A PROBABILISTIC APPROACH TO AMDF PITCH DETECTION

Goangshiuan S. Ying, Leah H. Jamieson, and Carl D. Michell

School of Electrical and Computer Engineering, Purdue University
West Lafayette, IN 47907-1285
Email: sgsying@ecn.purdue.edu lhj@ecn.purdue.edu cdm@research.att.com
URL: http://purcell.ecn.purdue.edu/~speechg

10.21437/ICSLP.1996-310

ABSTRACT

We present a probabilistic error correction technique to be used with an average magnitude difference function (AMDF) based pitch detector. This error correction routine provides a very simple method to correct errors in pitch period estimation. Used in conjunction with the computationally efficient AMDF, the result is a fast and accurate pitch detector. In performance tests on the CSTR (*Center for Speech Technology Research*) database, probabilistic error correction reduced the gross error rate from 6.07% to 3.29%.

1. INTRODUCTION

Fundamental frequency (F_0) as an acoustic correlate is strongly related to prosodic information of stress and intonation. A reliable pitch detection algorithm (PDA) is a very important component for measuring prosodic features. Many PDAs have been developed, with different PDAs focusing on different properties and features of the speech signal. However, many factors in the acoustic signal can cause the failure of PDAs. For example, in the source spectrum of the vocal fold vibrations during voicing, F_0 has stronger spectral energy than its harmonics during voicing since F_0 is the rate of the vocal fold vibrations. However, the vocal tract resonances, i.e., formants, act like a series of band pass filters, and may reduce the spectral energy of F_0 and enhance the energy of harmonics. The harmonics therefore dominate the spectral energy in the acoustic signal. This phenomenon causes PDAs to select the higher harmonics to be the estimate of F_0, resulting in estimation errors. For most of the PDAs, global error correction routines are needed to locate and correct errors in the local decisions made by PDAs. We present a new approach to global error correction that is based on probabilistic weighting of the candidate pitch estimates within each speech frame. We apply this correction method to an AMDF-based PDA. The experimental results show that probabilistic error correction is indeed successful in reducing the errors that occur during pitch detection.

2. AMDF-BASED PITCH DETECTION ALGORITHM

Many PDAs have been developed [1, 2, 3, 5, 6, 7, 8, 12, 13, 14]. The autocorrelation function and the average magnitude difference function (*AMDF*) [8, 12] are the two most frequently used time domain PDAs. The AMDF pitch detection algorithm is chosen in our study is because it has relatively low computational cost and is easy to implement.

Basic Algorithm: For each frame k, the short-term difference function AMDF is defined as follows:

$$AMDF_n(j) = \frac{1}{N}\sum_{i=1}^{N} |x_n(i) - x_n(i+j)|, 1 \le j \le MAXLAG$$

where $MAXLAG$ is the maximum number of AMDF values generated in each frame. The difference function is expected to have a strong local minimum if the lag j is equal to or very close to the fundamental period. For each frame, the lag for which the AMDF is a global minimum is a strong candidate for the pitch period of that frame.

One major disadvantage of the short-term AMDF approach PDA is that the magnitude of the principle minimum in each frame is highly influenced by the intensity variation and the background noise of the speech signal. To eliminate the effects of intensity variation and background noise, a standard pre-processing procedure can be applied to the signal before calculating the difference function [4]:

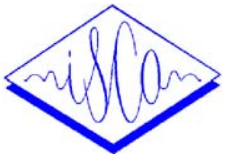
1. Lowpass filter the speech samples with 3dB attenuation at 600 Hz and 40 dB attenuation at 900 Hz.
2. Segment the filtered signal into frames.
3. Apply *center clipping* to each frame [8].

For each frame, the difference function is smoothed by a 5 point Hanning window.

