SAFE: A Statistical Approach to F0 Estimation under Clean and Noisy Conditions

Wei Chu, *Student Member, IEEE,* Abeer Alwan, *Fellow, IEEE*

Abstract

A novel Statistical Algorithm for F0 Estimation, SAFE, is proposed to improve the accuracy of F0 estimation under both clean and noisy conditions. Prominent Signal-to-Noise Ratio (SNR) peaks in speech spectra constitute a robust information source from which F0 can be inferred. A probabilistic framework is proposed to model the effect of noise on voiced speech spectra. Prominent SNR peaks in the low frequency band (0 - 1000 Hz) are important to F0 estimation, and prominent SNR peaks in the middle and high frequency bands (1000 - 3000 Hz) are also useful supplemental information to F0 estimation under noisy conditions, especially the babble noise condition. Experiments show that the SAFE algorithm has the lowest Gross Pitch Errors (GPE) compared to prevailing F0 trackers in white and babble noise conditions at low SNRs. Experimental results also show that SAFE is robust in maintaining a low Mean and Standard Deviation of the Fine Pitch Errors (MFPE and SDFPE) in noise. The code of SAFE is available at http://www.ee.ucla.edu/~weichu/safe .

I. Introduction

The source-filter model of speech production [1] assumes that speech signals can be modeled as an excitation signal filtered by a linear vocal-tract transfer function. The fundamental frequency (F0) is defined as the inverse of the period of the excitation signal during the voicing state [2] [3]. Accurate F0 tracking in quiet and in noise is important for several speech applications, such as speech coding, analysis and recognition.

Some F0 tracking algorithms are based on the source-filter theory of speech production and estimate F0 for voiced speech segments. They assume that F0 is constant and the vocal tract transfer function is time invariant within a short period of time, e.g, a frame of 10-20 milliseconds. These algorithms usually have two stages. The first stage consists of obtaining F0 candidates and the likelihood of voicing on a

 Supported in part by the NSF.

frame-by-frame basis. The second stage consists of using dynamic programming to decide the optimal F0 and voicing state for each frame.

The first stage can be classified into two categories: single-band and multi-band. In the single-band method, F0 candidates are extracted from one frequency band [2]. There are several methods to generate F0 candidates. SIFT [4] applies inverse filtering to voiced speech to obtain the excitation signal from which it estimates F0 by using autocorrelation. Cepstral-based methods (e.g., [5]) separate the excitation from the vocal tract information in the cepstral domain by using a homomorphic transformation; the interval to the first dominant peak in the cepstrum is related to the fundamental period. RAPT [6] and YAPPT [7] generate F0 candidates by extracting local maxima of the normalized cross correlation function which is calculated over voiced speech. Praat [8] calculates cross correlation or autocorrelation functions on the speech signal and regards local maxima as F0 hypotheses. TEMPO [9] obtains F0 candidates by evaluating the 'fundamentalness' of speech which achieves a maximum value when the AM and FM modulation magnitudes are minimized. YIN [10] uses the autocorrelation-based squared difference function and the cumulative mean normalized difference function calculated over voiced speech, with little post-processing, to acquire F0 candidates. Yegnanarayana et al. [11] obtain F0 candidates from exploiting the impulse-like characteristics of excitation in glottal vibrations. Finally, Le Roux et al. [12] simultaneously perform frame-wise F0 candidate generation and time-direction smoothing.

In the multi-band method, a decision module is usually used to reconcile the F0 candidates generated from different bands. Gold and Rabiner [13] use measurements of peaks and valleys of voiced speech as input to six separate functions whose values are then processed by an F0 estimator to obtain F0 candidates. Lahat et al. [14] calculate autocorrelation functions of the spectral magnitudes in different bands and then obtain F0 candidates by evaluating the local maxima of the functions. Sha et al. [15] detect F0 candidates by minimizing the values of sinusoid-based error functions calculated on 4 frequency bands: 25-100, 50-200, 100-400, and 200-800 Hz. These multi-band methods focus mainly on the low frequency bands.

The multi-band approach has also been used to apply Licklider's pitch perception theory [16] to F0 estimation. The irregular excitation signal may cause voiced speech to be aperiodic in some frequency bands [17]. It is hypothesized that the higher levels of auditory processing isolate groups of contiguous harmonics to infer the fundamental frequency from a selection of these groups. In this view, it is hypothesized that auditory nerves and the auditory brainstem are capable of using an autocorrelation mechanism to infer F0 over different frequency channels. de Cheveigne shows that integrating the values of AMDFs across different channels in the time domain can improve F0 estimation accuracy [18]. Wu

et al. [19] used correlograms to select reliable frequency bands, modeled F0 dynamics using a statistical approach, and then searched for the optimal F0 contour in an HMM framework.

These F0 candidate generation methods can also be applied to noisy conditions. Krusback et al. [20] use an autocorrelation function with confidence measures. Shimamura et al. [21] proposed a weighted autocorrelation function. Abe et al. [22] use the instantaneous frequency spectrum to enhance harmonics and suppress aperiodic components, which improves F0 estimation accuracy. Liu et al. [23] use joint time-frequency analysis to obtain robust adaptive representation of the speech spectrum from which important harmonic structures can be extracted. Nakatani et al. [24] use dominance spectra based on instantaneous frequencies to evaluate the magnitudes of the harmonics relative to background noise, and estimate F0 using only the reliable harmonics. Deshmukh et al. [25] use an aperiodicity, periodicity, and pitch detector to generate F0 candidates by calculating the AMDFs over different frequency channels in the spectral domain.

According to the experimental results in this study, some of the methods mentioned above can work well under relatively noise-free conditions. However, when the low-frequency band is contaminated by noise, an increase in F0 estimation errors is observed. Since it is possible that F0 harmonics in the middle or high frequency bands are not corrupted, it may be beneficial for an F0 estimation method to utilize these harmonics in determining F0. Current multi-band methods [14] [15] mainly retain F0 candidates obtained from the most reliable band, which is a ‘hard-decision’, while the Licklider’s pitch perception model uses an empirically-based ’soft-decision’ to merge the information from different bands [18]. Wu et al. [19] uses a ‘soft-decision’ approach to combine the information across bands. We propose a Statistical Algorithm for F0 Estimation (SAFE) which also utilizes a ‘soft-decision’ method. A data-driven approach is used to learn how the noise affects the amplitude and location of the peaks in the Signal-to-Noise Ratio (SNR) spectra of clean voiced speech. The likelihoods of F0 candidates are obtained by evaluating the peaks in the SNR spectrum using the corresponding models learned from different bands. It is worth noting that Ying et al. [26] use a probabilistic method to estimate F0 distribution in order to avoid local optima in F0 estimation. Wang et al. [27] modeled the between-frame F0 transitions in a statistical approach to improve both F0 estimation and unvoiced/voiced decision.

In the following sections, the statistical effects of noise on clean voiced speech spectra are studied. This relationship between the noise and information source for F0 estimation is modeled in a probabilistic framework. In testing, the posterior probabilities of the F0 candidates are then calculated. In the experimental section, the performance of the proposed method under different noise types and SNRs is compared with prevailing F0 estimation methods.

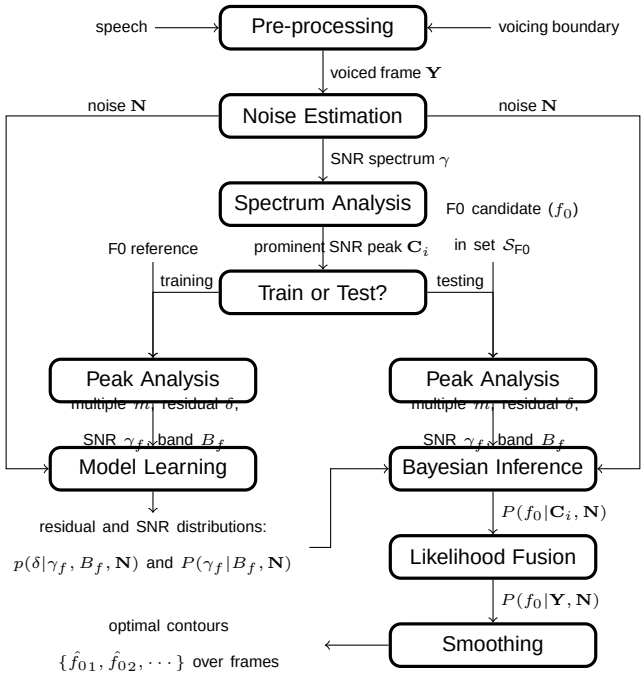


Fig. 1. A flowchart of SAFE.

II. SAFE: A Statistical Algorithm for F0 Estimation

A flowchart of SAFE is shown in Fig. 1. This paper focuses on estimating fundamental frequency (F0) values over voiced frames that may be corrupted by quasi-stationary noise. Suppose that the range of F0 in human speech is from f_{0min} to f_{0max}, and the frequency resolution of F0 estimation is Δ. Then $\mathcal{S}_{\text{F0}}$ is used to denote the set of all possible F0 values $\{f_{0min}, f_{0min} + \Delta, \cdots, f_{0max}\}$.

Given the power spectrum $\mathbf{Y}$ of a single observed noisy voiced frame under a stationary noise condition $\mathbf{N}$, the probability of f_0 being the fundamental frequency of that frame can be expressed as $P(f_0|\mathbf{Y}, \mathbf{N})$. The most likely estimate, denoted by $\hat{f}_0$, should be:

$$\hat{f}_0 = \arg\max_{f_0 \in \mathcal{S}_{\text{F0}}} P(f_0|\mathbf{Y}, \mathbf{N}). \tag{1}$$

Let $\mathbf{Y}_f$ and $\mathbf{N}_f$ denote the power spectrum of the noisy voiced frame and noise at frequency f, respectively. Then the $a\ posteriori$ SNR at frequency f denoted by γ_f is:

$$\gamma_f = 10 \log_{10} \frac{\mathbf{Y}_f}{\mathbf{N}_f}. \tag{2}$$

As quasi-stationary noise is assumed in this study, the noise spectrum for each utterance is estimated by averaging the initial 10 and final 10 frames of noisy speech. The frame shift is 10 ms, and the frame length is 40 ms.

