A post-processing algorithm for time domain pitch trackers

Philippe Specker

21 January 1983

Carnegie-Mellon University
Computer Science Department
Schenley Park
Pittsburgh, PA 15213

This research was sponsored in part by the National Science Foundation, Grant MCS-7825824 and in part by the Defense Advanced Research Projects Agency (DOD), ARPA Order No. 3597, monitored by the Air Force Avionics Laboratory Under Contract F33615-78-C-1551.

Table of Contents

List of Figures

List of Tables

Abstract

This paper describes a powerful post-processing algorithm for time-domain pitch trackers. On two successive passes, the post-processing algorithm eliminates errors produced during a first pass by a time-domain pitch tracker. During the second pass, incorrect pitch values are detected as "outliers" by computing the distribution of values over a sliding 80 msec window. During the third pass (based on artificial intelligence techniques), remaining pitch pulses are used as anchor points to reconstruct the pitch train from the original waveform. The algorithm produced a decrease in the error rate from 21% obtained with the original time domain pitch tracker to 2% for isolated words and sentences produced in an office environment by 3 male and 3 female talkers. In a noisy computer room errors decreased from 52% to 2.9% for the same stimuli produced by 2 male talkers. The algorithm is efficient, accurate, and resistant to noise. The fundamental frequency micro-structure is tracked sufficiently well to be used in extracting phonetic features in a feature-based recognition system.

1. Introduction

Feature-based speech recognition systems categorize phonetic events using acoustic/phonetic features extracted from the signal [1]. These features must be detected very precisely in order to produce low error rates. A necessary component of a feature-based system is a pitch tracker; decisions about phonetic events requires *precise* specification of the onset and offset of voicing, which requires the exact location of the pitch pulses in the waveform. For example, an important difference between the letter "p" and "b" is the duration of the consonant noise from burst onset to vowel onset. The location of the vowel onset must therefore be correctly identified with less than a few milliseconds error, and information about voicing is needed to help locate this point in an utterance. In addition to voicing, pitch micro-variations can give important clues to identify a sound.

The reliable measurement of pitch periods from the waveform is very difficult. Several factors limit the accuracy of time-domain pitch trackers. First, pitch can rapidly change, which results in a glottal pulse excitation train which is not exactly periodic. Second, the shape of the vocal-tract can significantly alter the glottal waveform so that the pitch pulses can be difficult to detect. Finally, voiced and unvoiced intervals can be similar when the amplitude of the pitch pulses is low. For these reasons, the reliability and accuracy of existing time-domain pitch trackers is highly variable [2].

The accuracy of existing pitch trackers is further limited by the fact that computations are usually done

locally, over a 10 or 20 msec window. Over periods of this duration, the speech waveform is not always stationary and is often characterized by "unusual " properties. For example, the primary pitch pulses can be smaller than adjacent secondary peaks, distorted by adjacent peaks or lost in the noise.

A possible solution to this problem is to use more global properties of the signal to direct the algorithm to those pulses which are most likely to be correct. Examination of speech waveforms reveals that globally (over 300 or 400 msec of speech), the speech waveform almost always displays the expected and usual characteristics, rather than the exceptional properties just described. Thus, when taking a more global perspective, unusual properties of the waveform can be readily detected[1]. It should therefore be possible to write a post-processing algorithm for time-domain pitch trackers, based on artificial intelligence techniques and including speech-specific rules, that evaluates the pitch train on a global base.

The implementation of our pitch tracker uses three successive passes. An existing time-domain pitch tracker is run as a first pass. Incorrect pitch values are detected as "outliers" by computing the distribution of values over a sliding 80 msec window (second pass). Pitch pulses which are close to the mean are then used as anchor points to find the correct pulses. These pulses are used to construct a chain of linked pulses. In this way, the correct pitch train is reconstructed from the original waveform (third pass). At this point, global decisions over 300 or 400 msec of speech are made in order to choose the right pitch chains.

2. Description of the Pitch Tracker

2.1 First pass : existing pitch tracker

An existing time domain pitch tracker is used as a first pass [3]. This pitch tracker consists of :

1 - a low pass filtering of the speech signal with a cutoff of 700 Hz.

