INTERSPEECH 2009 BRIGHTON

Robust F0 Estimation Based on Log-Time Scale Autocorrelation and Its Application to Mandarin Tone Recognition

Yusuke Kida, Masaru Sakai, Takashi Masuko, Akinori Kawamura

Corporate Research & Development Center, Toshiba Corporation, Japan

{yusuke.kida, masaru4.sakai, takashi.masuko, akinori.kawamura}@toshiba.co.jp

10.21437/Interspeech.2009-752

Abstract

This paper proposes a novel F0 estimation method in which delta-logF0 is directly estimated based on autocorrelation function (ACF) on a logarithmic time scale. Since peaks of ACFs of periodic signals have a specific pattern on the log-time scale and the period only affects the position of the pattern, delta-logF0 can be estimated directly from the shift of the peaks of the log-time scale ACF (LTACF) without F0 estimation. Then logF0 is estimated from the sum of LTACFs shifted based on delta-logF0. Experimental results show that the proposed method is more robust against noise than the baseline ACF-based method. It is also shown that the proposed method significantly improves the Mandarin tone recognition accuracy.

Index Terms: F0 estimation, autocorrelation, log-time scale, Mandarin speech recognition, tone recognition

1. Introduction

Robust F0 estimation is one of the most important problems in speech signal processing. For example, F0 features are used for tone recognition to distinguish a number of homophonic words in tonal languages such as Mandarin. Noise-robustness of F0 estimation is cruicial for practical applications of tone recognition including mobile phones and car-navigation systems.

A number of F0 estimation techniques have been proposed to date, most of which can be classified into two categories: a time-domain approach and a frequency-domain approach. Time-domain approaches are mainly based on the periodicity of speech. Autocorrelation function (ACF) and average magnitude difference function (AMDF) based methods are well known time-domain approaches [1, 2]. Frequency-domain approaches are based on the harmonicity of speech. For example, Iwano et al. proposed a frequency-domain method using Hough tranformation on time series cepstrum [3]. Tadokoro et al. also proposed the parallel connected comb filter based method [4]. In addition, some methods have been proposed to utilize both time and frequency-domain approaches [5]. These techniques perform satisfactorily in clean environments, however, the performance under noise conditions remains to be improved.

In this paper, we propose a novel F0 estimation method robust against noise. In the proposed method, since peaks of ACFs of periodic signals have a specific pattern on a log-time scale and the period only affects the position of the pattern, $\Delta \log \mathrm{F0}$ can be estimated directly from the shift of the peaks of the log-time scale ACF (LTACF) before F0 estimation. Then $\log \mathrm{F0}$ is estimated from the sum of LTACFs shifted based on $\Delta \log \mathrm{F0}$. Wang et al. proposed an F0 estimation method which is based on the fact that harmonics also show a specific pattern on a log-frequency scale [6]. However, our method differs in using time-domain ACF instead of frequency-domain harmonics. In addition, the proposed method utilizes the consistency of $\Delta \log \mathrm{F0}$ during a short time period in order to improve robustness against noise.

2. Baseline F0 Estimation Method

Figure 1 shows a block diagram of the baseline F0 estimation method, which is based in part on an autocorrelation-based F0 estimation technique proposed by Ghulam et al. [7]. In the baseline method, the input signal is decomposed into subband signals by a filterbank, and ACFs are calculated for each subband signal. Then the ACFs are normalized in order to reduce the influence of the spectral envelope, and a summary auto-correlogram [7] is obtained by summing up the normalized ACFs. Finally, the pitch period is obtained from a lag giving a maximum coefficient value of the summary auto-correlogram, and $\log \mathrm{F0}$ is obtained from the estimated pitch period. If the maximum coefficient value is lower than a threshold, the frame is classified as unvoiced.

3. Proposed Method

3.1. $\Delta \log$ F0 estimation from autocorrelation functions

Here we describe $\Delta \log \mathrm{F0}$ estimation from ACFs on the log-time scale of two adjacent frames without loss of generality.

On the log-time scale, since peaks of an ACF of a periodic signal with a period T appear at $\log T$, $\log T + \log 2$, $\log T + \log 3$, and so on, intervals between adjacent peaks are consistent regardless of the period, and the period only affects the position of a set of peaks on the log-time scale. That is, a set of peaks of log-time scale ACFs (LTACFs) always have a specific pattern and the pattern shifts according to the change of period.

