SOME COMPARISONS AMONG SEVERAL PITCH DETECTION ALGORITHMS

M. J. Cheng, L. R. Rabiner, A. E. Rosenberg and C. A. McGonegal
Bell Laboratories
Murray Hill, New Jersey 07974

ABSTRACT

A comparative performance study of seven pitch detection algorithms was conducted. A speech data base, consisting of eight utterances spoken by 3 males, 3 females, and 1 child was constructed. Telephone, close talking microphone, and wideband recordings were made of each of the utterances. For each of the utterances in the data base a "standard" pitch contour was semiautomatically measured using a highly sophisticated interactive pitch detection program. The "standard" pitch contour was then compared with the pitch contour that was obtained from each of the seven programmed pitch detectors. The algorithms used in this study were (1) a center clipping, infinite-peak clipping, modified autocorrelation method, (2) the cepstral method, (3) the SIFT method, (4) the parallel processing time domain method, (5) the data reduction method, (6) a spectral flattening LPC method, and (7) the AMDF method. A set of measurements was made on the pitch contours to quantify the various types of errors which occur in each of the above methods. Included among the error measurements were the average and standard deviation of the error in pitch period during voiced regions, the number of gross errors in the pitch period, and the number of voiced-unvoiced classification errors. For each of the error measurements, the individual pitch detectors could be rank ordered as a measure of their relative performance as a function of recording condition, and pitch range of the various speakers. Results are presented on rankings based on one category of errors.

I. INTRODUCTION

A pitch detector is an essential component in many speech analysis systems and almost all analysis-synthesis systems. A wide variety of algorithms for pitch detection have been proposed in the speech processing literature.$^{1-7}$ However, very little formal evaluation and comparison among the different types of pitch detectors has been attempted. There are a wide variety of reasons why such an evaluation has not been attempted. Among these are selection of a reasonable standard of comparison, collection of a comprehensive data base, choice of pitch detectors to be evaluated, and the difficulty in interpreting the results in a meaningful and unbiased way. This paper is a report on an attempt to provide a performance evaluation of 7 pitch detection algorithms.^{8} The evaluation was carried out on a NOVA 800 laboratory computer.

II. DATA BASE FOR EVALUATION

Figure 1 shows the pitch detectors which were chosen for evaluation. The choice of pitch detectors was based on practical considerations (i.e., availability of reasonably portable Fortran code) as well as the desire to choose a good cross section of the types of pitch detectors which have been described in the literature. Included in the study are two waveform pitch detectors (#4 and #5), two autocorrelation pitch detectors (#1 and #7), one spectral analysis pitch detector (#2) and two LPC hybrid pitch detectors (#3 and #6). The names in parentheses in Fig. 1 are the individuals (or group) who were the source of the Fortran code for the pitch detector. Due to space limitations we will not discuss the method of operation of these pitch detectors. The basic ideas in almost all the methods have been described in the literature.$^{1-7}$ It should be noted that all but one of the detectors [#4] incorporated some sort of voiced/unvoiced detector based on energy and/or zero crossing calculations.

Pitch Detectors

1. Modified autocorrelation analysis using clipping - AUTOC (Dubnowski [7])
2. Cepstrum method - CEP (Schafer [9])
3. Simplified Inverse Filtering Technique - SIFT (Markel [4])
4. Data reduction method - DARD (Miller [5])
5. Parallel processing method - PPROC (Rabiner [2])
6. Spectral equalization LPC method using Newton's transformation - LPC (Atal, unpublished)
7. Average Magnitude Difference Function - AMDF (NSA version [6])

Figure 1: The 7 pitch detectors used in this study.

Figure 2 shows the data base used in making the evaluation and comparisons among the 7 pitch detectors. A set of 7 speakers were chosen including a low pitch male (LM), 2 additional male speakers (M1, M2), 2 female speakers (F1, F2), a 4-year old child (C1), and a diplophonic speaker (D1). (For those unfamiliar with diplophonia, this is a condition where higher correlation exists between alternate glottal pulses than between successive glottal pulses. Thus accurate pitch detection on diplophonic speech is exceedingly difficult.)

