FAST AND ACCURATE PITCH DETECTION USING PATTERN RECOGNITION AND ADAPTIVE TIME-DOMAIN ANALYSIS

Dimitrios P. Prezas, Joe Picone, and David L. Thomson

AT&T Bell Laboratories
Naperville, Illinois 60566

ABSTRACT

A method of determining pitch and voicing information from speech signals is presented. The algorithm, which employs time-domain analysis and pattern recognition techniques, is fast and yields accurate pitch and voicing estimates. A search routine is employed to find periodicity in each of four signals derived from the speech waveform and the results are combined to form a pitch estimate. The voicing decision uses linear discriminant analysis, and declares speech frames voiced or unvoiced based on a weighted sum of 13 parameters. Performance comparisons with other pitch detectors are reported.

INTRODUCTION

The pitch detector described here uses a parallel processing algorithm to determine voicing and pitch estimates from speech [4]. Speech is first sampled at 8 khz and divided into 20 ms frames. The tenth-order LPC residual (also known as the LPC prediction error signal) is then found using the Burg method [2]. The positive residual is defined as the LPC residual with all negative samples set to zero. The negative residual is the LPC residual with the positive samples set to zero and the sign of the negative samples reversed. The positive and negative speech waveforms are defined similarly. A time-domain analysis method is employed to form a preliminary pitch estimate and voicing decision for each of these four waveforms. Pattern recognition techniques are used to make the final voicing decision and, if the frame is declared voiced, the final pitch estimate is derived from the pitch estimates found during the time-domain analyses. By using four waveforms, the final pitch estimate is more reliable than it would be if only one waveform were used. A block diagram of the system is shown in Figure 1.

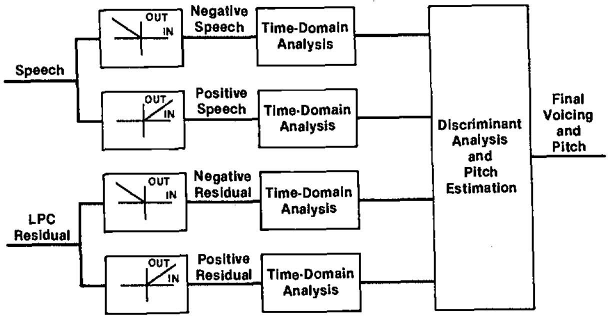

Figure 1. Pitch detection process.

TIME-DOMAIN ANALYSIS

Typically, periodicity appearing in the speech waveform also occurs in the residual. However, two cases demonstrate the need for both types of waveforms. Figure 2 shows a voiced speech waveform for which the pitch pulse locations are not easily observable. Since the formant structure is removed in the residual, the pitch pulses are readily located. Figure 3 illustrates the opposite situation where the residual appears noisy, yet the speech signal is clearly voiced. Therefore, the likelihood of finding the correct pitch is increased by using both residual and speech signals.

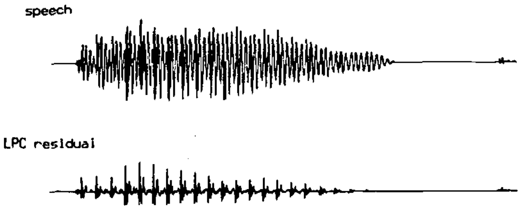

Figure 2. Example where the pitch period is easily found in the residual but not in the speech.

Pitch and voicing are determined by examining pitch pulses in the positive and negative residual and the

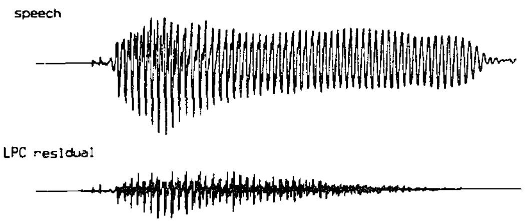

Figure 3. Example where the pitch period is easily found in the speech but not in the residual.