The SNR γ_f is a measure of the spectral magnitude at frequency f being contaminated by the noise. According to the source-filter theory of speech production, a voiced speech spectrum has a harmonic structure. Local SNR peaks (correspond to mainly harmonics) contain more information than valleys regarding F0. It is assumed that the information contained in the set of local SNR peaks $\{\mathbf{C}_1, \cdots, \mathbf{C}_M\}$ is sufficient to estimate F0, where M is the number of local SNR peaks. Thus, the posterior probability of f_0 is:

$$P(f_0|\mathbf{Y}, \mathbf{N}) = P(f_0|\mathbf{C}_1, \cdots, \mathbf{C}_M, \mathbf{N}). \tag{3}$$

In a ROVER system for automatic speech recognition [28], the posterior probabilities of a word from different sub-systems are combined with different weights. Inspired by ROVER, local SNR peaks can be assumed to be independent in inferring F0 given the noise shape and level. The overall posterior probability can be approximated as a weighted combination of posterior probabilities $P(f_0|\mathbf{C}_i, \mathbf{N})$:

$$P(f_0|\mathbf{Y}, \mathbf{N}) \approx \sum_{i=1}^{M} w_i P(f_0|\mathbf{C}_i, \mathbf{N}), \tag{4}$$

where w_i is the confidence measure of the i-th local SNR peak. If each local SNR peak is assumed to have an equal confidence score, then w_i is set to $1/M$. ($i = 1, 2, \cdots, M$)

If the distribution of f_0 given the noise, i.e., $P(f_0|\mathbf{N})$, is assumed to be uniformly distributed when prior information is not available, then $P(f_0|\mathbf{C}_i, \mathbf{N})$ can be obtained from the Bayesian rule:

$$P(f_0|\mathbf{C}_i, \mathbf{N}) = \frac{p(\mathbf{C}_i|f_0, \mathbf{N})}{\sum_{f_0 \in \mathcal{S}_{\text{F0}}} p(\mathbf{C}_i|f_0, \mathbf{N})}. \tag{5}$$

Let f denote the frequency of the local SNR peak $\mathbf{C}_i$. Because f is not usually equal to a multiple of f_0, f can be decomposed into a multiple m and a residual δ as follows:

$$m = [\frac{f}{f_0}], \quad \delta = \frac{f}{f_0} - m, \tag{6}$$

where $[\frac{f}{f_0}]$ denotes the nearest integer of $\frac{f}{f_0}$. Hence, the residual ranges from -0.5 to 0.5. If the fraction of $\frac{f}{f_0}$ is exactly 0.5, either rounding upwards or downwards does not change F0 estimation error rates in SAFE.

Given f_0 and noise $\mathbf{N}$, the local SNR peak $\mathbf{C}_i$ has the following attributes: multiple m, residual δ, $a\ posteriori$ SNR γ_f, and frequency band index B_f in which the frequency f is. In other words, the peak $\mathbf{C}_i$ resides in band B_f. The reason why f is not adequate on its own is because there are not

enough training samples for each frequency bin. Then we have:

$$\begin{aligned} p(\mathbf{C}_i|f_0,\mathbf{N}) &= p(m,\delta,\gamma_f,B_f|f_0,\mathbf{N}) \\ &= P(m|f_0,\mathbf{N})p(\delta|m,\gamma_f,B_f,f_0,\mathbf{N}) \\ &\quad p(\gamma_f|m,B_f,f_0,\mathbf{N})P(B_f|m,f_0,\mathbf{N}). \end{aligned} \tag{7}$$

We assume that the deviation of a local SNR peak from a multiple of f_0, caused by noise, will not exceed half f_0. Therefore, m is independent of the noise $\mathbf{N}$, i.e., $P(m|f_0,\mathbf{N}) = P(m|f_0)$. After the decomposition shown in (6), the residual δ can be assumed to be independent of m and f_0 given γ_f, B_f, and $\mathbf{N}$, i.e., $p(\delta|m,\gamma_f,B_f,f_0,\mathbf{N}) = p(\delta|\gamma_f,B_f,\mathbf{N})$. The local SNR γ_f is independent of m and f_0 given the band index B_f and noise condition $\mathbf{N}$, i.e., $p(\gamma_f|m,B_f,f_0,\mathbf{N}) = p(\gamma_f|B_f,\mathbf{N})$. Furthermore, $P(m|f_0)$ is assumed to be uniformly distributed. Since B_f can be assumed to be determined by m and f_0 regardless of noise, the Dirac function $P(B_f|m,f_0,\mathbf{N})$ is assumed to be equal to 1. Then we can have:

$$\begin{aligned} & p(\mathbf{C}_i|f_0,\mathbf{N}) \\ = \;& D_1 \cdot p(\delta|\gamma_f,B_f,\mathbf{N})p(\gamma_f|B_f,\mathbf{N}). \end{aligned} \tag{8}$$

where D_1 is a constant.

A. Prominent SNR Peaks

Before studying the distribution of the residual and local SNR peaks, it is important to select useful local SNR peaks for F0 estimation. Short and long-term smoothed SNRs denoted by $\gamma_f^{\mathtt{S}}$ and $\gamma_f^{\mathtt{L}}$ are obtained by smoothing γ_f with a Hamming window of length f_{0min} and f_{0max} in Hz, respectively. The Hamming window is used because of its relatively small side lobes. Since the short-term smoothing can reduce the number of false alarm local SNR peaks and retain F0 information, γ_f in (8) is replaced by $\gamma_f^{\mathtt{S}}$. To depict the relationship between the two smoothed SNRs, an SNR difference at the i-th local peak in $\gamma_f^{\mathtt{S}}$ denoted by ζ_i can be expressed as follows:

$$\zeta_i = \gamma_{f_i}^{\mathtt{S}} - \gamma_{f_i}^{\mathtt{L}}, \quad i = 1,\cdots,M^{\mathtt{S}}, \tag{9}$$

where $M^{\mathtt{S}}$ is the number of the local peaks in $\gamma_f^{\mathtt{S}}$. ζ_i is further normalized with respect to all the peaks in the frame as follows:

$$\bar{\zeta}_i = \frac{\zeta_i - \mu_\zeta}{\sigma_\zeta}, \quad (i = 1,\cdots,M^{\mathtt{S}}.), \tag{10}$$

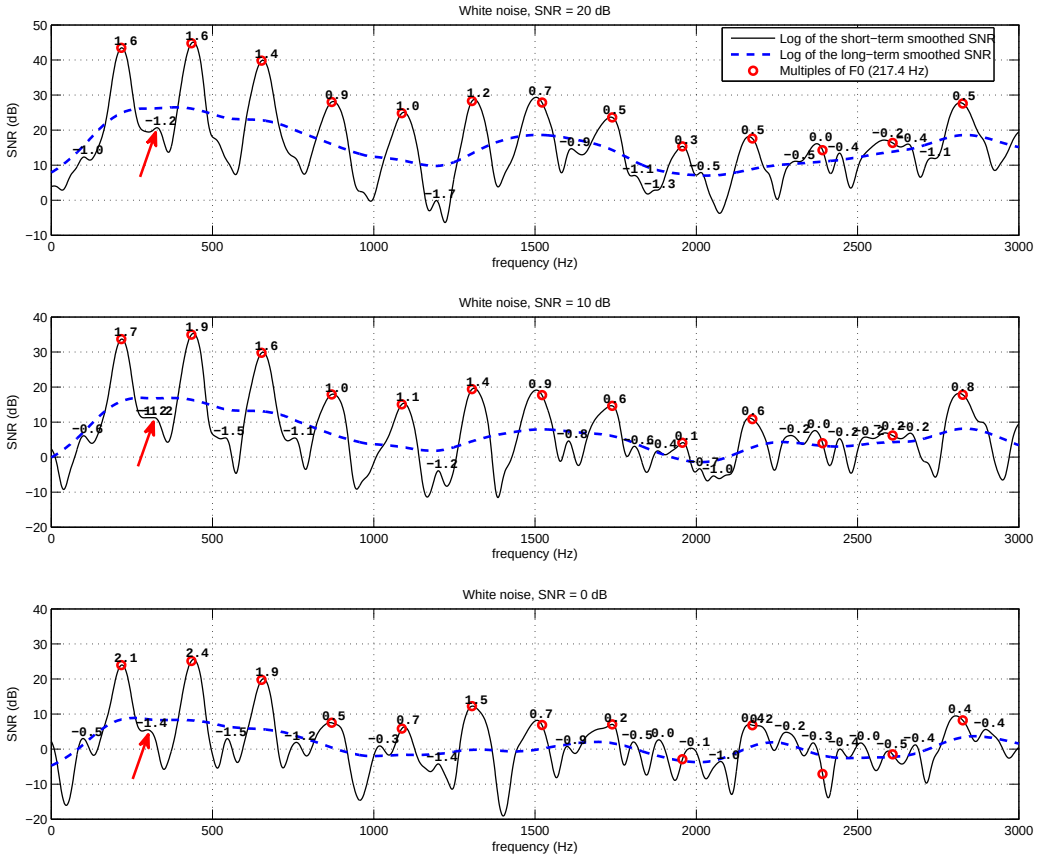


Fig. 2. The SNR spectrum of a voiced frame of a female speaker corrupted by different levels of additive white noise (20, 10 and 0 dB). The number on top of each peak of the short-term smoothed SNR is the value of the normalized difference SNR $\bar{\zeta}_i$ of that peak. Arrows around 300 Hz indicate peaks with a lower $\bar{\zeta}_i$ than their adjacent prominent peaks.

where μ_ζ and σ_ζ are the mean and standard deviation of the sequence ζ_i. The i_{th} local SNR peak ($\mathbf{C}_i$) is regarded as a $prominent\ SNR\ peak$ for F0 estimation only if $\bar{\zeta}_i$ is above a certain threshold. In this study, the threshold is empirically set to 0.33.

