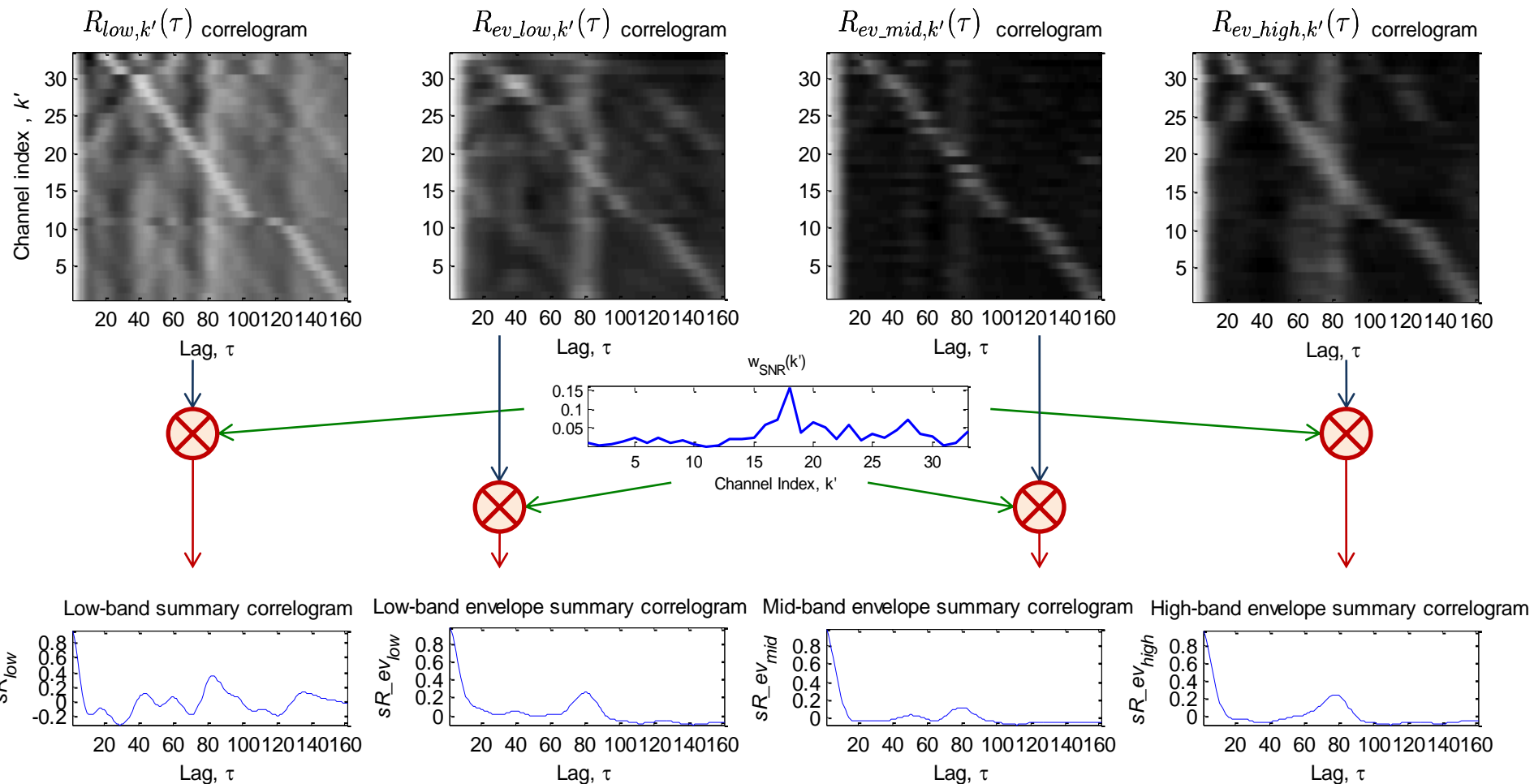
- IFFT on selected channels
- Hilbert envelopes are extracted :
($|analytic\ signal(t)|^2$)

