AN ALGORITHM FOR LOCATING FUNDAMENTAL FREQUENCY MARKERS IN SPEECH SIGNALS

Princy Dikshit, Stephen A. Zahorian and Shivram Nagulapati

Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, VA, USA

ABSTRACT

This paper describes an algorithm for determining pitch period markers in a continuous speech signal using prior knowledge of pitch values. The algorithm uses dynamic programming to determine optimal markers from a set of probable markers. Local costs are assigned based on amplitudes of local peaks, and transition costs are based on the closeness of peak spacings to the spacing predicted by the pitch period. In tests conducted with the Keele database, it was found that over 90% of pitch markers are within ±0.3ms of reference marks in the laryngograph signal recorded with the Keele acoustic data, and over 96% of markers are within ±1 ms of reference markers. One of the features of this algorithm is tolerance to errors in the pitch track. The algorithm can tolerate errors in the pitch track of approximately -10% to +60%, with very little degradation in performance. It also generates =1% extraneous peaks.

1. INTRODUCTION

In this paper, an algorithm is described and experimental results reported for identifying fundamental frequency (pitch) markers in a speech signal, given that pitch tracking has already been performed. This work is motivated by the observation that even very accurate pitch-tracking algorithms are generally based on some type of smoothing over time, such as that implicit in an autocorrelation calculation of speech data several pitch cycles long, and thus pitch tracks do not identify individual pitch cycles in the acoustic waveform. Since some work has shown that speech recognition, especially under conditions of moderate noise, can be improved by pitch-synchronous analysis, the present work is intended to provide the detailed cycle-by-cycle identification of pitch periods that would be needed for pitch synchronous analysis. Other potential applications of the markings of pitch period markers include analysis of jitter, prosody in speech [1], text-to-speech synthesis [2, 3], analysis of voice quality and pitch synchronous speech analysis [4].

A number of pitch marking algorithms have been proposed in the literature, many of which primarily extract certain signal attributes such as Glottal Closure Instants (GCIs), excitation moments from LPC analysis, zero crossings, etc. These algorithms have not proven to be sufficiently accurate [5], [2] for pitch synchronous analysis. To overcome the accuracy problem, another group of algorithms have been proposed which use dynamic programming to combine multiple sources of information ([5], [1], [6] and [2, 3]). The reported algorithms using dynamic programming have improved accuracy for identifying pitch markers but still are insufficient for many desired applications. Recently, algorithms have been proposed that combine dynamic programming with GCI detection [7] or artificial neural networks [1] and have been successful in identifying pitch markers accurately. The pitch marking algorithm presented in this paper is another attempt to use dynamic programming along with the pitch information of a signal to identify pitch markers more accurately.

The algorithm presented for identification of pitch cycle markers in the speech signal is based on combining information from the pitch track and peak locations in the acoustic signal. The algorithm is presented in some detail in the third section of this paper. The algorithm was experimentally evaluated using the Keele database, using markers obtained from the laryngograph signal as a reference. Although the pitch cycles are quite apparent in the laryngograph signal, still some care in signal processing is required to obtain the control markers. Error measures were defined for evaluating the accuracy of the marker locations in the speech signal, and experimental results are presented in section 4 of this paper.

The algorithm used for pitch tracking is the YAPT [8] algorithm developed by Kasi and Zahorian (2002). Summarizing briefly, the kernel processing in YAPT is the normalized cross correlation (NCRSS), which has been found to be a reliable indicator of pitch even in the presence of rapidly changing speech amplitudes. To improve the reliability of the NCRSS, an approximate but robust pitch track is also computed from the low frequency spectrogram of both the original signal and the absolute value signal. Multiple potential pitch candidates identified by the NCRSS are reduced to a single "lowest cost" pitch candidate for each speech frame using dynamic programming. Costs are based on peak amplitudes in the NCRSS, continuity considerations, and the approximate pitch track determined from the spectrogram.

2. THE ALGORITHM

This section describes the algorithm developed in this work to obtain pitch period markers on a cycle-by-cycle basis, based on prior knowledge of pitch values of the signal. The computed pitch values are used to approximate the pitch period, which in turn is used to determine a block and frame size for use in the marking algorithm. In each block, the peaks are identified in the acoustic signal that are potential pitch cycle markers. These peaks are grouped into frames and dynamic programming is used to determine those peaks whose locations appear to be the best candidates as pitch period markers.