Figs. 2 and 3 show the SNR spectra of a voiced frame of a female and a male speaker, respectively, corrupted by different levels of additive white noise (20, 10 and 0 dB). The number on top of each peak of the short-term smoothed SNR is the value of the normalized difference SNR $\bar{\zeta}_i$ of that peak. It can be seen that not all local SNR peaks reside in the vicinity of multiples of F0. Most false alarm or deviated peaks have a lower normalized SNR difference compared to the peaks near the multiples of F0. Take

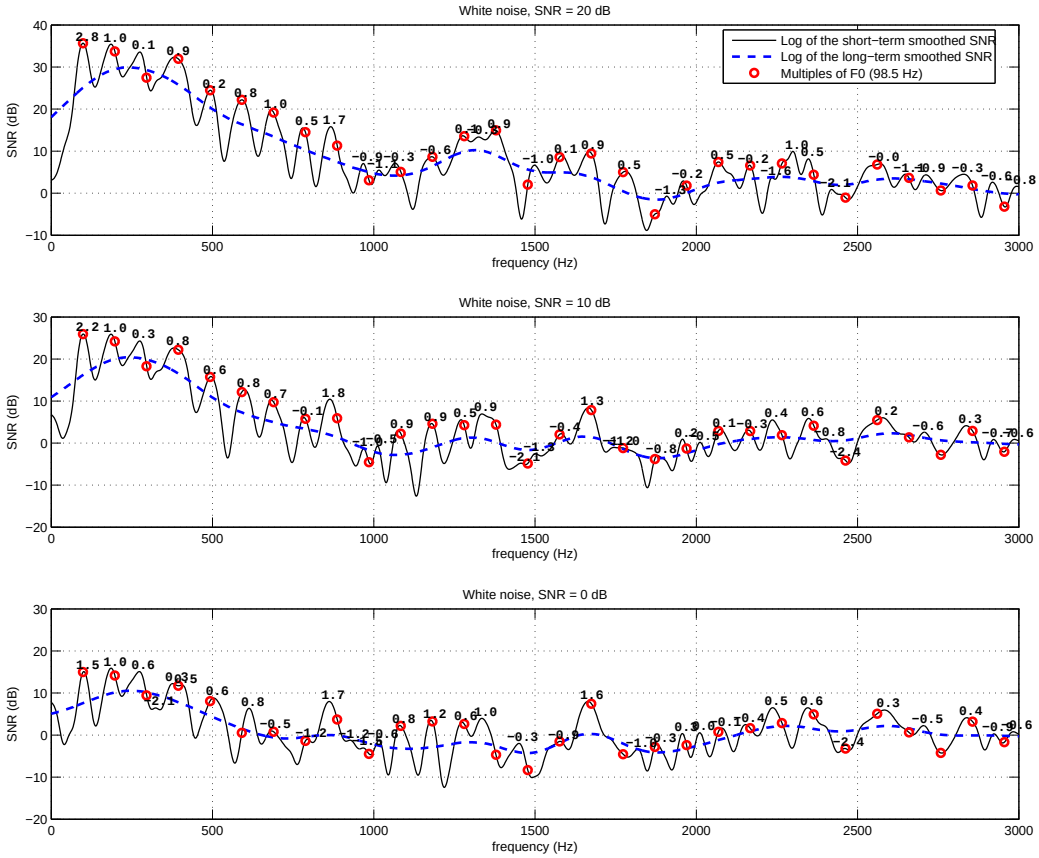


Fig. 3. The SNR spectrum of a voiced frame of a male speaker corrupted by different levels of additive white noise (20, 10 and 0 dB). The number on top of each peak of the short-term smoothed SNR is the value of the normalized difference SNR $\bar{\zeta}_i$ of that peak.

the false alarm local peaks around 300 Hz of the voiced frames in all panels of Fig. 2 for example. These peaks, indicated by arrows, have a lower $\bar{\zeta}_i$ than their adjacent prominent peaks in the three noise conditions.

As shown in Figs. 2 and 3, the lower a peak is than the long-term smoothed SNR, the more likely it is corrupted by the noise and shifted from its original location, and the less likely it is to be close to the multiples of F0. Hence, prominent SNR peaks which are less corrupted by the noise and less deviated from a multiple of F0 can provide reliable information for inferring F0s. When middle and high frequency bands are less corrupted by noise, it is possible that prominent peaks can exist in these

bands, e.g., the peaks around 2800 Hz in female voiced frames and the peaks around 1700 Hz in male voiced frames under 20, 10 and 0 dB SNR conditions. Retaining these prominent peaks and discarding non-prominent peaks might improve the performance of F0 estimation.

As mentioned above, only prominent peaks are used in (4), i.e., M is changed to the number of prominent SNR peaks.

B. Distribution of the Residuals

Recall that the residual δ is dependent on the local SNR value and the band index. To reduce the model complexity, it can be assumed that the distribution of the $p(\delta|\gamma_f, B_f, \mathbf{N})$ in (8) slightly changes when γ_f is rounded, i.e.,

$$p(\delta|\gamma_f, B_f, \mathbf{N}) \approx p(\delta|Q_{\gamma_f}, B_f, \mathbf{N}), \tag{11}$$

where Q_{γ_f} denotes the SNR bin which γ_f is rounded to. The intervals of the SNR bins in dB are spaced by 3.33 dB and are as follows: (-∞, 0], (0, 3.33], $\cdots$, (66.67, 70], (70, ∞).

The distributions of the residuals given different rounded SNR bins, frequency band index and noise conditions are shown in Fig. 4. Two white noise conditions: 20 and 0 dB SNRs are studied. This analysis is conducted over all the voiced frames in the KEELE corpus [29] with F0 ground truth values obtained from the simultaneously recorded laryngograph signal. In this study, three bands: 0-1000 Hz, 1000-2000 Hz, and 2000-3000 Hz, are employed to represent the low, middle, and high frequency bands, respectively. Note that all the residual distributions in Fig. 4 are derived only from the prominent peaks in the low frequency band. Most distributions are centered on zero, which means that these peaks can generate unbiased F0 estimates. It can be seen that under a certain noise condition, the higher the rounded SNR is, the smaller the variance of the residuals. Because having a smaller residual variance means that the frequencies of local SNR peaks are less likely to be affected by noise, local SNR peaks from higher SNR bins are more reliable for F0 estimation. Under 20 dB conditions, no prominent peak has a local SNR higher than 56.67 dB; under 0 dB condition, the local SNRs of all prominent peaks are below 36.67 dB.

A comparison of the distributions of the residuals of the prominent and non-prominent peaks is shown in Fig. 5 for the white noise condition with 0 dB SNR. In the low frequency band, prominent peaks can have a local SNR as high as 36.67 dB, while the local SNRs of non-prominent peaks are below 26.67 dB. Furthermore, the residuals of the non-prominent peaks with low local SNRs are mostly distributed away from zero, which means that it is difficult to infer F0 from these non-prominent peaks. Although the residuals of the non-prominent peaks with high local SNRs are distributed around zero, the distributions have larger variances compared to the residuals of the prominent peaks with the same local SNR.

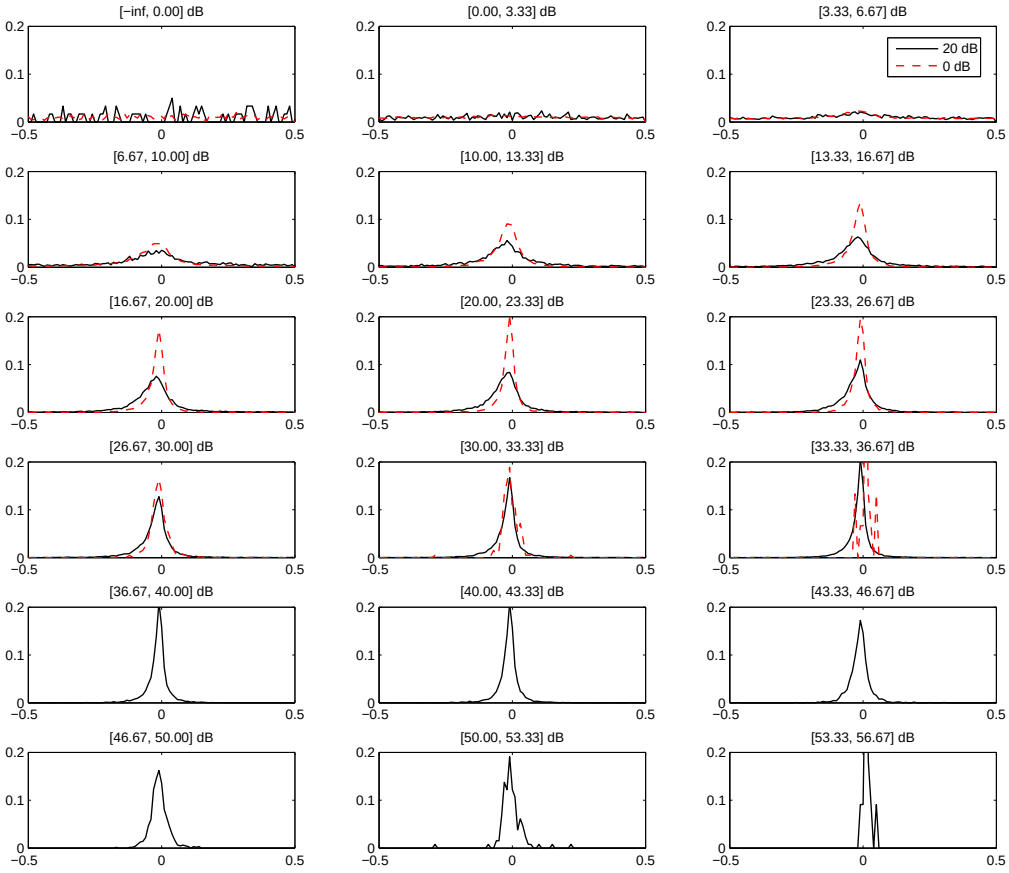


Fig. 4. The distributions of the residuals given different rounded local SNRs for a 3.33 dB interval at the low frequency band (0-1000Hz). Different white noise conditions (20 and 0 dB global SNRs) are shown. The horizontal axes are the residuals with a bin size of 0.01. The vertical axes are the probabilities of occurrences. The title on each sub-figure shows the interval of rounded local SNR Q_{γ_f}.

