A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches

*Xufang Zhao, Douglas O'Shaughnessy, Nguyen Minh-Quang
Institut National de la Recherche Scientifique, Université du Québec

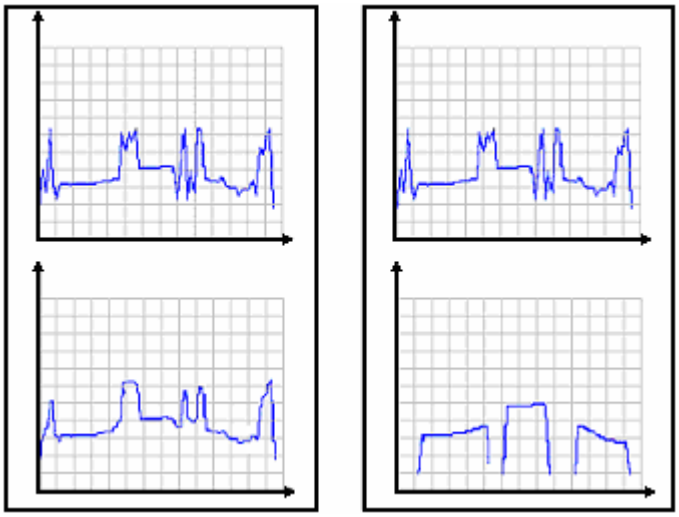
***Abstract*-Chinese is known as a syllabic and tonal language and tone recognition plays an important role and provides very strong discriminative information for Chinese speech recognition [1]. Usually, the tone classification is based on the F0 (fundamental frequency) contours [2]. It is possible to infer a speaker's gender, age and emotion from his/her pitch, regardless of what is said. Meanwhile, the same sequence of words can convey very different meanings with variations in intonation. However, accurate pitch detection is difficult partly because tracked pitch contours are not ideal smooth curves. In this paper, we present a new smoothing algorithm for detected pitch contours. The effectiveness of the proposed method was shown using the HUB4-NE [3] natural speech corpus.**

***Index Terms*—Acoustic signal analysis, Acoustic signal detection**

I. INTRODUCTION OF PREVIOUS WORKS

Traditionally, detailed tone features such as the entire pitch curve are used for tone recognition. Various pitch tracking algorithms based on different approaches have been developed for fast and reliable estimation of fundamental frequency. Generally, pitch analyses are performed in the time domain, frequency domain, or a combination of both domains [4]. [5] compared the performance of several pitch detection algorithms such as AMDF (Average Magnitude Difference Function [6]), SIFT (Simplified Inverse Filtering Technique [7]), and Cepstrum [8] in both the time domain and frequency domain.

By observing the original output F0 curves, it was found that for both autocorrelation and cepstrum approaches, tracked pitch contours were hardly ideal smooth curves; for example, the cepstrum approach often detected some sharp peaks and the autocorrelation approach often tracked octave jumping. Here, sharp peaks are abruptly changing on detected F0, and octave jumping means double frequency jumping; for example, from frequency P to $2P$. Both sharp peak and octave jumping will cause pitch detection errors. Therefore, it is necessary to smooth F0 curves after pitch tracking. There are several previous line smoothing algorithms, including McMaster's Distance Weighting Algorithm [9]. The simplest smoothing algorithm is the rectangular or unweighted sliding-average smoothing that simply replaces each point in the signal with the average of m adjacent points, where m is a positive integer called the smoothing width. Another algorithm, triangular smoothing, is like the rectangular smoothing, except that it implements a weighted smoothing function. Using triangular or rectangular smoothing methods, the noise is greatly decreased while the curve shape itself is hardly changed. Fig. 1 is an illustration example of triangle smoothing and pitches smoothing; the left column (a) is an example using triangular smoothing method, and the right column (b) is an example using the proposed pitch smoothing method in this paper. In Fig. 1 (a), the noise is greatly reduced after triangular smoothing, while the peak itself is kept unchanged, making it easier to measure the peak position, height, and width. However, for pitch contour smoothing, some sharp peaks should be discarded instead of smoothed. The classic approaches such as curve fitting, linear smoothing and median smoothing can get a good result, but not deal with several successive fundamental frequencies that are errors [10].