positive and negative speech waveforms. Locating pitch pulses begins by assuming that the maximum peak in the waveform being analyzed is a pitch pulse. It has been verified experimentally that if the frame is voiced this is nearly always the case. Next, the second largest peak which is separated from the maximum peak by at least a minimum pitch period (typically 2.0 ms) is located. This process continues until all peaks which are separated by at least 2.0 ms are found. Some of these peaks may be recognized as invalid pitch pulses based on their amplitudes. Suppose peaks A and B have no other peaks between them and that A is closer to the maximum peak in the waveform. Empirical data have shown that if A is a valid pitch pulse its amplitude is rarely less than 75% as great as the amplitude of a point directly above on a line drawn from the maximum peak to peak B [9]. Also, the amplitudes of valid pitch pulses are nearly always at least 25% as great as that of the maximum pulse. Pulses not meeting these criteria are deleted.

Once the pulses are located, a search determines if a regularly spaced subset exists. The distance from the largest pulse to every other pulse in the frame is tested as a possible pitch period. The test is successful if a pulse is found at every integral multiple of the pitch period from the largest pulse. Experimental evidence indicates that the true pitch period rarely changes by more than 1.25 ms between adjacent frames regardless of the pitch. Consequently, only pitch periods are tested which are within 1.25 ms of that found in the previous frame. If only one peak exists and the previous frame is voiced, the test is considered successful if the pulse is one pitch period away from the last pulse in the previous frame.

Since the human voice rarely produces exactly periodic pitch pulses, some tolerance must be allowed. Pulses are therefore considered to be evenly spaced if the distance between them varies no more than some distance Δ. Since the regularity of pitch pulses varies with pitch, Δ must adapt to the pitch of the speaker. The pitch period d (measured in milliseconds) of the most recent voiced frame is used as an estimate of the true pitch period. It has been determined that over 95% of all pitch pulses are within

$$\Delta = 0.44 + 0.033d \qquad (1)$$

milliseconds of their theoretical position [9]. If periodicity is found, the waveform being analyzed is declared voiced with a pitch period equal to the average distance between pulses. This process, depicted in Figure 4, is repeated for all four waveforms.

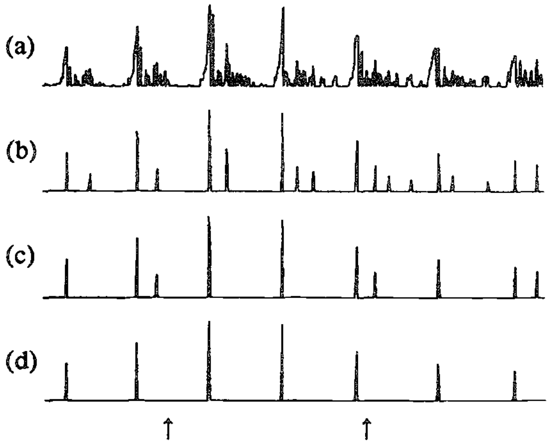

Figure 4. Time-domain analysis. Arrows indicate frame boundaries. Plots shown are (a) the positive residual, (b) set of pulses located at peaks in the residual, (c) after invalid pulses are deleted, and (d) the subset of regularly spaced pulses.

FINAL VOICING AND PITCH

Results from the time-domain analysis are used in combination with other parameters to form a final voicing decision and pitch estimate. Rather than rely entirely on a single function of the data, the voicing decision is made from a weighted sum of p parameters. Similar pattern recognition techniques have been demonstrated to yield reliable voicing decisions [1,8]. The weighted sum, denoted by y, is referred to as a discriminant variable. The set of parameters for a given frame is represented by a $p \times 1$ vector **x**. If the weight for each parameter is an element in the vector **a** then the discriminant variable is

