# Maximum Likelihood Pitch Estimation Using Sinusoidal Modeling

Vijay Mahadevan, Carol Y.Espy-Wilson
Institute for Systems Research & Department of Electrical & Computer Engineering,
University of Maryland, College Park, MD, USA
vijaymahad@gmail.com, espy@umd.edu

***Abstract*— An algorithm for optimal estimation of pitch frequency using a maximum likelihood formulation is presented. The speech waveform is modeled using sinusoidal basis functions that are harmonically tied together to explicitly capture the periodic structure of voiced speech. The problem of pitch estimation is casted as a model selection problem and the Akaike Information Criterion is used to estimate the pitch.**

***Keywords-voicing, pitch, fundamental frequency, sinusoidal modeling, maximum likelihood, Akaike information criteria***

## I. INTRODUCTION

The fundamental frequency of voiced speech is an important cue for almost all speech analysis- synthesis systems. Accurate and robust pitch estimation is necessary for several applications like speech coding, speaker recognition, speech recognition etc. Several pitch detection algorithms (PDA) have been proposed and a comparative study highlighting the problems and performance of these pitch detectors is presented in [1, 2].

Typically, pitch determination requires a search of different possible candidate frequencies over an analysis window. A cost function is defined for every pitch candidate and the estimated frequency is chosen to be the one that gives an optimum cost. For example, the autocorrelation based pitch detector can be formally viewed as minimizing, over possible pitch periods the mean squared error between the signal and its delayed version. It is essentially a measure of self-similarity [5] and we expect to observe peaks near the actual period. The maximum likelihood formulation of this problem was discussed in [3] which is related to the work by Steiglitz [4] who discussed the problem of pitch estimation by trigonometric curve fitting. In both these cases, an explicit model about the signal periodicity is imposed where the former takes place in the time domain with a similarity measure and the latter in the frequency domain with signal model.

In this paper, we present a statistical method for pitch tracking by using a generalization of the discrete Fourier transform representation. It can also be viewed as a special case of sinusoidal speech model where all the sinusoidal components are assumed to be harmonically related i.e. the integer multiples of the fundamental frequency. The system outputs a pitch estimate for every frame that is detected to be voiced. We follow a metric that estimates the local signal to noise ratio (SNR) and decide on the voicing probability [5]. The voice activity detection is an integral part of the algorithm which is measured by the goodness of the model fit to the observation. The statistical method for pitch tracking presented in this paper follows the maximum likelihood estimation of the parameters. We follow a regression framework and decide on the pitch frequency using the Akaike Information Criteria (AIC). A related work to the problem formulation is given in [6] where maximum a posteriori (MAP) estimation is used in pitch tracking.

The two sources of errors in the performance of a pitch detector originate from voice activity detection (VAD) and pitch estimation. The pitch insertion and deletion errors are used to measure the performance of the VAD and the pitch substitution errors account for the gross error in the pitch estimates [2]. Different kinds of post processing schemes like median filtering and dynamic programming are often used to remove discontinuities in the pitch tracks. The discontinuities arise from pitch doubling or halving errors or any of the above mentioned errors. Paul Bagshaw's database [13] was used for evaluation. The results are reported for both raw pitch estimates and the post-processed pitch values using median filtering. An extensive comparison of the performance with several algorithms which were evaluated on the same database is presented.

We consider three principal parts of the mathematical model presented in this paper i.e. the conceptual, analytic and computational aspects in sections II, III and IV. The voicing decision block is outlined in section V and the experimental results and discussion are presented in section VI. Finally, section VII concludes the paper with our future work.

## II. SIGNAL MODEL

### A. Motivation

For a stationary speech signal, pitch can be defined as the perception of a fundamental frequency of a pure harmonic template which optimally fits successive harmonic component pattern of the speech signal [7]. We follow a signal model that explicitly captures the periodic structure of the speech signal. This approach towards estimating pitch is referred as Harmonic Structure Matching Pitch Estimation (HSMPE) [8]. In our work, we explicitly model the time domain signal using sinusoidal basis functions that are harmonically tied together.

