Time Delay Estimation and Adaptive Frame Length Iterations for Noise Robust Pitch Extraction

Ramon E. Prieto, Sora Kim.
Electrical Engineering Department
Stanford University
621 Escondido Road, Apt 449, Stanford
CA 94305, USA

Abstract: - We describe a pitch tracking method based on a time delay estimation technique. Given two frames of voiced speech and it's time delay, we will run a linear regression on the unwrapped phase of the quotient of the spectrum of both frames. A weighted linear regression will allow us to avoid the effect of phases corrupted by spectral leakage and noise. Iterations adapting the frame length will allow us to have a better time resolution, avoiding inaccuracies that could come from pitch doubling and jitter.

Key-Words: - Pitch detection, noise robustness, time delay estimation.

1 Introduction

Many methods for tracking the pitch period of a voiced segment of speech have been proposed ([1]). Some of the most widely accepted ones are the Cepstrum method ([2]) and and the SIFT method([3]). In both of them the analysis frame used is several times the pitch period itself, causing jitter and pitch doubling to be a problem. Other methods like the autocorrelation function method and (ACF, [4]) and the average magnitude difference function method (AMDF, [5]) keep the formant structure of the signal, making the pitch estimation hard in the presence of high energy, high frequency harmonics.

In [6], we presented a method based on a time delay estimation technique and we called it linear regression of the phase. We assumed two different frames of a sampled voiced speech signal:

$$x_1(n) = s(n) \qquad 0 \le n \le N-1 \qquad (1)$$

$$x_2(n) = s\left(n + \frac{T+\delta}{T_s}\right) \qquad 0 \le n \le N-1 \qquad (2)$$

and we defined our problem as to find the period T given that we know the frames $x_1(n)$, $x_2(n)$, and the time delay between them: $T+\delta$. If we assume that the time length of both frames is T (i.e. both frames are pitch synchronous, $\frac{T}{T_s} = N$) we can use the $Circular\ Shift\ of\ a\ Sequence$ property of the DFT as stated in [7]:

$$\frac{X_2(k)}{X_1(k)} = e^{i2\pi k \frac{\delta}{T_s N}} \qquad 0 \le k \le N-1 \qquad (3)$$

where X_1 and X_2 are the DFT coefficients of x_1 and x_2 respectively. The unwrapped phase of the term in formula 3 is bilinear in k and δ. Then, given x_1 and x_2, we can calculate the unwrapped phase of formula 3 and make a linear regression vs k. The result gives us δ which gives us the value of T. Although this is an unrealistic case, it is shown in [6] that the use of non-pitch synchronous frames will give us an accurate answer as far as we use the right unwrapping method and we use the right weighting scheme in the linear regression.

Work has been done on the field of phase Unwrapping. Phase unwrapping has been used to to calculate the Complex Cepstrum ([8],[9]). Other methods have been proposed to unwrap the phase of one dimensional digital signals, out of which, the most widely used is [10]. Other methods are [11] and [12]. In the first one, a better indicator to detect zeros very close to the Z unit circle was developed. However, our case doesn't deal with zeros in the interval defined as critical in that work. In the second, the use of Sturm Sequences was proposed, but more significant digits than available in double precision floating-point representations are needed. We have compared the following phase unwraping methods:

1.1 Basic Unwrapping (BU)

This method adds 2π or -2π to the phase of all the frequency bins greater or equal than q if the difference between the phase of the frequency bins q and q-1 is lower than $-\pi$ or greater than π respectively. The method starts from frequency bin $q = 1$ until it reaches frequency bin $q = N - 1$.

1.2 Slope Forced Unwrapping (SFU)

At frequency bin q, we calculate the slope of the line that departs from frequency bin zero to frequency bin $q - 1$. An estimate of the phase at q will be calculated using that slope, and the actual phase at frequency bin q will be unwrapped around that estimate. Since we want only reliable frequency bins to modify the estimated slope, the slope will be recalculated only in the frequency bins where the magnitude is greater or equal than ϵ times the maximum magnitude in the spectrum.

