A Precision, Large Scale, Anti-Noise Method for Correction of Speech Fundamental Frequency

Mang Zhao, Xiangning Chen, Yun Ge, Shuang Yu, and Ying Chen

***Abstract*—The speech fundamental frequency (F0) is an important speech parameter which has significant effect in domains such as voiceprint analysis, speaker recognition, voice conversion and etc. The existing extraction methods of speech fundamental frequency are not satisfactory in terms of accuracy due to the limitations of the methods themselves and noise so that correction is essential for high-quality analysis, recognition, conversion. The effect of existing correction methods rely heavily on the accuracy of first extraction, and are sensitive to noise. This article proposes a new correction method which makes use of cross-correlation of original signal and estimated signal to convert the difference between their frequencies to phase difference, thereby gains the accurate fundamental frequency by measuring this phase difference. Experiments indicate that the proposed method is effective when the estimated frequency ranges from 70% to 190% of the actual frequency, and keeps the average error within -50dB (0.35%) when signal to noise ratio (SNR) reduces to 3dB.**

***Index Terms*—Speech fundamental frequency, precision correction, large scale, anti-noise, phase difference.**

I. Introduction

The speech fundamental frequency (F0) [1] is an important speech parameter which has significant effect in domains such as voiceprint analysis [2], speaker recognition [3], voice conversion [4], [5] and etc. Now lots of speech analysis/synthesis [6], [7] frameworks use fundamental frequency as an important model parameter such as STRAIGHT [8]-[10]. The STRAIGHT express speech as fundamental frequency and spectral envelope and can easily manipulate these parameters and synthetize the manipulated speech as well. The accuracy of fundamental frequency estimation immediate influences the performance of the model.

The speech fundamental frequency fluctuates due to short-time stationarity of speech signal [1], resulting in high computational complexity and inaccuracy in extraction process. Japanese scholar H. Kawahara has developed four types of F0 extractors [11]-[14] for STRAIGHT model, and all of them committee to improve the accuracy of first extraction thus sacrificing computation time. For all this, the precision is still not enough so that corresponding correction methods are proposed. These correction methods mostly inherit thoughts of the corresponding extractors therefore relying heavily on the accuracy of first extraction, and are susceptible to noise.

This article proposes a new method which makes use of cross-correlation of actual signal and estimated signal to convert the difference between their frequencies to phase difference, thereby gains the accurate fundamental frequency by measuring this phase difference. This method overcomes the shortcoming of overmuch dependent on the accuracy of first extraction, and resists the interference of the noise very well.

Manuscript received September 27, 2013; revised November 15, 2013. This work was supported in part by National Nature Science Foundation of China (81371638), Fundamental Research Funds for the Central Universities (1106021034), Jiangsu Provincial Nature Science Foundation of China (BE101258, BK2011393, BY2012186). The authors would like to thank Miss Ruoyu Mao and Miss Xiaoxia Cheng for their helpful discussions and suggestions.

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing, China (e-mail: zhaomangzheng@ gmail.com, shining@nju.edu.cn, geyun@nju.edu.cn).

II. Existing Extraction and Correction Methods of Speech F0

The method based on fixed point analysis [11] works as follows: First, use a band-pass filter bank that consists of filters equally spaced along the log frequency axis to separate the signal; second, find the fixed points that may be the fundamental frequency by hilbert transform; third, use carrier-to-noise (C/N) ratio to determine the final point that represents the fundamental frequency.

$$\omega(t) = \frac{d[\arg[x(t) + jHilbert(x(t))]]}{dt}$$
$$\sigma^2(t) = c_a\left(\frac{\partial \omega_c(t,\lambda)}{\partial \lambda}\right)^2 + c_b\left(\frac{\partial^2 \omega_c(t,\lambda)}{\partial t \partial \lambda}\right)^2 \quad (1)$$

$$c_a = \left[\int_{-\infty}^{\infty} \left(\delta \frac{dg(\lambda)}{d\lambda}\Big|_{\lambda=\delta}\right)^2 d\delta\right]^{-1}$$
$$c_b = \left[\int_{-\infty}^{\infty} \left(\delta^2 \frac{dg(\lambda)}{d\lambda}\Big|_{\lambda=\delta}\right)^2 d\delta\right]^{-1} \quad (2)$$

The principle of correction is the same as extraction: search around the fixed point to find the best point which has the lowest C/N ratio. Fig. 1 (left) shows that although this correction method has a good performance in noise, it easily causes errors when the first extraction result deviates from the accurate value too far.