Assume that the amplitude of the peaks are unity. Then an LTACF at frame t with period T_t can be represented by sum of unit impulse functions $\delta(\cdot)$ as follows:

$$A_t(\tau) = \sum_n \delta((\log T_t + \log n) - \tau), \tag{1}$$

and be related to an LTACF $A_{t-1}(\tau)$ at time $t-1$ with a shift s_t of the set of peaks between frame $t-1$ and t as follows:

$$A_t(\tau) = A_{t-1}(\tau + s_t), \tag{2}$$

where

$$s_t = \log T_t - \log T_{t-1}. \tag{3}$$

The shift s_t corresponds to the first difference of the period $\Delta \log T_t$ between these two frames, and $\Delta \log \mathrm{F0}_t$ at frame t

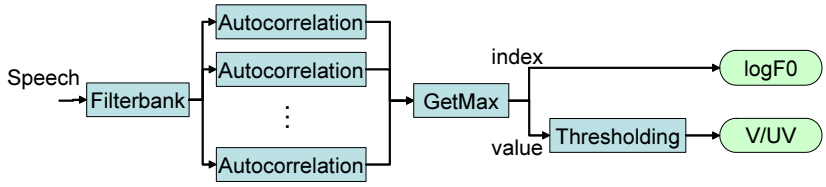

Figure 1: Block diagram of baseline F0 estimation method

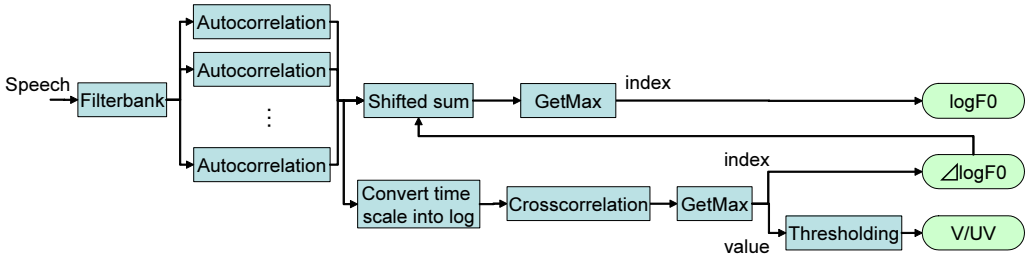

Figure 2: Block diagram of proposed F0 estimation method

is obtained from $\Delta \log T_t$ using the following relationship:

$$\begin{aligned}\Delta \log \mathrm{F0}_t &= \log \mathrm{F0}_t - \log \mathrm{F0}_{t-1} \\ &= -(\log T_t - \log T_{t-1}) \\ &= -\Delta \log T_t. \end{aligned} \tag{4}$$

The shift s_t can be estimated from the cross-correlation function between two neighboring LTACFs. Thus, $\Delta \log \mathrm{F0}_t$ can be obtained without F0 estimation.

3.2. Algorithm of proposed method

Figure 2 shows a block diagram of the proposed F0 estimation method. In the proposed method, $\Delta \log \mathrm{F0}$ is estimated first, and then $\log \mathrm{F0}$ is estimated using $\Delta \log \mathrm{F0}$.

The proposed method calculates summary auto-correlograms as well as the baseline method, and converts them from linear-time scale to log-time scale. Then cross-correlation functions are calculated from the summary auto-correlograms on the log-time scale. Figure 3 shows a schematic diagram of the calculation of the cross-correlation functions. As seen in the figure, the proposed method calculates cross-correlation functions between two frames with a certain distance, and then sums up a time-series of neighboring cross-correlation functions. It can be assumed that $\Delta \log \mathrm{F0}$ is consistent during a short time period, and hence a cross-correlation peak is enhanced by summing up the time-series of neighboring cross-correlation functions.