The speech data base consisted of the 4 monosyllabic words hayed, hod, heed and hoed as well as the 4 sentences shown in Fig. 2. The recording conditions consisted of simultaneous recordings of these utterances over both a close talking microphone (M) and a standard telephone connection (T) over the local Murray Hill PBX. In addition, high quality microphone recordings (W) were made of the utterances.

Speakers

1. Low pitch male - LM
2. Male speaker #1 - M1
3. Male speaker #2 - M2
4. Female speaker #1 - F1
5. Female speaker #2 - F2
6. Child - C1
7. Diplophonic speaker - D1

Data Base

1-4. Hayed, Hod, Heed, Hoed
5. We were away a year ago
6. I know when my lawyer is due
7. Every salt breeze comes from the sea
8. I was stunned by the beauty of the view

Recording Conditions

1. Close talking microphone - M
2. Standard telephone transmission - T
3. High quality microphone - W

} simultaneous recordings

Figure 2: The data base used to test the 7 pitch detectors.

III. MEASUREMENT OF STANDARD PITCH CONTOUR

In order to be able to make quantitative measurements of each of the pitch detectors it was necessary to be able to define the standard pitch for each utterance. Since the entire error analysis was based on the standard pitch contour a sophisticated method for obtaining this pitch contour was required. Figure 3 shows a block diagram of a semi-automatic method which was used for obtaining the standard pitch contour of an utterance.[10] The speech signal, s(n), sampled at a 10 kHz rate, was processed to give three simultaneous displays, for each section of speech. For two of the displays the speech was first lowpass filtered by a sharp cutoff linear phase digital filter with cutoff frequency of 900 Hz.

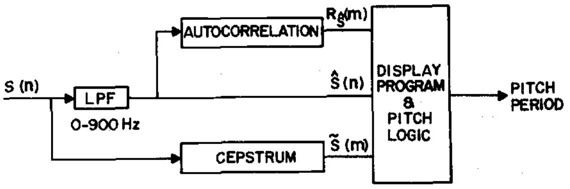

Figure 3: Block diagram of the system used to obtain the standard pitch contour.

The first display used in the semi-automatic method was the lowpass filtered speech signal. The second display was the autocorrelation function of the lowpass filtered waveform. The third display was the cepstrum of the wideband speech waveform. The choice of these three displays was dictated by the desire to obtain three reasonably independent estimates of the pitch period for each section of speech.

Figure 4 shows a typical frame of the semi-automatic pitch detector output for a section of voiced speech. For the waveform display the user marked an estimate of the pitch period directly on the waveform. For the autocorrelation and cepstrum displays the program found the maximum value over a specified range and indicated this to the user who could then change the marker to a different place if an error is indicated. The three estimates of pitch period are shown on each frame and the program chose the median of these as the estimate of pitch for this frame. The user could change this value if an error is indicated.

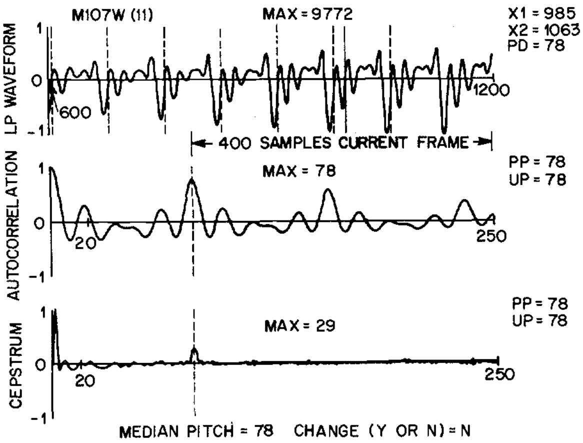

Figure 4: Typical frame of the semi-automatic analysis during a voiced segment.