1.3 Linear Regression Slope Forced Unwrapping (LRSFU)

The most widely used method for phase unwrapping is [10], and a less general version of it was implemented in [13]. For intermediate estimate at frequency bin q, frequency bins 0 to $q - 1$ are used to perform a linear regression. The calculated slope is used to predict an estimate of the phase of frequency bin q, unwrapping the actual phase around that estimate. The value ϵ was used in the same way as in section 1.2.

After unwrapping the phase, we want to apply a linear regression to the unwrapped phase of formula 3. The problem is stated as:

$$\min_{\delta} \left\| W^{\frac{1}{2}} \left(\Phi - \frac{2\pi\delta}{T_s N} Q \right) \right\| \qquad \text{problem} \qquad (4)$$

$$\widehat{\delta} = \frac{T_s N \left(Q^T W \Phi \right)}{2\pi \left(\left\| W^{\frac{1}{2}} Q \right\|^2 \right)} \qquad \text{solution} \qquad (5)$$

$$\xi = \frac{\left\| W^{\frac{1}{2}} \left(\Phi - \frac{2\pi\widehat{\delta}}{T_s N} Q \right) \right\|}{\left\| 1^T W^{\frac{1}{2}} \right\|} \qquad \text{error} \qquad (6)$$

where W is an NxN diagonal matrix with the weights as the diagonal elements. Q is a vector containing the frequency bin indexes $q = 0$ to $q = N - 1$, and Φ is the vector containing the unwrapped phase of each of the frequency bins of formula 3.

We know that the phase of frequency bins with low magnitude will be more susceptible to be corrupted by both spectral leakage and white noise. For these reasons we proposed the following weighting scheme:

$$W_{q,q} = |X_2(q)|^{\mu} \qquad (7)$$

where μ is a real number greater than one to emphasize the frequencies with high amplitude over the ones with low amplitude.

After defining the phase unwrapping and weighting schemes, we have to design an iterative method to find the pitch period from scratch. In section 2 we design two iterative methods to find the pitch. One that finds an initial estimate of the pitch using a long window (20ms), and another that finds an accurate estimate of the pitch using an adaptive window length, avoiding pitch doubling and jitter effects. In section 3 we find the thresholds used to classify a speech segment as voiced or unvoiced. We also evaluate the performance of our method with both clean and noisy speech, comparing it to the performance of the cepstrum method ([2]) and the autocorrelation method ([14]).

2 Iterative Method for Extracting the Pitch

We saw in [6] that the estimated pitch $\widehat{T}$ and the regression error ξ calculated from formulas 4, 5 and 6 would be able to tell us that either $\widehat{T} = 0$ or $\widehat{T}$ is the pitch (given that $T + \delta$ is between 0 and 1.5 times the pitch period). It would also tell us that $\widehat{T}$ is a wrong estimate of the pitch period if the regression error ξ is too high. Using this information we can perform several iterations of linear regressions of the phase, fixing the position of frame x_1, and shifting frame x_2 regarding the estimated pitch $\widehat{T}$ and regression error ξ of the last linear regression iteration. This method is what we call Iterative Linear Regressions of the Phase (ILRP).

Figure 1 gives a glimpse of the ILRP method. The algorithm starts with a frame length of 20ms. It assumes the beginning of x_1 fixed, and the beginning of x_2 initially shifted 1ms after the beginning of x_1. ($T + \delta = 1ms = shift$). Then, the algorithm begins. At any iteration, a linear regression of the phase will be performed using formulas 3, 4, 5, 6 and 7. As a result, $\widehat{T}$ and $\widehat{\delta}$ will be calculated. More iterations will be performed until the regression error ξ and the estimated $\widehat{\delta}$ are below the thresholds, in which case, $\widehat{T}$ will be the output of the algorithm. In case the number of iterations goes above a maximum threshold the algorithm outputs that a pitch couldn't be found. After each iteration, the beginning of x_2 is intended to be shifted to a position that gets closer to one pitch period after the beginning of x_1.