2 - pitch estimation by correlating properties (duration, amplitude, energy) of each positive peak to successive positive peaks. This operation is repeated for negative peaks.

After this first pass an average of 21% of the pitch pulses in an office environment, and 52% in a noisy computer room, are missed or incorrectly labeled (see Table 3-3 and 3-5). This is due to the fact that speech is non-stationary, so that corresponding successive peaks in the waveform can be very different and not comparable by correlation.

---

[1] Our studies of waveforms indicate that, more than 95% of the time, for different sounds and speakers, the number of "usual" pitch pulses in a long utterance is greater than the number of "unusual" pulses. By "usual" pitch pulses, we mean pulses that are larger than secondary peaks and are therefore easily detected as pitch pulses by an expert.

Figure 2-1 and 2-2 gives an example of the pitch values (in Hz) after the first pass. The two figures represent 100 msec of speech spoken, respectively, by a male and a female speaker. In these figures, major peaks which are not accompanied by a number have been missed by the pitch tracker. It can be seen that pulses have been missed or mislabeled in Figure 2-1 (in this case secondary pulses have been labeled as pitch pulses), while numerous pitch pulses have been missed in Figure 2-2.

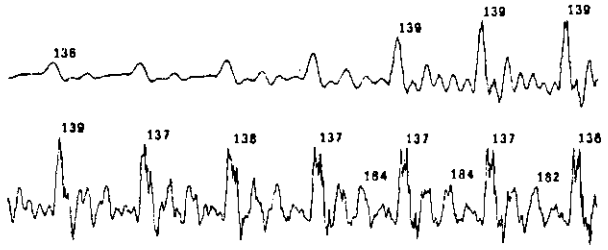

Figure 2-1: Pitch values (in Hz) after the first pass (1)

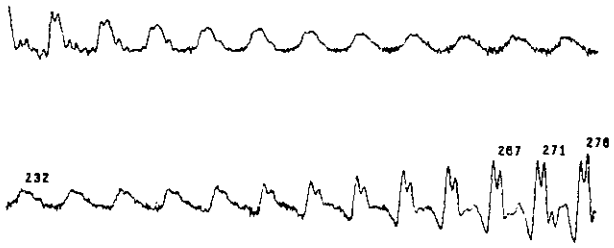

Figure 2-2: Pitch values after the first pass (2)

2.2 Second pass : detection and removal of deviant pitch values

After the first pass it is necessary to remove the incorrect pitch values so that they can be replaced by correct values during the third pass. A moving 80 msec window is used to compute a distribution of the local pitch values, estimated from the positive and negative peaks in the signal. The duration of the window must not be too long, because pitch can vary by as much as 40% within 150 msec. Figure 2-3 displays the histogram of pitch values computed for an 80 msec window within the 100 msec of speech shown in Figure 2-1. It can be seen in Figure 2-1 that the correct values are in the range 130-140 Hz.

If the 80 msec window includes an insufficient number of pitch values (as in Figure 2-2) a crude pitch estimation, computed over 300 msec, is used to define a lower and upper bound of the expected pitch values, for instance 0.6 and 1.4 times the average pitch in this interval.

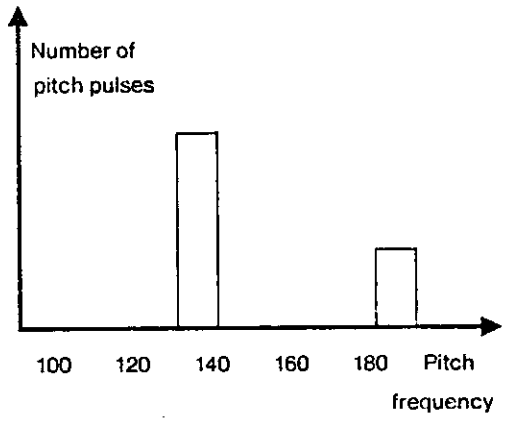

Figure 2-3: Distribution of pitch values for a 80 msec window of speech

Once the average pitch in a window and the expected pitch range have been estimated (the typical variation around the mean is 15 Hz), the deviant values are detected as "outliers" (when they are outside this expected range) and eliminated.

