SAFE: a Statistical Algorithm for F0 Estimation for Both Clean and Noisy Speech

Wei Chu and Abeer Alwan

Speech Processing and Auditory Perception Laboratory
Department of Electrical Engineering
University of California, Los Angeles

Supported in part by the NSF

Outline

- Introduction
- System Description
- Experiments
- Conclusion

F0 Estimation

Assumption

- F0 values remain constant within a voiced frame (20-40 ms)

Objective

Accurately estimate F0 using clean or noisy voiced frames.

Current F0 estimators

Two-stage estimation:

1. Generate F0 candidates for each frame:

- Single-band: usually the low frequency band (0-1000 Hz)
- Multi-band: deterministic approaches
  - Hard decision: only information from the most reliable band
  - Soft decision: combine information from different bands

The estimation methods in each band:

- NCCF in Get_F0 [1]; NCCF/NACF in Praat [2]; instantaneous frequency in TEMPO [3]; AMDF in YIN [4]

2. Generate optimal F0 contour over frames

Dynamic programming based on F0 candidate likelihoods

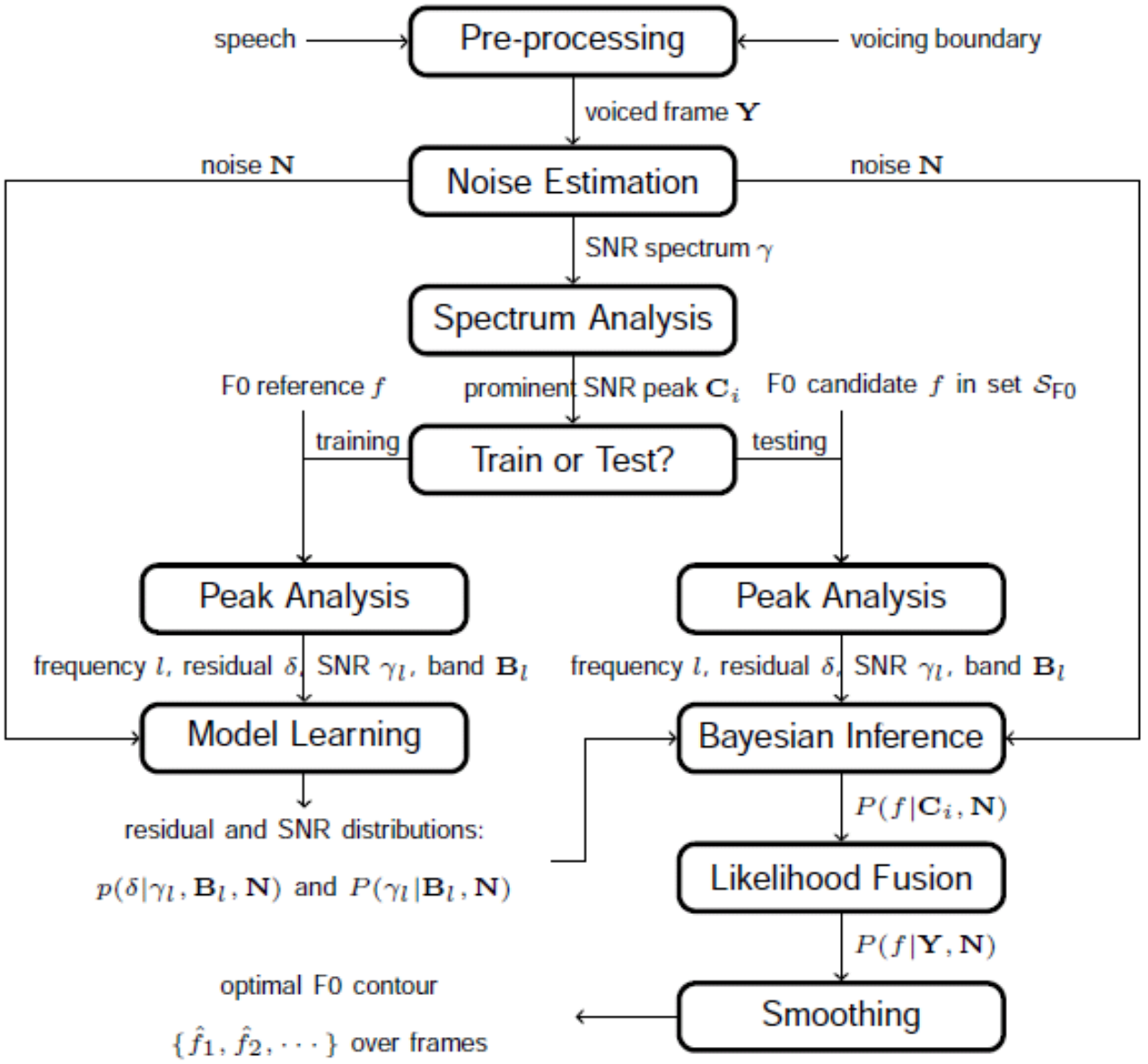
These methods have been evaluated mostly using clean speech.