The semi-automatic pitch detector was run on a frame-by-frame basis with each frame being 100 samples or 10 msec in duration. The analysis time for an experienced user was about 30 minutes to process 1 second of speech (i.e., 100 frames). For the data base used here a total of 60 hours of computer processing was used in this analysis.

An analysis of the results obtained from this semi-automatic pitch detector across several users on the same utterances showed the method to be highly reliable.

IV. EVALUATION RESULTS

The entire data base of speech utterances was used as input to each of the 7 pitch detectors as well as the semi-automatic method. Figure 5 shows a typical set of pitch contours for one utterance. The curve at the upper left is the result of the semi-automatic analysis. By comparing the standard contour with the pitch contours obtained from each pitch detector, it can be seen that several types of errors can occur in the pitch detection process. First a frame which is voiced can be classified as unvoiced, or vice versa. If a voiced frame is classified as voiced then two types of pitch period error can occur. One type of error is the gross pitch error in which the estimated pitch period is significantly different from the standard pitch period. The threshold for such errors is a 10 sample (or 1 msec) difference from the standard contour. These

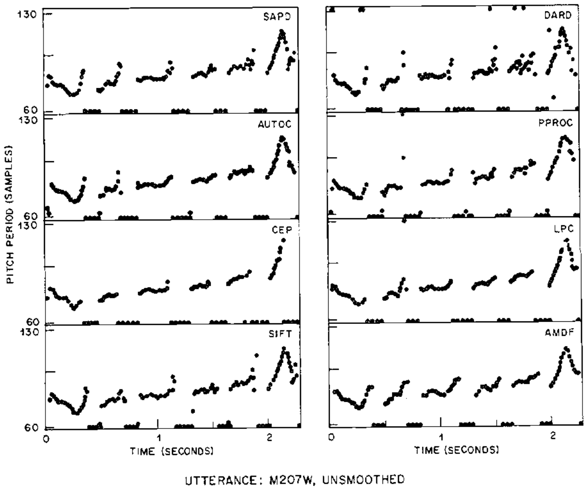

Figure 5: Typical set of pitch contours including the standard contour (SAPD), and the pitch contours from each of the 7 pitch detectors.

errors can be due to pitch period doubling effects, etc. This type of error is quantified by counting the number of occurrences for each utterance. The second type of error is the fine error in which there is a small discrepancy between the pitch period from a given detector and the standard pitch period. In this case the mean and standard deviation of the error is tabulated.

It is possible to detect and correct some or all the analysis errors using a nonlinear smoother.[11] Such a smoother was applied to the pitch contour output of each detector. (It should be noted that the AMDF algorithm incorporated a nonlinear smoother directly.) By way of example, Fig. 6 shows the resulting pitch contours obtained by nonlinearly smoothing the pitch contours of Fig. 5. The overall similarity among pitch contours is quite evident in this figure.

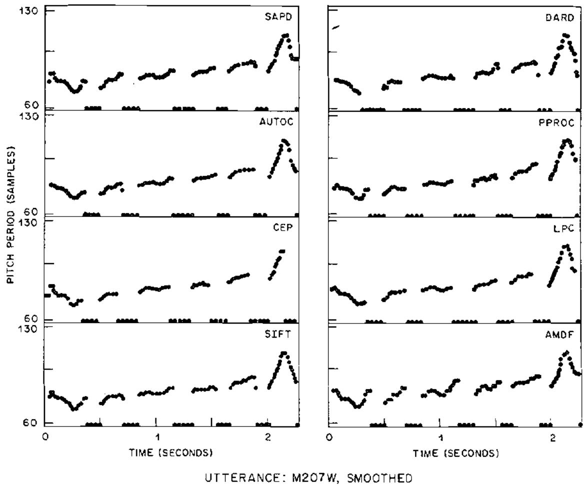

Figure 6: Nonlinearly smoothed versions of the pitch contours of Fig. 5.