(a) (b)

Fig. 1: An illustration of triangular smoothing and proposed pitch smoothing approaches

Basically, our smoothing approaches are based on discarding some error values firstly, and then estimating reasonable values for these points. [11] presented a new approach to smooth fundamental frequency for Mandarin tone recognition. Compared with this smoothing algorithm, the smoothing algorithm in [11] was only suitable for autocorrelation F0 detection. [11] utilized two thresholds, but they did not explain clearly how these two thresholds were obtained and what were the exact threshold values in their system. In this paper, the smoothing algorithm for the autocorrelation approach only used one threshold for both forward and backward comparison. Moreover, [11] utilized 600 isolated syllables for training, and another 500 isolated syllables for testing, so it looks like the system in [11] was designed for isolated tone recognition. Our work was based on the objective to recognize tones for natural and continuous speech.

1-4244-1449-0/07/$25.00 © 2007 IEEE

II. AUTOCORRELATION AND CEPSTRUM APPROACHES FOR F0 DETECTION

The F0 (fundamental frequency) contour contains tone information. For example, if the prosody of a Mandarin tone is from low to high, then in general, its pitch pattern is also from low to high. Basically, there are two categories of approaches for pitch tracking [5]. One category is in the time domain, and the other category is in the frequency domain. Time-domain analysis could use some time-related features such as ZCR (Zero-Crossing Rate), peak picking, and autocorrelation. Frequency domain analysis could apply, for example, to cepstrum and harmonic matching. These two kinds of approaches have their advantages and disadvantages, respectively; for example, frequency-domain approaches generally have higher accuracy than time-domain methods, but they need more computation. The general method of fundamental frequency estimation is to take a portion of the signal and to find the dominant frequency of repetition. Difficulties for reliable pitch detection arise from the following facts: (*a*) speech signals are not always periodic, for example, the plosive consonants of "*b"*, "*d"*, and "*g"*, (*b*) speech signals may be contained environment noise, or could be even with periodic signals of different fundamental frequencies, (*c*) speech signals that are periodic may be changing in fundamental frequency over the time of interest, and (*d*) speech signals that are periodic with interval *P* are also periodic with interval *2P*, *3P* etc, so we need to find the smallest periodic interval or the highest fundamental frequency.

A. Autocorrelation F0 Detection Approach

Although a large number of different methods has been proposed for detecting pitch, the autocorrelation pitch detector is still one of the most robust and reliable of pitch detectors [11]. Autocorrelation and energy spectrum is a Fourier transform pair. Autocorrelation preserves information about harmonic and formant amplitudes in speech signals, while ignoring phase; we use autocorrelation because phase is less important perceptually and carries much less communication information than spectral magnitude. Actually, our ears are not very sensitive for speech phases. The autocorrelation function is a special case of the cross-correlation function. Assume a speech signal is *s(n)*; its autocorrelation R_{ss} is:

$$R_{ss}(k) = \sum_{m=-\infty}^{\infty} s(m)s(m-k) \qquad (1)$$

The autocorrelation measures the similarity of the signal and its time delay. By summing the products of a speech signal and a delayed signal of itself, the autocorrelation is large if at some delay the two signals have similar waveforms. The range of summation is usually limited, and dividing by the number of summed samples could normalize the function.

The short-time autocorrelation function is obtained by windowing *s(n)* and then using autocorrelation:

$$R_n(k) = \sum_{m=-\infty}^{\infty} s(m)w(n-m)s(m-k)w(n-m+k) \qquad (2)$$

In F0 determination, $R_n(k)$ is needed for *k* near the estimated number of samples in a pitch period; if no suitable prior F0 estimate is available, $R_n(k)$ is calculated for *k* from the shortest possible period. Narrowband smoothing applies long windows on signals and smoothes the signal in time over a few pitch periods, which is good for F0 estimation. The window size could be estimated by this way: average F0 values are 132 Hz and 223 Hz for males and females [12]; for F0 estimation, windows typically contain at least two pitch periods, so pitch analysis uses a long window that is often 30 to 50 msec..

An autocorrelation pitch detector calculates the cross-correlation for each block signal with its time delayed signal. In fact, each speech block is measured for similarity with its time delayed signal. If a section of a speech signal is periodic, its autocorrelation function will reach the maximum value at the location of pitch periods. For example, if the pitch period is *P*, the maximum value of its autocorrelation function occurs when the time delay equals *0*, *P*, *2P*, etc.

B. Cepstral F0 Detection Approach

A reliable way of obtaining an estimate of the dominant fundamental frequency for long, clean, stationary speech signals is to use the cepstrum, and the cepstrum pitch detector performed much better on lower pitch speakers than on higher pitch speakers [5]. The cepstrum is a Fourier analysis of the logarithmic amplitude spectrum of the signal. If the log amplitude spectrum contains many regularly spaced harmonics, the Fourier analysis of the spectrum will show a peak corresponding to the spacing between the harmonics: i.e, the fundamental frequency. Effectively, we are treating the signal spectrum as another signal, and then looking for periodicity in the spectrum itself.