### B. Mathematical Formulation

We start with the basic Fourier series representation of a stationary periodic signal. The windowed speech waveform is

978-1-4244-9799-7/11/$26.00 ©2011 IEEE

represented by a sum of sinusoidal functions with fixed amplitudes, frequencies and phases [9]. This approach can be viewed as a generalization of the discrete Fourier transform i.e. the period of the signal is arbitrary and not necessarily equal to the length of the signal. This framework was used in [10] in the name of regressive discrete Fourier series and it is well known in the statistical literature as least squares spectral analysis. Under this condition, the windowed speech signal *s[n]* is represented as,

$$s[n] = \sum_{k=1}^{M(f_0)} (a_k \cos(2\pi f_0 k n + \varphi_k)) + \varepsilon[n] \quad (1)$$

where $1 \le n \le N$, $a_k$, $\varphi_k$ and $f_0$ represent the amplitude, phase and fundamental frequency and $\varepsilon[n]$ represents the residual error from the model. Equation 1 can be compactly written in matrix form as,

$$\mathbf{s} = \mathbf{A}(f_0) * \boldsymbol{\gamma} + \boldsymbol{\varepsilon} \quad (2)$$

$$\boldsymbol{A}(f_0) = \begin{pmatrix} e^{1i\omega_0} & e^{1i2\omega_0} \dots & e^{1iM\omega_0} \\ e^{ni\omega_0} & e^{ni2\omega_0} \dots & e^{niM\omega_0} \\ e^{Ni\omega_0} & e^{Ni2\omega_0} \dots & e^{NiM\omega_0} \end{pmatrix}$$

where the matrix **A** contains complex exponentials at the multiples of $\omega_0 = 2\pi f_0$ and is of size NxM($f_0$). The harmonic amplitude and phase information is captured in $\boldsymbol{\gamma}$. The residual error is assumed to be additive white Gaussian noise with zero mean and covariance matrix $R = \sigma^2 I$. Hence the unknown parameters in the model are $f_0$, $\boldsymbol{\gamma}$ and $\sigma^2$ which we wish to estimate from the observed signal.

### C. Maximum Likelihood Estimation

The likelihood of observing the data given the parameters is,

$$g(\boldsymbol{s}|f_0, \sigma^2, \boldsymbol{\gamma}) \sim N(\boldsymbol{A}(f_0) * \boldsymbol{\gamma}, \sigma^2 \boldsymbol{I}) \quad (3)$$

and the log-likelihood function $L(\boldsymbol{\theta})$ with $\boldsymbol{\theta} = [\sigma^2, \gamma, f_0]$ containing all the unknown parameters is given by,

$$L(\sigma^2, \boldsymbol{\gamma}, f_0) = \frac{N}{2} \ln\left(\frac{1}{2\pi\sigma^2}\right) - \frac{1}{2\sigma^2}([\boldsymbol{s} - \boldsymbol{A}(f_0)\boldsymbol{\gamma}]^H[\boldsymbol{s} - \boldsymbol{A}(f_0)\boldsymbol{\gamma}]) \quad (4)$$

The maximum likelihood parameter estimate is found by maximizing (4),

$$\hat{\boldsymbol{\theta}} = \arg max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{\theta}) \quad (5)$$

The log-likelihood function is non-linear in $f_0$ and the usual optimization methods will yield local maxima. However, the parameter space for $f_0$ is restricted to the possible pitch frequency for humans and therefore we do a global brute force approach for estimating $f_0$. To do so, we fix $f_0 = f_0'$ and observe that the optimization problem is quadratic in $\boldsymbol{\gamma}$ and the solution is given by Moore-Penrose pseudo inverse of $\mathbf{A}(f_0')$ denoted as $\boldsymbol{A}^+(f_0') = (\mathbf{A}(f_0')^{\mathbf{T}} * \mathbf{A}(f_0'))^{-1} * \mathbf{A}(f_0')$. The well known optimal estimates is noted below for $\boldsymbol{\gamma}$ and $\sigma^2$,

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{A}^+(f_0') * \boldsymbol{s} \quad (6)$$

$$\widehat{\sigma^2} = [s - \mathbf{A}(f_0') * \hat{\boldsymbol{\gamma}}]^{\mathbf{T}}[s - \mathbf{A}(f_0') * \hat{\boldsymbol{\gamma}}]/\boldsymbol{n} \quad (7)$$

