Performance Evaluation of IFAS-based Fundamental Frequency Estimator in Noisy Environment

Dhany Arifianto, Takao Kobayashi

Interdisciplinary Graduate School of Science and Engineering
Tokyo Institute of Technology
4259 Nagatsuta, Midori-ku, Yokohama, Japan, 226-8502
{dany.arifianto, takao.kobayashi}@ip.titech.ac.jp

Abstract

In this paper, instantaneous frequency amplitude spectrum (IFAS)-based fundamental frequency estimator is evaluated with speech signal corrupted by additive white gaussian noise. A key idea of the IFAS-based estimator is the use of degree of regularity of periodicity in spectrum of speech signal, defined by a quantity called harmonicity measure, for band selection in the fundamental frequency estimation. Several frequency band and window length selection methods based on harmonicity measure are asessed to find out better performance. It is shown that the performance of the IFAS-based estimator is maintained at constant error rate about 1% from clean speech data up to 15 dB and about 11% at 0 dB SNR. For both female and male speakers, the IFAS-based estimator outperforms several well-known methods particularly at 0 dB SNR.

1. Introduction

Time-varying characteristic of speech signal makes accurate estimation of its fundamental frequency (F_0) difficult to achieve. Additionally, noisy environment usually lowers the performance of the estimator because of obscure structure of harmonic components. Numerous methods for the fundamental frequency estimation for overcoming this problem, in either time-domain or frequency-domain, have been proposed in the texbooks and papers [1]-[3], including references cited therein.

The notion of instantaneous frequency (IF), originated from time-frequency analysis, has the ability to describe the frequency at any instant time which is defined as phase derivative with respect to time. Abe *et.al* [4] reported a harmonic tracking method based on IF derived from bandpass filter representation and showed its effectiveness in F_0 estimation. Furthermore, the authors introduced an idea of instantaneous frequency amplitude spectrum (IFAS) and showed its application to F_0 estimation which achieved superior performance in both clean and noisy speech data to cepstrum method [5]. In [6], it was also reported the use of wavelet-based IF analysis where the carrier-to-noise ratio was used to determine a fixed point in IF to filter output relation to estimate F_0. An approach of F_0 estimation based on IF with multi-talker noise and spectral distortion was reported by defining a measure called degree of dominance which essentially an evaluation of magnitude of harmonic components using prescribed relation to the other components in STFT [7]. Recently, we proposed a refinement of IFAS-based F_0 estimation method in which a quantity called harmonicity measure is introduced and lower error rate was obtained for clean speech compared to the original IFAS-based method, even without F_0 contour smoothing [8].

In this paper, we present performance evaluation of the IFAS-based F_0 estimator in noisy environment. The instantaneous frequency amplitude spectrum is defined from short-time Fourier transform of a signal as a function in time and frequency and a better representation showing harmonic structure than STFT amplitude spectrum can be obtained. The key idea of the method is the use of the harmonicity measure which provides degree of regularity of periodicity. Optimum frequency band and/or window length are chosen in such a way that the harmonicity measure is maximized to obtain accurate and reliable F_0 estimates. Then, we obtain the F_0 estimates after calculating the likelihood using IFAS.

The IFAS is revisited briefly in the second section, then subsequently followed by harmonicity measure definition [8]. The third section elaborates implementation of the explained procedures for both F_0 estimation of additive white noise to speech signal with several schemes of frequency band and window length selection. To demonstrate its effectiveness, the results of the proposed method with several experimental conditions and performance comparison are discussed respectively.

2. IFAS and Harmonicity Measure

Let $x(t)$ and $X(\omega, t)$ be a function which represents speech signal and its short-time Fourier transform (STFT), respectively. The STFT of $x(t)$ is written in the form

$$\begin{aligned} X(\omega, t) &= e^{-j\omega t} \int_{-\infty}^{\infty} w(\tau - t) x(\tau) e^{-j\omega(\tau - t)}\, d\tau \quad (1) \\ &= e^{-j\omega t} G(\omega, t), \quad (2) \end{aligned}$$

where $w(t)$ is an analysis window function. If the Fourier transform of $w(t)$ is a lowpass function, then $G(\omega, t)$ will be the output of a bandpass filter whose impulse response is $w(-t)e^{j\omega t}$ [9]. The output $G(\omega, t)$ contains only non-negative frequency components which consequently implies that it is analytic.