To avoid covering several pitch periods in one 20ms window, and to approximate the method to the ideal pitch synchronous case treated in section 1, a variation of the ILRP method is implemented. This variation is used after the first pitch period has been successfully found

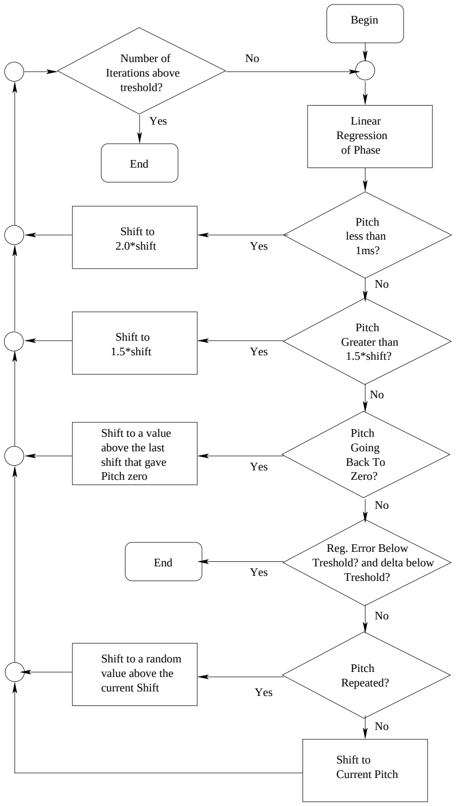
Begin
Number of Iterations above treshold?
No
Yes
End
Linear Regression of Phase
Pitch less than 1ms?
Yes
Shift to 2.0*shift
No
Pitch Greater than 1.5*shift?
Yes
Shift to 1.5*shift
No
Pitch Going Back To Zero?
Yes
Shift to a value above the last shift that gave Pitch zero
No
Reg. Error Below Treshold? and delta below Treshold?
Yes
End
No
Pitch Repeated?
Yes
Shift to a random value above the current Shift
No
Shift to Current Pitch

Fig. 1. Flow Chart describing the ILRP method.

by ILRP. This variation sets the frame length to the last pitch period found. The beginning of x_2 is initially shifted a last $\widehat{T}$ found after the beginning of x_1. This method is called Adaptive Frame Length Iterative Linear Regression of the Phase (AFLILRP).

The threshold for ξ in ILRP is 0.2. The threshold for $|\widehat{\delta}|$ in ILRP is 0.5ms. The threshold for ξ in AFLILRP is 1.5.The $|\widehat{\delta}|$ threshold for AFLILRP is 0.2ms. The value for μ in ILRP is 6, while the μ value used for AFLILRP is 1. The reason we chose these thresholds is stated in section 3. The maximum number of DFTs that can be executed in both AFLILRP and ILRP is 10. For both SFU and LRSFU we used $\epsilon = 0.2$ ([6]).

To avoid further errors, we will assume that the pitch cannot occur over silence (low energy) and that consecutive pitch periods must not be too different. This difference is given by the following similarity condition:

$$0.85\widehat{T}_{j-1} \leq \widehat{T}_j \leq 1.15\widehat{T}_{j-1} \qquad (8)$$

where j and j-1 are consecutive periods found by consecutive AFLILRP iterations.

The steps followed to classify the frame x_1 as voiced or unvoiced and to extract the pitch period at the location of x_1 are performed by the following two states machine:

1. Unvoiced Segment. Go to next frame. If the energy is high apply an iteration of ILRP. If a pitch period is found apply one AFLILRP iteration to the same frame. If both pitch periods follow the similarity condition, go to state 2. Else, restart 1.