A more detailed description of the algorithm consists of the steps given below:

1. Block creation process.
2. Peak picking process.
3. Peak organization into frames.
4. Dynamic programming.
5. Post-processing of pitch markers.

A 'block' of the speech signal can contain any of the four types of signal: unvoiced (u), unvoiced followed by voiced (u-v), totally voiced (v), or voiced followed by unvoiced (v-u). A fundamental assumption is made that blocks are short enough in time (typically about 5 pitch periods, or about 30-60 ms) that normally at most one transition in voicing is made. Identification of pitch periods in the case of (v) regions of speech is generally relatively easy and accurate, as compared to regions with transitions between voiced and unvoiced intervals (u-v, v-u). The algorithm is used mainly to identify pitch cycles in the (v) regions of speech, but does attempt to identify pitch period makers in the voiced portions of (u-v, v-u) regions, and even "examines" portions of the "u" regions for these cases.

2.1. Block creation process

For all regions (v, u-v, v-u), the size of a block is a fixed multiple (typically 5) of the average pitch value in the (v) region of the block, typically about 60ms long. This method insures that a minimum of 3 complete pitch cycles are contained in the block. If the end of the preceding block was also voiced, the block begins at the location of the last pitch marker found in the preceding block. For the case of u-v blocks, the pitch is determined as the average value in the v region, the block starting point is 2 pitch periods prior to the u-v transition. For the case of v-u regions, the block is determined as for v regions, but the v-u transition point is noted for later use in processing.

2.2. Peak identification process

Identification of pitch cycles requires the use of landmarks that will mark the beginning and the end of each pitch cycle. In this algorithm, signal peaks are assumed to be those landmarks. A peak is considered as a candidate if it is the largest peak within a window of width 1.0 nominal pitch periods wide, centered about the peak. An examination of many seconds of speech data indicated that this process was able to identify nearly all pitch period markers, and tended to minimize the extraneous markers found. A cost value was then associated with each peak. In particular larger peaks are the best candidates for markers, and thus have the lowest costs, as given by

$$\text{Local cost} = 1 - \text{normalized amplitude.} \qquad (1)$$

Note that peaks are also located in the unvoiced regions, but are considered differently than are the peaks in the voiced regions, as described in a later section.

2.3. Organization of peaks into frames

In order to conveniently use dynamic programming, the peaks are grouped in terms of overlapping frames, with frames chosen so that ultimately one cycle marker should be found in each frame. With frames appropriately chosen, and assuming that the pitch tracking is reasonably accurate, dynamic programming can be used to find the best fitting markers to the pitch track. However, there may still be errors if the actual pitch periods deviate "too much" from the nominal pitch period. This is related in part to inaccuracy of the pitch tracking and also to the sometimes rapidly varying pitch periods within a block.

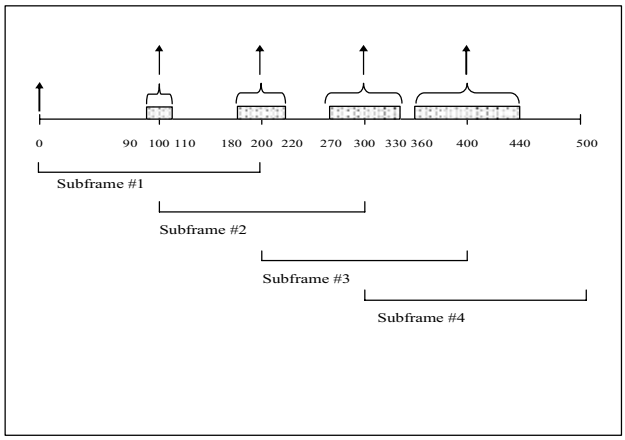

Figure 1: Illustration of frames within a block, and possible errors. The shaded regions represent the interval within which a pitch marker will lie assuming there is a possible ±10% error with respect to the assumed pitch period of 100 sample points. Thus the possibility of a pitch marker location outside the assumed area grows as the frame number increases in a block. This figure assumes a pitch marker (last one found in previous block) is located at the beginning of the block (a voiced region).