The instantaneous frequency at frequency ω and at instant time t is defined by

$$\begin{aligned}\lambda(\omega, t) &= \frac{\partial}{\partial t}\arg[G(\omega, t)] \\ &= \omega + \frac{\partial}{\partial t}\arg[X(\omega, t)]. \qquad (3)\end{aligned}$$

The following expression will be used to calculate instantaneous frequency

$$\frac{\partial}{\partial t}\arg[X(\omega, t)] = \frac{a\frac{\partial b}{\partial t} - b\frac{\partial a}{\partial t}}{a^2 + b^2}, \qquad (4)$$

$$\frac{\partial}{\partial t}[X(\omega, t)] = \int_{-\infty}^{\infty} -\psi(\tau - t)e^{-j\omega\tau}x(\tau)\, d\tau, \qquad (5)$$

where $X(\omega, t) = a + jb$ and $\psi(t)$ is the derivative of analysis window $w(t)$ with respect to time.

In the following, it is considered that all derivations are at instant t, and t will be omitted for notation simplicity. Let $S(\lambda_0)$ be the IFAS at the instantaneous frequency λ_0 defined by the following equation [5]

$$S(\lambda_0) = \lim_{\Delta\lambda \to 0} \frac{1}{\Delta\lambda} \int_{\Omega_0} |G(\omega)|\, d\omega, \qquad (6)$$

where $\Omega_0 = \{\omega | \lambda_0 \leq \lambda(\omega, t) \leq \lambda_0 + \Delta\lambda\}$.

We define a transform function

$$\eta(F) = \alpha^{-\frac{\beta}{F}} \int_{\lambda_l}^{\lambda_u} S(\lambda)\Lambda(\lambda, F)\, d\lambda, \qquad (7)$$

where α and β are real constants and

$$\Lambda(\lambda, F) = \begin{cases} 0, & \lambda/F < \pi \\ \frac{1}{2}\big(\cos(\lambda/F) + 1\big), & \lambda/F \geq \pi. \end{cases} \qquad (8)$$

In (7), λ_l and λ_u are lower and upper bounds of IF band respectively. If the signal is periodic and $S(\lambda)$ shows harmonic structure with a fundamental frequency, F_0, then $\eta(F)$ has local maxima at the frequencies $F = F_0/n,\ n = 1, 2, \ldots$. Hence, the value of $\eta(F)$ can be considered to be likelihood where the fundamental frequency of the signal will be F. In (7), the term $\alpha^{-\beta/F}$ is a weighting constant to give priority to higher fundamental frequencies. The interval of the integral $[\lambda_l, \lambda_u]$ in (7) is selected such that the reliability of the likelihood becomes higher.

Consider an interval $[\lambda_l, \lambda_u]$ on the IF axis λ, and let Ω be a set of intervals on the frequency axis such that $\lambda_l \leq \lambda(\omega) \leq \lambda_u$. A harmonicity evaluation function is defined as follows

$$\xi_{\lambda_l,\lambda_u}(F) = \frac{1}{m(\Omega)} \int_{\Omega} C(\lambda(\omega), F)\, d\omega, \qquad (9)$$

where $m(\Omega)$ be the measure of Ω in Lebesgue's sense, i.e., the total length of intervals, and

$$C(\lambda(\omega), F) = \begin{cases} 0, & \lambda(\omega)/F < \pi/2 \\ \cos(\lambda(\omega)/F), & \lambda(\omega)/F \geq \pi/2. \end{cases}$$

We also define harmonicity measure in the instantaneous frequency domain by

$$P_{\lambda_l,\lambda_u} = \max_F \xi_{\lambda_l,\lambda_u}(F). \qquad (11)$$

The degree of regularity of periodicity in spectrum, called harmonicity measure, lies somewhere between $-1 \leq P_{\lambda_l,\lambda_u} \leq 1$.

Fundamental frequency contour continuity can be acquired by maximizing the harmonicity evaluation function. Therefore, this method removes the requirement of post-processing to refine F_0 estimates which also means to reduce computational complexity. Furthermore, estimates error like F_0 halving or doubling is automatically eliminated although it may rapidly fluctuate.