Figure 2-4 shows the pitch values obtained after the second pass, given the initial values of Figure 2-1.

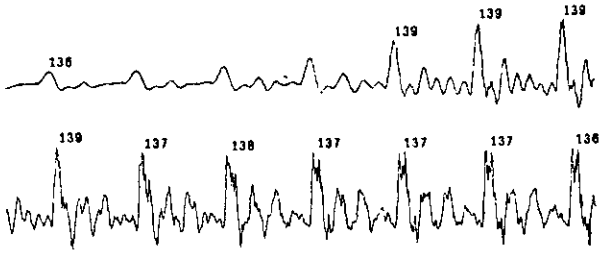

Figure 2-4: Pitch values after the second pass

2.3 Third pass : reconstruction of the pitch train

During the third pass the pitch values remaining after the second pass are used as anchor points to reconstruct, forward and backward, the pitch train.

2.3.1 Chaining

Most of the time, numerous pitch pulses are correctly labeled in a voiced segment of speech after the second pass. Nevertheless, for didactic reasons, suppose that only one pitch value is left in a voiced segment after the second pass (Figure 2-5) (in fact, this situation can effectively happen if the pitch pulses are rapidly changing).

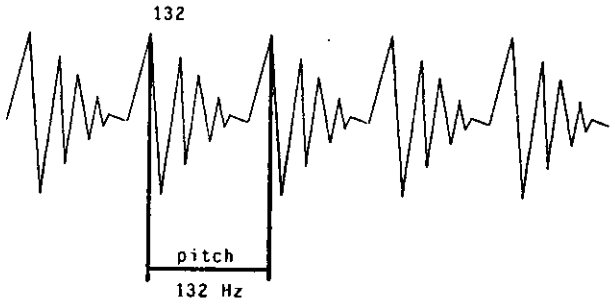

Figure 2-5: Reconstruction of the pitch train (1)

This single value is now used to find the next pitch pulse. For this we assume that the pitch variation is not larger than 15% from one pitch pulse to the next (In fact, this hypothesis is true only when the pitch is slowly changing. We will see later how to handle fast changing pitch pulses). The algorithm tries to find a maximum in the waveform starting at the place where the pitch pulse would be if the pitch frequency was constant, then examines an expanding interval around this point, up to a 15% variation from the original pitch value (Figure 2-6). If a maximum is found, it is taken as the new pitch pulse and the exact pitch value (for instance 138 Hz in the current example) is computed using the number of samples from the preceding pulse to the current one.

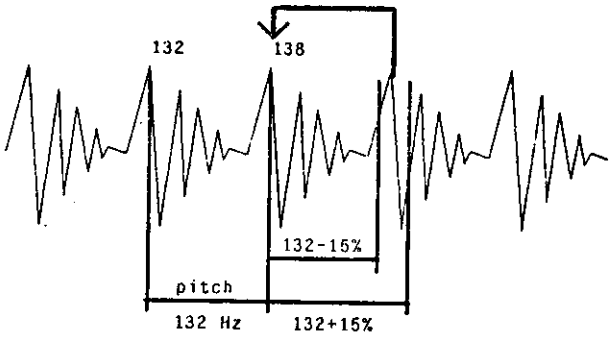

Figure 2-6: Reconstruction of the pitch train (2)

The operation is then repeated for the next pulse using the average pitch of the two last pulses as the expected pitch value (Figure 2-7). The chain construction process stops if no maximum is found in the 15% variation interval allowed (i.e. if the pitch variation is too large to be captured by the algorithm or if an unvoiced segment is reached).

The forward pitch train for the whole voiced segment of speech is reconstructed this way. Similarly, a "backward" pitch train is reconstructed from the same anchor point using an analogous algorithm. During these two operations the successive pitch pulses are **linked in a chain** (Figure 2-8). This operation will allow global decisions over 200 or 300 msec of speech to be made at a later stage of processing.

