A Spectral-Temporal Method for Pitch Tracking

Stephen A. Zahorian, Princy Dikshit, Hongbing Hu

Department of Electrical and Computer Engineering
Old Dominion University, Norfolk, VA 23529, USA.
szahoria@odu.edu, pdiks001@odu.edu, hhuxx001@odu.edu

Abstract

In this paper, a new spectral/temporal method is described for robust pitch tracking for both high quality and telephone speech. A previous version of this algorithm was presented as YAAPT (Kasi and Zahorian, 2002) [10]. In the current paper, a novel method is presented for spectral pitch tracking, using nonlinear processing to partially restore the potentially missing fundamental frequency. A frequency domain modified autocorrelation is used to determine the spacing between harmonic peaks in the spectrum. The frequency domain spectral track is then used to refine time-domain pitch candidates obtained using the "NCCF or Normalized Cross Correlation" reported by Talkin [1]. Dynamic programming is used to find the "best" pitch track among all the candidates, using both local and transition costs. The algorithm was evaluated using the Keele pitch extraction reference database.

Index Terms: speech analysis, pitch tracking, dynamic programming

1. Introduction

Numerous studies show the importance of prosody for human speech recognition, but only a few automatic systems actually combine and use fundamental frequency (F0) or pitch as it commonly called. Combined with other acoustic features, prosody can be used to significantly increase the performance of automatic speech recognition (ASR) systems [2]. A big stumbling block remains the lack of robust algorithms for F0 tracking. F0 is especially important for ASR in tonal languages such as Mandarin speech, for which pitch patterns are phonemically important [5]. Other applications for accurate F0 tracking include devices for speech analysis, transmission, synthesis; speaker recognition; speech articulation training aids for the deaf ([4], [6]), and foreign language training.

An important consideration for any speech-processing algorithm is performance using telephone speech, due to the many applications of ASR in this domain [3]. However, since the fundamental frequency is often weak or missing for telephone speech, and the signal distorted and noisy and overall degraded in quality, pitch detection for telephone speech is especially difficult [3].

Many pitch detection algorithms have been reported, using a variety of techniques and with varying degrees of accuracy (see [7], [8] for summary). In a 2004 paper [11], Nakatani uses the harmonic dominance spectrum and reports high accuracy of pitch tracking for noisy speech. In our own past work [10], YAAPT was introduced and high accuracy was reported. However, more extensive testing with additive noise to both studio quality and telephone speech indicated an unacceptable drop in the accuracy of the tracking. A comprehensive examination of the YAAPT algorithm for speech signals with additive noise indicated that much of the problem was due to errors in the spectral pitch tracking used. Therefore, in this paper, the focus is on the new methods developed for spectral pitch tracking.

2. Algorithm

The entire F0 tracking algorithm is summarized in the flow chart given in Figure 1. The general strategy for pitch tracking is the same as that used in YAAPT.

The details of determining F0 candidates from the NCCF, the intelligent peak picking, use of multiple candidates with computed merits, and the use of dynamic programming to determine the final lowest cost pitch path were described in [10]. In the remainder of this section, the focus is on the illustration of using the nonlinear (squared) speech signal to restore missing harmonics, and especially on the method used to identify F0 from the spectral information.