3. Algorithm

The algorithm of IFAS-based F_0 estimation can be summarized as follows,

1. Analyze the input signal $x(t)$ using STFT to obtain its spectrum $X(\omega)$.
2. Calculate the instantaneous frequency $\lambda(\omega)$ by using (3) and (4).
3. Select an IF band $[\lambda_l, \lambda_u]$ which maximizes the measure of harmonicity in the IF-domain P_{λ_l,λ_u} in (11).
4. Calculate the $\eta(F)$ of the selected IF band $[\lambda_l, \lambda_u]$ and determine $F = F_0$ which maximizes $\xi(F)$ in (9).

The implementation of described algorithm in discrete time-domain, DFT is used for calculation of the short-time Fourier transform. The STFT $X(\omega)$ and the instantaneous frequency $\lambda(\omega)$ are calculated at the frequency of $f_k = kF_s/N$, where F_s is sampling frequency and N is DFT size. In the IF calculation, it sometimes occurs that the IF has a meaningless value which means the nonexistence of frequency component within the passband of the bandpass filters centered at each frequency bin. Consequently, if the value of the obtained IF $\lambda(2\pi f_k)$ at the k-th frequency bin (i.e k-th bandpass filter) does not exist in the passband, the value is excluded from the evaluation of $\xi_{\lambda_l,\lambda_u}(F)$ and $\eta(F)$.

4. Results and Discussion

For experimental purpose, NAIST-CREST clean speech database which contains continuous speech and its coresponding Electroglottograph (EGG) waveforms uttered by 14 male and 14 female speakers is used for performance assessment. We selected, with no particular order, three Japanese sentences from the database for evaluation, 84 sentences in total. Noisy speech is obtained by adding white gaussian noise with prescribed SNR. The whole experimental setup can be referred to Table 1.

For F_0 reference, pitch periods of the speech signal were extracted from the corresponding EGG waveform automatically and then the errors were corrected by eye inspection and hand

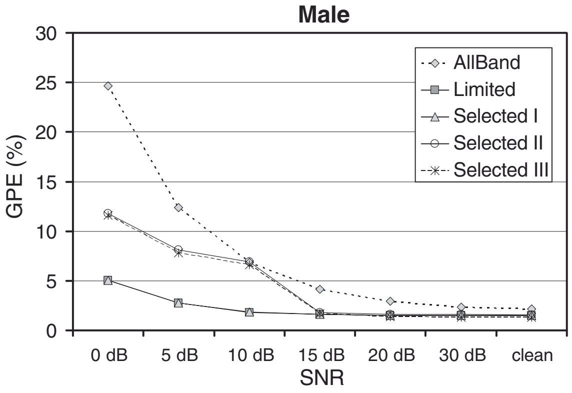

Figure 1: Performance of IFAS-based F_0 for male speakers

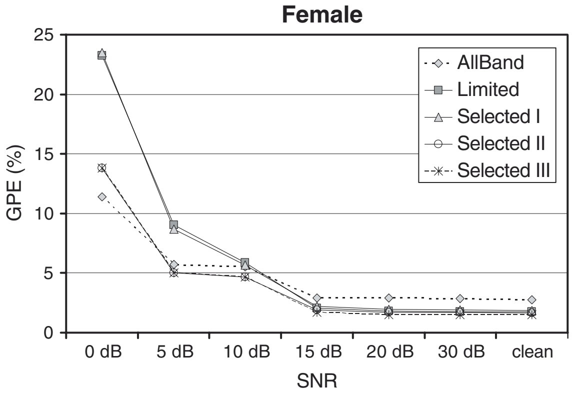

Figure 2: Performance of IFAS-based F_0 for female speakers

labeling. For objective evaluation to measure estimator accuracy, what so-called Gross Pitch Error (GPE) is used. If the error between estimated F_0 and the reference is greater than 10%, it is classified as gross error.