$pSNR$ -weighted Summary Correlograms

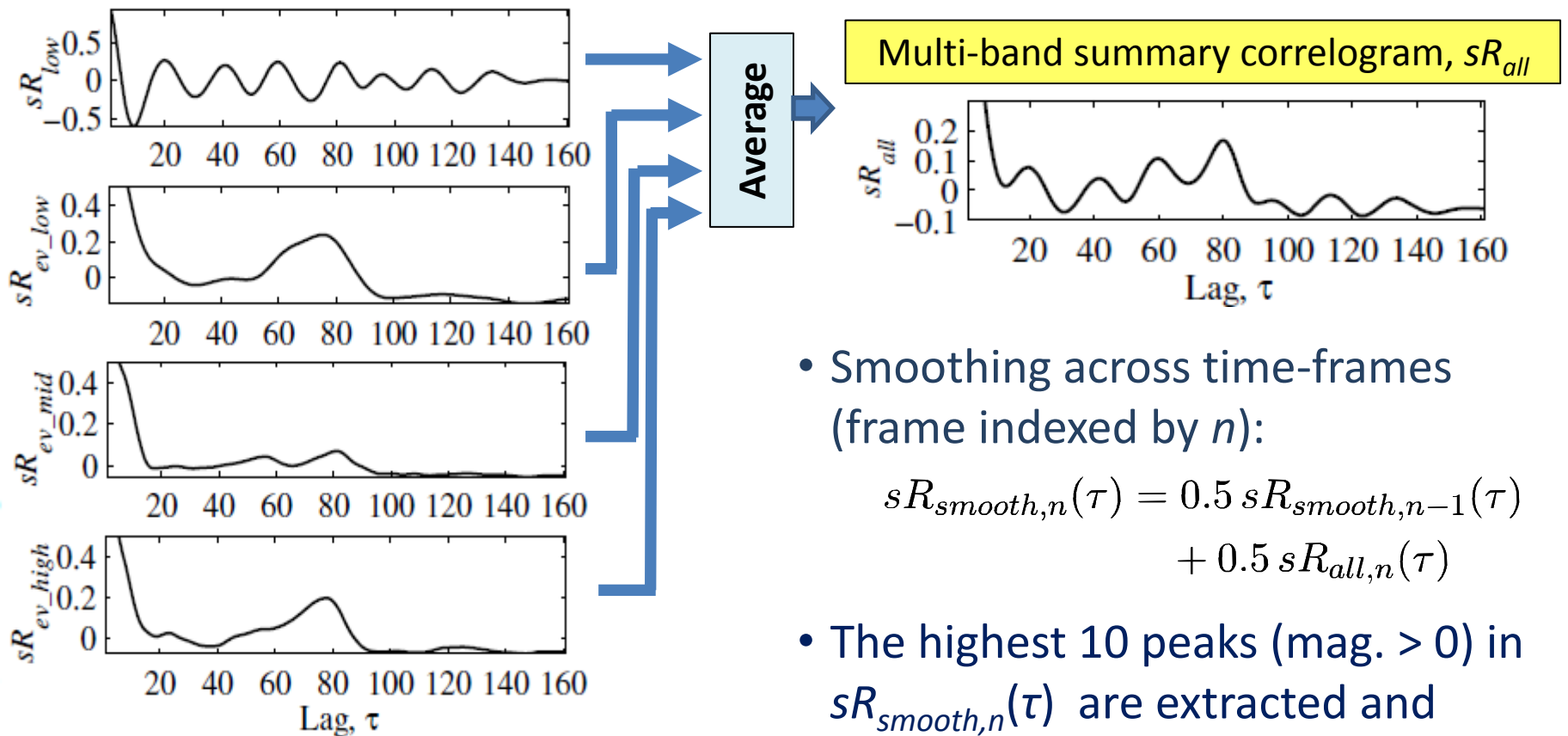


$$w_{SNR}(k') = \frac{\overline{pSNR}(k') - 1}{\sum_{k'} [\overline{pSNR}(k') - 1]}$$