The estimated signal $\hat{s}$ is given by the projection of the observation on the space spanned by the columns of $\mathbf{A}(f_0')$,

$$\hat{\boldsymbol{s}} = \mathrm{P}_{\mathbf{A}(f_0')} * \mathbf{s} \quad (8)$$

$$\mathrm{P}_{\mathbf{A}(f_0')} = \mathbf{A}(\mathbf{A}^{\boldsymbol{T}}\mathbf{A})^{-1}\boldsymbol{A}|_{f_0'} \quad (9)$$

The maximized value of the log-likelihood function ignoring the additive constants is then given by,

$$L(\hat{\theta}) = \frac{N}{2} \ln\left(\frac{1}{\widehat{\sigma^2}}\right) \quad (10)$$

The problem formulation is reduced to minimizing the residual sum of squares. The column space of $\mathbf{A}(\frac{f_0}{2})$ is a superset of $\mathbf{A}(f_0)$ and therefore the residual error variance will follow $\widehat{\sigma^2}_{f_0/2} \le \widehat{\sigma^2}_{f_0}$. It can be seen that choosing $f_0$ that maximizes $\mathbf{L}(\theta)$ in (10) will result in pitch halving error almost always when $\frac{f_0}{2}$ is in the parameter space. This should come as no surprise as we are simply doing a regression on the data using different models indexed by $f_0$. Therefore we need a tradeoff on the number of parameters used to describe the model i.e. the complexity of the model and the goodness of fit from the model. This is achieved using the AIC described in the next section.

## III. Model Selection

The AIC model selection stems from the Kullback- Leibler (K-L) information loss [11, 12]. It follows an information theoretic approach to choose the best model from a set of candidates. In our case, the different models are indexed by the fundamental frequency. The tradeoff between the model complexity and the goodness of fit as given by AIC is,

$$AIC(model) = -2 * (Max.value\ of\ the\ likelihood/model) + 2 * number\ of\ parameters\ in\ the\ model \quad (11)$$

$$AIC(f_0) = N \ln\left(\widehat{\sigma^2}(f_0)\right) + 2 * M(f_0) \quad (12)$$

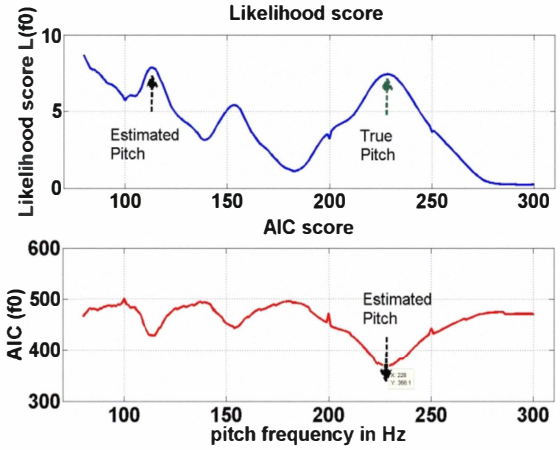


Figure 1. Illustration of pitch halving error (top) Likelihood score and (bottom) AIC score

We have the maximized log-likelihood value using the templates of projection matrices indexed by $f_0$. The number of parameters in the model is equal to the number of regressors used i.e. the dimension of the harmonic coefficients M($f_0$). We choose the $f_0$ that gives the lowest AIC score. A scenario illustrating the pitch halving error through ML model selection which is corrected using AIC information criteria is show in Fig.1. The algorithm provides high resolution in estimating the pitch frequency as we are not restricted to work with integer periods with resolution dictated by the sampling interval. The effect of pitch resolution in computational complexity is analyzed in the following section.

## IV. Computational Complexity

It should be noted that other minimum mean squared error methods based on similarity measures like the autocorrelation, cross correlation [17, 19] and the Average Magnitude Difference Function (AMDF) [20] require O(N) computations for every candidate pitch period (brute force approach) and therefore a total of $O(T * N)$ computational load. For methods that transform the time domain signal to the frequency domain like cepstrum [14], harmonic product spectrum [16] and sub-harmonic to harmonic ratio [21] require $O(N * logN)$ computations.