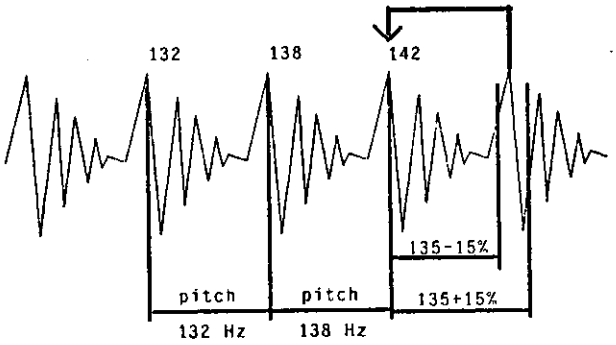

Figure 2-7: Reconstruction of the pitch train (3)

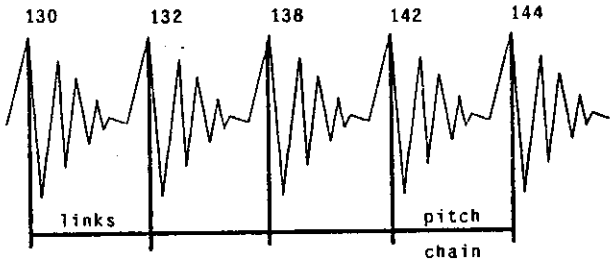

Figure 2-8: Reconstruction of the pitch train (4)

The chain construction process is repeated for all pitch values remaining after the second pass. If labeled pitch pulses are captured during this process, the pitch value, approximate after the second pass, is set to the exact value. The process stops if a pitch pulse already linked in a chain is encountered. In this way, only one chain of comparable pitch pulses is built if several pulses are correctly located in a voiced segment after the second pass.

2.3.2 Global computation

During the first phase of the third pass (chaining), chains of linked pitch pulses are created. Each chain describes an ensemble of pitch pulses with the same properties (i.e. slowly changing pitch values , comparable shapes ...). At this point a global computation over a few hundred msec of speech must be made, using the existing pitch chains, to eliminate errors due to local computation and to include the unusual pulses (i.e., those with fast changing pitch values, different shapes ...) in order to reconstruct exactly the whole pitch train in the utterance.

2.3.2.1 Parallel pitch chains

After the first phase of the third pass, it occasionally happens (especially for female speakers) that two parallel chains of pitch pulses describe the same segment in the signal (Figure 2-9). This is due to the fact that pitch estimation is strictly local during the first pass (by correlation of successive comparable peaks), so pitch values can be incorrectly located on secondary peaks in the waveform instead of the actual pitch pulse. A chain of secondary peaks can therefore be created if the conditions are favorable (i.e. if the secondary peaks are stable in the waveform).

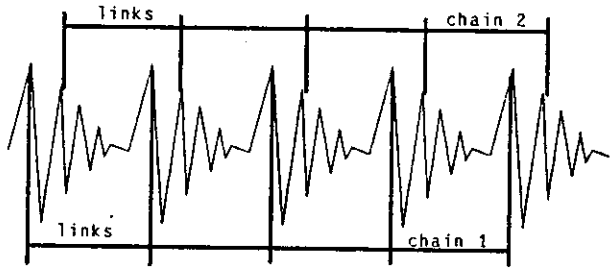

Figure 2-9: Example of parallel pitch chains

At this point the parallel chains are compared on the basis of global considerations about the waveform (number of waveform maximums included in a chain) over a time interval of 200 or 300 msec of speech, using adjacent chains as anchor point. The chain most likely to be right is chosen as the correct pitch chain including the actual pitch peaks.

2.3.2.2 Adjacent chain connection

If the pitch variations are too large, the pitch chain construction process will not be able to include all peaks in a train. Two independent and adjacent chains will be built (Figure 2-10). This situation often occurs in vowel-nasal sequences for instance.

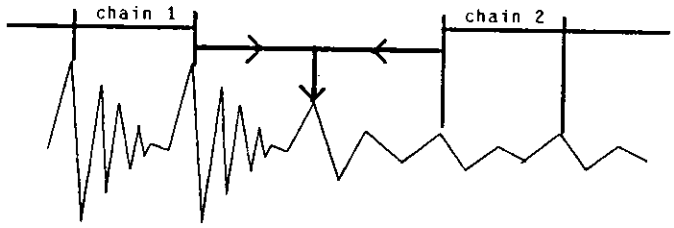

Figure 2-10: Connection of two adjacent pitch chains

A mechanism of chain connection can be implemented, using the pitch values at the end of both chains to estimate the position of the undetected pitch pulses and construct a unique chain.