Pitch Period Estimation: In most AMDF-based PDAs, the lag for which the magnitude of the difference function is a global minimum is chosen as the pitch period estimate for that frame. In our modified AMDF PDA, in addition to the lag with global minimum, a set of candidates for the pitch period in a frame is selected and will be used to determine the final pitch period estimate in the later process. Three terms are used to define the candidates for the pitch period:

- *local_min*: The locations of the difference function at which the magnitudes are the local minima in a frame.
- *marker*: local_min which satisfy the AMDF pattern constraints. Markers are the candidates for the pitch period. There can be several markers in a frame.
- *period marker*: The marker whose magnitude is the global minimum among all markers in a frame.

From observing the shape of the AMDF for the different sounds of speech, a distinct pattern difference can be found



4th International Conference on Spoken Language Processing (ICSLP 96)
Philadelphia, PA, USA
October 3-6, 1996

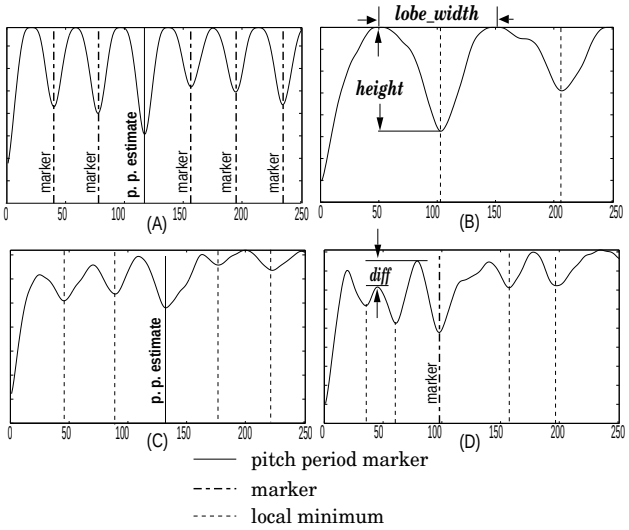


Figure 1: AMDF and markers for a frame of (A) a vowel, (B) a voiceless consonant, (C&D) the end of a vowel segment before a consonant. In (C) and (D), no pitch period estimate is made. Horizontal axes are in lags.

between the AMDF of sonorants and obstruents. For vowel segments, the local maxima of the difference function typically have similar values and the duration or number of lags between two adjacent local maxima is shorter (Figure 1 A). The difference function for an obstruent does not have these characteristics. It is very easy to use pattern matching visually to determine which AMDF pattern represents a voiced segment and which represents a voiceless consonant. A set of constraints is designed here to utilize the inherent patterns of the difference function. To be a qualified marker, the local_min needs to satisfy this set of constraints. With the constraints, we can reduce the number of markers and also decrease the chance of choosing a pitch period in a segment where no pitch period is present.

Five features are used to set up the constraints. For the i^{th} local minimum, $local\_min_i$, we define:

- $global\_max$ = largest value in the difference function in each frame.
- $peak\_ratio = \frac{\text{local maximum}}{\text{global\_max}}$
- $height_i = \min(left\_height_i, right\_height_i)$.
- $diff_i = |left\_height_i - right\_height_i|$.
- $lobe\_width$ = distance between right and left local maxima.

To be a marker, the i^{th} candidate needs to satisfy:

1. $peak\_ratio \geq \mathbf{0.8}$
2. $height_i \geq \mathbf{0.3} \times global\_max$
3. $diff_i \leq \mathbf{0.1} \times global\_max$
4. $lobe\_width_i \leq \mathbf{100}$ lags

The values of the parameters in the constraint rules were decided by running the evaluation experiment on the CSTR database[1]. The values are selected to achieve the best performance in reducing both the error in the voiced/voiceless decision and the gross error rate.

Once markers have been selected, the initial estimate of the pitch period for each frame is chosen to be the marker with the AMDF global minimum. A global distribution of the initial estimates of period markers is then established.