Curve-fitting or Gaussian mixture modeling can be used to model the distributions of the residuals; however, it is important to control the number of parameters in the model which enables training with limited data and prevent model over-fitting. A *Doubly truncated Laplacian distribution*, denoted by $p(\delta|\mu, b)$, is used for modeling $p(\delta|Q_{\gamma_f}, B_f, \mathbf{N})$, i.e. the distribution of residuals given the rounded SNR

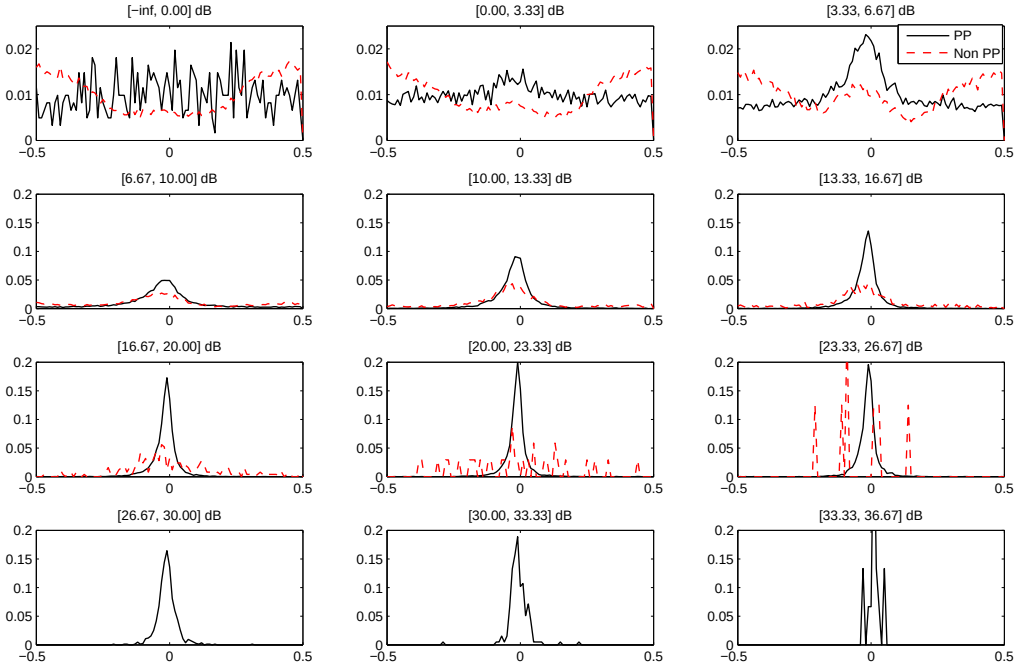


Fig. 5. A comparision of the distributions of the residuals of prominent SNR peaks (PP) and non-prominent SNR peaks (Non PP) given different rounded local SNRs at the low frequency band (0-1000Hz). The noise condition is white noise at 0 dB global SNR. The horizontal axes are the residuals with a bin size of 0.01. The vertical axes are the probabilities of occurrences. The title on each sub-figure shows the interval of rounded local SNR Q_{γ_f}.

bin, band index and noise condition:

$$p(\delta|\mu,b) = \begin{cases} \dfrac{A}{2b} e^{-\dfrac{|\delta-\mu|}{b}} & -\dfrac{1}{2} \le \delta \le \dfrac{1}{2} \\ 0 & \text{otherwise} \end{cases}, \tag{12}$$

where μ and b represent the mean and the variance, respectively. A is set to $(1-e^{-1/(2b)})^{-1}$ to ensure that $\int_\delta p(\delta|\mu,b) = 1$. Hence, only two free parameters (μ, b) need to be estimated.

Given a sequence of residuals $\{\delta_1, \cdots, \delta_N\}$ denoted by $\boldsymbol{\delta}$, (suppose all the residuals are independent and identically distributed,) we have:

$$p(\boldsymbol{\delta}|\mu,b) = \prod_{i=1}^{N} p(\delta_i|\mu,b). \tag{13}$$

Let $\alpha = 1/(2b)$ and $\mathcal{L}(\boldsymbol{\delta}|\mu, \alpha) = \log p(\boldsymbol{\delta}|\mu, b)$. Then:

$$\begin{aligned} & \mathcal{L}(\boldsymbol{\delta}|\mu, \alpha) \\ = & \sum_{i=1}^{N} \log p(\delta_i|\mu, b) \\ = & N \log \alpha - N \log(1 - e^{-\alpha}) - 2\alpha \sum_{i=1}^{N} |\delta_i - \mu|. \end{aligned} \tag{14}$$

Under the maximum-likelihood criterion, the estimated mean and variance denoted by $\hat{\mu}$ and $\hat{b}$ (or $\hat{\alpha}$) should maximize the joint probability $p(\boldsymbol{\delta}|\mu, b)$ which is equivalent to maximize $\mathcal{L}(\boldsymbol{\delta}|\mu, \alpha)$.

Since $\partial^2 \mathcal{L}/\partial \mu^2 = -2\alpha \sum_{i=1}^{N} \delta(\delta_i - \mu) \leq 0$ when $\alpha > 0$, $\mathcal{L}$ achieves its maximum when $\partial \mathcal{L}/\partial \mu = 0$ for any α, i.e.:

$$-2\alpha \sum_{i=1}^{N} \text{sgn}(\delta_i - \hat{\mu}) = 0. \tag{15}$$

Since $\partial^2 \mathcal{L}/\partial \alpha^2 = 1/(e^{\alpha} - 1) - 1/\alpha^2 < 0$ when $\alpha > 0$, $\mathcal{L}$ achieves its maximum when $\partial \mathcal{L}/\partial \alpha = 0$ and $\mu = \hat{\mu}$, i.e.:

$$\frac{N}{\hat{\alpha}} - \frac{N}{e^{\hat{\alpha}} - 1} - 2 \sum_{i=1}^{N} |\delta_i - \hat{\mu}| = 0. \tag{16}$$

To solve $\hat{\mu}$, let $\tilde{\boldsymbol{\delta}} = \{\tilde{\delta}_1, \cdots, \tilde{\delta}_N\}$ denote the sorted sequence of the sequence $\boldsymbol{\delta}$ in an ascending order, we have one feasible solution of $\hat{\mu}$:

$$\hat{\mu} = \begin{cases} \tilde{\delta}_{\frac{N+1}{2}} & N \text{ is odd} \\ \frac{1}{2}(\tilde{\delta}_{\frac{N}{2}} + \tilde{\delta}_{\frac{N}{2}+1}) & N \text{ is even} \end{cases}. \tag{17}$$

Note that when N is even, any value between $\tilde{\delta}_{\frac{N}{2}}$ and $\tilde{\delta}_{\frac{N}{2}+1}$ can satisfy (15). As shown in (17), the number of residuals that are greater than $\hat{\mu}$ is equal to the number of residuals that are less than $\hat{\mu}$. (16) can be simplified as:

$$\frac{N}{\hat{\alpha}} - \frac{N}{e^{\hat{\alpha}} - 1} - 2 \sum_{i=1}^{N} |\delta_i| = 0. \tag{18}$$

Although there is no close-form solution to (16), Newton's method can be used to search for $\hat{\alpha}$. Note that $\hat{b} = 1/(2\hat{\alpha})$. When a bin with a high rounded SNR does not have training instances, no effort of running the mean and variance solvers is spared. In case of some unseen residuals might have higher SNRs, the mean is set to 0, and the variance is set to a small value, e.g., 0.01.

There is one similarity between SAFE and Wu et al.'s method [19]: the use of Laplacian distribution for data modeling. The meaning and range of the modeled random variables are different. SAFE models

the residual derived from the prominent peak in the SNR spectrum. The residual ranges from -0.5 to 0.5. Wu et al.'s method models the time lag derived from the peak in the correlogram. The time lag ranges from $-\infty$ to ∞.

The logarithms of the averaged estimated variances of the residual distributions for different bands are shown in Fig. 6. Averaging is across all noise levels: clean, 20 dB, 10 dB, 5 dB, 0 dB, -5 dB. The noise type is white noise. It can be seen that the variance of the lower frequency band at a certain rounded SNR bin is smaller than the counterpart of the higher frequency band. When the variance of the estimated residual distribution is small given a frequency band, it means that the probability of accurately estimating F0 in that band is high. As mentioned above, it is still possible to use the prominent peaks lying in the middle and high frequency bands to improve F0 estimation. Note that the higher the rounded local SNR, the smaller the variance is.

In Fig. 7, the estimated means of the residual distributions for different bands under clean and noisy conditions are compared. The noise types are white and babble noise. The means under noisy conditions at different SNRs (20 dB, 10 dB, 5 dB, 0 dB, -5 dB) are averaged. It can be seen that the estimated means are not exactly equal to zero under both clean and noisy conditions if local SNR is less than 55 dB. F0 estimation actually benefits from learning a Laplacian distribution with a non-zero mean which better fits the real distribution of the data.

C. Distribution of the local SNRs

In the previous section, local SNRs of the prominent peaks are rounded. It can be assumed that this rounding does not significantly change the $p(\gamma_f|B_f, \mathbf{N})$ in (8), i.e.:

$$p(\gamma_f|B_f, \mathbf{N}) \approx D_2 P(Q_{\gamma_f}|B_f, \mathbf{N}), \tag{19}$$

where D_2 is a constant. The distribution can be learned by using a histogram-like approach based on the training set.

The distributions of the rounded local SNRs of the prominent peaks under different bands and noise conditions are shown in Fig. 8. The distribution under noisy conditions at different SNRs (20 dB, 10 dB, 5 dB, 0 dB, -5 dB) are averaged. It can be seen that the peaks of noisy speech are more likely to be distributed in bins with low SNRs compared to clean speech, which can be one of the reasons why estimating F0 values is difficult under noisy conditions. For either clean or noisy condition, the rounded local SNRs of the prominent peaks from the low frequency band are also more likely to be concentrated in high SNR bins compared to the middle and high frequency bands.