2.3.2.3 Capture of endpoint pitch pulses

At the beginning and the end of a voiced segment of speech the pitch values often dramatically change (30% or more over a few pitch pulses) during the onset and offset of voicing. These isolated pulses (at the beginning and end of pitch chains) must be captured separately (Figure 2-11) using specialized procedures anchoring on the existing chains.

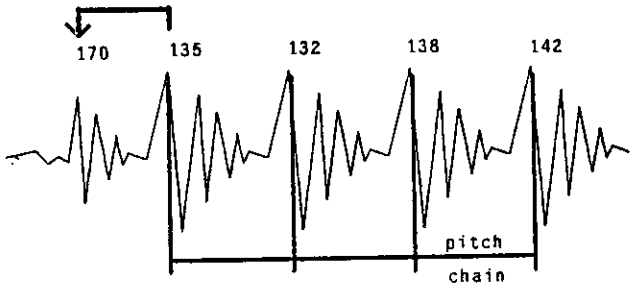

Figure 2-11: Capture of a pitch pulse at the beginning of a voiced segment

2.3.3 Third pass summary

During the third pass, the pitch train is reconstructed, anchoring on the pitch pulses correctly located after the second pass. One pitch pulse properly labeled in a voiced segment of speech is usually enough to reconstruct the entire segment pitch train, so that every pitch pulse has been exactly located in the utterance. Because of this characteristic the post-processing algorithm is little affected by noise. In this case the number of pitch pulses detected during the first pass decreases but most of the time at least one pulse is located in a voiced segment and the reconstruction process can be used. However there is no way to do so if no pitch pulse has been located in a segment during the first pass.

Figure 2-12 and 2-13 shows an example of the pitch values after the third pass. By comparison to the original values after the first pass (Figure 2-1 and 2-2), we can see that the pitch pulses have been either correctly located during the reconstruction process or have had their pitch value corrected.

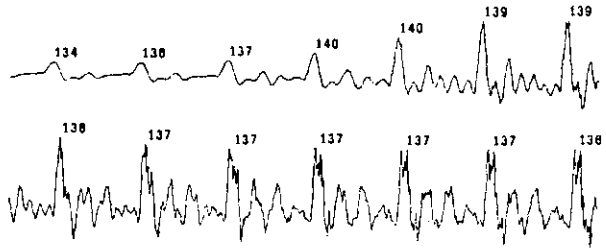

Figure 2-12: Pitch values after the third pass (1)

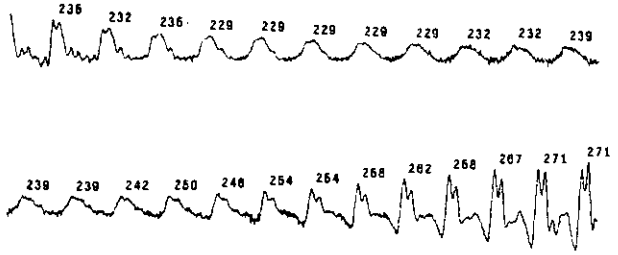

Figure 2-13: Pitch values after the third pass (2)

3. Evaluation of the pitch tracker algorithm

3.1 Definition of the error rate

Most pitch trackers estimate the pitch in successive 10 msec intervals (see Rabiner et al [2]). The usual way to evaluate these algorithms is to manually check the speech signal for the presence or absence of voicing, as well as the pitch in voiced intervals. The error rate is then computed for different categories as the ratio of the number of intervals in that category correctly labeled to the total number of intervals in that category. For instance, the voiced-to-unvoiced error rate gives the accuracy of correctly classifying voiced intervals and is computed as the ratio of the number of voiced intervals taken as unvoiced to the total number of voiced intervals.