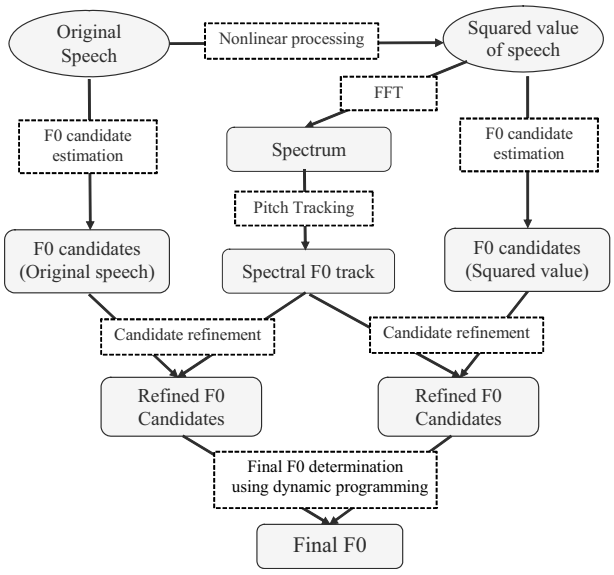

Figure 1. Flow chart of spectral/temporal pitch tracker

2.1. Restoration of missing fundamental via nonlinear operations

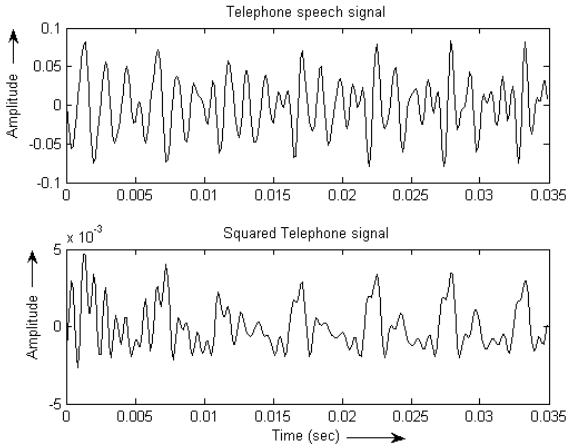

Figure 2. The telephone speech signal (top panel) and squared telephone signal (bottom panel) for one frame.

The underlying concept used for non-linear processing is that a periodic sound is characterized by the spectrum of its harmonics. To illustrate the principle, consider a speech signal consisting of only three spectral components:

$$x(t) = a_1 \cos(\omega t) + a_2 \cos(2\omega t) + a_3 \cos(3\omega t) \cdot$$

For x(t), the first term represents the fundamental frequency ω = 2πF0 and the rest of the cosine terms have frequencies that are harmonics of the fundamental.

The above representation is best suited for clean/studio quality speech signals as the fundamental and its harmonics are very prominent. Telephone quality signals where the fundamental is either very weak or absent can be approximated as

$$y(t) = b_2 \cos(2\omega t) + b_3 \cos(3\omega t)$$

It can be shown that the fundamental frequency (F0) re-appears for the telephone quality signal by squaring the signal and applying some basic trigonometric properties.

Squaring the telephone signal results in

$$y^2(t) = {b_2}^2 \cos^2(2\omega t) + {b_3}^2 \cos^2(3\omega t) + 2 b_2 b_3 \cos(2\omega t) \cos(3\omega t)$$

After applying trigonometric identities,

$$y^2(t) = \frac{\left[{b_2}^2 + {b_3}^2\right]}{2} + b_2 b_3 \cos(\omega t) + \frac{{b_2}^2}{2} \cos(4\omega t) + b_2 b_3 \cos(5\omega t) + \frac{{b_3}^2}{2} \cos(6\omega t)$$

As can be observed the fundamental (F0) re-appears in the squared version of the telephone signal. This is further illustrated in Figure 3.

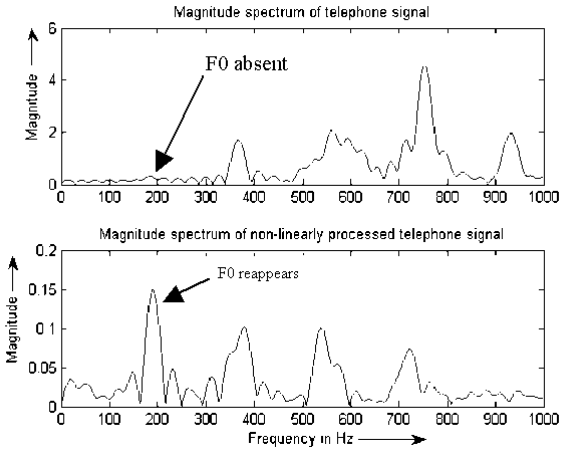

Figure 3. The magnitude spectra for the telephone and non-linearly processed signal. The top panel illustrates that F0 is absent in the telephone signal and the bottom panel shows the reappearance of F0 for the nonlinear signal. The spectra were computed from the signals shown in Figure 2.