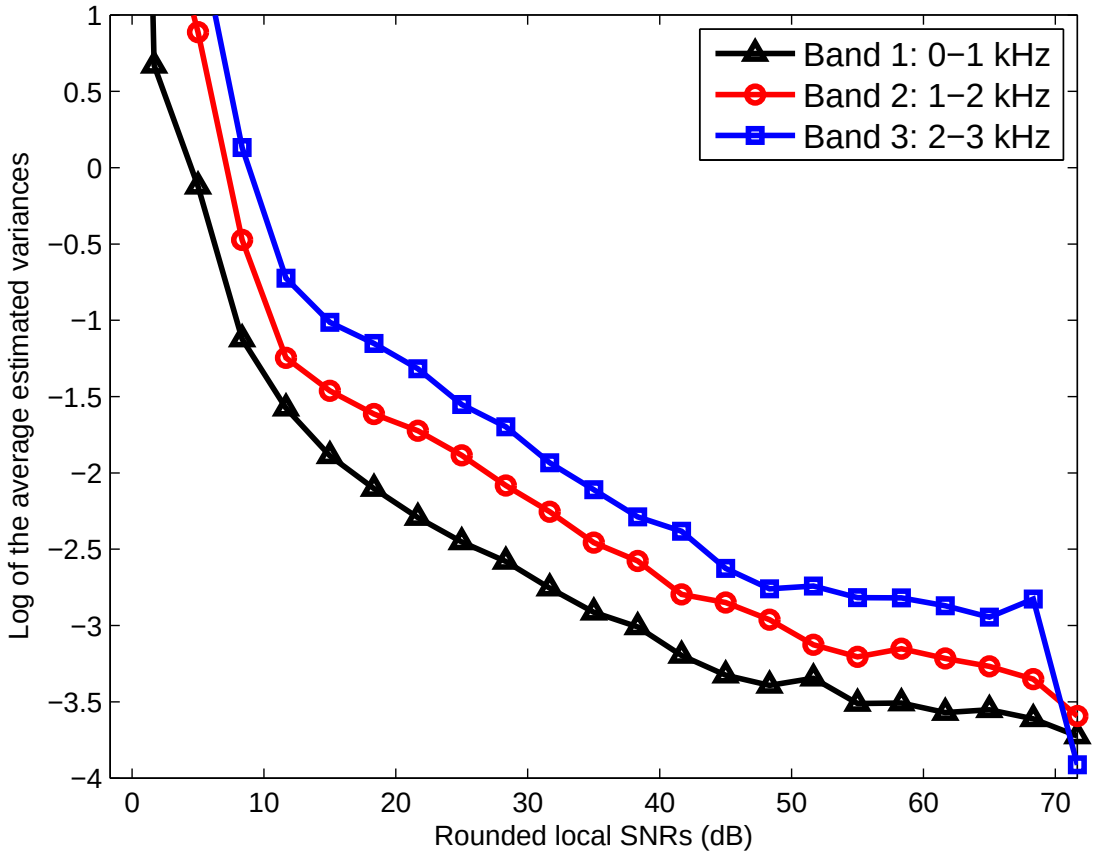


Fig. 6. A comparison of the log of the averaged estimated variances of the residual distributions under different frequency bands (low, middle, high). The noise condition is white noise. Estimated variances from different noise levels (clean, 20 dB, 10 dB, 5 dB, 0 dB, -5 dB) are averaged.

D. Post-Processing

For an utterance, the posterior probabilities, $P(f_0|\mathbf{Y},\mathbf{N})$, for each frame are obtained by calculating (4). Then, a dynamic programming approach, the same as that used in RAPT, was used to smooth the tracked F0 contour and to allow octave jumps at a certain cost [6]. A brief description of the dynamic programming is as follows.

The objective of dynamic programming is to search for an F0 contour that minimizes an objective function. Given an F0 contour, the objective function is defined as a summation of the frame-level local cost and transition cost functions. The local cost function for a certain frequency at one frame is inversely related to the F0 likelihood value. The inter-frame F0 transition cost function is defined under

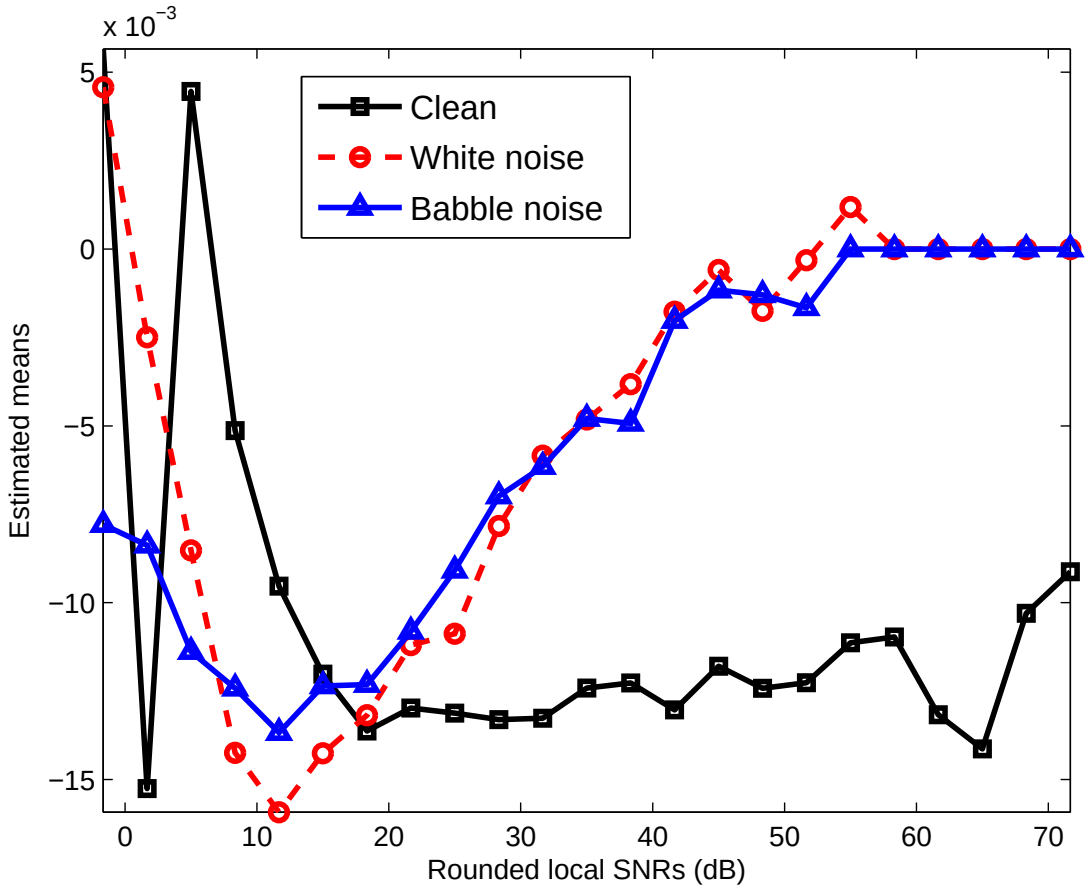


Fig. 7. A comparison of the averaged estimated means of the distributions of residuals under different noise conditions using the KEELE corpus. Estimated means from different noise levels (clean, 20 dB, 10 dB, 5 dB, 0 dB, -5 dB) and different frequency bands (low, middle, high) are averaged.

4 conditions: voiced-to-voiced (V→V), unvoiced-to-unvoiced (U→U), voiced-to-unvoiced (V→U), and unvoiced-to-voiced (U→V). In the V→V condition, the cost function is defined as an increasing function of inter-frame proportional frequency change, but allows for octave jumps at some specifiable cost. In the U→U condition, the cost function is defined as 0. In the V→U or U→V conditions, the cost function is defined as a combination of a spectral stationarity function and the inverse function of the Itakura distortion [30].

The focus of the proposed method is to reduce F0 estimation error under both clean and noisy conditions. However, voicing boundaries can affect the results of F0 tracking [31]. Hence, each F0 tracking algorithm is forced to estimate F0 values over all the voiced frames regardless of the SNRs.

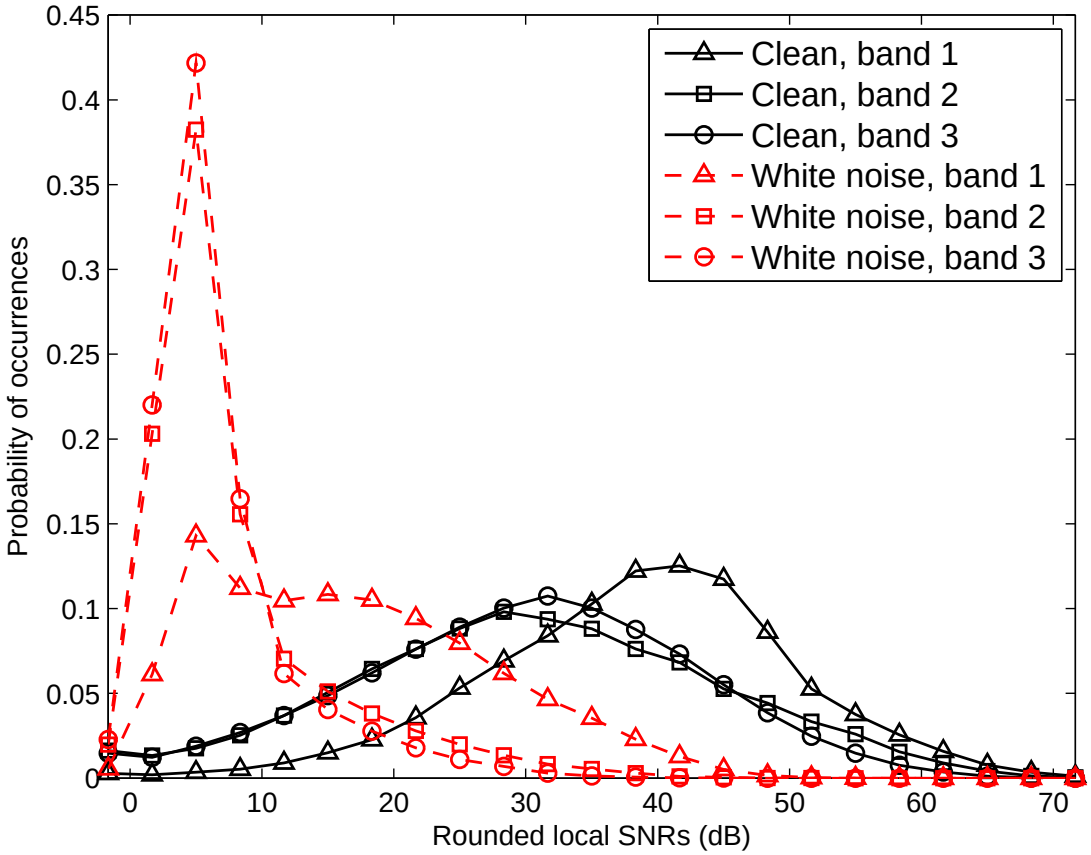


Fig. 8. A comparison of the distributions of rounded local SNRs under different frequency bands (low-1, middle-2, high-3). The noise condition is white noise. The distributions under different noise levels (20 dB, 10 dB, 5 dB, 0 dB, -5 dB) are averaged.