The method based on TANDEM spectrum [13], [14] works as follows: First, calculate the TANDEM spectrum and the smooth spectral envelope of the signal; second, divide the TANDEM spectrum by the spectral envelope to gain an analytic signal in which the dominant periodic components will be represented as salient peaks; third, search the unique

peak from the inverse Fourier transform of the analytic signal which represents the fundamental frequency.

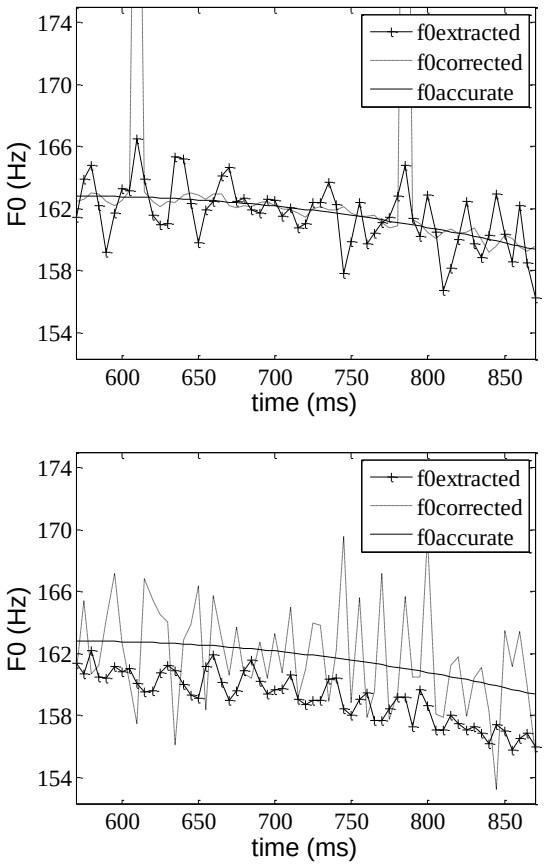

Fig. 1. Extraction and correction results of testing signal's F0 by fixed-point (left), tandem-spectrum (right) method, SNR=5dB. f0extracted, f0corrected, f0accurate present extraction, correction and accurate trajectories respectively

$$P_P(\omega) = \frac{P_T(\omega)}{P_{TST}(\omega)} - 1$$
$$r(\tau) = \int_{-\infty}^{\infty} w_B(\omega) P_P(\omega) e^{j\omega\tau} d\omega \tag{3}$$

The correction method makes use of the mathematical features of the window which is used in calculating the spectrum to gain the difference between current frequency point and the peak point. Fig. 1 (right) shows that this method performs badly when the noise is strong.

III. Proposed Correction Method

A. Basic Principle

Assume f_0 and f_c represents the accurate fundamental frequency and the first extraction result respectively. Be regardless of the noise for the moment, then define a signal $r(t)$ as (5) which is similar to cross-correlation of signal $x_0(t)$ and $x_c(t)$:

$$x_0(t) = \sum_k a_k \sin(2\pi k f_0 t + \varphi_k) + noise$$
$$x_c(t) = \sin(2\pi k f_c t) \tag{4}$$

$$\begin{aligned} r(t) &= \int_0^{T_c} x_c(\tau) x_0(\tau + t) d\tau \\ &= \sum_k a_k [(\int_0^{T_c} \sin(2\pi k f_c \tau) \sin(2\pi k f_0 \tau) d\tau) \times \cos(2\pi k f_0 t + \varphi_k) \\ &\quad + (\int_0^{T_c} \sin(2\pi k f_c \tau) \cos(2\pi k f_0 \tau) d\tau) \times \sin(2\pi k f_0 t + \varphi_k)] \end{aligned} \tag{5}$$