2. Voiced Segment. Go to next frame. If energy is high, apply AFLILRP. If a pitch period is found and the last two pitch periods found follow the similarity condition restart state 2. Else, apply another AFLILRP iteration to the next frame. If pitch period is found restart state 2. Else, start in state 1.

3 Results

3.1 Test Data

For the results in this section we used 164 seconds of speech among 5 male speakers and 196 seconds of speech among 12 female speakers. The method used to label each utterance is similar to the one used in [15]. For each speech file we recognize either a negative or positive maximum that is easy recognize in all the voiced segments. The time difference between two adjacent maximums (or minimums) is in fact the pitch period which is stored in our reference database.

3.2 Estimation Accuracy vs. Regression Error (ξ) and Delta ($|\delta|$)

To make the measurements of the three plots in figure 2, we used 82 seconds of male speech. For both ILRP and AFLILRP and for each unwrapping method we chose the beginning of each pitch period as the beginning of frame x_1 and performed a linear regression of the phase positioning x_2 for each δ in the interval $[-0.4T, 0.4T]$. Then we stored the resulting estimation error $\left(\left|\frac{T-\widehat{T}}{T}\right|\right)$, regression error ($\xi$) and estimated delta ($\widehat{\delta}$). Lets give each of these results a data index i and define:

$$C(\delta, \widehat{\delta}, \xi) = \{i \mid \left|\frac{\delta_i}{T_i}\right| \leq \delta, \left|\frac{\widehat{\delta_i}}{T_i}\right| \leq \widehat{\delta}, |\xi_i| \leq \xi\} \qquad (9)$$

$$E = \sum_{i \in C\left(\delta, \widehat{\delta}, \xi\right)} \frac{\left|\widehat{T}_i - T_i\right|}{T_i M} \qquad (10)$$

where M is the number of elements in the set $C(\delta, \widehat{\delta}, \xi)$. E is defined in formula 10 as the mean error of all the estimation errors in set C.

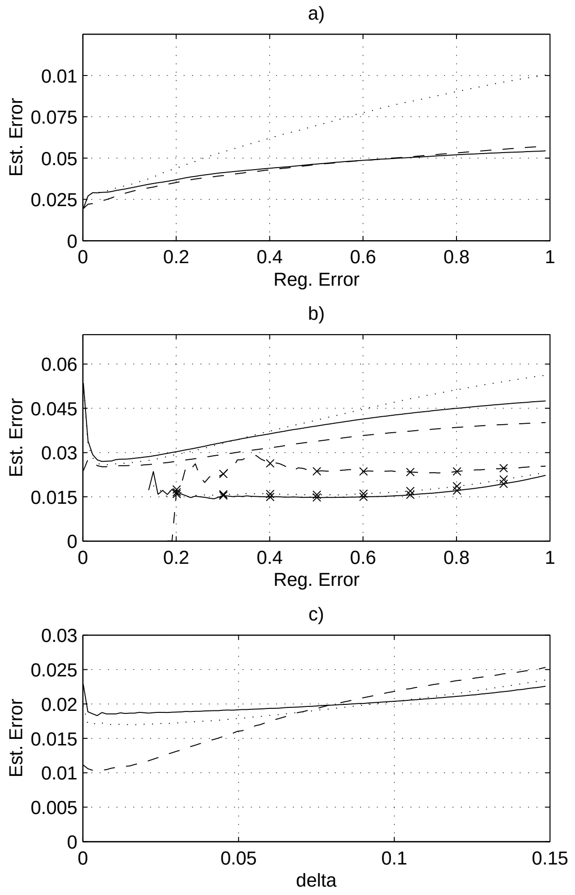

Fig. 2. a) Mean of estimation errors vs regression error ξ for the set $C(0.4, 0.7, \xi)$ using ILRP. b) Mean of estimation errors vs regression error ξ for the set $C(0.15, 0.6, \xi)$ using AFLILRP. c) Mean of estimation errors vs $|\delta|$ in for the set $C(\delta, 0.6, 1)$ using AFLILRP. Dashed, dotted and continuous lines stand for BU,SFU and LRSFU respectively. Lines with crosses are for $\mu = 1$ while lines without crosses are for $\mu = 6$.