The F0 trackers (RAPT, Praat, TEMPO, WWB) also output voiced/unvoiced decisions. If the ground truth and the F0 tracker agree that a frame is voiced or unvoiced, the F0 value is not changed. If a ground truth unvoiced frame is assumed to be voiced, the F0 value is set to be 0. If a ground truth voiced frame N_c is assumed to be unvoiced, f_{0N_c} is estimated by using an interpolation-based method:

$$f_{0N_c} = f_{0N_l} + \frac{N_c - N_l}{N_r - N_l}(f_{0N_r} - f_{0N_l}), \tag{20}$$

where N_l and N_r denote the left and right closest frame to the current frame N_c among the frames that both the ground truth and F0 tracker agree to be voiced. One exception of this interpolation is that if frame N_c is in the first or last assumed unvoiced segment by the F0 tracker in a ground truth voiced segment, the f_{0N_c} is set to be either f_{0N_r} or f_{0N_l} depending on whether the right or left frame is closer.

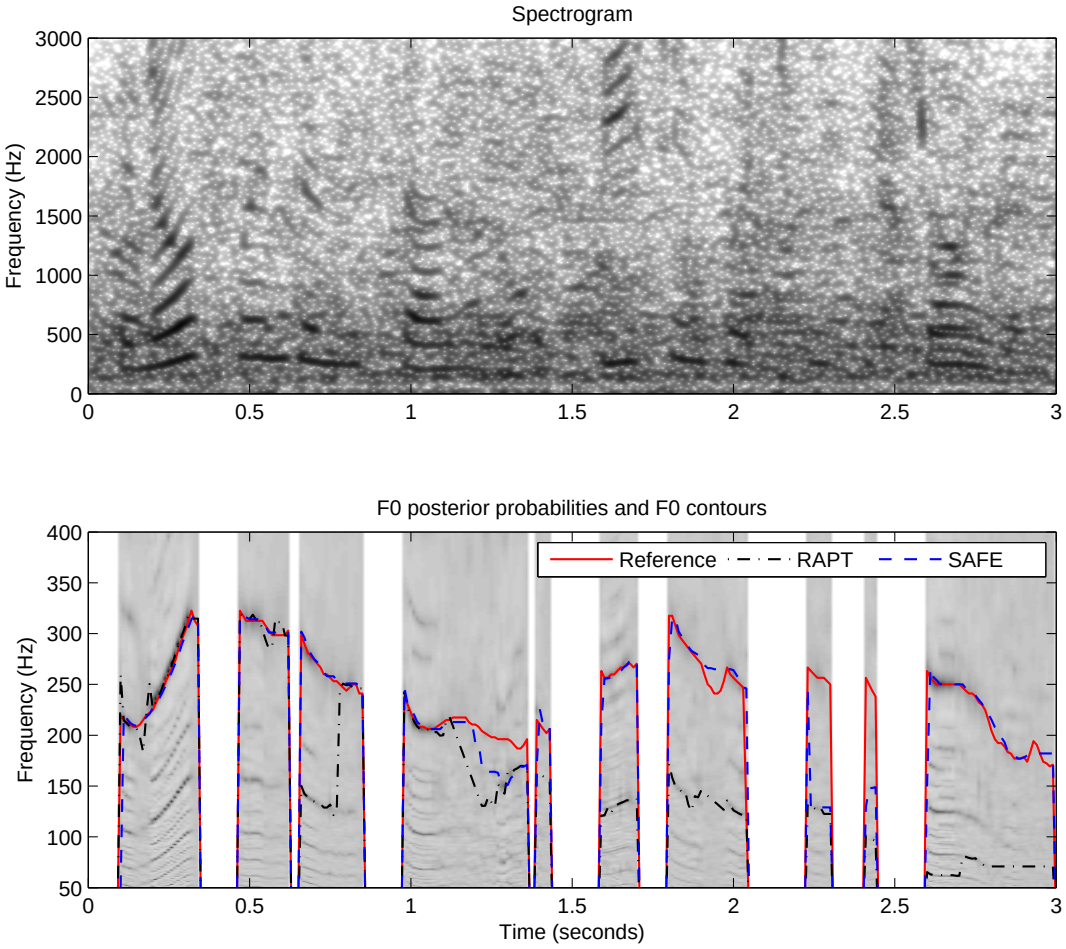


Fig. 9. The spectrogram, F0 posterior probabilities from SAFE, and F0 contours from RAPT and SAFE of a segment of an utterance from the second female speaker (f2nw0000) in the KEELE corpus under babble noise condition at 0 dB SNR.

An example of F0 estimation made with SAFE is shown in Fig. 9. The segment corresponds to the beginning of the utterance of the second female speaker in the KEELE corpus. The noise condition is babble noise at 0 dB SNR. Each vertical strip in the bottom panel shows the F0 posterior probabilities over the voiced frame. The darker a point is, the higher the probability that F0 corresponds to that frequency. Since RAPT has the lowest GPE among all the F0 estimators, and SAFE uses the same cost function as RAPT for the dynamic programming-post processing, only the tracked F0 of RAPT and SAFE are shown in Fig. 9. It can be observed from the spectrogram in the top panel that the babble noise is mostly concentrated on the low frequency band. The babble noise can corrupt the harmonic structure of the

voiced frame by suppressing or shifting the spectral peaks in the original clean speech, or by inserting new harmonic structures of the periodic babble noise into the original spectrum. The new added strong harmonic structures from babble noise may cause estimation errors. For some regions in which the target speech has high energy at high frequencies, e.g., around 1.6 s, the prominent peaks in the middle and high frequency bands, which are less affected by noise, can be used to infer the F0 value.

III. EXPERIMENTS

Three error metrics: Gross Pitch Error (GPE), Mean of the Fine Pitch Errors (MFPE), and Standard Deviation of the Fine Pitch Errors (SDFPE) are used [2] to evaluate the performance of F0 estimation algorithms. Let N_{VV} denote the number of the frames that both the F0 tracker and the ground truth consider to be voiced, VV means 'both voiced'; and N_{GE} represent the number of frames for which

$$|\frac{f_{0i,estimated}}{f_{0i,reference}} - 1| > \epsilon, \quad i = 1, \cdots, N_{VV}, \tag{21}$$

where i is the frame index, and ϵ is a threshold which is typically 20%, GE means 'gross error'. The number of remaining frames, denoted by N_{FE}, is equal to $N_{VV} - N_{GE}$. FE means 'fine error'.

GPE is defined as:

$$GPE = \frac{N_{GE}}{N_{VV}} \times 100\%. \tag{22}$$

MFPE denoted by μ_{FPE} is defined as:

$$\mu_{FPE} = \frac{1}{N_{FE}} \sum_{i \in \mathcal{S}_{FE}} (f_{0i,estimated} - f_{0i,reference}), \tag{23}$$

where $\mathcal{S}_{FE}$ denotes the set of all the frames in which no gross error occurs.

SDFPE denoted by σ_{FPE} is defined as:

$$\sigma_{FPE} = \frac{1}{N_{FE}} \sum_{i \in \mathcal{S}_{FE}} f_{0i,estimated}^2 - \mu_{FPE}, \tag{24}$$

where MFPE and SDFPE are used to measure the bias and precision of the F0 estimation when no gross estimation error is occurred.

In this section, we compare the GPE, MFPE, and SDFPE using the KEELE [29] and CSTR [32] corpora. The 5 minute 37 seconds KEELE corpus contains a simultaneous recording of speech and laryngograph signals for a phonetically-balanced text which was read by 5 male and 5 female speakers. The 5 minute 32 seconds CSTR corpus is composed of laryngograph and speech signals from one male and one female speaker. Each speaker read 50 sentences in the CSTR corpus. Ground truth F0s were

obtained by running an autocorrelation method on the laryngograph signal in addition to some manual correction.

Speech signals are downsampled from 20000 Hz to 16000 Hz for both corpora. Noise is artificially added to the corpora to test the robustness of the F0 trackers under different noise conditions. The program FaNT [33] with the default command line option (-u -m snr_8khz) was used to employ white and babble noise segments from the NOISEX92 [34] corpus to the speech signals to generate utterances with SNR of 20, 10, 5, 0, and -5 dB. The white noise is acquired by sampling high-quality analog noise generator. The babble noise is acquired by recording 100 people speaking in a canteen with room radius over 2m.

The parameters of SAFE are as follows: FFT size is 16384; frequency resolution is 1 Hz; frame length and step size are 0.04 and 0.01 seconds, respectively; f_{0min} and f_{0max} are 50 and 400 Hz, respectively; the lengths of the short-term and long-term windows for spectrum smoothing are 50 and 400 in Hz, respectively. A peak is regarded as a prominent peak if the normalized difference SNR $\bar{\zeta}_i$ is greater than an empirically determined threshold of 0.33; the ranges of the low, middle, and high frequency bands are 0-1, 1-2, and 2-3 kHz, respectively; local SNRs of the peaks are rounded to the nearest value in the following sequence $10r/3$, where $r = 0, 1, \cdots, 21$. The weighting factors in (4) are all set to the reciprocal of the number of the prominent peaks in that frame.