$$y = \mathbf{a}^t \mathbf{x} \, . \qquad (2)$$

If y is greater than a given threshold, the frame is declared unvoiced, otherwise it is declared voiced. The optimum weight vector is determined using a training set of speech where the correct voicing is known. The training set is divided into two groups consisting of n_1 unvoiced frames and n_2 voiced frames. The weight vector is chosen to maximize the ratio of the between-groups sum of squares to the within-groups sum of squares, a criterion suggested by Fisher [3]. This measure will be defined. Let $\mathbf{X}_1$ be an $n_1 \times p$ matrix where each row contains the parameters for an unvoiced frame. Likewise let $\mathbf{X}_2$ be an $n_2 \times p$ matrix of parameters for all voiced frames. The parameter means for voiced and unvoiced frames are denoted by vectors $\bar{\mathbf{x}}_1$ and $\bar{\mathbf{x}}_2$, respectively. The between-groups sum of squares is defined as

$$\mathbf{a}^t\mathbf{B}\mathbf{a} = \mathbf{a}^t\left[\frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)^t\right]\mathbf{a}\,, \qquad (3)$$

where $\mathbf{B}$ is the between-groups sum of squares and products matrix. The within-groups sum of squares is expressed as

$$\mathbf{a}^t\mathbf{W}\mathbf{a} = \mathbf{a}^t(\mathbf{X}_1^t\mathbf{X}_1 - n_1\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^t + \mathbf{X}_2^t\mathbf{X}_2 - n_2\bar{\mathbf{x}}_2\bar{\mathbf{x}}_2^t)\mathbf{a}\,, \qquad (4)$$

where $\mathbf{W}$ is the within-groups sum of squares and products matrix. A weight vector is determined which maximizes $\mathbf{a}^t\mathbf{B}\mathbf{a}/\mathbf{a}^t\mathbf{W}\mathbf{a}$. It can be shown [5] that this ratio is maximum when

$$\mathbf{a} = \mathbf{W}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\,. \qquad (5)$$

Once $\mathbf{a}$ has been determined, a voiced/unvoiced threshold is chosen. Increasing the threshold decreases the probability of VU errors (voiced frames declared unvoiced) but increases the probability of UV errors (unvoiced frames declared voiced). For the training set of 3746 voiced and 1654 unvoiced speech frames spoken by three females and three males, maximum voice quality was attained when the threshold was chosen to minimize the total number of VU and UV errors.

Out of 49 parameters tested for their value as voicing classifiers 13 were finally chosen. The selection was based partly on the change in error rate when the parameter was used. Another measure of effectiveness used was the variance of each parameter multiplied by its corresponding weight. If this product was large compared to that of the other parameters, then the parameter was considered likely to be a useful classifier.

The parameters selected to be used as classifiers are:

1) The log of the speech power. The speech power is normalized by dividing it by the average power in previous voiced frames. The average power is estimated using an exponentially decaying average with a time constant of approximately two seconds and is updated only when the present frame is declared voiced.

2) The log of the LPC gain, defined as the speech power divided by the residual power.

3-6) The first four reflection coefficients found during LPC analysis.

7) The number of waveforms for the present frame which are declared voiced by the time-domain analysis.

8) The number of waveforms for the previous frame which are declared voiced by the time-domain analysis.

9) The number of waveforms for the next frame which are declared voiced by the time-domain analysis.

10) The difference between the closest two out of five pitch period estimates. The five estimates are pitch periods from the four time-domain analyses and the pitch period from the most recent voiced frame. The usefulness of this parameter is increased by allowing its maximum value to be 0.625 ms. This maximum value is also used if no periodicity is found for the present frame.

11) The total number of regularly spaced pulses found by the four time-domain analyses in the present frame. All regularly spaced pulses are counted even if the corresponding waveform is declared unvoiced.

12) Parameter 11 multiplied by the pitch period from the most recent voiced frame.

13) The number of waveforms for the present frame containing at least three regularly spaced pulses.

The final pitch period is estimated by finding the median of thirteen estimates. These estimates are those found during the four analyses for the present frame and for both adjacent frames, and the pitch period declared for the most recent voiced frame. Thus, if none of the analyses find a pitch yet the frame is declared voiced, the most recent pitch is used. This median pitch estimate is exceptionally robust and insensitive to spurious errors.