What's New in SAFE

A statistically-based soft-decision multi-band method.

- Extract information from voiced speech;
- Estimate F0 value in a statistical approach:
  - Evaluate the likelihood of a frequency to be close to F0 reference;
    - Training: the frequency is the F0 reference;
    - Testing: the frequecy is one of the F0 candidates
  - Estimate a model from the extracted information and F0 reference (training);
  - Decode the F0 according to the extracted information and learned model (testing).
- Obtain robustness by considering noise effect on the extracted information and learned model.

System Flowchart



Maximum Likelihood-Based F0 Decoding

Given a voiced frame's power spectrum **Y** corrupted by the noise power spectrum **N**, the maximum likelihood estimation of F0 is:

$$\hat{f} = \arg\max_{f \in \mathcal{S}_{\mathrm{F0}}} P(f|\mathbf{Y}, \mathbf{N}) \tag{1}$$

where $\mathcal{S}_{\mathrm{F0}} = \{f_{min}, f_{min} + \Delta, \cdots, f_{max}\}$.

Note

- **N** is estimated by using initial and final frames in the utterance.
- F0 candidates with likelihoods are generated in each frame for subsequent dynamic programming.

Extract information for F0 estimation

Obtain Prominent SNR Peaks

SNR Spectrum

$\gamma_l = 10 \log_{10} \mathbf{Y}_l / \mathbf{N}_l$, where l denotes the frequency.

Two smoothed SNR spectra

- Short-term smoothed SNR spectrum: γ_l^{S}
- Long-term smoothed SNR spectrum: γ_l^{L}

SNR difference

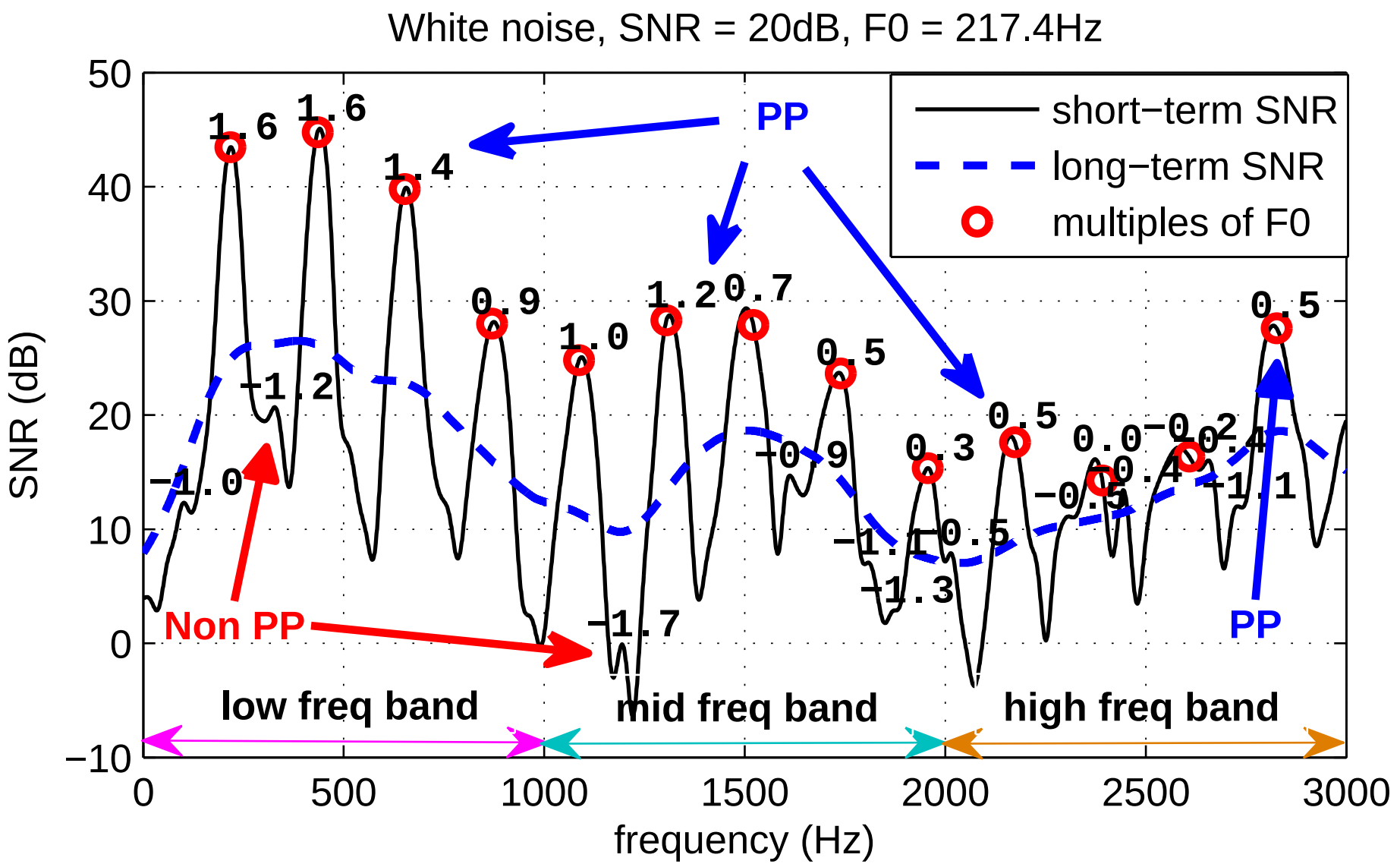
$\zeta_i = \gamma_{l_i}^{\mathsf{S}} - \gamma_{l_i}^{\mathsf{L}}, \quad i = 1, \cdots, M \quad$ (M: # of peaks in γ_l^{S}).