For the KEELE corpus, a 5-fold cross-validation scheme is applied. For each fold under a certain noise level, the speech of one male and one female speaker are used for testing, the residual and SNR models are trained from the remaining speech and its ground truth. Since 54% of the KEELE corpus is voiced speech, if the frame step size is 0.01 seconds, each fold has about 14000 frames for training. Since there are 23 rounded local SNR bins, if each voiced frame has 10 prominent peaks on average, each residual model has about 6000 samples for training. Because some bins with high SNRs might have fewer training instances, e.g., 5% of the average - 300 samples, it is still possible to robustly train a doubly-truncated Laplacian distribution with only two free parameters.

A comparison of the GPEs of RAPT, Praat, TEMPO, YIN, Wu et al.'s method (WWB), and SAFE on the KEELE corpus is shown in Table I. Note that Yegnanarayana et al.'s [11] results are not included, because silence was added to the KEELE corpus in their experiments. There are two configurations of Praat: autocorrelation (default) or cross-correlation. The cross-correlation configuration is used, since it consistently provided better results. The default settings were used for RAPT, Praat, TEMPO, YIN, and WWB, except that the voicing thresholds were optimized. The implementation of WWB was provided by Prof. Dan Ellis and his group at Columbia University. Three configurations of SAFE were compared: standard (SAFE), only with information from the low frequency band as the prevailing F0 tracking

TABLE I

THE GPE (%) OF THE RAPT, PRAAT, TEMPO, YIN, WWB, AND SAFE USING THE KEELE AND CSTR CORPORA. **LFB**: ONLY THE LOW FREQUENCY BAND (0-1000 HZ) IS USED. μ**=0**: A ZERO MEAN IS USED IN THE DOUBLY TRUNCATED LAPLACIAN DISTRIBUTION. BOLD NUMBERS REPRESENT THE LOWEST GPE IN EACH COLUMN.

SNR (dB)	Clean	20	10	5	0	-5
		KEELE White Noise				
RAPT	**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
WWB	4.22	4.27	5.21	5.57	6.42	8.87
SAFE (LFB)	3.13	3.09	3.74	4.39	4.72	6.29
SAFE ($\mu = 0$)	3.00	3.04	3.38	3.71	4.10	5.16
SAFE	2.98	3.01	3.35	**3.66**	**4.06**	**5.01**
		KEELE Babble Noise				
RAPT		**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
WWB		6.76	12.48	21.20	32.84	55.40
SAFE (LFB)		3.23	6.01	10.21	20.64	47.21
SAFE ($\mu = 0$)		3.14	4.75	7.68	16.23	39.62
SAFE		3.10	**4.72**	**7.44**	**15.88**	**39.23**
		CSTR White Noise				
RAPT	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
WWB	2.75	3.00	4.00	4.83	5.35	7.64
SAFE (LFB)	2.49	2.52	2.97	3.49	3.93	4.14
SAFE ($\mu = 0$)	2.40	2.41	2.69	**3.10**	**3.24**	**3.68**
SAFE	2.45	2.46	2.73	3.25	3.34	3.76
		CSTR Babble Noise				
RAPT		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
WWB		4.82	14.15	30.09	49.05	66.00
SAFE (LFB)		2.69	5.37	9.97	23.59	63.20
SAFE ($\mu = 0$)		2.61	**4.14**	**7.73**	**19.32**	**57.17**
SAFE		2.63	4.23	8.23	20.74	59.54

algorithms (SAFE (LFB)), and with zero mean residual estimation (SAFE ($\mu = 0$)). It can be seen that all F0 trackers have GPEs lower than 3.5% in quiet. All algorithms suffer from performance degradation when the SNR drops. As expected, it is more difficult to accurately estimate F0 in the babble noise condition compared to the white noise condition with the same SNR. The SAFE algorithm has the lowest GPE when the SNR is at or below 5 dB under white noise, or at or below 10 dB under babble noise. It can be concluded from Table I that discarding information from middle and high frequency bands can cause an increase in GPE, especially for babble noise which is usually concentrated at low frequencies. Forcing the means of the estimated residual distributions to be zero can also result in an increase in GPE. The performance of the SAFE algorithm using non-prominent peaks is not tested, because not only the non-prominent peaks are negative factors in F0 estimation, but also the proposed doubly truncated Laplacian distribution is not supposed to fit the saddle-like distributions of non-prominent peaks as shown in Fig. 5.

To determine the generalizablity of SAFE, the model trained from the KEELE corpus is used for the CSTR corpus. According to the performances of the F0 algorithms shown in the Table I, it can be seen that F0 estimation for the CSTR corpus is easier under white noise, but harder under babble noise compared to the KEELE corpus. Although there is mismatch between the KEELE and CSTR corpora, SAFE still has the lowest GPE under low SNR conditions for both. The mismatch can explain why SAFE ($\mu = 0$) has a lower GPE compared to the standard SAFE. Thus, it may be more appropriate to use SAFE ($\mu = 0$) when prior information of the testing set is not available.

The MFPEs for the KEELE and CSTR corpora are shown in Table II. It can be seen that the best configuration of SAFE has less than 1 Hz MFPEs under all noise conditions. Other F0 trackers have less than 3 Hz MFPEs under most noise conditions. Note that 3 Hz is only 1.2% of the average of all possible F0s which is 225 Hz. That means all F0 trackers do not make significantly biased F0 estimation under clean and most noisy conditions. For the KEELE corpus, the means of the residuals are slightly less than zero most of the time as shown in Fig. 7. Thus, the standard SAFE which considers the bias is supposed to have slightly lower F0 estimation than the zero mean version of SAFE. Due to the mismatch between KEELE and CSTR corpora, the negative bias causes the MFPEs of the standard SAFE to be more deviated from zero compared to the zero mean version of the SAFE on the CSTR corpus.

The SDFPEs on KEELE and CSTR corpus are shown in Table III. It can be seen that the SDFPEs of SAFE are slightly higher (1-2 Hz) than other F0 estimators under some conditions. Since the MFPE and SDFPE are calculated over the frames in which the F0 tracker does not have gross F0 estimation errors (less than 20% gross error), the number of frames for calculating the SDFPE over different F0

TABLE II

THE MFPE (HZ) OF THE RAPT, PRAAT, TEMPO, YIN, WWB, AND SAFE USING THE KEELE AND CSTR CORPORA. **LFB**: ONLY THE LOW FREQUENCY BAND (0-1000 HZ) IS USED. μ**=0**: A ZERO MEAN IS USED IN THE DOUBLY TRUNCATED LAPLACIAN DISTRIBUTION.

SNR (dB)	Clean	20	10	5	0	-5
		KEELE White Noise				
RAPT	0.79	0.60	0.60	0.32	-0.18	-1.87
Praat	0.19	0.21	-0.14	0.67	-1.93	-4.08
TEMPO	0.41	0.36	0.27	0.08	-1.26	-2.16
YIN	0.55	0.56	0.54	0.53	0.53	0.43
WWB	2.86	2.87	2.74	2.67	2.35	2.05
SAFE (LFB)	-0.40	-0.40	-0.43	-0.47	-0.66	-0.61
SAFE ($\mu = 0$)	0.15	0.34	0.34	0.28	0.04	-0.05
SAFE	-0.36	-0.46	-0.50	-0.57	-0.72	-0.86
		KEELE Babble Noise				
RAPT		0.74	0.47	0.23	-0.35	-0.24
Praat		0.24	0.21	0.05	0.16	0.50
TEMPO		0.34	-0.06	-1.19	-0.09	1.22
YIN		0.66	0.83	0.93	1.11	1.03
WWB		2.66	2.34	1.95	1.35	0.89
SAFE (LFB)		-0.42	-0.52	-0.51	-0.33	0.10
SAFE ($\mu = 0$)		0.21	0.04	-0.12	-0.19	-0.12
SAFE		-0.49	-0.65	-0.78	-0.71	-0.47
		CSTR White Noise				
RAPT	-0.06	-0.27	-0.22	-0.31	-0.59	-2.07
Praat	-0.77	-0.78	-0.97	-1.34	-2.79	-4.71
TEMPO	-0.85	-0.73	-0.76	-0.97	-1.21	-2.66
YIN	-0.39	-0.40	-0.44	-0.47	0.60	-0.62
WWB	2.73	2.67	2.49	2.34	2.19	1.93
SAFE (LFB)	-1.28	-1.32	-1.39	-1.40	-1.45	-1.53
SAFE ($\mu = 0$)	-0.78	-0.53	-0.50	-0.53	-0.62	-0.81
SAFE	-1.39	-1.43	-1.46	-1.49	-1.59	-1.69
		CSTR Babble Noise				
RAPT		-0.19	-0.34	-0.18	-0.35	-0.14
Praat		-0.79	-0.72	-0.44	-0.30	0.13
TEMPO		-0.54	-0.71	-0.77	0.51	0.69
YIN		-0.36	-0.14	-0.06	0.04	0.28
WWB		2.05	1.55	1.24	0.77	0.34
SAFE (LFB)		-1.40	-1.45	-1.39	-1.13	-0.47
SAFE ($\mu = 0$)		-0.65	-0.78	-0.81	-0.92	-0.42
SAFE		-1.46	-1.55	-1.52	-1.40	-0.69

TABLE III

THE SDFPE (HZ) OF THE RAPT, PRAAT, TEMPO, YIN, WWB, AND SAFE USING THE KEELE AND CSTR CORPORA. **LFB**: ONLY THE LOW FREQUENCY BAND (0-1000 HZ) IS USED. μ**=0**: A ZERO MEAN IS USED IN THE DOUBLY TRUNCATED LAPLACIAN DISTRIBUTION.