Figure 2 shows that unwrapping methods BU and LRSFU are the most accurate for finding the pitch in ILRP. We can also see that, having 0.2 as a threshold for ξ will give us an expected error of less than 0.05 times the actual period. Figure 2 b) shows that the AFLILRP method behaves more accurately for $\mu = 1$ than for $\mu = 6$ that means that we don't need aggressive weights when the frame length is almost pitch synchronous. Figure 2 c) shows that, it is better to use Basic Unwrapping for AFLILRP only in situations where δ is known to be small.

From figure 2 we can justify the threshold values used for ILRP and AFLILRP stated in section 2.

Table 1. Pitch Estimation Performance For Different Phase Unwrapping Methods In ILRP And AFLILRP

Male Data, Clean Speech						
Measure	1-1	1-2	2-1	2-2	ceps	ac
GPE(%)	10.60	7.33	10.04	6.58	9.94	3.09
V-UV(%)	30.66	10.31	29.82	9.67	6.27	9.84
UV-V(%)	2.54	4.06	2.56	3.77	9.81	11.62
GEC(%)	2.73	2.73	2.76	3.02	6.00	7.95
FPEAV(ms)	0.096	0.057	0.049	0.059	0.047	0.13
FPESD(ms)	0.29	0.30	0.30	0.30	0.32	0.40
FFTV	5.1	4.75	5.0	4.6	2.0	- -
FFTUV	13.65	13.53	11.0	10.9	2.0	- -

Female Data, Clean Speech						
Measure	1-1	1-2	2-1	2-2	ceps	ac
GPE(%)	11.68	10.73	10.12	8.99	4.87	5.60
V-UV(%)	9.43	8.11	6.96	6.68	5.07	4.51
UV-V(%)	2.59	3.66	6.57	6.86	6.34	12.14
GEC(%)	1.84	2.44	2.77	1.91	2.59	6.55
FPEAV(ms)	-0.014	0.008	0.001	0.012	0.021	0.045
FPESD(ms)	0.17	0.144	0.15	0.15	0.18	0.19
FFTV	4.97	4.49	4.42	4.36	2.0	- -
FFTUV	14.00	13.71	13.35	9.75	2.0	- -

3.3 Performance

Table 1 shows the performance measure in each row for the different phase unwrapping methods in each column. Number 1 stands for SFU and 2 stands for LRSFU. For example, method 2-1 means LRSFU in ILRP and SFU in AFLILRP. We also compared the the performance of our method with the Cepstrum pitch detection method ([2]) and the Autocorrelation method ([14]). The performance measures we used are the ones used in [15] and [16]. We added the measures FFTV and FFTUV that stand for the number of fourier transforms that had to be executed to decide that a frame is voiced or unvoiced respectively.

We can see that 2-2 is the method that performs the best in terms of GPE. In temrs of V-UV and UV-V, 2-2 performs the best for male data, while it performs almost the same as 1-2 for female data. In terms of GEC, FPEAV and FPESD, 1-2 and 2-2 performs almost the same. However, 2-2 is faster and more efficient in finding out if a segment is voiced or unvoiced. For this reason we will use 2-2 for our noise analysis and for a comparison with the cepstrum method and autocorrelation method.