As shown in Fig. 1, frames are typically 2*pitch periods long and are spaced 1 pitch period apart. Since the first frame in a block is synchronized to begin at the last pitch period marker found in the preceding block (in the case of a voiced region), the expected location of each pitch period marker is at the center of each frame. However, as shown in Fig. 1, if the actual period deviates by as much as 10% from the value obtained from the pitch track, after only 4 frames, the range of locations for pitch period markers is nearly as wide as the entire frame.

2.4. Use of dynamic programming

After determining the peaks and grouping them into frames, two matrices are formed. One matrix (A) stores the *LOCAL COST* values (as mentioned above) and the other matrix (B) stores the *TRANSITION COST* values. The entries in matrix B are linearly proportional to the squared difference between peak locations in consecutive frames, as compared to the expected pitch periods. Thus the transition costs increase as the spacing deviate from the expected spacing based on the pitch track.

The transition cost described above is defined as follows:

$$\text{TransitionCost} = \left(\frac{\text{Est.pitch period} - (\text{I-J})}{\text{Est.pitch period}} \right)^2 \qquad (2)$$

Where, I = candidate peak location in i^{th} frame,
J = candidate peak location in $i\text{-}1^{th}$ frame,
and Est.pitch period is the estimated pitch period.

Dynamic programming is then used to find the lowest weighted

cost path through the A and B matrices. The results of this path are the most likely pitch markers over the interval processed.

2.5. Post processing

The post processing step is used to take a closer look at peaks found from the preceding step, which lie in regions of speech classified as unvoiced. Markers that are found in unvoiced region are retained only if they have large amplitude values. Normally, markers found in unvoiced regions with amplitudes greater than 1/5 (v-u) or 1/10 (u-v) of the largest amplitude peak in the block are retained, whereas smaller peaks are eliminated. The bottom panel in figure 2 shows the final pitch markers identified in a sample speech signal.

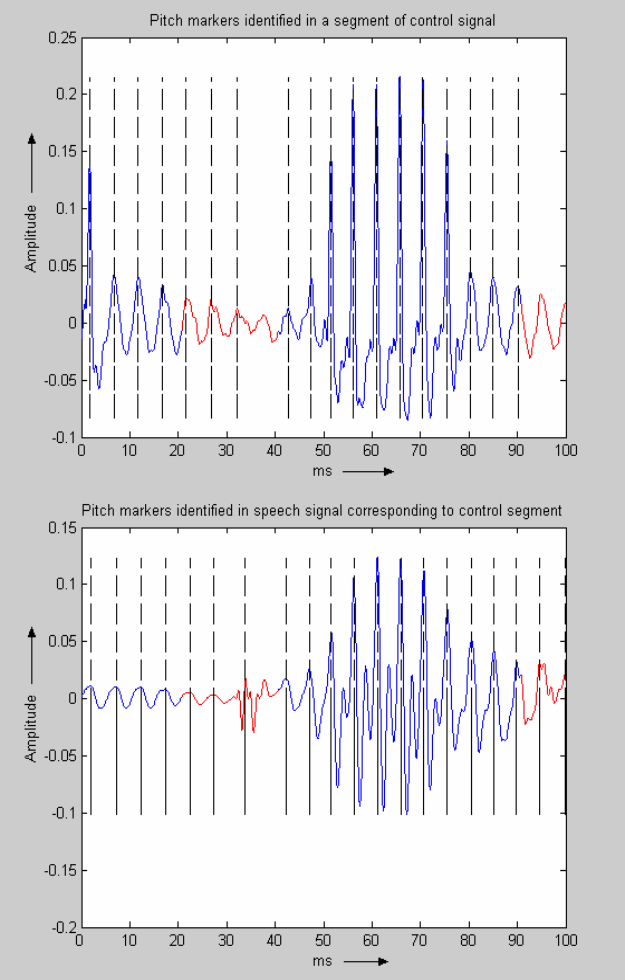

Figure 2: Top panel shows the pitch markers (black dashed lines) identified in control signal. Bottom panel shows pitch markers identified in speech signal. Note signal in both the panels corresponds to v-u-v-u transitions.

3. EXPERIMENTAL VERIFICATION

3.1 Database

The database used for evaluating the cycle-by-cycle marking algorithm is the Keele database [9] which consists of 10 sentences (5 male speakers, 5 female speakers) sampled at 20 kHz, each about 30 seconds long. This particular database has been designed for evaluations of pitch tracking routines, since it is reasonably large and varied, and also has a simultaneously recorded laryngograph signal for which the pitch periods are quite apparent. The database is also supplied with a reference pitch track, which can be used for evaluation of algorithms.