In the problem of pitch estimation we are essentially solving a system of linear equations through projection templates. The storage complexity of these templates requires a memory space of the order (Big- O notation) $O(T * N^2)$ where T denotes the cardinality of the $f_0$ parameter search space. The number of computations done per candidate model is $O(N^2)$ and therefore for T models we have a total of $O(T * N^2)$. The algorithm can be easily scaled to meet the computational requirements with a tradeoff on the accuracy of the pitch estimates. By computing the pitch frequency in the first voiced frame, gradient search techniques can be used to estimate the fundamental frequency in the successive frames. There can be various strategies to efficiently search the pitch grid starting from a coarse resolution and then tuning it to a finer resolution according to the required level of accuracy. Fig.2 illustrates the computational time required to process a signal of length 1.35$s$ sampled at 8 kHz at 10$ms$ frame rate in 3GHz Intel processor. The computational time further scales with the sampling frequency of the signal. If we down sample the signal by a factor of L, the computational complexity scales by a factor of $L^2$ i.e. the load for T models is $O\left(T * \left(\frac{N}{L}\right)^2\right)$.

## V. Voicing Detection

Voice activity detection is an integral part of the algorithm which is measured by the goodness of the model fit to the observation. The estimated speech signal $\hat{s}$ and the residual ε can be used to arrive at a measure of local SNR as follows,

$$\boldsymbol{\varepsilon} = \boldsymbol{s} - \hat{\boldsymbol{s}} \tag{13}$$

$$SNR = 10log_{10}\left(\frac{\hat{s}^T\hat{s}}{\varepsilon^T\varepsilon}\right) \tag{14}$$

The voicing decision can be based on the SNR level and one approach indicated in [12] is,

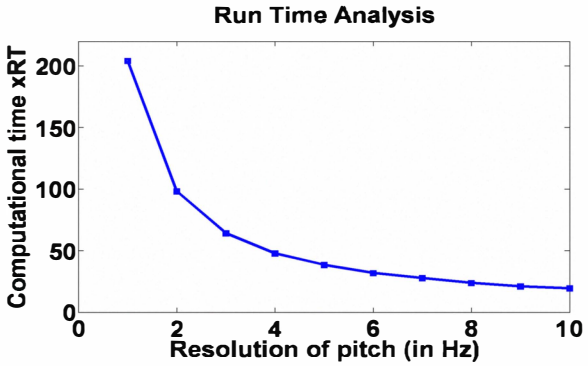


Figure 2. Computational time analysis

$$\left\{P_v = \begin{vmatrix} 1, SNR > 10dB \\ \frac{1}{6}(SNR - 4), 4dB \le SNR \le 10dB \\ 0, SNR < 4dB \end{vmatrix}\right\} \tag{15}$$

Hence we can see that the algorithm can provide an optimal speech enhancement by using the reconstructed speech signal. This can be further refined by introducing time-frequency gain manipulation in the reconstruction process to improve the SNR of the recovered signal.

## VI. Experimental Results

### A. CSTR Database

Performance evaluation is done on the publicly available database provided by the Center for Speech Technology Research at University of Edinburgh, Scotland, UK. The database includes 50 sentences each from a male and female speaker. The database was biased towards utterances containing voiced fricatives, nasals, liquids and glides, since PDAs generally find these difficult to analyze [13]. The analysis window length was fixed at 25ms at 20 kHz sampling frequency and a frame rate of 6.4ms was followed. The pitch range analyzed was between 80-400Hz for both male and female speakers. There was no pre-processing stage to filter the speech signal.

### B. Performance Comparison

The $f_0$ value from the laryngeal frequency contour was used as the reference. Every $f_0$ value in the reference file had a time label which was used to align the estimated pitch value ($P_{est}$) with the reference pitch ($P_{ref}$). A nearest neighbor interpolation was used to compare the two pitch values at the time label where the algorithm estimated the pitch. The error measures computed for performance evaluation are the same as specified in [13]. When the estimated and reference pitch represent voiced speech, we have two error measures namely, gross errors and fine errors. The gross error high (GEH) is counted if $P_{est} > 1.2 * P_{ref}$ and gross error low (GEL) is counted if $P_{est} < 0.8 * P_{ref}$ for the duration when both represent voiced speech. Net gross error (GE) is the sum of GEL and GEH. Fine errors in pitch estimation are defined on the frames where $|P_{est} - P_{ref}| \le 0.2 * P_{ref}$. The duration of unvoiced or silent regions incorrectly classified as voiced by the PDA is noted as *unvoiced in error*. This result is