The cross-correlation function betweeen two frames at a distance d is calculated as follows:

$$C_t(\eta) = \frac{\sum_{\tau=0}^{\tau_{\max}-\eta} A_t(\tau) A_{t-d}(\tau+\eta)}{\sum_{\tau=0}^{\tau_{\max}-\eta} A_t(\tau) A_{t-d}(\tau)}, \tag{5}$$

where $\tau_{\max}$ denotes the maximum index of the LTACFs. Then the weighted sum of a time-series of neighboring cross-correlation functions is calculated as follows:

$$\overline{C}_t(\eta) = \sum_{n=-L+d}^{L} W_n C_{t+n}(\eta), \tag{6}$$

where L denotes the length of a window determining the range of the summed cross-correlation function. The weights W_n are given as follows:

$$W_n = \begin{cases} \sum_{y=-n+\frac{d-1}{2}+1}^{(L-d)} y & -(L-d) \le n \le \frac{d-1}{2}, \\ \sum_{y=n-\frac{d-1}{2}}^{(L-\frac{d-1}{2})} y & \frac{d-1}{2}+1 \le n \le L, \end{cases} \tag{7}$$

where d is assumed to be odd. $\Delta \log T_t$ is estimated from the peak of $\overline{C}_t$ as follows:

$$\Delta \log T_t = -\frac{1}{d} \operatorname*{argmax}_{\eta} \overline{C}_t(\eta), \tag{8}$$

and $\Delta \log \mathrm{F0}_t$ is obtained from $\Delta \log T_t$ by equation (4). If the difference between maximum and minimum values of the summed cross-correlation function are lower than a threshold, the frame is classified as unvoiced.

Based on the estimated $\Delta \log \mathrm{F0}$, $\log \mathrm{F0}$ is estimated from the sum of shifted LTACFs. As noted before, sets of peaks of LTACFs of periodic signals have a specific pattern and only shift on the log-time scale according to the change of period. Hence, the sets of peaks can be enhanced by summing up LTACFs shifted based on $\Delta \log \mathrm{F0}$ as follows:

$$\overline{A}_t(\tau) = \sum_{n=-N^-}^{N^+} A_{t+n}(\tau + S_t(n)) \quad (0 \le \tau \le \tau_{\max}), \tag{9}$$

where N^- and N^+ are the number of frames to be added before and after the current frame, and are chosen not to exceed a voiced section that the current frame belongs to. The shifts of LTACFs $S_t(n)$ are defined as follows:

$$S_t(n) = \begin{cases} -\sum_{i=n+1}^{0} \Delta \log T_{t-i} & (n < 0), \\ 0 & (n = 0), \\ \sum_{i=1}^{n} \Delta \log T_{t+i} & (n > 0). \end{cases} \tag{10}$$

Finally, $\log T_t$ is converted to $\log F0_t$.

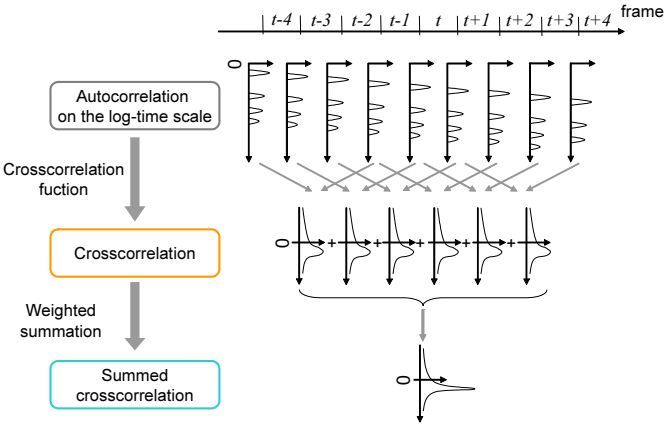

Figure 3: Schematic diagram of cross-correlation functions

Table 1: Gross Error Rates (GER) (%)

	0dB	5dB	10dB	15dB	25dB
Baseline	41.3	26.2	16.0	10.6	7.0
Proposed	25.1	15.0	11.0	8.8	7.5