3. GLOBAL ERROR CORRECTION

In most PDAs, the pitch extractor uses only local information to determine the estimate of the period period for each frame of speech samples. The estimates can be incorrect. Errors occur when PDAs fail to select the correct pitch period candidate. For those PDAs that obtain the pitch period estimate by selecting one of the candidates for the pitch period in each frame, incorrect estimates occur when the PDAs fail to select the correct candidate. A global error correction routine is required for the pitch detection system to locate the incorrect estimates and correct the errors. Various techniques for global error correction have been proposed [4, 11, 14], ranging from Reddy's logical decision-based approach[11], to Secrest and Doddington's dynamic programming and pattern matching approach [14]. Linear and non-linear smoothing procedures are often used as global correction routines to enforce continuity of the pitch contour[4].

3.1. Probabilistic Approach

In our algorithm, a probabilistic approximation approach is used. Global error correction is performed on a sentence-by-sentence basis. Using the set of local initial estimates, the distribution of initial pitch period estimates for an entire sentence is obtained. This distribution is then approximated with a normal distribution. The *heights* for all the markers in each frame are weighted with the value of the normal distribution. Finally, the new pitch period estimate is selected based on the weighted *heights* of the markers. The marker with the largest *height* is the new period marker in a frame. Markers located around the mean of the distribution will have a stronger *height* value than the ones away from the center of the distribution. The procedure, summarized as follows, is performed twice.

1. Approximate the initial pitch period estimates with a normal distribution (Figures 2 A and B).
2. Weight the heights of the markers in each frame with the value of the distribution.
3. Select the new pitch period marker based on the weighted markers in each frame.

Figure 2 (A and B) shows that the normal distribution is a good approximation of the initial pitch period estimate for both speakers. The normal distribution makes a better pitch period estimate approximation for female speech than for male speech because, for female speech, the distribution

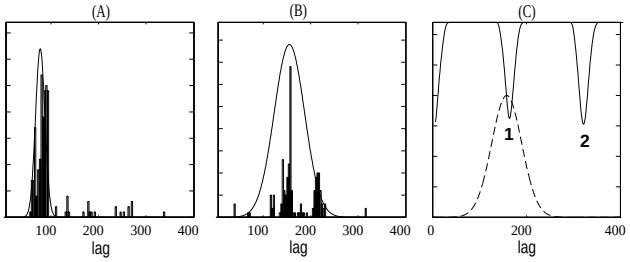


Figure 2: Typical histogram and the distribution approximation of the initial pitch period estimate for (A) a female utterance, (B) a male utterance. (C) AMDF for a voiced frame from a male speaker. The dashed line shows the normal distribution approximation.

of the pitch period estimates has smaller variance. Figure 2 (C) shows how the global error correction can help the PDA avoid selecting the period marker incorrectly. In this figure, the solid line is the AMDF function of a voiced frame of a male speaker and the dashed line is the normal distribution approximation of the pitch period of the whole utterance. Marker 2, $lag \approx 330$, would be assigned as the pitch marker initially since it is the global minimum in that frame. However, according to the distribution, marker 1 makes a better candidate for the pitch marker and marker 2 actually represents the sub-harmonic of the fundamental frequency. After weighting the height of each marker with the normal distribution, the weighted height of marker 2 is much smaller than the height of marker 1. Therefore marker 1 is assigned as the pitch marker.

4. PERFORMANCE EVALUATION

4.1. CSTR Database

Evaluation is done on a database [1] provided by the Center for Speech Technology Research at University of Edinburgh, Scotland, UK [1]. This database includes 50 sentences from two speakers, one male and one female. A laryngeal frequency (F_X) contour provides reference F_0 values. The sampling rate for the speech signal is 20kHz using a 16-bit A/D converter. To be consistent with the experiments done at *CSTR*, a frame length of 38.4 msec is used and a new frame is processed every 6.4 msec. At these rates, approximately 16,400 pitch period estimates are made. As described in Bagshaw's paper [1], the speech was pre-processed with a low pass filter with -3dB at 600Hz and -85dB at 700Hz in his experiment, which is different than the filter used in this experiment.