where $x_0(t)$ and $x_c(t)$ represents the original signal and assumed signal respectively, and T_c represents $1/f_c$. By Further consolidation, we can gain the following equations:

$$S_k = \int_0^{T_c} \sin(2\pi k f_c \tau) \sin(2\pi k f_0 \tau) d\tau$$
$$C_k = \int_0^{T_c} \sin(2\pi k f_c \tau) \cos(2\pi k f_0 \tau) d\tau \tag{6}$$

$$r(t) = \sum_k a_k \sin(2\pi k f_0 t + \varphi_k + \Delta\varphi_k) \tag{7}$$

The difference between the gained signal $r(t)$ and the original signal $x_0(t)$ is just a phase shift which can be simplified as (8). Note that this simplification is accurate, without approximation.

$$\begin{aligned} \Delta\varphi_k &= \arctan(\frac{S_k}{C_k}) \\ &= \arctan(\frac{\frac{\sin(2\pi k(f_c - kf_0)t)|_0^{T_c}}{f_c - kf_0} - \frac{\sin(2\pi k(f_c + kf_0)t)|_0^{T_c}}{f_c + kf_0}}{-\frac{\cos(2\pi k(f_c - kf_0)t)|_0^{T_c}}{f_c - kf_0} - \frac{\cos(2\pi k(f_c + kf_0)t)|_0^{T_c}}{f_c + kf_0}}) \\ &= \arctan\frac{\sin(2\pi k f_0 T_c)}{\cos(2\pi k f_0 T_c) - 1} = \pi k f_0 T_c - \pi/2 \end{aligned} \tag{8}$$

Then a low-pass filter is used upon the signal $x_0(t), r(t)$ to gain $\tilde{x}_0(t), \tilde{r}(t)$ which only have fundamental component. Due to the phase response upon the two signals is exactly the same, the final phase shift shows as (9):

$$\Delta\varphi_0 = \pi f_0 T_c - \pi/2 \tag{9}$$

$$\varphi(t) = \arg[x(t) + jHilbert[x(t)]] \tag{10}$$

The instantaneous phase of a signal can be calculated by Hilbert transform as (10). Due to the fact that the valid phase scope of the operation $\tan^{-1}$ is $0 \sim \pi$, we can correct the frequency by the proposed method when f_c ranges from $0.67f_0$ to $2f_0$ theoretically.

B. Anti-Noise-Interference Ability

The autocorrelation of the original signal shows as (11) where $x_{0t}(t)$ represents the no-noise part of the original signal, $noise(t)$ represents the noise, and T_w represents the autocorrelation window width. The wider the window is, the better the anti-noise-interference effect we gain.

$$x_{0c}(t) = \int_0^{T_w} [x_{0t}(\tau) + noise(\tau)][x_{0t}(\tau + t) + noise(\tau + t)] d\tau \tag{11}$$

The signal $x_{0c}(t)$ and $x_0(t)$ have different amplitude and phase, however their frequency is the same. We can use $x_{0c}(t)$ to correct the frequency by the proposed method because this

method is independent of the amplitude and phase of signal. This is a perfect character which other methods do not have.

C. System Model and Description

The proposed correction method works as Fig. 2 shows: first, use autocorrelation of the original signal to decrease noise interference; second, use cross-correlation of original signal and estimated signal to convert the difference between their frequencies to phase difference; third, calculate the phase of $x_{0c}(t)$ and $r(t)$ after low-pass filtering respectively; lastly, use the phase difference to gain the correction fundamental frequency.