For male data, 2-2 performs clearly better than cepstrum. 2-2 performs considerably better than autocorrelation in the UV-V, GEC and FPESD measures. For female data, cepstrum performs slightly better than 2-2 in the V-UV and UV-V measures, while considerably better in the GPE measure. In terms of GEC and FPESD,

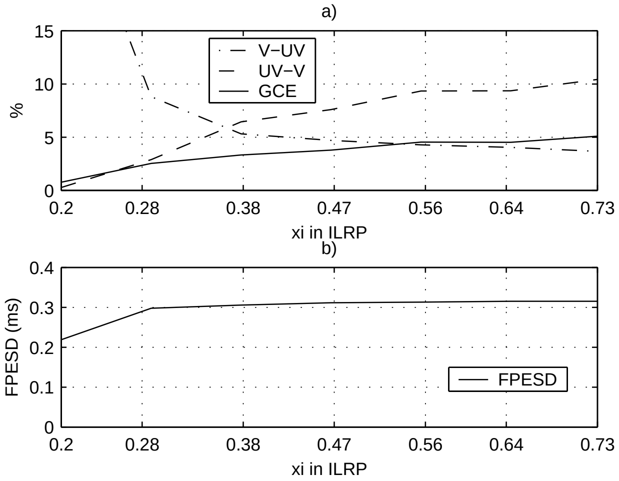

Fig. 3. a) V-UV, UV-V and GEC performance measures vs ξ in ILRP for male speech. b) FPESD performance measure vs ξ in ILRP for male speech.

2-2 performs better than cepstrum. The autocorrelation method performs considerably worse than 2-2 in the UV-V, GEC and FPESD measures. The high UV-V measure in the autocorrelation method makes it hard to make a comparison regarding V-UV and GPE.

In summary, our method, 2-2, performs better always in terms of GEC and FPESD, while it performs similarly or better in the rest of the measures depending on if the data is from male or female.

3.4 Performance over noisy speech

Since the noise in a signal will be reflected in the linear regression error ξ of the phase of our method, we can adjust the tresholds discussed in section 2 to improve the performance of our method over nosy data. Figure 3 shows the performance measures over male and female data at 5db SNR. We run our pitch detection method (2-2, LRSFU-LRSFU) with the following set of tresholds:

1. ILRP: $\xi = 0.20$,$\delta = 0.50ms$. AFLILRP: $\xi = 1.50$,$\delta = 0.20ms$.

2. ILRP: $\xi = 0.28$,$\delta = 0.67ms$. AFLILRP: $\xi = 1.67$,$\delta = 0.29ms$.

3. ILRP: $\xi = 0.38$,$\delta = 0.83ms$. AFLILRP: $\xi = 1.83$,$\delta = 0.38ms$.

4. ILRP: $\xi = 0.47$,$\delta = 1.00ms$. AFLILRP: $\xi = 2.00$,$\delta = 0.47ms$.

5. ILRP: $\xi = 0.56$,$\delta = 1.20ms$. AFLILRP: $\xi = 2.17$,$\delta = 0.56ms$.

Table 2. Pitch Estimation For 5db Nosy Speech

Male Data, SNR = 5db			
Measure	Our Method	Cepstrum	Autocorr.
V-UV(%)	5.31	18.43	0.71
UV-V(%)	6.42	14.55	15.64
GEC(%)	3.33	11.47	8.23
FPEAV(ms)	0.051	0.027	0.12
FPESD(ms)	0.31	0.33	0.40

Female Data, SNR = 5db			
Measure	Our Method	Cepstrum	Autocorr.
V-UV(%)	6.63	10.80	0.48
UV-V(%)	7.01	15.53	38.35
GEC(%)	7.22	18.17	8.16
FPEAV(ms)	-0.004	0.038	0.039
FPESD(ms)	0.20	0.20	0.20

6. ILRP: $\xi = 0.64$,$\delta = 1.30ms$. AFLILRP: $\xi = 2.33$,$\delta = 0.64ms$.

7. ILRP: $\xi = 0.73$,$\delta = 1.50ms$. AFLILRP: $\xi = 2.50$,$\delta = 0.73ms$.

Figure 3 shows that we can reach excellent V-UV and UV-V measures without hurting the GEC and FPESD measures at 5db SNR. We can also see that the V-UV and UV-V measures are highly sensible to parameter variations, while GEC and FPESD are less sensible to parameter variation. Table 2 shows the superior performance of our method compared to the cepstrum and autocorrelation method. The flexibility of our method gives us great advantages for tuning up parameters in noisy situations.