4.2. Performance Comparison of PDAs

The evaluation is done by calculating the gross error rate for each of the PDAs in the experiment. The F_0 value from the reference laryngeal frequency contour is represented by

[1] The authors would like to thank Dr. Bagshaw for providing the database and for allowing us to use his results in this paper.

PDAs	Male		Female		Overall gross error (%)
	Gross errors		Gross errors		
	high (%)	low (%)	High (%)	Low (%)	
CPD	4.09	0.64	0.61	3.97	4.65
FBPT	1.27	0.64	0.60	3.55	3.10
HPS	5.34	28.2	0.46	1.61	18.27
IPTA	1.40	0.83	0.53	3.12	2.98
PP	0.22	1.74	0.26	3.20	2.76
SRPD	0.62	2.01	0.39	5.56	4.41
eSRPD	0.90	0.56	0.43	0.23	1.03
mAMDF	**3.35**	**5.19**	**1.22**	**14.8**	**6.07**
mAMDFp	**1.94**	**2.33**	**0.63**	**2.93**	**3.29**

Table 1: Comparison of 8 different PDAs.

$F_{X[REF]}$, and that from the PDA estimate is represented by $F_{0[PDA]}$. Based on Bagshaw's definition, a gross error is counted when $F_{0[PDA]}$ is more than 20% higher or lower than $F_{X[REF]}$ when they both represent voiced speech [1]. Table 1 shows the gross error results of seven PDAs in Bagshaw's experiment and two in our study. The PDAs are

- Cepstrum pitch determination (CPD)[6]
- Feature-based pitch tracker (FBPT)[7]
- Harmonic product spectrum (HPS)[13]
- Integrated pitch tracking algorithm (IPTA)[14]
- Parallel processing method (PP)[3]
- Super resolution pitch determinator (SRPD)[5]
- Enhanced version of SRPD (eSRPD)[1]
- Modified AMDF-based PDA without error correction (mAMDF)
- Modified AMDF-based PDA with probabilistic error correction (mAMDFp)

In Table 1, the results for the first seven PDAs were provided by Dr. Bagshaw. The last two lines of the table compare the performance of the raw AMDF PDA with the results using probabilistic correction. The reduction in error rate from 6.09% to 3.29% is very promising.

Figures 3 shows a comparison of $F_{X[REF]}$ and $F_{0[PDA]}$ contours extracted by mAMDFp. No separate procedure of voiced/voiceless/silence classification was applied to the PDA. Voiced and voiceless classification of a frame is done by detecting the existence of $F_{0[PDA]}$. The frame is voiced if $F_{0[PDA]}$ is not zero; voiceless/silence otherwise. The figures show that most of the voiced/voiceless classification errors happen at the beginning and the end of each voiced segment. Since the PDA is a short-term time-domain approach PDA and the frame size is six times the frame rate, it is very difficult to have pitch synchronous frames so that the PDA will know where exactly voicing starts. Therefore, the voiced/voiceless errors at both ends of the voiced segments can be ignored if the number of consecutive errors is small, e.g., ≤ 3, which corresponds to 50% of a frame.

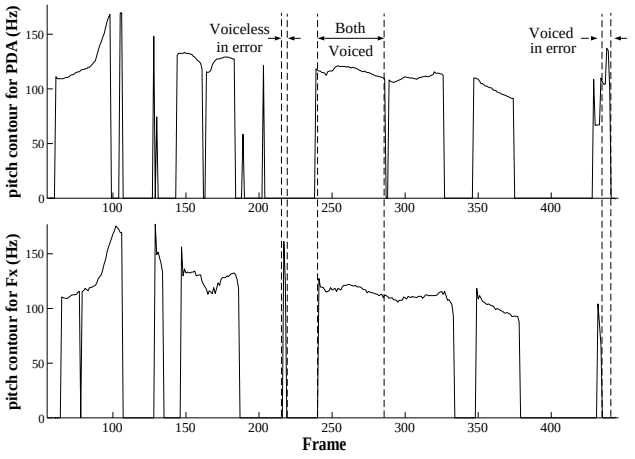


Figure 3: Comparison of asynchronous $F_{X[REF]}$ and $F_{0[PDA]}$ contours on male speech