p SNR-weighted Summary Correlograms



Time-smoothed Multi-band Summary Correlogram

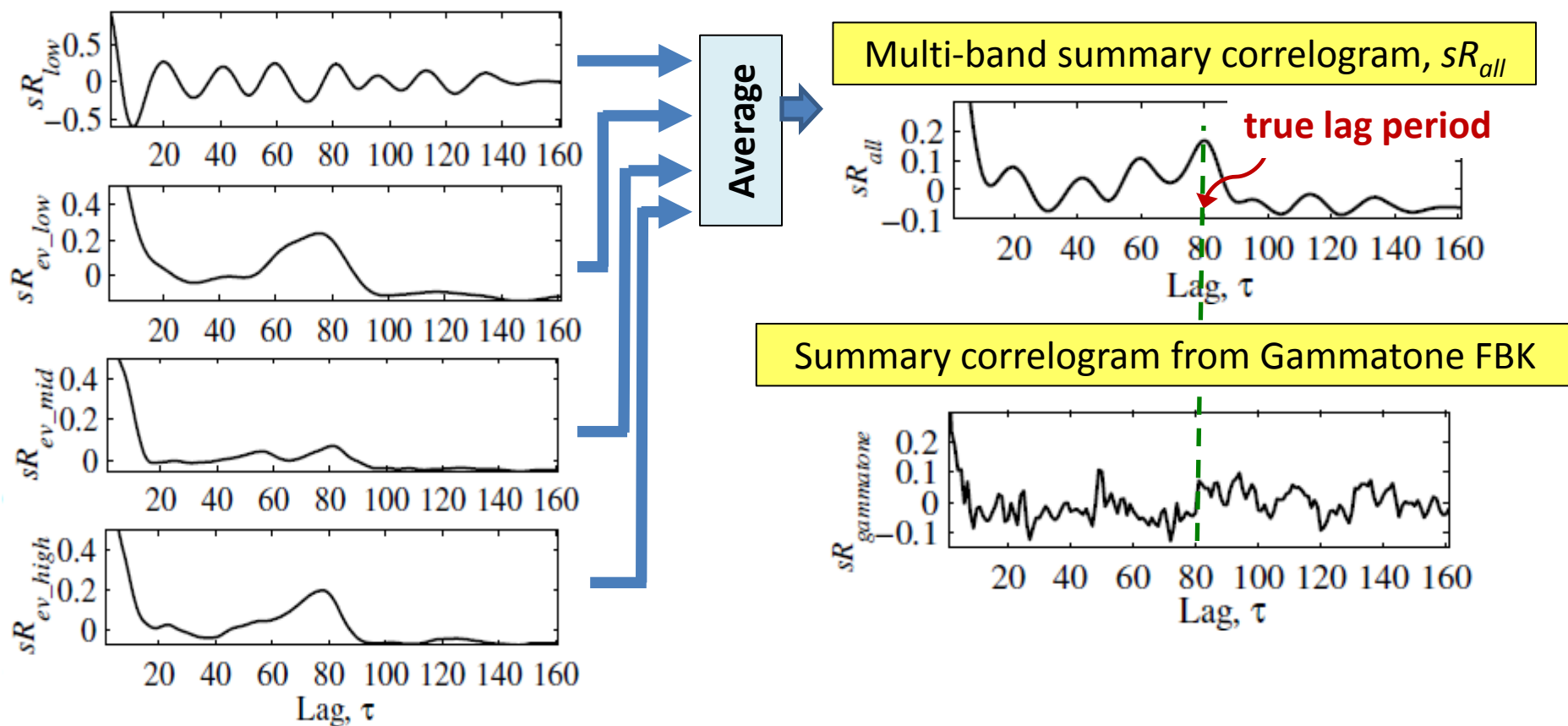


- Smoothing across time-frames (frame indexed by n):

$$sR_{smooth,n}(\tau) = 0.5 sR_{smooth,n-1}(\tau) + 0.5 sR_{all,n}(\tau)$$

- The highest 10 peaks (mag. > 0) in $sR_{smooth,n}(\tau)$ are extracted and passed into a dynamic prog. algo. to obtain final F0 contour