This procedure for computing the error rate is not well adapted to the current pitch tracker since all pitch pulses in the utterance are supposed to be exactly located by the post-processing algorithm. We prefer to define the error rate as the number of mislabeled pitch pulses to the total number of pitch pulses in the utterance. The mislabeled pulses include missed pulses (not finding a pitch pulse when there is one), false alarms (finding a pitch pulse when there is none) and incorrect pitch values (pitch errors larger than 2 Hz by

comparison to pulses manually labeled on a waveform display). The overall error rate is the sum of these three types of errors.[2]

3.2 Stimuli

The stimuli, taken from Rabiner et al [2], consisted of the four mono-syllabic isolated words and four sentences shown in Table 3-1.

Words	Sentences
Hayed	We were away a year ago
Heed	I know when my lawyer is due
Hod	Every salt breeze comes from the sea
Hoed	I was stunned by the beauty of the view

Table 3-1: Stimuli

The isolated words and sentences were spoken by :

- three male and three female speakers in an office environment (40 dB average signal-noise ratio) using a noise-canceling microphone.
- two male speakers in a noisy computer room (20 dB average signal-noise ratio) using a low quality microphone.

3.3 Results

3.3.1 Office environment

Table 3-2 gives pitch characteristics for the stimuli spoken by six speakers in an office environment : the total number of pitch pulses, the minimum, maximum and average pitch values for each speaker.

Table 3-3 gives the different error rates (computed under the conditions explained in paragraph 3.1) for isolated words and sentences recorded by six speakers in an office environment, before and after post-processing (i.e. after the first and the third pass). We can notice a substantial drop in the overall error rate, accompanied by a slight increase in false alarms. This increase in false alarms is caused by the presence of

---

[2] At this point a theoretical problem appears : the peaks labeled as false alarms are not included in the count of the total number of pitch pulses in the utterance, since false alarms are defined as finding pitch pulses where there are none. Mathematically they should then not be included in the mislabeled pulses. Nevertheless the error rate can only be incremented by doing so and the false alarms are not numerous. In such conditions we think that the error rates previously defined gives a better overall view of the pitch tracker accuracy.

	Total number of pitch pulses	Minimum pitch	Maximum pitch	Average pitch
Male speakers				
ap	848	103	200	141
fa	581	58	114	86
rt	997	92	198	132
Female speakers				
bf	1126	137	250	184
bg	1689	148	390	235
tp	1584	143	291	223

Table 3-2: Speakers pitch characteristics (office environment)

noise peaks in unvoiced intervals that are occasionally labeled as voiced peaks by the endpoint pitch pulse algorithm (see paragraph 2.3.2.3). This happens when random noise peaks are located in the prolongation of pitch chains (with a small pitch variation from one pulse to the next one).

	Missed pulses		False alarms		Incorrect pitch		Overall error rate	
Male speakers	before	after	before	after	before	after	before	after
ap	17.3%	0.8%	0%	0%	10.0%	0.5%	27.3%	1.3%
fa	23.9%	1.2%	0%	1.4%	0.1%	0.5%	24.0%	3.1%
rt	16.2%	1.1%	0%	0.2%	1.0%	0.0%	17.2%	1.3%
Female speakers								
bf	12.7%	0.5%	0%	0.1%	5.7%	0.2%	18.4%	0.8%
bg	11.9%	2.2%	0.2%	0%	13.2%	1.2%	25.3%	3.4%
tp	10.2%	1.6%	0%	0%	4.2%	0.3%	14.4%	1.9%
All speakers	15.4%	1.3%	0.0%	0.3%	5.7%	0.4%	21.1%	2.0%

Table 3-3: Error rates before and after post-processing (office environment)

3.3.2 Noisy computer room

Table 3-4 gives pitch characteristics for the stimuli spoken by two male speakers in a noisy computer room : the total number of pitch pulses, the minimum, maximum and average pitch values for each speaker.

Table 3-5 gives the different error rates for isolated words and sentences recorded by two male speakers in a noisy computer room , before and after post-processing (i.e., after the first and the third pass). We can see that after post-processing the missed pulses and incorrect pitch error rates are comparable in an office environment and in a noisy computer room. The false alarm error rate is higher under noisy conditions because random peaks are more likely to be labeled as pitch pulses.