accumulated over all the utterances for a speaker and noted as a percentage of total unvoiced (or silent) duration. Similarly, we have *voiced in error* for the duration of voiced speech that are erroneously classified as unvoiced. The statistics of the absolute deviation in the fine pitch errors are reported in mean and population standard deviation (p.s.d). Table 1 shows the results for the seven PDAs in Bagshaw's experiment[1], modified AMDF with probabilistic error correction and sub-harmonic to harmonic ratio approaches. The list of PDAs used in the comparison is,

- Cepstrum pitch determination (CPD)[14]
- Feature-based pitch tracker (FBPT)[15]
- Harmonic product spectrum (HPS)[16]
- Integrated pitch tracking algorithm (IPTA)[17]
- Parallel processing method (PP)[18]
- Super resolution pitch determinator (SRPD)[19]
- Enhanced version of SRPD (eSRPD)[13]
- Modified AMDF-based PDA with probabilistic error correction (mAMDFp) [20]
- Pitch determination algorithm based on sub-harmonic to harmonic ratio (SHR) [21]
- Maximum likelihood pitch detection (ML-AIC)
  - Raw pitch results (raw)
  - Post-processed by median filter (filtered)

The results for the first 7 PDAs are taken from [13] where eSRPD was shown to perform superior to the rest. The raw pitch estimates from the ML-AIC algorithm were post-processed with a 5 point median filter. The results from Table 1 indicate that the performance of the algorithm is comparable to or better than most of the PDAs listed.

The GEL values for ML-AIC are quite high as compared to GEH. The explanation for such bias in error is due to model over fitting. Detailed analyses on these errors on ML-AIC (raw) reveal that 75.86% of the GEL for male and 76.74% of the GEL for female occur due to pitch halving or sub multiple error i.e. $|z * P_{est} - P_{ref}| \leq 0.2 * P_{ref},\ z \in \{2,3,4\}$ . Most of the deletion errors (voiced in error) occur in the first few frames or last few frames of a voiced segment. When three frames in boundary of a voiced segment were excluded from the analysis, the deletion errors dropped to 3.51% for male and 4.99% for female. Overall the results for the raw pitch estimates indicate that the performance of the algorithm is comparable to (eSRPD) or better than most of the methods in gross errors and fine pitch errors. Median filtering reduced the insertion and deletion errors to some extent. The tradeoff for reduction in VAD errors is reflected in fine error measures. The mean absolute deviation and p.s.d show an increase in their values after smoothing. Figures 3 and 4 compare the reference pitch with the estimated pitch contour for a male and a female speaker respectively. The reference pitch values were linearly interpolated in the voiced segments at the frame rate followed in the algorithm. The post processed pitch estimates are shown in blue.