The cepstrum is so-called because it turns the spectrum inside out. The cepstrum has units of quefrency, and peaks in the cepstrum (which relate to periodicities in the spectrum) are called rahmonics. To render the cepstrum suitable for digital algorithms, the DFT must be used in place of the general Fourier transform in Equation (3):

$$c_d(n) = \frac{1}{N}\sum_{k=0}^{N-1} \log|X(k)|\, e^{j2\pi kn/N} \quad . \qquad (3)$$

To obtain an estimate of the fundamental frequency from the cepstrum we look for a peak in the quefrency region corresponding to typical speech fundamental frequencies.

III. THE PITCH SMOOTHING METHOD

For the autocorrelation method, the detected F0 values doubled from time to time. The reason for this is that the maximum values occur at twice the delay time *P*. To correct this kind of error, each F0 value was compared with its immediately preceding and following values. If a F0 value had a big difference with its preceding F0 or following F0, it was marked as a "jump" point. Next, if the F0 value was the approximate double of any of its neighbors' F0 values, it was divided by two; otherwise, it was deleted as a bad value and

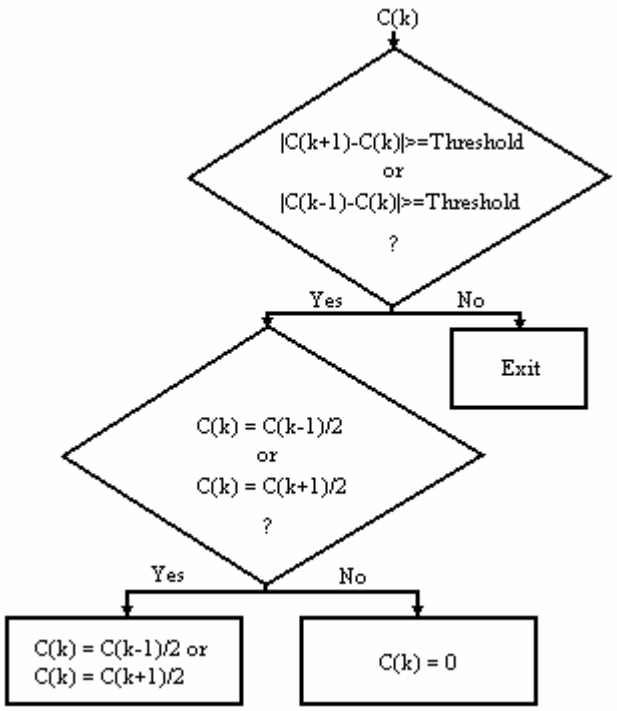


Fig. 2: The smoothing algorithm for an autocorrelation F0 contour

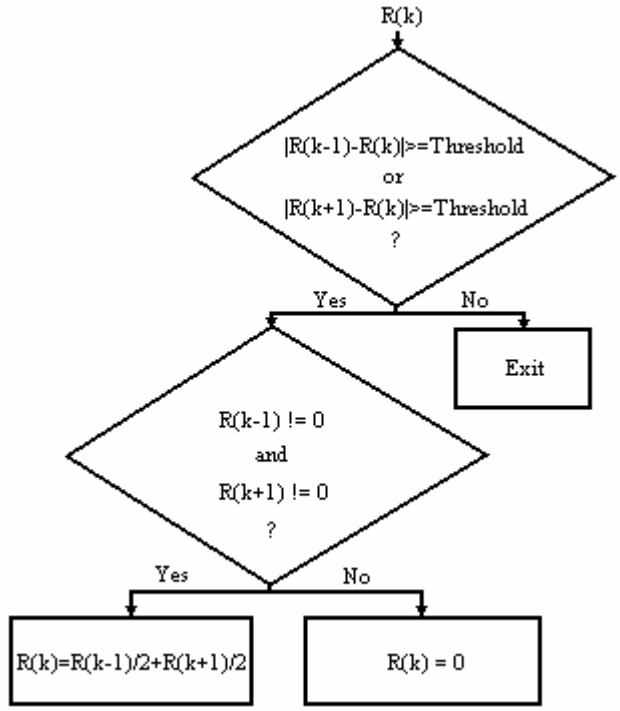


Fig. 3: The smoothing algorithm for the cepstrum F0 contour

forced to zero. Assuming that *C(k)* represents the autocorrelation F0 contour, the smoothing processing is in Fig. 2 and the threshold was determined to be 30 Hz by experiments.