4.1. IFAS Performance

Figure 1 and Figure 2 show evaluation results of the proposed method with a number of frequency band and window length selection methods. The condition for *AllBand* case, represented by dotted lines, the lower bound λ_l was set to zero while the upper bound $\lambda_u/2\pi$ is 8 kHz. For the *Limited* band condition, denoted by square, λ_l is zero and $\lambda_u/2\pi$ is 600 Hz. *Selected I* means λ_l is zero and $\lambda_u/2\pi$ is moving starting from 600 Hz up to 2 kHz with 100 Hz increment. In those cases, the window length is fixed to 500 points. For the *Selected II* case, denoted by circle, we use 400, 450, 500, 600, 800, 1000 samples windows then the window which maximizes the harmonicity measure value is selected. In the case of *Selected III* with variable window length, shown with dashed lines, F_0 candidates are taken from previous consecutive 7 frames with the lowest and the highest frame values elimination. Within these remaining 5 frames, pitch-lags are averaged then multiplied by 4 to provide a window length candidate. If this window length is lower than 400 samples length, the last is used instead.

Database	NAIST-CREST with EGG
F_0 reference	Extracted from EGG
Number of speakers	14 males and 14 females
Sentences	3 sentences for each speaker
Sampling frequency	16 kHz
Window type	Blackman
Window shift	1 msec
F_0 search range	40 - 400 Hz
α, β in eq.(7)	10, 8 Hz

Table 1: Summary of experimental parameters

It can be seen that, in general, the estimator performs satisfactorily. From clean speech up to 15 dB SNR, the error rate remains same. Although from 10 dB to 0 dB the estimator performance is lowered, the GPE value is about 10% for male and 13% for female at 0 dB SNR. In these figures, the *Selected III* shows superior performance for both speaker groups and similarly to the *Selected II*. For both speaker groups, the *AllBand* case has the largest error rates for clean speech up to the noisy speech due to the use of wider frequency band. Particularly for male speakers, *Limited* and *Selected I* show better performance with relatively stable accuracy even in the very noisy environment with only about 5% error rate. On the other hand, *Limited* and *Selected I* performance are the lowest for female groups down to about 20%. The discrepancy of the GPE rates for male and female are significant in fixed window length cases. From the figures, it is demonstrated that the estimator has better accuracy in the cases of frequency band and window length are not fixed. Furthermore, the estimator seems to be gender independent since the error dissimilarity is relatively small.

4.2. Performance Comparison

We used the latest version of an open-source speech analysis tool called *Wavesurfer* [10] and a MATLAB based software called STRAIGHT-TEMPO (hereafter called TEMPO) [6] for comparison after minor modifications. Since performance evaluation is conducted with 1 ms frame shift, slight modification is necessary. For the *Wavesurfer*, the window shift is adjusted to 1 ms instead of its default 10 ms. Within the *Wavesurfer*, there are available analysis options which consists of ESPS-based pitch tracking using normalized cross correlation smoothed by dynamic programming and AMDF (Average Magnitude Difference Function) [1].

The same database of speakers and references are employed to evaluate ESPS, AMDF and TEMPO, then these results are compared against each other. The evaluation procedure is slightly modified where usually the F_0 reference compared directly to the assessed F_0. Firstly, we determine voiced region of respective evaluated methods by marking the voiced

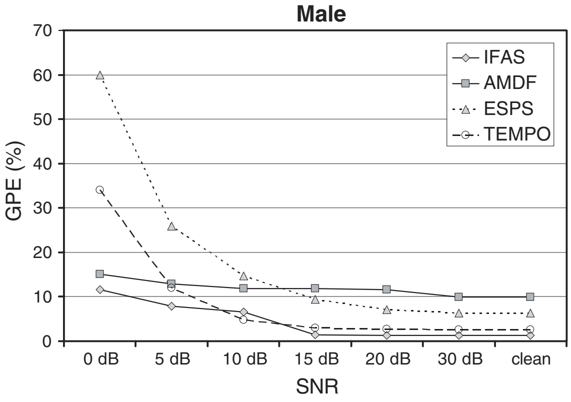

Figure 3: Performance comparison for male speakers

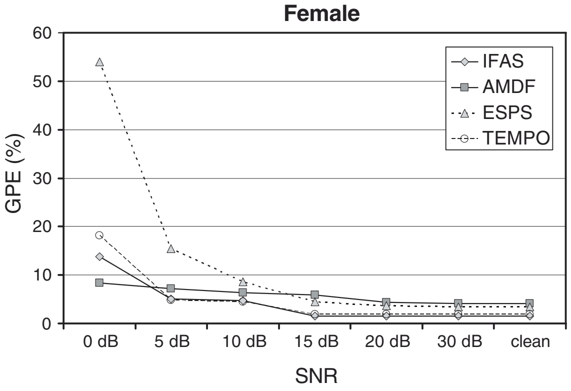

Figure 4: Performance comparison for female speakers

onset and its end on the existence of fundamental frequency. F_0 of reference and evaluated methods are compared only within these matched voiced region, otherwise counted out. This can guarantee the fairness of the comparison.