| PDA | Unvoiced in error (%) | Voiced in error (%) | Gross Errors (%) | | Net GE (%) | Absolute deviation (Hz) | |
|---|---|---|---|---|---|---|---|
| | | | High | Low | | Mean | p.s.d |
| *Male* | | | | | | | |
| CPD | 18.11 | 19.89 | 4.09 | 0.64 | 4.73 | 2.94 | 3.60 |
| FBPT | 3.73 | 13.9 | 1.27 | 0.64 | 1.91 | 1.86 | 2.89 |
| HPS | 14.11 | 7.07 | 5.34 | 28.15 | 33.49 | 3.25 | 3.21 |
| IPTA | 9.78 | 17.45 | 1.40 | 0.83 | 2.23 | 2.67 | 3.37 |
| PP | 7.69 | 15.82 | 0.22 | 1.74 | 1.96 | 2.64 | 3.01 |
| SRPD | 4.05 | 15.78 | 0.62 | 2.01 | 2.63 | 1.78 | 2.46 |
| eSRPD | 4.63 | 12.07 | 0.90 | 0.56 | 1.46 | 1.40 | 1.74 |
| mAMDFp | - | - | 1.94 | 2.33 | 4.27 | - | - |
| SHR | - | - | 1.29 | 0.78 | 2.07 | - | - |
| **ML-AIC (raw)** | **8.69** | **7.59** | **0.21** | **0.44** | **0.65** | **1.60** | **1.92** |
| **ML-AIC (filtered)** | **5.68** | **6.48** | **0.18** | **0.86** | **1.04** | **1.77** | **2.33** |
| *Female* | | | | | | | |
| CPD | 31.53 | 22.22 | 0.61 | 3.97 | 4.58 | 6.39 | 7.61 |
| FBPT | 3.61 | 12.16 | 0.60 | 3.55 | 4.15 | 5.40 | 7.03 |
| HPS | 19.10 | 21.06 | 0.46 | 1.61 | 2.07 | 4.59 | 5.31 |
| IPTA | 5.70 | 15.93 | 0.53 | 3.12 | 3.65 | 4.38 | 5.35 |
| PP | 6.15 | 13.01 | 0.26 | 3.20 | 3.46 | 6.11 | 6.45 |
| SRPD | 2.35 | 12.16 | 0.39 | 5.56 | 5.95 | 4.14 | 5.51 |
| eSRPD | 2.73 | 9.13 | 0.43 | 0.23 | 0.66 | 4.17 | 5.13 |
| mAMDFp | - | - | 0.63 | 2.93 | 3.56 | - | - |
| SHR | - | - | 0.75 | 1.69 | 2.44 | - | - |
| **ML-AIC (raw)** | **4.26** | **14.4** | **0.06** | **2.02** | **2.08** | **3.96** | **4.37** |
| **ML-AIC (filtered)** | **2.05** | **13.91** | **0.04** | **1.86** | **1.90** | **4.02** | **4.5** |

Table 1: PDA evaluation for male speech (top) and female speech (bottom)

## VII. CONCLUSION

The results indicate the superior performance of the algorithm in comparison with several existing PDAs. Raw pitch estimates indicate high level of accuracy. We observe that there is a great potential to reduce the computational time through intelligent search techniques. The use of AIC for regularization mitigates some of the pitch halving error problems but there still remains significant contribution of these errors. This suggests the use of prior information to enforce continuity on the tracks as well as other post processing schemes which can be done by allowing suitable latency. Our future work will be directed towards,

- Testing the robustness of the algorithm in the presence of noise.
- Exploring regularization methods to reduce the pitch halving errors.
- Improving the computational performance.

In summary we have presented a statistically optimal framework for high resolution pitch estimation and signal enhancement.

## ACKNOWLEDGMENT

The research was supported by the NSF grant IIS-0812509.

[1] The authors would like to thank Dr.Bagshaw for providing the database and evaluation results . http://www.cstr.ed.ac.uk/research/projects/fda/

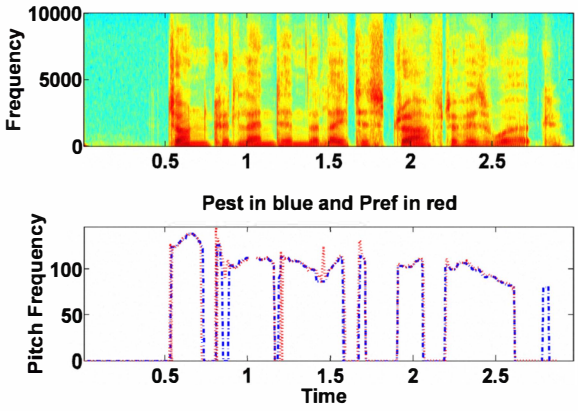


Figure 3. (top) Spectrogram and (bottom) comparison of $P_{est}$(blue) with $P_{ref}$(red) for male