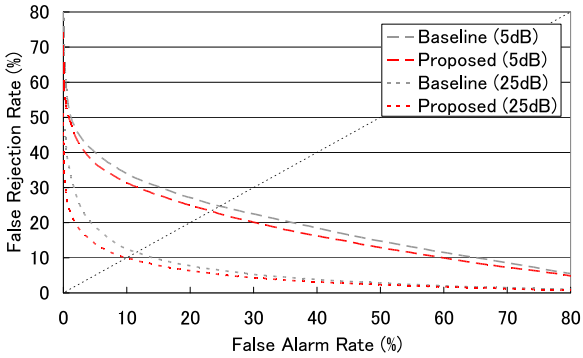

Figure 4: V/UV classification accuracy

4. F0 Estimation Experiments

4.1. Task and conditions

We conducted two experiments using *keele pitch database* [8] to compare the baseline and proposed methods; an F0 estimation experiment and a voiced/unvoiced classification experiment. To make the noisy data, we added in-car noise recorded under highway conditions with SNRs of 0, 5, 10, 15 and 25dB. The sampling frequency was 16000Hz, the frame length was 40ms and the frame shift was 8ms. Each input signal was decomposed into sub-band signals using eleven FIR Hamming filters of order 61 with center frequencies uniformly spaced on the Bark scale between 0 Hz and 1750 Hz. The maximum ACF lag τ_{max} was set to 320. The LTACFs were obtained by re-sampling the linear-time scale auto-correlograms using a sync function at 1024 sample points uniformly located on the log-time scale between 2.5ms and 20ms, which corresponds to 40 and 320 samples (lags) on the linear-time scale, respectively. The distance d between frames for calculating the cross-correlation functions was set to 3, and the window length L determining the range of the summed cross-correlation functions was set to 4. The maximum number of frames before and after the current frame used to calculate the shifted sum of LTACFs was set to 25. The minimum and the maximum values of F0 were 50Hz and 400Hz, respectively.

4.2. F0 estimation

In the F0 estimation experiment, accuracy is evaluated on frames labelled as voiced. Gross error rate (GER), which is the percentage of voiced frames in which the estimated values of F0 deviate from the references by more than 20%, was used for evaluation. Table 1 shows the GERs of the baseline and the proposed methods. From Table 1, it can be seen that the proposed method shows better performance than the baseline method under all but the quietest noise condition. The GERs of the proposed method under lower SNR conditions were significantly improved from the baseline method.

4.3. Voiced/unvoiced classification

In the voiced/unvoiced (V/UV) classification experiment, the frame-based false alarm rate (FAR) and false rejection rate (FRR) were used as evaluation measures. FAR is the percentage of unvoiced frames incorrectly classified as voiced frames, and FRR is the percentage of voiced frames incorrectly classified as unvoiced frames. The experimantal results under two SNR (5, 25dB) conditions are shown in Figure 4. The horizontal axis corresponds to FAR, and the vertical axis corresponds to FRR. The operating curve is plotted by varying the threshold value. Under both SNR conditons, the proposed method shows better performance than the baseline method.

To achieve optimal V/UV classification performance in real environments, the threshold should be tuned according to the noise condition. However, dynamic tuning of the V/UV threshold is difficult because estimation of noise conditions, such as SNR, is difficult to do in real environment. Hence, it is required for V/UV classification techniques to perform sufficiently regardless of noise condition without dynamic tuning of the threshold. To evaluate robustness against change of SNR, we also conducted a V/UV experiment using a fixed threshold value for various SNR conditions. Two thresholds were used for the evaluation: one is tuned to give an equal error rate (EER) for 5dB and the other for 25dB. We refer to the former and the latter threshold as "θ:5dB" and "θ:25dB," respectively. Table 2 and 3 show results of the baseline and the proposed methods. For the baseline method, the FAR on higher SNR using θ:5dB was much higher than that using θ:25dB, while FRR on lower SNR using θ:25dB was much higher than that using θ:5dB. This indicates that the V/UV threshold must be tuned according to SNR to achieve sufficient performance with the baseline. On the other hand, the proposed method showed a smaller performance gap between the two threshold values than the baseline method. This demonstrates that the proposed method is more robust against changes of SNR than the baseline method.

5. Tone Recognition Experiments

5.1. Task and conditions

We conducted a Mandarin tone recognition experiment given toneless syllable transcriptions. In this experiment, neutral tone

Table 2: V/UV classification error rates of the baseline method (%)

	0dB	5dB	10dB	15dB	25dB
FAR (θ:5dB)	21.7	24.1	27.5	33.1	46.4
FRR (θ:5dB)	41.1	25.0	13.8	7.5	3.1
FAR (θ:25dB)	0.5	1.0	2.4	4.6	11.3
FRR (θ:25dB)	72.5	51.5	33.9	21.8	11.8

Table 3: V/UV classification error rates of the proposed method (%)

	0dB	5dB	10dB	15dB	25dB
FAR (θ:5dB)	24.7	23.2	21.3	19.7	15.0
FRR (θ:5dB)	34.9	23.3	15.5	11.0	7.8
FAR (θ:25dB)	16.6	15.3	14.2	12.9	10.0
FRR (θ:25dB)	40.9	27.7	18.6	13.8	10.0

was not considered because it was not included in the test utterances. Test data were composed of three-syllable human names, and the total number of utterances were about 2,000 from 40 speakers. We added the in-car noise described in the previous section to make noisy data.