5. CONCLUSION

The probabilistic error correction is quite different from most other techniques that have been used to correct pitch estimate errors [11, 14]. In contrast to both linear and nonlinear filtering techniques [10], the preservation and use of markers ensures that the final pitch estimate for a frame is in fact drawn from one of the AMDF candidates for the frame's pitch. No logic-based decisions are used, as required in [11]. The probabilistic error correction involves significantly less computation than the dynamic programming and dynamic pattern matching technique in [14]. As shown in Figure 4, the technique successfully removes anomalous discontinuities in the overall pitch contour.

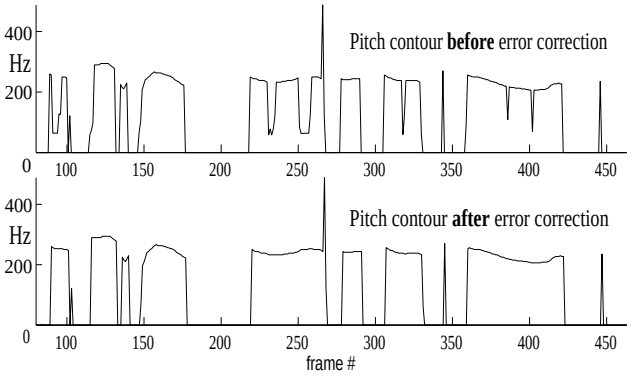


Figure 4: Pitch contour for a sentence spoken by a female speaker before and after probabilistic error correction.

The gross error count of this this modified AMDF PDA with probabilistic error correction is still higher than other PDAs' performance. However, initial experiments at extending this technique indicate that the gross error count can be reduced further by using a finer approximation of the distribution. New experiments are being conducted that combine two distributions, one for the whole utterance and the other for a local region. By including both distributions, we expect to better capture the dynamics of regional pitch changes, while preserving the desirable properties of the probabilistic method.

6. REFERENCES

1. P.C. Bagshaw. *Automatic prosody analysis.* PhD thesis, University of Edinburgh, Scotland, UK, 1994.
2. J.J. Dubnowski, H.L. Shaffer, and L.R. Rabiner. Real-time digital hardware pitch detector. *IEEE Trans. ASSP*, ASSP-24:2–8, 1976.
3. B. Gold and L.R. Rabiner. Parallel processing technique for estimating pitch period of speech in the time domain. *JASA*, 46(2, part 2):442–448, 1969.
4. W.H. Hess. *Pitch determination of speech signals: Algorithms and Devices.* Springer-Verlag, Heidelberg, Germany, 1983.
5. Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. ASSP*, ASSP-39:40–48, Jan. 1991.
6. A.M. Noll. Cepstrum pitch determination. *JASA*, 41(2):293–309, 1967.
7. M.S. Phillips. A feature-based time domain pitch tractor. *JASA*, 77-S9-S10(A), 1985.
8. L.R. Rabiner. On the use of autocorrelation analysis for pitch detection. *IEEE Trans. ASSP*, ASSP-25:24–33, 1977.
9. L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal. A comparative performance study of several pitch detection algorithms. *IEEE Trans. ASSP*, ASSP-24(5):399–418, July 1976.
10. L.R. Rabiner and B. Gold. *Theory and application of digital signal processing.* Prentice Hall, Englewood Cliffs, NJ, 1975.
11. D.R. Reddy. Pitch period determination of speech sound. *CACM*, 10:343–348, 1967.
12. M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley. Average magnitude difference function pitch extractor. *IEEE Trans. ASSP*, ASSP-22:353–362, Oct. 1974.
13. M.R. Schroeder. Period histogram and product spectrum: New methods for fundamental frequency measurement. *JASA*, 43(4):829–834, 1968.
14. B.G. Secrest and G.R. Doddington. Postprocessing techniques for voice pitch trackers. In *Proc. ICASSP*, pages 172–175, 1982.