Error analyses were carried out on the entire set of pitch contours obtained by processing the data base of Fig. 2 through each of the pitch detectors of Fig. 1. Since it is not possible to present all the results here, we will instead concentrate on presenting the results of one error category - i.e., the average number of gross pitch errors for each speaker and each pitch detector. Figure 7 shows a ranking of each of the pitch detectors, for each speaker (averaged across transmission conditions and utterances) based on the average number of gross errors.* The ranking scores vary from 1 (the best performance) to 5 (the worst performance). The rankings are absolute numbers based on a histogram of the total number of gross errors across all conditions. It can be seen from this figure that each pitch detector performed better for some speakers (i.e., range of pitch variation) than for others. The overall rankings of each pitch detector (i.e., the sum of the rankings over the speakers) are given at the bottom of Fig. 7. The ranking in the rightmost columns of Fig. 7 is a measure of the difficulty of detecting pitch for a given speaker. The lower the score, the easier it is to detect pitch.

	Pitch Detector							
	AUTOC	CEP	SIFT	DARD	PPROC	LPC	AMDF	Total
LM	4	1	2	3	4	3	4	21
M1	2	1	1	3	2	1	1	11
M2	3	1	2	4	3	2	3	18
F1	1	4	2	2	2	1	1	13
F2	1	1	2	3	2	1	1	11
C1	1	3	5	3	3	3	3	21
Overall Score	12	11	14	18	16	11	13	

Figure 7: Performance of the 7 pitch detectors based on the average number of gross pitch errors.

Finally, Fig. 8 shows the execution time for each pitch detector on the NOVA 800 per second of speech input. It can be seen that the pitch detectors with the best rankings generally require the most computation time.

DARD	5 sec
PPROC	7.5 sec
AMDF	50 sec
AUTOC	120 sec
SIFT	250 sec
LPC	300 sec
CEP	400 sec

Figure 8: Speed of execution of pitch detectors on the NOVA 800.

V. SUMMARY

In summary a fairly extensive performance evaluation of 7 pitch detection algorithms was made. The results showed a number of dimensions in which comparisons among the pitch detectors could be made. The overall conclusion has been that no single pitch detector was uniformly superior to the others across all speakers and recording conditions.

* Speaker D1 was eliminated because no pitch detector worked adequately on her data.

REFERENCES

1. A. M. Noll, "Cepstrum Pitch Determination," J. Acoust. Soc. Am., Vol. 41, No. 2, pp. 293-309, February 1967.

2. B. Gold and L. R. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Am., Vol. 46, No. 2, pp. 442-448, August 1969.

3. M. M. Sondhi, "New Methods of Pitch Extraction," IEEE Trans. on Audio and Elect., Vol. AU-16, No. 2, pp. 262-266, June 1968.

4. J. D. Markel, "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. on Audio and Elec., Vol. AU-20, No. 5, pp. 367-377, December 1972.

5. N. J. Miller, "Pitch Detection by Data Reduction," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, No. 1, pp. 72-79, February 1975.

6. M. J. Ross et al., "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-22, No. 5, pp. 353-362, October 1974.

7. J. J. Dubnowski, R. W. Schafer, and L. R. Rabiner, "Real-Time Digital Hardware Pitch Detector," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-24, No. 2, April 1976.

8. M. J. Cheng, "A Comparative Performance Study of Several Pitch Detection Algorithms," MIT M.S. Thesis, June 1975.

9. R. W. Schafer and L. R. Rabiner, "System for Automatic Analysis of Voiced Speech," J. Acoust. Soc. Am., Vol. 4, No. 2, pp. 634-648, February 1970.

10. C. A. McGonegal, L. R. Rabiner and A. E. Rosenberg, "A Semiautomatic Pitch Detector," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, No. 6, pp. 570-574, December 1975.

11. L. R. Rabiner, M. R. Sambur and C. E. Schmidt, "Applications of a Nonlinear Smoothing Algorithm to Speech Processing," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. ASSP-23, No. 6, pp. 552-557, December 1975.