The results are shown in Fig. 3 for male and Fig. 4 for female speakers, where the IFAS is represented by diamond, square for AMDF, dotted line for ESPS and dashed line for TEMPO respectively. In overall case, the IFAS-based fundamental frequency estimator performance exceeds ESPS, AMDF and TEMPO method respectively. In Fig. 3, the error rate of IFAS and TEMPO are almost similar upto 5 dB SNR. From clean to 15 dB SNR in Fig. 4, the error rates of IFAS and TEMPO are similar. However, GPEs of TEMPO at 0 dB are about 30% for male and 18% for female respectively. The noise seems to affect the estimator accuracy in particular with respect to gender. IFAS error rates at 0 dB SNR are about 10% for male and 13% for female. Only AMDF method seems to maintain its consistency of the error rate for both speaker groups, with about 15% for male and only 8% on female at 0 dB, despite its lower accuracy compared to IFAS method or TEMPO. At 0 dB SNR, ESPS has the most error about 60% for both speaker groups. The ESPS and IFAS take similar computational time to completely estimate F_0 while AMDF and TEMPO requires more than two times longer of computational time.

5. Conclusions

We have presented evaluation results of the IFAS-based fundamental frequency estimator with respect to additive white gaussian noise. A central concept in the proposed method is a quantity which provides degree of regularity of periodicity, called harmonicity measure, which is used for frequency band and window length selection to obtain reliable F_0 estimates. In this investigation, the gross error rate of estimated F_0 is about 1% from clean to 15 dB SNR and down to about 11% at 0 dB SNR. Performance of the proposed method was also compared to ESPS, AMDF and TEMPO. It has been demonstrated that the IFAS-based fundamental frequency estimator outperforms ESPS, AMDF and TEMPO method respectively. We are investigating the proposed method performance under real noise with similar framework to evaluate its robustness.

6. References

[1] W. Hess, Pitch Determination of Speech Signals, Springer Verlag, Berlin, 1983.

[2] M.Cooke, S. Beet and M.Crawford, Eds., Visual Representation of Speech Signals, John Wiley & Sons, New York, 1993.

[3] P. Veprek and M. S. Scordilis, “Analysis, enhancement and evaluation of five pitch determination techniques,” Speech Communication, vol.37, pp.249-270, 2002.

[4] T. Abe, T. Kobayashi, S. Imai, “Harmonics estimation based on instantaneous frequency and its application to pitch determination of speech”, IEICE Trans., Information and Systems, vol.E78-D, No.9, pp. 1188-1194, 1995.

[5] T. Abe, T. Kobayashi, and S. Imai, “Robust pitch estimation with harmonic enhancement in noisy environment based on instantaneous frequency,” Proc. 4th ICSLP, pp.1277-1280, Philadelphia, USA, 1996.

[6] H. Kawahara, H. Katayose A. de Cheveigne, R. D. Patterson, “Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F_0 and periodicity,” Proc. EUROSPEECH’99, pp.2781-2784, 1999.

[7] T. Nakatani, and T. Irino, “Robust fundamental frequency estimation against background noise and spectral distortion,” Proc. ICSLP-2002, vol. 3, pp. 1733–1736, Denver, USA, Sep., 2002.

[8] T. Tanaka, T. Kobayashi, D. Arifianto, T. Masuko, ”Fundamental frequency estimation based on instantaneous frequency amplitude spectrum”, Proc. ICASSP, vol-I, pp.329-332, Orlando, USA, May 2002.

[9] P.P. Vaidyanathan, Multirate Systems and Filter Banks, Prentice Hall, New Jersey, 1993.

[10] `http://www.speech.kth.se/wavesurfer/`