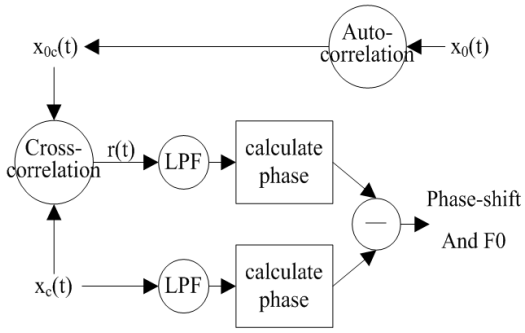

Fig. 2. Working process of the proposed method

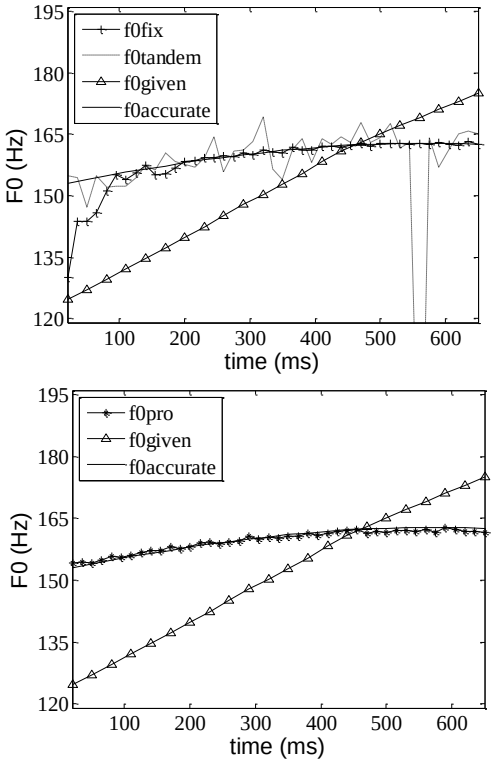

Fig. 3. Correction results of testing signal's F0 trajectory with the same given estimated trajectory by fixed-point, tandem-spectrum (left) and the proposed (right) method, SNR=5dB. f0given, f0accurate, f0fix, f0tandem, f0pro represent the given, accurate trajectories and which are gained by the three methods.

IV. Experiment Results

A. Overall Contrast

Fig. 3 shows the correction F0 results of a testing signal with a given estimated frequency trajectory by the fixed-point method, TANEM-spectrum method (left) and the proposed method (right) respectively while SNR equals to 5dB. The fixed-point method lose efficacy when the deviation of estimated frequency is too big. And the TANDEM-spectrum method is so much sensitive to noise that the correction result deviates more further from the accurate trajectory. The proposed method well overcomes the above shortcomings.

B. Impact of Extraction Deviation

Fig. 4 shows the average fundamental frequency errors after correction of a testing signal with a series of given estimated frequency trajectories by the fixed-point method, TANEM-spectrum method and the proposed method respectively while SNR equals to 15dB. Assume it is effective when the error is below -30dB (3.2%), then the effective range of f_c/f_0 of the fixed-point method and the TANDEM-spectrum method is [0.85, 1.15] and [0.75, 1] respectively. The proposed method performs well while f_c/f_0 ranges from 0.7 to 1.8.

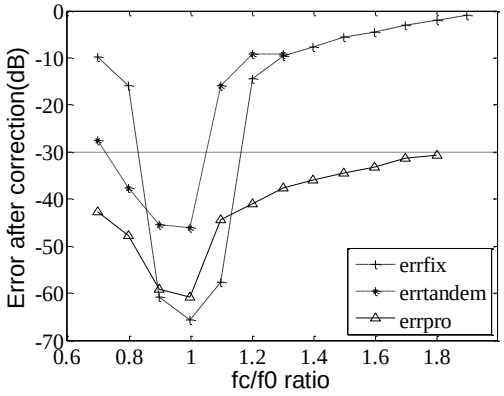

Fig. 4. Correction errors of testing signal's F0 trajectories with different given estimated trajectories, SNR=15dB. errfix, errtandem, errpro represent the results by the three methods respectively.

This large-scale character provide a train of thought: we can use a very simple extraction method with low computation complexity to gain a rough fundamental frequency trajectory, and then correct the result by the proposed method. The error can be kept as low as -60dB (0.1%) after double corrections by the proposed method.

C. Impact of Noise Interference

Table I shows the correction frequency errors of a testing signal with a given estimated frequency trajectory by the fixed-point method, TANEM-spectrum method and the proposed method respectively under different SNRs. The proposed method without autocorrelation module is sensitive to noise, the same as or even more than the TANDEM-spectrum method.