Normalized SNR difference

$\bar{\zeta}_i = (\zeta_i - \mu_\zeta)/\sigma_\zeta, \quad i = 1, \cdots, M.$

An SNR peak is prominent, if $\bar{\zeta}_i >$ a preset threshold.

Examples of Prominent SNR Peaks (PP) and Non-PPs



F0 Candidate Likelihood Fusion and Bayesian Inference

Assume prominent SNR peaks $\mathbf{C}_i$, $i = 1, \cdots, M$, contain sufficient information regarding F0:

$$P(f|\mathbf{Y}, \mathbf{N}) = P(f|\mathbf{C}_1, \cdots, \mathbf{C}_M, \mathbf{N}) \tag{2}$$

Assume that the set of prominent SNR peaks are independent in inferring the F0 given noise $\mathbf{N}$

$$P(f|\mathbf{Y}, \mathbf{N}) = \sum_{i=1}^{M} w_i P(f|\mathbf{C}_i, \mathbf{N}) \tag{3}$$

Currently, the confidence measure w_i is set to be $1/M$. Suppose $p(f|N)$ is uniformly distributed, we have:

$$P(f|\mathbf{C}_i, \mathbf{N}) = \frac{p(\mathbf{C}_i|f, \mathbf{N})}{\sum_{f \in \mathcal{S}_{\text{F0}}} p(\mathbf{C}_i|f, \mathbf{N})} \tag{4}$$

Peak Analysis in Training and Testing

Let the frequency f denote:

- the F0 reference in training
- an F0 hypothesis among possible values in testing

The local SNR peak $\mathbf{C}_i$ with a frequency l is represented by the following properties:

- the multiple m: $m = \lceil l/f \rceil$
- the residual δ: $\delta = l/f - m$
  - If $l = 503$ and $f = 100$, then $m = 5$ and $\delta = 0.03$.
- the *a posteriori* SNR γ_l
- the frequency band $\mathbf{B}_l$ where the frequency l resides

Then, we have:

$$\begin{aligned} p(\mathbf{C}_i|f,\mathbf{N}) &= p(m,\delta,\gamma_l,\mathbf{B}_l|f,\mathbf{N}) \\ &= D \cdot p(\delta|\gamma_l,\mathbf{B}_l,\mathbf{N})p(\gamma_l|\mathbf{B}_l,\mathbf{N}) \end{aligned} \tag{5}$$

Estimate a model from the extracted information and F0 reference

Model Learning

To reduce the model complexity,

$$\text{Residual:} \quad p(\delta|\gamma_l, \mathbf{B}_l, \mathbf{N}) \approx p(\delta|\mathbf{Q}_{\gamma_l}, \mathbf{B}_l, \mathbf{N}) \tag{6}$$

$$\text{Local SNR:} \quad p(\gamma_l|\mathbf{B}_l, \mathbf{N}) \approx p(\mathbf{Q}_{\gamma_l}|\mathbf{B}_l, \mathbf{N}) \tag{7}$$

where $\mathbf{Q}_{\gamma_l}$ denotes the SNR bin which γ_l is rounded to.
Assume the residual δ and local SNR $\mathbf{Q}_{\gamma_l}$ are i.i.d..