2.2. Calculation of approximate pitch track from spectrum of squared signal

In addition to pitch tracking in the time domain, pitch tracking in the frequency domain is also widely used. In particular, the spectrum of a periodic signal consists of a series of peaks at the fundamental frequency and its harmonics. In this paper, the pitch track obtained from the spectrum plays an important role to refine the pitch candidates estimated from the waveform, since some the time domain extracted candidates are often in error, at least for noisy telephone speech.

An obvious way of determining the pitch from the spectrum is the extraction of the spectral peak at the fundamental frequency. This requires that the peak at the first harmonic be present and identifiable, which is often not the case, especially for noisy telephone speech. To achieve a more noise robust pitch track, a (frequency domain) autocorrelation type of function that considers multiple harmonics is used.

A function that takes into account multiple harmonics is:

$$y(k) = \sum_{i=-WL/2}^{WL/2} \prod_{n=1}^{N} f(nk + i),$$

where f(i) is the spectrum of the signal and WL, N indicate window length and the number of harmonics respectively.

The basic idea is that for each k (k is a frequency index), y(k) represents the extent to which the spectrum has high amplitude at integer multiples of that k. The use of a window, empirically determined to be approximately 40 Hz, makes the calculation less sensitive to noise, while still resulting in prominent peaks for y(k) at the fundamental frequency. The calculation is performed only for $k_{F0\_\min} < k < k_{F0\_\max}$.

Experiments were conducted to determine the best value for the number of harmonics, N. Empirically, it appeared that $N = 3$ resulted in the most prominent peaks in y(k) for voiced speech, and thus was used for the results given in this paper.

Figure 4 shows the spectrum (top panel) and the output of this autocorrelation type of function (bottom panel). Compared to the small peak at the fundamental frequency around 220 Hz in the spectrum, a very prominent peak is observed in y(k).

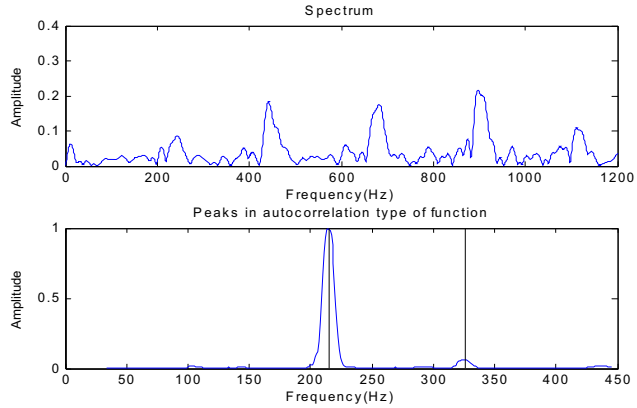

Figure 4. Peaks in autocorrelation type of function

For the peak picking from the spectrum, the same procedure is used as for F0 candidate estimation in the time domain [10]. Merit is assigned to each candidate according to its magnitude. For the examples shown in Figure 4, two candidates were chosen as the pitch candidates, as marked by lines.

To reduce pitch halving or doubling, additional logic is used to insert extra pitch candidates for some frames. In particular, if all candidates found by peak picking are larger than some threshold (typically 150 Hz), an additional candidate is inserted at half the frequency of the highest-ranking candidate. Similar logic is used to insert pitch candidate for frames where all the pitch candidates are very low.

Figure 5 shows an example of the additional logic to reduce pitch doubling. In the bottom panel of the autocorrelation function, only one pitch candidate P1 was chosen by the peak picking, and in fact this candidate was at twice F0. However, as indicated by the dotted line, a new pitch candidate P2 whose value is half of P1 was produced by the additional logic.

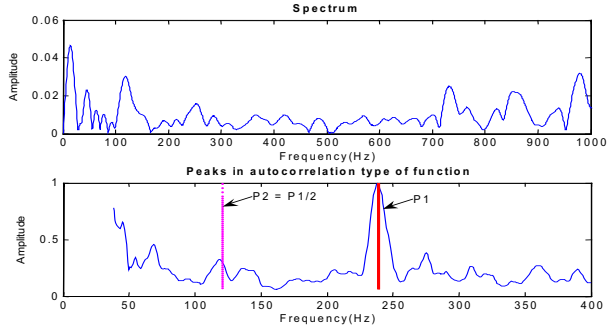

Figure 5. Candidate insertion to reduce pitch doubling

Given the candidates obtained from the processes above, dynamic programming is used to choose the best pitch track from the spectral candidates. Experiments (section 3) indicate that the spectral pitch track is quite good, but not quite a good as the one obtained by combining the spectral and NCCF tracks.