After the F0 contour was gotten using the cepstral approach, it was found that the results were not satisfactory. In the smoothing phase, for each F0 value, we compared it with its immediately preceding F0 value and following F0 value. If this value had a big difference with its preceding F0 or following F0, it was marked as a "jump" point. Next, we checked if its neighbors' values were non-zero (zero was used for unvoiced frames). If both the immediately preceding and following F0 values were non-zero, we modified the jump point to the average value of its preceding and following F0 values; otherwise, we marked it as a bad value and set it to zero. Assume *R(k)* is the cepstrum F0 contour; the smoothing algorithm is in Fig. 3.

After we got the smoothed cepstral F0 contour and autocorrelation F0 values, they were merged together to obtain the final F0 contour. The combination rule is: at each sampling point, if both cepstral and autocorrelation methods have non-zero values, we use their mean value as the F0; otherwise, take the non-zero value as the final F0 value. Assume *R(k)* is autocorrelation F0, and *C(k)* represents cepstral F0. The combined F0 contour *F(k)* is: if both *C(k)* and *R(k)* are non-zero, *F(k)* is the mean of *C(k)* and *R(k)*; if any one of *C(k)* and *R(k)* is zero, the *F(k)* value is the non-zero one; otherwise, *F(k)* equals zero.

IV. Experimental Results

To confirm this smoothing algorithm, a hundred continue Mandarin utterances were tested. The experimental testing data were also taken from broadcast news uttered in Mandarin from the HUB-4NE database, including 4 female voices and 4 male voices, 856 Mandarin syllables. Speech signals were monaural sound and sampled at a rate of 16 kHz. These utterances are non-noise speech, and the average speech rate is 226 msec. per syllable. Fig. 4 shows one of the Matlab simulations of F0 contour smoothing. The top row shows a Mandarin speech signal. The label contained both Mandarin Pin Yin and tone. Digits 1, 2, 3, and 4 stand for the Mandarin tone 1, tone 2, tone 3, and tone 4. In Fig. 4, the second and the third subplots were smoothed F0 contours, and we could find that it did not contain doubled F0 and was smoother than the original un-smoothed one.

To score our smoothing algorithms, we counted the total number of sharp peaks in the cepstrum approach and the total number of octave jumps in the autocorrelation approach before smoothing and after smoothing. Tab. 1 showed the statistic results. Through the comparison of before smoothing and after smoothing, we found that most of F0 detection errors were deleted and final F0 contours were smoothed.

	Before smoothing	**After smoothing**
Sharp peak number	610	0
Octave jump number	322	0

Tab. 1: The comparison of the sharp peak number and the octave jump number before F0 smoothing and after F0 smoothing