SNR (dB)	Clean	20	10	5	0	-5
		KEELE White Noise				
RAPT	4.41	4.50	4.75	5.54	6.62	9.92
Praat	3.69	3.71	4.89	6.05	8.96	12.74
TEMPO	5.04	5.19	5.84	7.25	9.43	11.52
YIN	4.45	4.47	4.60	4.82	5.21	5.59
WWB	5.65	5.59	5.61	5.75	6.02	6.82
SAFE (LFB)	5.63	5.62	5.63	5.70	5.99	6.48
SAFE ($\mu = 0$)	5.48	5.49	5.51	5.54	5.95	6.43
SAFE	5.53	5.56	5.62	5.67	6.07	6.50
		KEELE Babble Noise				
RAPT		4.85	5.96	6.83	8.57	9.39
Praat		3.85	4.86	5.79	7.03	8.86
TEMPO		5.92	9.06	12.01	13.29	11.79
YIN		4.71	5.30	5.81	6.76	7.94
WWB		5.62	6.08	6.61	7.17	8.15
SAFE (LFB)		5.60	6.09	6.67	7.65	8.40
SAFE ($\mu = 0$)		5.56	6.04	6.64	7.48	9.35
SAFE		5.60	6.07	6.70	7.58	9.34
		CSTR White Noise				
RAPT	5.49	5.78	6.02	6.57	7.92	10.67
Praat	6.04	6.09	6.54	7.56	10.22	14.38
TEMPO	6.76	7.28	7.74	8.55	10.41	13.29
YIN	6.28	6.29	6.35	6.46	6.68	6.75
WWB	6.86	6.83	6.79	6.90	7.09	7.61
SAFE (LFB)	8.10	8.00	7.97	7.93	7.92	8.31
SAFE ($\mu = 0$)	7.85	7.81	7.82	7.80	7.89	8.19
SAFE	7.89	7.85	7.74	7.71	7.87	8.19
		CSTR Babble Noise				
RAPT		5.84	6.86	7.47	7.84	8.78
Praat		6.06	6.36	6.45	6.85	7.87
TEMPO		7.76	10.86	13.96	14.98	12.50
YIN		6.33	6.25	5.96	5.59	6.25
WWB		6.14	5.84	5.85	5.74	6.46
SAFE (LFB)		8.03	8.12	8.19	8.34	7.09
SAFE ($\mu = 0$)		7.64	7.97	8.19	8.55	7.62
SAFE		7.59	7.91	8.16	8.49	7.49

trackers under the same noise condition is different. It is known that F0 estimation accuracy is higher over less noisy frames [31]. Given a certain noise condition, if an estimator only correctly estimates F0 over a few frames that have high frame-level SNRs, it could have relatively low MFPE and SDFPE, but a high GPE. Therefore, having higher MFPEs or SDFPEs does not necessarily mean that SAFE is less accurate in F0 estimation.

IV. Conclusions

Prominent Signal-to-Noise Ratio (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 the additive noise on the clean spectra given F0. 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 robust in maintaining low Mean and Standard Deviation of the Fine Pitch Errors.

V. Acknowledgment

The authors would thank Prof. Hideki Kawahara for providing the TEMPO package, Dr. Georg Meyer for providing the KEELE corpus, and Prof. Dan Ellis and his group for providing the implementation of Wu et al.'s method.

References

[1] Gunnar Fant, *Acoustic Theory of Speech Production*, Mouton, 1960.

[2] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[3] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, 1983.

[4] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367–377, 1972.

[5] A.M. Noll, "Cepstrum pitch determination," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293–309, 1967.

[6] D. Talkin, "Robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp. 497–518, 1995.

[7] K. Kasi and S. Zahorian, "Yet another algorithm for pitch tracking," *ICASSP*, 2002, vol. 1, pp. 361–364.

[8] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[9] H. Kawahara, H. Katayose, A de Chevigne, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," *EUROSPEECH*, 1999, vol. 6, pp. 2781–2784.

[10] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, 2002.

[11] B. Yegnanarayana and K.S.R. Murty, "Event-based instantaneous fundamental frequency estimation from speech signals," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[12] J. Le Roux, H. Kameoka, N. Ono; A. de Cheveigne, and S. Sagayama, "Single and multiple F0 contour estimation through parametric spectrogram modeling of speech in noisy environments," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1135–1145, 2007.

[13] B. Gold and L.R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *The Journal of the Acoustical Society of America*, vol. 46, no. 2B, pp. 442–448, 1969.

[14] M. Lahat, R. Niederjohn, and D. Krusback, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 35, no. 6, pp. 741–750, 1987.

[15] F. Sha, J. Burgoyne, and L. Saul, "Multiband statistical learning for F0 estimation in speech," *ICASSP*, 2004, vol. 5, pp. 661–664.

[16] J.C.R. Licklider, "A duplex theory of pitch perception," *Experientia*, vol. 23, no. 4, pp. 128–134, 1951.

[17] P. Hedelin and D. Huber, "Pitch period determination of aperiodic speech signals," *ICASSP*, 1990, vol. 1, pp. 361–364.

[18] A. de Cheveigne, "Speech F0 extraction based on Licklider's pitch perception model," *ICPhS*, 1991, pp. 218–221.

[19] M. Wu, D. Wang, and G.J. Brown, "A multipitch tracking algorithm for noisy speech," *IEEE Trans. on Speech and Audio Processing*, vol. 11, no. 3, pp. 229–241, 2003.

[20] D. Krusback and R. Niederjohn, "An autocorrelation pitch detector and voicing decision with confidence measures developed for noise-corrupted speech," *IEEE Trans. on Signal Processing*, vol. 39, no. 2, pp. 319–329, 1991.

[21] T. Shimamura and H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 7, pp. 727–730, 2001.

[22] T. Abe, T. Kobayashi, and S. Imai, "Robust pitch estimation with harmonics enhancement in noisy environments based on instantaneous frequency," *ICSLP*, 1996, pp. 1277–1280.

[23] D.-J. Liu and C.-T. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 6, pp. 609–621, 2001.

[24] T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," *JASA*, vol. 116, no. 6, pp. 3690–3700, 2004.

[25] O. Deshmukh, C.Y. Espy-Wilson, A. Salomon, and J. Singh, "Use of temporal information: Detection of periodicity, aperiodicity, and pitch in speech," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 776–786, 2005.

[26] G.S. Ying, L.H. Jamieson, and C.D. Michell, "A probabilistic approach to AMDF pitch detection," *ICSLP*, 1996, vol. 2, pp. 1201–1204.

[27] Y.R. Wang, I.J. Wong, and T.C. Tsao, "A statistical pitch detection algorithm," *ICASSP*, 2002, vol. 1, pp. 357–360.

[28] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," pp. 347–354, 1997.

[29] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," *EUROSPEECH*, 1995, pp. 837–840.

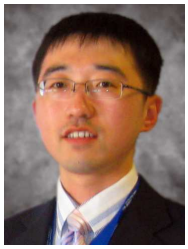
[30] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[31] W. Chu and A. Alwan, "Reducing F0 frame error of F0 tracking algorithms under noisy conditions with an unvoiced/voiced classification frontend," *ICASSP*, 2009, pp. 3969–3972.

[32] P.C. Bagshaw, S.M. Hiller, and M.A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," *EUROSPEECH*, 1993, vol. 2, pp. 1003–1006.

[33] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *ASR2000*, 2000, pp. 181–188.

[34] A.P. Varga, H. J. M. Steeneken, M. Tomlinson, and D. Jones, "The NOISEX-92 study on the effect of additive noise on automatic speech recognition," *Technical report, DRA Speech Research Unit*, 1992.



Wei Chu (S'06) will earn his Ph.D. degree in Electrical Engineering from the University of California, Los Angeles at the end of 2011. His thesis work is on noise robust signal processing for human pitch tracking (http://www.ee.ucla.edu/~weichu/safe) and bird song recognition (http://www.ee.ucla.edu/~weichu/bird). In 2007, he earned his Master's degree in Electronic Engineering from the Tsinghua University, where he developed a real-time speech recognition system with a non-speech input rejection frontend on chip. Since 2005, he has collaborated with researchers from Microsoft Research Redmond and Asia, Disney Research, Rosetta Stone, Mitsubishi Electronic Research Laboratory, and Intel on audio-visual and noise robust speech recognition, kid speech analysis and recognition, pronunciation modeling, acoustic modeling, and speaker detection and clustering. Since 2007, He has been serving as a reviewer for IEEE Transactions on Audio, Speech, and Language Processing; Computer Speech and Language; ICASSP; Interspeech; and Automatic Speech Recognition and Understanding workshop. Currently, his research interests include speech recognition, speech perception, statistical signal processing, and machine learning.

Mr. Chu is the receipient of UCLA Continuous Student Fellowship (2008) and University Fellowship (2007), Tsinghua University Scholarship (2006), and IEEE and ISCA travel grants (2009-2010). He is also a member of Tau Beta Pi.

Abeer Alwan (M’85-SM’00-F’08) received her Ph.D. in EECS from MIT in 1992. Since then, she has been with the Electrical Engineering Department at UCLA where she is now a Full Professor. She established and directs the Speech Processing and Auditory Perception Laboratory at UCLA (http://www.ee.ucla.edu/~spapl). Her research interests include modeling human speech production and perception mechanisms and applying these models to improve speech-processing applications such as noise-robust automatic speech recognition. She is the recipient of the NSF Research Initiation and Career Awards, the NIH FIRST Career Development Award, the UCLA-TRW Excellence in Teaching Award, and the Okawa Foundation Award in Telecommunications.

Dr. Alwan is an elected member of Eta Kappa Nu, Sigma Xi, Tau Beta Pi, and the New York Academy of Sciences. She served, as an elected member, on the Acoustical Society of America Technical Committee on Speech Communication, on the IEEE Signal Processing Technical Committees on Audio and Electroacoustics and on Speech Processing. She was a member and then Chair of the Flanagan Speech and Audio Processing Award Committee. She is a member of the Editorial Board of Speech Communication and was a co-editor-in-chief of that journal, was an Associate Editor (AE) of the IEEE Transactions on Speech, Audio, and Language Processing, and is an AE for the Journal of the Acoustical Society of America. Dr. Alwan is a Fellow of the IEEE, the Acoustical Society of America, and the International Speech Communication Association (ISCA). She was a 2006-2007 Fellow of the Radcliffe Institute for Advanced Study at Harvard University, and a Distinguished Lecturer for ISCA.