Residual: $p(\delta|\mathbf{Q}_{\gamma_l}, \mathbf{B}_l, \mathbf{N})$

- Modeled as a doubly truncated Laplacian distribution
- Maximum likelihood-based parameter estimation

Local SNR: $p(\gamma_l|\mathbf{B}_l, \mathbf{N})$

- Learned by using a histogram-like approach

Data Set

KEELE corpus [5]

- The total length: 5 min 37 sec
- 5 male and 5 female speakers
- Used in both training and testing: 5-fold cross-validation

CSTR corpus [6]

- The total length: 5 min 32 sec
- 1 male and 1 female speaker
- Used only as a testing set

- Both databases include simultaneous recordings of speech and laryngograph signals.
- We added white and babble noise to the corpora with SNR of 20, 10, 5, 0, and -5 dB.

Error Metric

Gross Pitch Error (GPE) [7]

$$GPE = \frac{N_{F0E}}{N_{VV}} \times 100\% \tag{8}$$

N_{VV}: The number of frames which both the F0 tracker and the ground truth consider to be voiced
N_{F0E}: The number of frames for which

$$\left| \frac{F0_{i,estimated}}{F0_{i,reference}} - 1 \right| > 20\% \tag{9}$$

Note

In this paper, only estimate F0 on voiced frames.

Experiment Settings

SAFE's parameters

- F0 estimation resolution: 1 Hz
- Frame length and step size: 0.04 and 0.01 seconds
- F0 range: from 50 to 400 Hz
- normalized difference SNR threshold: 0.33
- Low, middle, and high frequency bands: 0-1, 1-2, and 2-3 kHz
- Local SNRs of the peaks are rounded to the nearest value in the following sequence $10r/3$, $r = 0, 1, \cdots, 21$.

Other F0 estimators' parameters

- default parameters
- voicing thresholds are optimized for each

GPEs (%) (KEELE used for training and testing, 5-fold CV)

SNR (dB)	Clean	20	10	5	0	-5
		KEELE White Noise				
Get_F0	**2.62**	**2.69**	**3.10**	4.09	7.69	17.83
Praat	3.22	3.16	4.28	6.11	11.53	30.91
TEMPO	2.98	3.41	4.27	5.57	12.79	22.64
YIN	2.94	2.94	3.20	3.96	6.70	14.48
SAFE (LFB)	3.13	3.09	3.74	4.39	4.72	6.29
SAFE	2.98	3.01	3.35	**3.66**	**4.06**	**5.01**
		KEELE Babble Noise				
Get_F0		**2.87**	7.19	15.99	29.76	58.40
Praat		3.18	8.33	17.97	35.26	54.06
TEMPO		4.69	13.99	26.98	43.98	65.15
YIN		3.27	8.89	19.71	36.75	57.35
SAFE (LFB)		3.23	6.01	10.21	20.64	47.21
SAFE		3.10	**4.72**	**7.44**	**15.88**	**39.23**

GPEs (%) (KEELE used for training, CSTR used for testing)

SNR (dB)	Clean	20	10	5	0	-5
		CSTR White Noise				
Get_F0	2.45	2.46	3.04	3.94	6.73	17.72
Praat	2.27	2.27	2.99	4.35	11.84	27.54
TEMPO	2.27	2.29	2.87	5.07	11.64	31.65
YIN	**2.25**	**2.25**	**2.36**	3.34	5.20	12.33
SAFE (LFB)	2.49	2.52	2.97	3.49	3.93	4.14
SAFE	2.40	2.41	2.69	**3.10**	**3.24**	**3.68**
		CSTR Babble Noise				
Get_F0		2.86	8.36	24.41	46.41	64.52
Praat		2.65	10.55	27.15	46.32	64.24
TEMPO		3.56	15.24	33.10	54.43	66.38
YIN		**2.36**	10.09	27.53	51.15	68.22
SAFE (LFB)		2.69	5.37	9.97	23.59	63.20
SAFE		2.61	**4.14**	**7.73**	**19.32**	**57.17**

Conclusion

- Prominent SNR peaks constitute a simple and an effective information source for F0 inference under both clean and noisy conditions
- The statistical framework of F0 estimation is promising in modeling the effect of additive noise on clean speech spectra.
- In addition to low frequencies, middle and high frequency bands (1-3 kHz) provide supplemental useful information for F0 inference.
- The proposed SAFE algorithm is more effective in reducing the GPE compared to prevailing F0 trackers especially at low SNRs and babble noise conditions.

Thanks for coming!

Q & A