2.3. Final pitch determination

The result of the processing described in section 2.2 is a pitch track computed from the spectrum. This track was found to be quite robust with respect to pitch doubling or halving, but only approximates the pitch track found from the NCCF time domain processing. As illustrated in Figure 1, and described in more detail in [10], the spectral pitch track was used to guide the final tracking of the NCCF candidates. This spectral track is used to increase the merit of NCCF pitch candidates close to the spectral track, and reduce the merit of NCCF peaks far from the spectral track. A two-sided triangular weighting function was used, with a width of approximately 100 Hz. NCCF pitch candidates outside the window were eliminated.

The 3 panels in Figure 6 illustrate the spectral tracking.

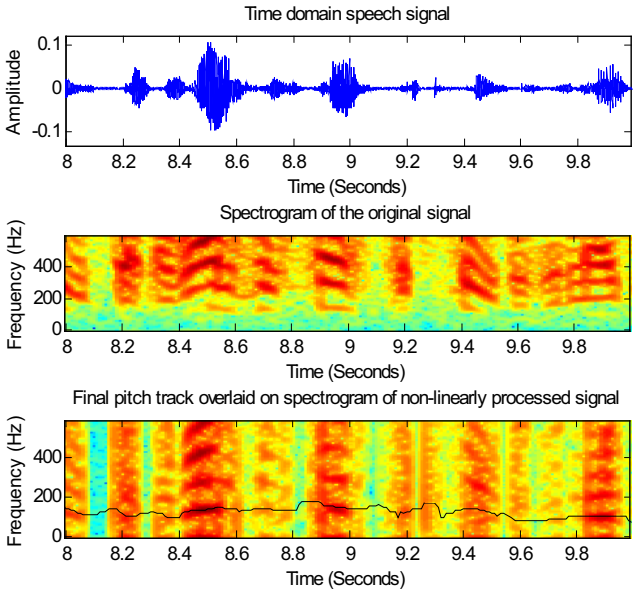

Figure 6. Illustration of the spectral pitch tracking

3. Experimental Evaluation

We evaluated the temporal/spectral pitch tracking using the Keele pitch extraction database [9]. This high quality speech (20 kHz sampling rate) contains 5 male and 5 female speakers, each speaking for about 35 seconds. The telephone version of this data, obtained from the spoken language systems group at MIT, was transmitted through a telephone channel and re-sampled at 8 kHz.

Although the Keele database includes what should be a very reliable control (which we call control C1), inspection of this track showed several instances of what appeared to be F0 halving. Therefore, we formed a second control (C2) by

applying the algorithm described in this paper to the first-differenced laryngograph signal. For some of the evaluations, errors were computed with respect to both C1 and C2.

Of the many error measures that can be used to quantify F0 tracking accuracy, we used only the gross error in the voiced sections to evaluate the tracking method reported in this paper. The gross errors (G_err) were computed as the percentage of frames such that the pitch estimate of the tracker deviates significantly (typically 20%) from the pitch estimate of the reference. The measure is based only on those frames for which the reference indicates voiced frames.

Tests were conducted with both studio quality speech and telephone speech, using both original versions, and versions with additive Gaussian noise to give an overall signal to noise ratio of 5 dB, as described in [11].

Table 1. Pitch tracking errors for various conditions

	Control	Studio clean(%)	Studio 5dB(%)	Tel clean(%)	Tel 5dB(%)
YAAPT	C1	4.26	7.62	8.14	17.85
YAAPT*	C1	1.59	1.99	2.69	4.48
Spectral method	C1	4.23	4.45	6.52	6.95
NCCF	C1	3.58	4.52	8.00	16.61

* Using control C1 for the spectral track

Table 1 gives the performance of YAAPT, YAAPT using control C1 as the spectral track, and individual performances of non-linear spectral method described in section 2 and the temporal method alone (NCCF). The second line in the above table give the upper limit on expected performance, if the spectral track is error free. Comparison of the first and last line shows that the time domain NCCF is better than YAAPT.