	Total number of pitch pulses	Minimum pitch	Maximum pitch	Average pitch
mq	1027	100	213	170
rc	902	94	195	136

Table 3-4: Speakers pitch characteristics (noisy computer room)

	Missed pulses		False alarms		Incorrect pitch		Overall error rate	
	before	after	before	after	before	after	before	after
mq	25.7%	1.1%	0.8%	0.5%	25.3%	0.2%	51.8%	1.8%
rc	35.3%	0.3%	2.0%	3.3%	16.2%	0.3%	53.5%	3.9%
All speakers	30.5%	0.7%	1.4%	1.9%	20.7%	0.3%	52.6%	2.9%

Table 3-5: Error rates before and after post-processing (noisy computer room)

3.4 Computational considerations

The pitch tracker, including the post-processing algorithm, is efficient. The first pass uses a low-pass digital filtering of the signal and correlation of successive peaks properties, operations that are not very time expensive and approximately linearly dependent on sampling rate. The second pass computes histograms of pitch values, also a simple operation. The third pass generates, and operates on, pitch chains, operations that can be implemented in an efficient way. Therefore, it should be possible to optimize the pitch tracker to work in real time on a 1 Mips computer.

4. Utilization of the post-processing algorithm

The results show that, 98% of the time, the post-processing algorithm produces a train of pulses in which every pitch pulse in the waveform is exactly located. In these cases, all pitch values are correct; there are no fine pitch errors. Therefore, the fundamental frequency micro-variations in speech are tracked sufficiently well to be used in extracting phonetic features in a feature-based recognition system. An example of feature extraction made possible by a precise pitch train is the localization of transitions between adjacent vowels. A sudden pitch variation often occurs as the vocal tract begins to move from one stable position to the next one.

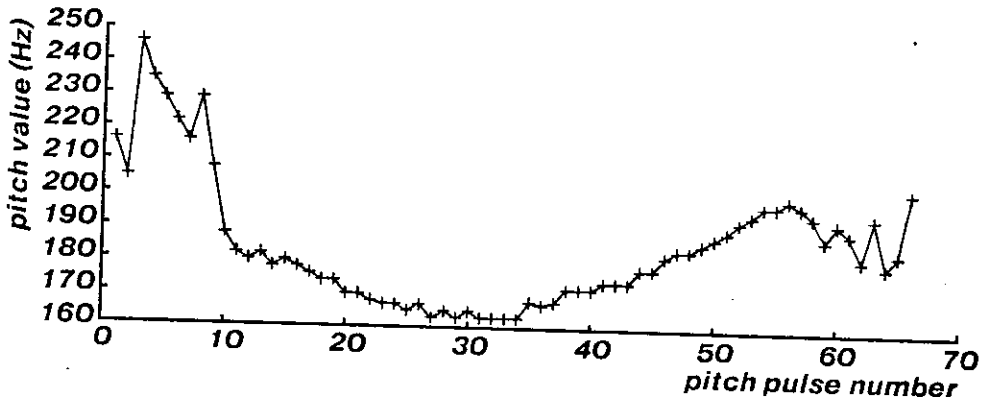

Figure 4-1: Pitch contour for the letter "a" spoken by a female speaker

For instance, Figure 4-1 and Figure 4-2 show respectively the pitch contour and the detailed pitch values for the letter "a" spoken by a female talker. We can see a sudden increase of the pitch value for the pitch pulse number 35 (with a jump from 163 to 168 Hz), corresponding to the beginning of the transition from the vowel [e] to the vowel [i]. We can also notice the unstable or "unusual" pitch pulses at the beginning and end of the utterance during the onset and the release of the vocal cords.

Although data produced by the post-processing algorithm has shown that vocal tract shape changes are strongly correlated with pitch changes, these changes are not consistent in amplitude and direction from one speaker to the other. Exact pitch values can give important clues for feature-based speech recognition systems, but are difficult to use properly.