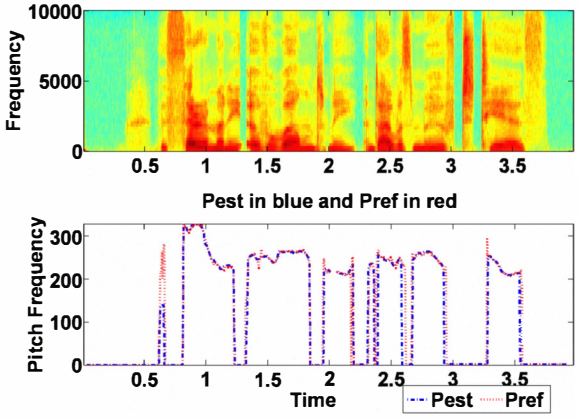


Figure 4. (top) Spectrogram and (bottom) comparison of $P_{est}$(blue) with $P_{ref}$ $(red)$ for female

## REFERENCES


[1] W. J. Hess, Pitch Determination of Speech Signals – Algorithms and Devices. Berlin, Germany: Springer, 1983.

[2] L. R. Rabiner, M. J. Cheng, A. E. Rosenberg, and C. A. McGonegal, "A comparative study of several pitch detection algorithms," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-24, pp. 399-418, Oct. 1976.

[3] J.D. Wise, J.R. Caprio, and T.W. Parks, "Maximum likelihood pitch estimation," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP -24, no.5, pp. 418-423, Oct 1976.

[4] K. Steiglitz, G. Winham, and J. Petzinger, "Pitch extraction by trigonometric curve fitting," IEEE Trans. Acoust. Speech, Signal Processing (Special Issue on 1974 Arden House Workshop on Digital Signal Processing) (Corresp), vol ASSP-23, pp. 321-323, June 1975.

[5] T. F. Quatieri, "Discrete-time speech signal processing – principles and practice," Delhi, India: Pearson Education, Inc., 2002.

[6] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model," IEEE Trans. Acoust., Speech and Audio Processing, vol.12, no. 1, pp 76-87, Jan 2004.

[7] J. L.Goldstein, "An optimum processor for the central information of pitch of complex tones," J. Acoust. Soc. Amer., vol. 54, pp. 1496-1516, 1973.

[8] Y. Gong and J. Haton, "Time domain harmonic matching pitch estimation using time-dependent speech modeling," IEEE Trans. Acoust., Speech, and Signal Processing, vol. ASSP-35, no. 10, Oct 1987.

[9] R. J. McAulay and T. F. Quatieri, "Speech analysis-synthesis based on a sinusoidal representation," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-34, pp. 744-754, 1986.

[10] J. R. F. Arruda, "A robust one-dimensional regressive discrete fourier series," Mechanical Systems and Signal Processing, vol. 24, Issue 3, pp 835-840, Apr 2010.

[11] C.R. Rao, H. Toutenburg, Shalabh, C. Heumann, "Linear models and generalizations – least squares and alternatives," Berlin, Germany: Springer-Verlag, 2008.

[12] K. P. Burnham and D. Anderson, "Model selection and multimodel inference: A practical information theoretic approach," New York, Springer.

[13] P. C. Bagshaw, "Automatic prosody analysis," PhD thesis, University of Edinburg, Scotland, UK, 1994.

[14] A. M. Noll, "Cepstrum pitch determination," Journal of the Acoustical Society of America, 41(2):293-309, 1967.

[15] M. S. Phillips, "A feature-based time domain pitch tracker," Journal of the Acoustical Society of America, 77:S9-S10(A), 1985.

[16] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental frequency measurement," Journal of the Acoustical Society of America, 43(4):829-834, 1968

[17] B. G. Secrest and G.R. Doddington, "An integrated pitch tracking algorithm for speech systems," In Proc. IEEE ICASSP-83, pp 1352-1355, 1983.

[18] B. Gold and L. Rabiner, "Parallel processing technique for estimating pitch period of speech in time domain," Journal of the Acoustical Society of America, 46(2, part 2):442-448, 1969.

[19] Y. Medan, E. Yair and D. Chazan, "Super resolution pitch determination of speech signals," IEEE Trans. Signal Processing, ASSP-39(1):40-48, 1991.

[20] G. S. Ying, L. H. Jamieson and C. D. Mitchell, "A probabilistic approach to AMDF pitch detection," Spoken Language, 1996, ICSLP 96. Proceedings., Fourth International Conference on, vol 2 no., pp. 1201-1204, 3-6 Oct 1996.

[21] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," the 6th International Conference of Spoken Language Processing, Beijing, China, 2000, 4, pp 676-679