Table 2. Results with new method, combining spectral and temporal information

Error threshold	Control	Studio clean(%)	Studio 5dB(%)	Tel clean(%)	Tel 5dB(%)
10%	C1	5.61	8.04	9.91	15.66
10%	C2	3.93	5.75	8.21	13.74
20%	C1	2.97	3.92	5.25	6.93
20%	C2	1.51	2.10	3.81	5.36
40%	C1	2.32	2.40	2.23	3.25
40%	C2	0.79	0.76	1.59	1.72

The results of the new method tested against error thresholds of 10%, 20%, and 40% are shown in Table 2. The pitch tracks obtained from studio quality and telephone speech were evaluated with controls C1 and C2. Comparing the 20% error results for control C1 with the results in Table 1 shows that the new method is significantly improved over YAAPT, especially for telephone speech, but still does not match the upper bound on performance indicated in the second row of Table 1.

4. Summary

In this paper, a new pitch-tracking algorithm has been developed which combines multiple information sources to enable accurate robust F0 tracking. The multiple information sources include peaks selected from the normalized cross correlation of both the original and squared signal and smoothed pitch tracks obtained from spectral information. These multiple information sources are combined using experimentally determined heuristics and dynamic programming. An analysis of errors indicates better performance for both high quality and telephone speech than previously reported performance for pitch tracking with the same data and the same conditions. The routines mentioned in this paper are available from the first author of this paper as MATLAB functions.

5. Acknowledgements

This work was partially supported by JWFC 900.

6. References

[1] D. Talkin, "A Robust Algorithm For Pitch Tracking," Speech Coding and Synthesis, pp-495-518, 1995.

[2] M. Ostendorf and K. Ross, "A Multi-Level Model for Recognition of Intonation Labels," Computing Prosody, Y. Sagisaka, N. Campbell and N. Higuchi (Eds.), 291-308, Springer-Verlag, NY: 1997.

[3] Chao Wang and Stephanie Seneff, "Robust Pitch Tracking For Prosodic Modeling In Telephone Speech," ICASSP'00, Turkey.

[4] Pc Bagshaw, SM Miller and MA Jack, "Enhanced Pitch Tracking and the processing of the F0 contours for computer aided intonation teaching," Proc. EUROSPEECH'93, Berlin, pp. 1003-1006.

[5] Eric Chang, Jianlai Zhou, Shou Di, Chao Huang, Kai-Fu Lee, "Large Vocabulary Mandarin Speech Recognition with Different Approaches In Modeling Tones," ICSLP'00, Beijing.

[6] S. A. Zahorian, A. Zimmer, and B. Dai, "Personal Computer Software Vowel Training Aid for the Hearing Impaired," ICASSP'98, pp. VI-3625-3628, Seattle, Washington.

[7] Eric Mousset, William A. Ainsworth, Jose A.R. Fonollosa, "A comparision of several recent methods of fundamental frequency and voicing decision estimation," ICSLP'96, pp. 1273-1276, Philadelphia.

[8] Rabiner Lawrence R., Cheng, Michael, J. Rosenberg, Aaron E. And McGonegal, Carol A., "A comparative performance Study of several pitch detection algorithms," IEEE Transaction on Acoustics, Speech and Signal Processing, Vol. ASSP-24, 399-417, No-5, Oct'76.

[9] F.Plante, G.Meyer, and W. A. Ainsworth, "A pitch extraction reference database," EUROSPEECH'95, Madrid, pp. 837-840.

[10] Kavita Kasi, Stephen A. Zahorian, "Yet another algorithm for pitch tracking," ICASSP'02, pp. 361-364. Orlando.

[11] Tomohiro Nakatani, Toshio Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," J. Acoust. Soc. Am., Vol. 116, No. 6, pp. 3690-3700, December 2004.