PERFORMANCE

The performance of this pitch detector was compared to several others using a 200 second, 58 speaker database of speech with reference pitch contours [7]. A perceptually-weighted, objective measure was used to compare pitch and voicing from the pitch detector to the reference pitch contours [7]. This measure has been shown to have a high correlation with subjective speech quality on studio quality speech data [6,7,10].

Table 1 gives the performance of each time-analysis compared to the final output. The percent errors shown are based on comparisons between the pitch detector output and the actual voicing determined every 10 ms (twice per frame). A gross pitch (GP) error occurs when a frame is correctly declared voiced but with an inaccurate pitch estimate. The value of using a combination of these estimates is shown from the relatively poor individual performance. In Table 2, the performance of the time-domain pitch detector is compared to that of several well known algorithms. It should be noted that for perceptually-weighted scores in the neighborhood of one and below, it becomes difficult to detect audible differences between pitch detectors. Also, the database consists of short, carefully spoken sentences. The performance of the pitch detectors may be different during actual conversations.

Table 1. Perceptually-weighted performance of individual pitch detectors and final output. Percent errors with respect to all frames are shown in parentheses.

Signal	GP	VU	UV	Total
- speech	0.07	4.34	0.04	4.46
		(16.2)	(1.3)	(17.5)
+ speech	0.03	2.58	0.06	2.66
		(11.5)	(1.3)	(12.8)
- residual	0.10	13.80	0.01	13.90
		(42.5)	(0.2)	(42.7)
+ residual	0.05	9.14	0.01	9.19
		(31.6)	(0.2)	(31.8)
Composite	0.11	0.09	0.10	0.30
		(1.2)	(3.2)	(4.3)

Table 2. Perceptually-weighted performance of several pitch detectors evaluated in [6] compared to the time-domain method.

Pitch Detector	GP	VU	UV	Total
Gold Rabiner	0.25	4.08	0.56	4.90
Cepstral	0.39	1.62	1.85	3.86
Integrated Correlation	0.23	0.38	0.65	1.29
Time-domain	**0.11**	**0.09**	**0.10**	**0.30**

CONCLUSION

Discriminant analysis and parallel time-domain analysis used together yield accurate pitch and voicing information at a modest computational load. This pitch detection method has been implemented as part of a real-time, 2.4 kbits/sec LPC vocoder on a single Texas Instruments TMS32020 digital signal processor. In this system, the time-domain analysis and discriminant analysis runs in 31% of real time.

REFERENCES

[1] B. S. Atal, and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-24, no. 3, pp. 201-212, June 1976.

[2] J. Burg, "A new analysis technique for time series data," *Proc. NATO Advanced Study Institute on Signal Proc.*, Enschede Netherlands, 1968.

[3] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pt. 2, pp. 179-188, 1936.

[4] B. Gold and L. R. Rabiner, "Parallel processing techniques for estimating pitch periods of speech in the time domain," *The Journal of the Acoustical Society of America*, vol. 46, no. 2, pp. 442-448, 1969.

[5] K. V. Mardia, *Multivariate Analysis,* Academic Press, London, 1979.

[6] B. G. Secrest and G. R. Doddington, "An integrated pitch tracking algorithm for speech systems," *Proceedings IEEE Conference on Acoustics, Speech, and Signal Processing,* pp. 1352-1355, April 1983.

[7] B. G. Secrest and G. R. Doddington, "Postprocessing techniques for voice pitch trackers," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* pp. 172-175, April 1982.

[8] L. J. Siegel, "A procedure for using pattern classification techniques to obtain a voiced/unvoiced classifier," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. ASSP-27, no. 1, pp. 83-89, February 1979.

[9] R. Sukkar, *A parallel processing pitch detector for LPC,* M.S. Thesis, Illinois Institute Of Technology, Chicago, 1985.

[10] V. R. Viswanathan and W. H. Russell, "New objective measures for the evaluation of pitch extractors," *Proceedings IEEE Conference on Acoustics, Speech, and Signal Processing,* pp. 411-414, March 1985.