3.2. Control markers

The laryngograph signal mentioned above was used to compute control markers so that the algorithm presented in this paper could be evaluated quantitatively. After extensive empirical testing via visual inspections of the waveform, it appeared that a heuristic algorithm was more reliable for locating the markers in the laryngograph signal than the more complex method described above. Note that since the primary method given in this paper was developed explicitly for acoustic speech signals, and since the larygngograph signal is substantially different from the acoustic signal, it does not seem particularly surprising that a different method for identifying markers was more suited to the control signal. The best method found was to first-order difference the signal, lowpass filter it at 1 kHz, and then to locate prominent peaks within a spacing approximately equal to the pitch period in the supplied reference pitch track. Figure 2 illustrates the algorithm, showing the markers obtained for both the control signal (top panel) and speech signal (bottom panel).

3.3. Algorithm for error determination

In order to compute errors, the following two processing steps were first performed.

1. The first step was to locate speech markers closest to each pair of consecutive control markers (control markers located at C_i and C_{i+1}, speech markers located at S_i and S_{i+1}).
2. The difference between the speech markers from step 1, $T_S = S_{i+1} - S_i$ was computed and the difference between the control markers, $T_C = C_{i+1} - C_i$, was computed. The differences between these differences, $T_D = T_S - T_C$, was considered as the error. If no speech marker was found close to either or both of the control markers, a default value (typically average pp) was used to define the error. Note that the difference of differences was used as an error measure, rather than absolute differences between markers in the control and speech signal, since there appeared to be delays between the two signals.

4. EXPERIMENTAL RESULTS

Evaluations were conducted using two different pitch tracks. For one case, the reference pitch track provided in the Keele database was used and for the other case, the YAPT-generated pitch track was used. In both cases, however, the supplied reference pitch track was used to compute the reference markers from the control. It was found that the marking accuracy in the speech signal was slightly higher using the YAPT generated pitch track for the acoustic signal pitch marking, so results are shown only the YAPT-computed pitch track as a graph (Fig. 3). Plotted are the percentage of pitch cycles for which the error is less than or equal to a certain deviation, as given on the x axis. From Fig. 3, it can be observed that the accuracy is considerably higher for the female speakers than for the male speakers. The overall marking accuracy is quite high, as over 90% of all marks are extremely

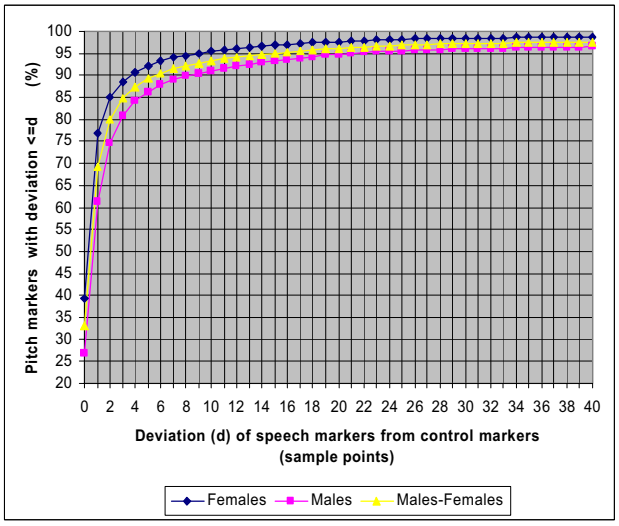

Figure 3: Accuracy of pitch tracking using YAPT pitch track.

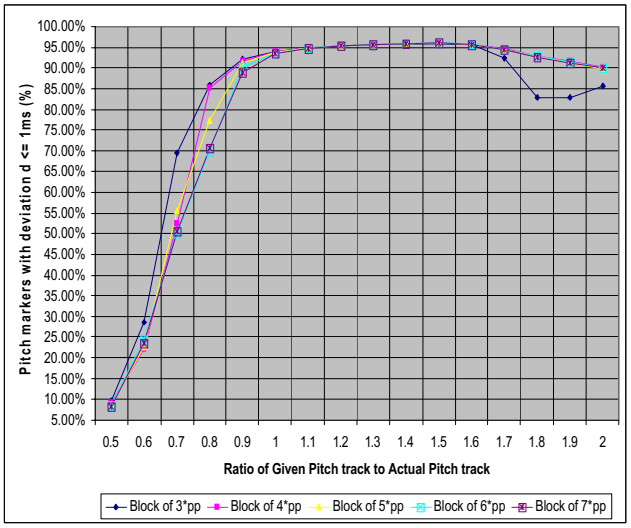

Figure 4: Robustness of the algorithm to errors in pitch track used for identifying pitch markers.