Comparison of our Multi-band Summary Correlogram to that from Gammatone FBK



Summary correlograms of a voiced telephone speech frame corrupted with babble noise at 5 dB SNR

Performance Evaluation :

Generating Noisy Speech Database

- Keele : clean reference pitch database [6]
 - Phonetically balanced story read by 5 male and 5 female adults (approx. 6 min), containing simultaneous recording of speech and laryngograph signals in quiet, with F0 ground-truth provided.
 - Speech signals are down-sampled to 8 kHz, then white and babble noise are artificially added at various SNRs to simulate noisy speech.
 - We also generated a noisy 8 kHz telephone speech version, which pre-filtered speech and noise with ITU G.712 characteristic filter

Performance Evaluation: Error Measure and Algorithms for Comparison

- Gross Pitch Error (GPE)

$$GPE = \frac{N_{F0err}}{N_V} \times 100\%$$

where N_V is the number of voiced frames in the ground truth, while N_{F0err} is the number of these frames with

$$\left| \frac{F0_{n,estimate}}{F0_{n,reference}} - 1 \right| > 0.2$$

- F0 estimation algorithms used for comparison
 - RAPT used in *Wavesurfer* [7], YIN [8], and Rouat's time-freq domain F0 estimator [3].

GPE Results

Averaged GPE (%) for clean and noise-corrupted Keele database

8 kHz Keele	Clean	White		Babble		
SNR (dB)	-	10	0	10	5	0
RAPT	2.58	3.56	12.06	6.90	12.81	26.28
YIN	2.68	3.45	11.79	8.69	18.58	36.74
Rouat	2.56	6.46	20.38	9.64	18.19	32.83
Proposed	2.64	2.73	6.28	4.81	9.42	21.87
G.712-filtered	Clean	White		Babble		
	-	10	0	10	5	0
RAPT	6.92	6.33	14.27	10.97	17.1	28.65
YIN	6.95	9.15	21.75	16.23	27.55	44.49
Rouat	6.25	14.46	32.03	14.46	21.53	32.03
Proposed	5.88	4.97	9.22	8.26	12.75	24.97

Conclusion and Future Work

- The proposed F0 estimation algorithm is effective in enhancing the accuracy of F0 estimation in the presence of noise
 - Usage of low-freq band envelope, multi-band comb FBKs, $pSNR$ -based channel selection and weighting function, and time-smoothing contributed to noise-robustness.
- Extend the algorithm to perform F0 tracking
 - Do both F0 estimation and voicing detection

References

- [1] J. C. R. Licklider, "A duplex theory of pitch perception," *Experientia*, 1951.
- [2] M. Slaney, and R. Lyon, "A perceptual pitch detector," *In Proc. ICASSP*, 1990.
- [3] J. Rouat, Y. Liu, and D. Morissette, "A pitch determination and voiced/unvoiced decision algorithm for noisy speech," *Speech Comm.*, 1997.
- [4] D. J. Hermes, "Measurement of pitch by subharmonic summation," *JASA*, 1988.
- [5] A. Cheveigne, "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," *JASA*, 1993.
- [6] F. Plante et al, "A pitch extraction reference database," in *EuroSpeech*, 1995.
- [7] D. Talkin, "Robust algorithm for pitch tracking," *Speech Coding & Synthesis*, 1995.
- [8] A. Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, 2002.