TABLE I: Correction Errors of Testing Signal's F0 Trajectories with the Same Estimated Trajectory and Different SNR (%)

SNR(dB)	15	10	5	3
Fixed-point	0.15	0.17	0.16	0.16
Tandem-spectrum	0.51	0.95	1.5	2.1
Proposed--no autocorrelation	0.30	0.71	X	X
Proposed	0.28	0.29	0.32	0.35

Autocorrelation is an effective measure to decrease noise interference. Further, the perfect character that the proposed

method is independent of the amplitude and phase of signal makes it possible to design many other methods which can decrease the noise interference and just do not change the frequency of signal.

V. Conclusion

This article proposes a new method to correct the speech fundamental frequency trajectory. This correction method performs well when the estimated-to-accurate frequency ratio f_c/f_0 ranges from 0.7 to 1.8, so makes it practicable to use a very simple extraction method with low computation complexity. And double corrections can keep the error as low as -60dB. Further, this correction method allows pre-autocorrelation or other modules which can decrease noise interference to be executed before the main process. Experiments indicate that this method is effective.

Acknowledgment

This work was supported by the National Nature Science Foundation of China (81371638), Fundamental Research Funds for the Central Universities (1106021034), Jiangsu Provincial Nature Science Foundation of China (BE101258, BK2011393, BY2012186). The authors would like to thank Miss Ruoyu Mao and Miss Xiaoxia Cheng for their helpful discussions and suggestions.

References

[1] J. Deller, J. Proakis, and J. Hansen, *Discrete-time processing of speech signals*, Hoboken, US: Wiley, 2000, ch. 2.

[2] Y. Wang, Y. H. Wang, and T. N. Tan, "Combining fingerprint and voiceprint biometrics for identity verification: an experimental comparison," in *Proc. ICBA 2004, LNCS 3072*, pp. 663-670, 2004.

[3] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models," *Digital Signal Processing*, vol. 10, issues 1–3, pp. 19–41, January 2000.

[4] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88*, vol. 1, pp. 655 – 658, April 1988.

[5] D. G. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, issue 2, pp. 127–138, February 1995.

[6] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am*, vol. 87, issue 2, pp. 820-857, 1990.

[7] D. G. Childers and C. K. Lee, "Vocal quality factors: Analysis, synthesis, and perception", *J. Acoust. Soc. Am*, vol. 90, issue 5, pp. 2394-2410, 1991.

[8] H. Kawahara, M. Katsuse, and A. Cheveigné, "Restructuring speech representations using apitch-adaptive time-frequency smoothing and an instantaneous frequency based F0 extraction," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[9] H. Kawahara, "STRAIGHT, exploration of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoust. Sci. Techno*, vol. 27, pp. 349–353, 2006.

[10] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0 and aperiodicity estimation," in *Proc. ICASSP 2008 IEEE*, 2008, pp. 3933–3936.

[11] H. Kawahara, H. Katayose, A. Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," in *Proc. EUROSPEECH'99, ESC*, 1999, vol. 6, pp. 2781–2784.

[12] H. Kawahara, A. Cheveigné, H. Banno, T. Takahashi, and T. Irino, "Nearly defect-free F0 trajectory extraction for expressive speech modifications based on STRAIGHT," in *Proc. Interspeech 2005 ISCA*, 2005, pp. 537–540.

[13] M. Morise, T. Takahashi, H. Kawahara, and T. Irino, "Power spectrum estimation method for periodic signals virtually irrespective to time window position," *IEEE Trans, IEICE*, vol. J90-D12, pp. 3265–3267, 2007.

[14] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sadhana*, vol. 36, issue 5, pp. 713-727, Oct 2011.

Mang Zhao was born in Yangzhou, Jiangsu, China on March 13, 1989. He received Bachelor degree in electrical engineering from Nanjing University, China, in 2011. Now he is studying for Master degree in Nanjing University, China.

Xiangning Chen works in Nanjing University, China. He received Ph.D degree in communication from Tsinghua University, China, in 2001.

Yun Ge works in Nanjing University, China. He received Ph.D degree in biomedicine from Southeast University, China, in 2001.