close (within 0.3ms) to the marks in the laryngograph signal, and over 96% of the pitch marks are quite close (within 1 ms) to the control marks. These results were obtained with a local cost/ transition cost ratio of .75. In general it was observed that results did not change much as the ratio was varied between 0 (i.e., local costs not used) to 1 (local and transition costs equally weighted). However, accuracy did degrade substantially if local costs were weighted much more heavily than transition costs.

Another set of experiments was conducted to determine the robustness of the algorithm to errors in the pitch track used for estimating speech markers, with results shown in Fig. 4. The numbers on the x-axis represents the amount of error present in pitch track used w.r.t the reference pitch track; a value of 1 implies no error, a value of 2 implies 100% error in track (pitch doubling). This experiment used the reference pitch track instead of the YAPT pitch track for estimating speech markers as, presumably, this supplied track is a better standard. The tests were conducted using different block sizes (3*pp-7*pp). The algorithm was found to have an error tolerance range of -10% to +100% (in the frequency domain) or -50% to +11.1% (in the time domain) when considering 90% or more of the deviations'd' to be =1ms. For this tolerance range, the number of false positives generated by the algorithm ranged from a minimum of 0.04% (7*pp long block, -10% error in track) to a maximum of 17.87% (3*pp long block, +100% error in track). Upon considering only the instances when the number of extraneous peaks generated is =1%, the algorithm has a tolerance range of -10% to +60% (frequency domain), for block sizes (3*pp-7*pp).

5. SUMMARY

In this paper we have presented an algorithm to determine pitch markers in the speech signal based on a prior knowledge of pitch values. Experiments have shown that the algorithm results in about 96% of the total pitch periods having a small error (less than 1 ms) and about 90% of frames having an extremely small error (less than .3 ms). The algorithm has also been shown to be robust to errors in the pitch track used for identifying pitch markers. It also generates =1% extraneous peaks.

6. ACKNOWLEDGEMENTS

This work was partially supported by NSF grant BES-9977260.

7. REFERENCES

[1]. Harbeck S., Kießling A., Kompe R., Niemann H. and Nöth E, "Robust pitch period detection using dynamic programming with an ANN cost function", *Proc. EUROSPEECH*, Madrid, vol. 2, pp. 1337-1340, September 1995.

[2]. V.Colotte and Y Laprie, "Higher precision pitch marking for TD-PSOLA", *Proceedings of XI European Signal ProcessingConference (EUSIPCO)*, Toulouse, 2002.

[3]. Laprie, Yves and Colotte, Vincent, "Automatic pitch marking for speech transformations via TD-PSOLA", *European Signal Processing Conference (EUSIPCO),* Rhodes, 1998.

[4]. Moulines, E. and Charpentier, F., "Pitch-Synchronous Waveform Processing Techniques for Text-To-Speech Synthesis Using Diphones", *Speech Communication,* 9: 453-467, 1990.

[5]. Veldhuis, Raymond, "Consistent pitch marking", *International conference on Speech language Processing*, vol.3, 207-210, 2000.

[6]. V. Goncharoff and P. Gries, "An Algorithm for Accurately Marking Pitch Pulses in Speech Signals", *Proc. IASTED'98 International Conference on Signal and Image Processing*, Las Vegas, NV, October 1998.

[7]. Kounoudes, A., Naylor, P.A. & Brookes, M, "The DYPSA Algorithm for estimation of Glottal closure instants in voiced speech", *Proc IEEE Int. Conf. Acoust., Speech, Signal Processing,* pp I.349-I.352, May 2002.

[8]. Kavita, Kasi, and Zahorian, S., "Yet Another Algorithm For Pitch Tracking", *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Orlando, Fl, May 2002.

[9]. F.Plante, G.Meyer, and W. A. Ainsworth, "A pitch extraction reference database", *EUROSPEECH*, Madrid, pp. 837-840, 1995.