A simpler usage of the pitch tracker has been made for a feature based, speaker-independent, isolated letter recognition system built at CMU [1]. In this system the pitch tracker has been modified to return an accurate estimation of the pitch value every 3 msec. Many of the feature extraction algorithms used in this system depend upon the location of voicing in the utterance. For instance the letters "a" and "h" can be distinguished by considering only the ratio of the number of voiced 3 msec slices to the total number of 3 msec slices in the sound. For "a" this ratio will be high, while low for "h". Examination of histograms of the ratio values for "a" and "h" reveals that there is no overlap between the two distributions of values for 4 tokens of each letter spoken by 40 speakers. Other features can be extracted by using the onset or offset of voicing as anchor points in an utterance. Finally, the average pitch in an utterance has been found to be useful for speaker normalization. The error rate of the existing system, in a speaker independent mode, is about 10% in an office environment.

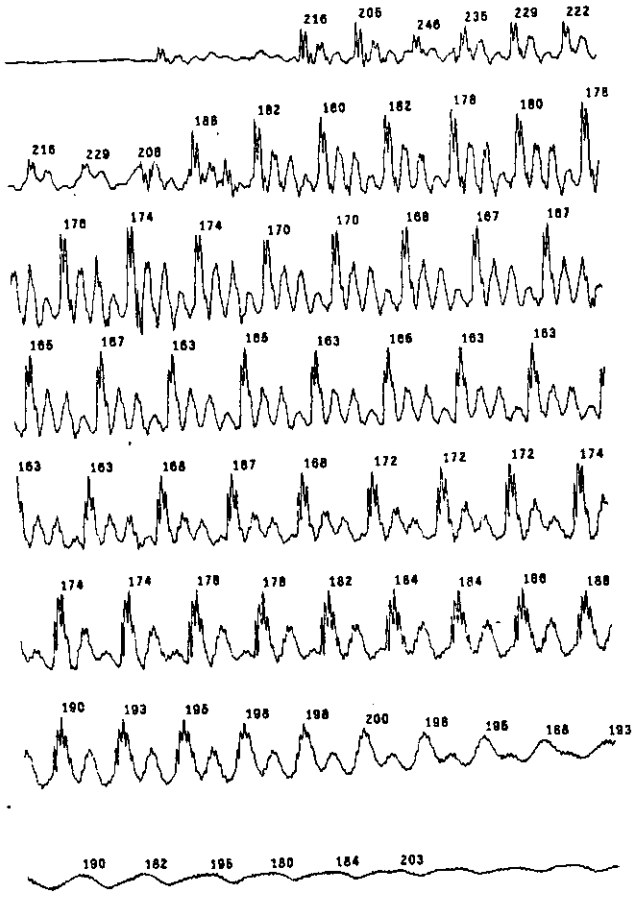

Figure 4-2: Detailed pitch values for the letter "a" spoken by a female speaker

5. Conclusions

In this paper we present a post-processing algorithm for time-domain pitch trackers. This algorithm is powerful because global processing over 300 or 400 msec of speech is used to estimate the pitch train, by building chains of pitch pulses. The algorithm is little affected by noise because a few pulses correctly located in a voiced segment of speech by the time domain pitch tracker are sufficient to reconstruct the pitch train for the whole segment. After the post-processing, all the pitch pulses are exactly located in the waveform with high accuracy (2% mislabeled pulses in an office environment).

Acknowledgments

The author would like to thank Prof. R. Reddy for many discussions and suggestions on all aspects of this work; and Dr. R. Cole for his careful study of the paper and valuable comments.

References

1. R.Cole, R.Stern, S.Brill, M.Phillips, A.Pilant, P. Specker. Feature-Based, Speaker-Independent, Isolated Letter Recognition. ICASSP 83 Proceedings, IEEE ASSP, , 1983, pp. ??.

2. L.R.Rabiner, M.J.Cheng, A.E.Rosenberg, C.A.McGonegal. "A Comparative Performance Study of Several Pitch Detection Algorithms." *IEEE Transactions on Acoustics, Speech, Signal Processing ASSP-24*, 5 (October 1976), 399-418.

3. W.H.Tucker, R.H.T.Bates. "A Pitch Estimation Algorithm for Speech and Music." *IEEE Transactions on Acoustics, Speech, Signal Processing ASSP-26*, 6 (December 1978), 597-604.