4 Conclusions

We have described a method that uses a time delay estimation technique and phase information to detect the pitch frequency of a speech signal. We have found that using the right phase regression method along with the right weights will find faster and more accurately the pitch period. We have also described an iterative method to find the pitch period from scratch at every frame in the waveform. We have shown that long windows (20ms) with dramatic weights will give a fast finding of a rough estimate of the pitch, while pitch synchronous windows with fair weights will give a more accurate finding of the pitch. Finally, experimental results have shown low fine pitch errors and gross error counts compared to the cepstrum and autocorrelation methods. Experimental results have also shown the flexibility of our method to attain better results in 5db SNR speech.

5 References

[1] W.J. Hess, Pitch Determination of Speech Signals, Berlin, Germany: Springer-Verlag, 1983.

[2] A.M. Noll, “Cepstrum Pitch Determination,” J. Acoust. Soc. America. Vol 41, pp. 293-309, 1967.

[3] J.D. Markel, “The SIFT algorithm for fundamental frequency estimation,” IEEE Trans. Audio Electroacust. vol. AU-20, pp 367-377, Dec. 1972.

[4] L.R. Rabiner, “On the use of autocorrelation analysis for pitch detection,” IEEE Trans. Acoust, Speech, Signal Processing, vol. ASSP-25, no. 1, pp. 24-33, Feb 1977.

[5] M.J. Ross, H.L. Shaffer, A. Cohen, R. Freudberg, and H. J. Manley, “Average magnitude difference function pitch extractor,” IEEE Acoust, Speech, Signal Processing, vol. ASSP-22, pp. 353-362, Oct. 1974.

[6] R.E. Prieto, S. Kim, “Robust Pitch Tracking using Linear Regression of the Phase,” The Ninth Australian International Conference on Speech Science and Technology, Melbourne, Australia, 2002.

[7] A.V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, Englewood Cliffs, NJ: Pretince-Hall,1989.

[8] J. M. Tribolet and T.F. Quatieri, “Computation of the complex cepstrum”, Programs for Digital Signal Processing, Digital Processing Comitee of the IEEE ASSP Sociate, Eds. New York: IEEE Press. 1979, Ch. 7.

[9] L.R. Rabiner and R.W. Schafer, Digital Processing of Speech Signals, Pretince Hall, 1978.

[10] J.M. Tribolet, “A new phase unwrapping algorithm,” in IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-25, pp. 170-177, 1977.

[11] H. Al-Nashi, “Phase Unwrapping of Digital Signals,” in IEEE Trans. Acoust., Speech, Signal Processing, vol. 37, pp. 1693-1702.

[12] R. McGowan and R. Kuc, “A direct relation between a signal time series and its unwrapped phase,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-30, pp. 719-726, Oct. 1982.

[13] Michael S. Brandstein, John E. Adcock, and Harvey F. Silverman, “A Practical Time-Delay Estimator for Localizing Speech Sources with a Microphone Array,” Computer, Speech, and Language, Vol. 9, April 1995, pp. 153-169.

[14] B. G. Secrest, G. R. Doddington, ”An integrated pitch tracking algorithm for speech systems,”

[15] D.-J. Liu, C.-T. Lin, ”Fundamental Frequency Estimation Based on the Joint Time-Frequency Analysis of Harmonic Spectral Structure,” IEEE Trans. Speech and Audio Processing, vol. 9, pp. 609-621, Sep. 2001.

[16] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. Mcgonegal, “A comparative performance study of several pitch detection algorithms,” IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 399-417, Oct. 1976. Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, 1983, pp. 1352-1355.