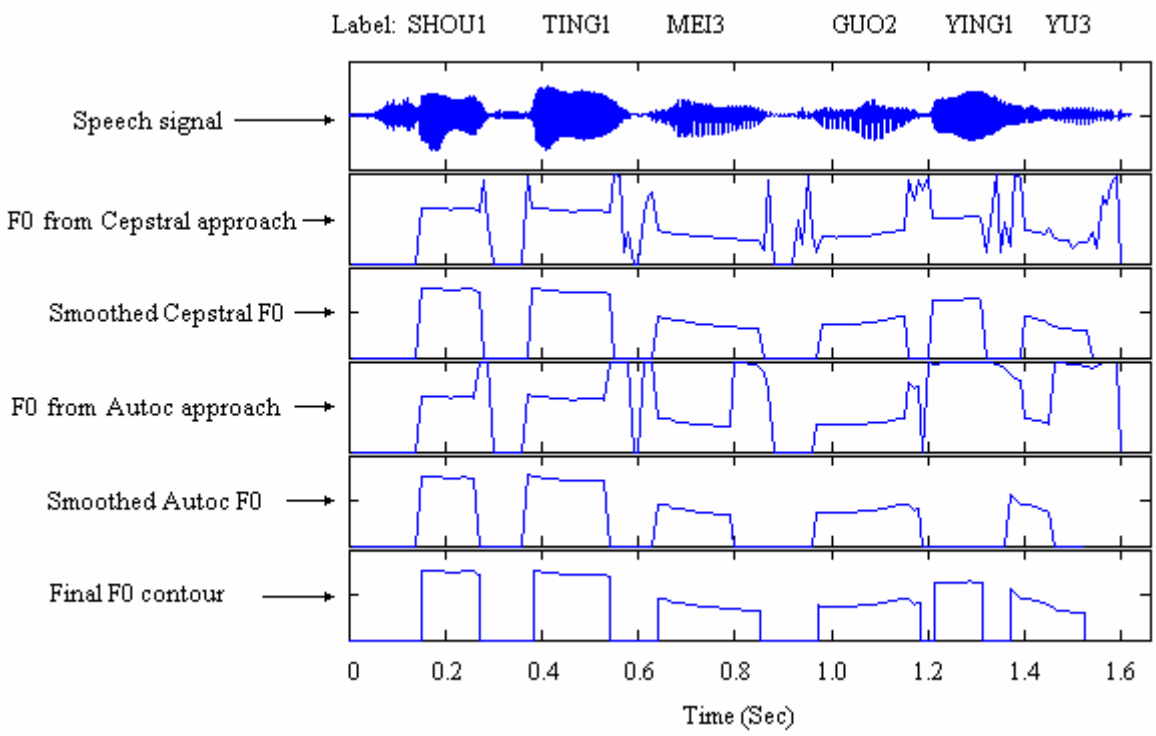


Fig. 4: A typical example of F0 contour smoothing

V. CONCLUSION

As we pointed out at the beginning of this article, pitch curves carry lots of information for Mandarin tones. In this paper, we aimed to smooth pitch contours detected by two F0 tracking approaches, autocorrelation and cepstrum. The autocorrelation approach created lots of doubled F0 detection values because a speech signal with period *P* also could meet the maximum autocorrelation value at the period of *2P*. To smooth the autocorrelation F0 tracking, we compared F0 detection values with their immediate previous values and next values to determine the possible jumped points; these possible octave jumped F0 values were forced to divide by two. The cepstral approach generated some noisy F0 values; these wrong F0 values were deleted and reconstructed by means of their immediate previous values and next values. Finally, autocorrelation F0 values were combined with cepstral F0 values to get smooth F0 contours. Smoothed F0 contours could be applied to build Mandarin tone models.

REFERENCES

[1] L. S. Lee, "Voice Dictation of Mandarin Chinese," in *IEEE Signal Processing Magazine*, Vol. 14, pp. 63-101, 1997.

[2] Guoliang Zhang, Fang Zheng, & Wenhu Wu, "Tone Recognition of Chinese Continuous Speech," in *Proceedings of International Symposium on Chinese Spoken Language Processing*, pp. 207-210, 2000.

[3] Linguistic Data Consortium, "1997 Mandarin Broadcast News Speech (HUB4-NE)," ISBN: 1-58563-125-6, 1998.

[4] Hynek Bořil and Petr Pollák, "Direct Time Domain Fundamental Frequency Estimation of Speech in Noisy Conditions," in *Proceedings of EUSIPCO 2004 (European Signal Processing Conference)*, Vol. 1, pp. 1003-1006, 2004.

[5] Lawrence R. Rabiner, Michael J. Cheng et al., "A Comparative Performance Study of Several Pitch Detection Algorithms," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. ASSP-24, NO. 5, pp. 399-418, Oct. 1976.

[6] M. J. Ross, H. L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, "Average Magnitude Difference Function Pitch Extractor," in *IEEE Transactions of Acoustic., Speech, Signal Processing*, Vol. ASSP-24, pp. 2-8, 1976.

[7] J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," in *IEEE Transactions Audio Electroacoust*., Vol. AU-20, pp. 367-377, 1972.

[8] R. W. Schafer and L. R. Rabiner, "System for Automatic Formant Analysis of Voiced Speech," in *Journal of Acoustic Society Amer.*, Vol. 47, pp. 634-648, 1970.

[9] McMaster, R.B. and Shea, S., "Generalisation in Digital Cartography," in *The Association of American Geographers*, pp. 71-91, 1992.

[10] W Gang, OYJ Zheng, "Chinese 4-tone Recognition Based on Analysis of the Nonlinear Trace of the Pitch Period with Neural Networks," in *Proceedings of ICIIPS' 92*, pp.214-217, 1992.

[11] Liu Jun, Xaoyan Zhu, and Yuping Luo, "An Approach to Smooth Fundamental Frequencies in Tone Recognition," in *Communication Technology Proceedings of ICCT' 98*. Vol. 1, pp. 5-9, 1998.

[12] T. Shimamura and H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech," in *IEEE Transactions on Speech and Audio Processing,* Vol. 9, pp. 727–730, 2001.