The acoustic features were composed of two streams: spectral feature stream and tonal feature stream. A spectral feature was a 39 dimention vector consisting of 12 cepstral coefficients, log energy, and their first and second order derivatives. A tonal feature was a 3 dimension vector consisting of $\log$ F0, its first and second order derivatives. In the proposed method, estimated $\Delta \log$ F0 were used instead of the first derivative of $\log$ F0, and $\Delta\Delta \log$ F0 was obtained as the first derivative of $\Delta \log$ F0. To reduce speaker dependency, $\log$ F0 is normalized by subtracting a moving average calculated in a short time window with length of 25 frames before and after the current frame. $\log$ F0 and its derivatives are set to uniform random values with the range of the corresponding parameters for unvoiced frames. The frame length was 25ms for the spectral feature, and 40ms for the tonal feature. The frame shift was 8ms for both features.

Initials and tonal-finals (ITFs) were adopted as acoustic units. The training data consisted of 81,969 utterances. Tri-ITFs were modeled by three-state left-to-right multi-stream hidden Markov models (HMMs). HMM states were clustered independently for each stream using a decision tree based context clustering technique. The number of states were about 2,000 and 500 and the number of mixtures were 32 and 4 for spectral and tonal feature streams, respectively. The threshold of V/UV was set to θ:5dB for training, while both θ:5dB and θ:25dB were tested in the evaluation.

5.2. Results

Experimental results are shown in Table 4. Table 4 compares tone error rates between tonal features based on the baseline and the proposed methods. In almost all SNR conditions regardless of V/UV threshold, the proposed method shows better performance than the baseline method. The difference is especially marked for lower SNR conditions with the proposed method degrading gracefully as SNR decreases. As before, the performance gap of the baseline method between two thresholds was much larger than that of the proposed method, because V/UV classification accuracy of the baseline method was significantly affected by threshold unlike the proposed method.

Table 4: Tone error rates (%)

	0dB	5dB	10dB	15dB	25dB
Baseline (θ:5dB)	40.3	26.6	22.5	22.8	25.9
Proposed (θ:5dB)	29.5	22.0	20.0	18.8	18.1
Baseline (θ:25dB)	58.3	33.3	20.9	17.7	19.2
Proposed (θ:25dB)	28.7	20.7	19.3	18.5	17.4

6. Conclusions

This paper proposed a novel F0 estimation method based on log-time scale autocorrelation function (LTACF) in which $\Delta \log$ F0 is estimated before estimation of $\log$ F0. Results of an F0 estimation experiment, a voiced/unvoiced (V/UV) classification experiment, and a Mandarin tone recognition experiment showed that the proposed method performed much better than a baseline method especially under lower SNR conditions, and that the threshold for V/UV classification of the proposed method is more robust to change of SNR condition than the baseline method.

Our future work includes further improvement of tone recognition accuracy and evaluation of the proposed method under various noise conditions.

7. Acknowledgements

This work was supported by the Ministry of Economy, Trade and Industry, Japan.

8. References

[1] D. Talkin, "A robust algorithm for pitch tracking(RAPT)," in *Speech Coding and Synthesis*, pp.497–518, 1995.

[2] A. de Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, Vol.111, No.4, pp.1917–1930, 2002.

[3] K. Iwano, T. Seki and S. Furui, "Noise robust speech recognition using F0 contour extracted by hough transform," *ICSLP*, pp.941–944, 2002.

[4] Y. Tadokoro, T. Morita and M. Yamaguchi, "Pitch detection of musical sounds noticing minimum output of parallel connected comb fi lters," *TENCON*, pp.380–383, 2003.

[5] C. Shahnaz, W.-P. Zhu and M. O. Ahmad, "Robust pitch estimation at very low SNR exploiting time and frequency domain cues," *ICASSP*, pp.389–392, 2005.

[6] C. Wang and S. Seneff, "A study of tones and tempo in continuous Mandarin digit strings and their application in telephone quality speech recognition," *ICSLP*, pp.695–698, 1998.

[7] M. Ghulam, T. Fukuda, J. Horikawa and T. Nitta, "A noise-robust feature extraction method based on Pitch-Synchronous ZCPA for ASR," *ICSLP*, pp.133–136, 2004.

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