euronoise

Pitch Tracking using the Generalized Harmonicity Indicator

Darren Haddad[a], Andrew Noga[b] and Tappan Sarkar[c]

[a]Air Force Research Lab, 525 Brooks Road, Rome, NY 13441, USA
[b]525 Brooks Road, Rome, NY 13441, USA
[c]323 Link Hall, Syarcuse University, Syracuse, NY 13244-1240, USA
darren.haddad@rl.af.mil

Abstract

For many audio applications, a process is required to obtain an accurate estimate of the fundamental and harmonics of periodic sections of an audio signal. More generally, any digital version of a periodic signal can potentially have an associated fundamental frequency component, along with harmonics which are frequency components located at integer multiples of the fundamental. In this description, the focus will be on audio applications and speech applications in particular, without loss of generality to applications outside the speech and audio domains.

For speech, tracking and assessment of fundamental and harmonic frequencies can be a key step in accomplishing such tasks as automated speaker identification, speech data compression, pitch alteration and natural sounding time compressions and expansions. Linguists and speech therapists also use such tracking and assessment for prosodic analyses and training.

Introduction

Various methods of fundamental and harmonic frequency tracking have been proposed and developed, but most have been based on other low resolution techniques such as FFT and cepstral analyses [1]. This is as opposed to using super-resolution frequency estimation as provided by the Matrix Pencil (MP) technique. The prior art in the area of super-resolution speech fundamental determination consists of the "super resolution pitch determinator" (SRPD) [2] and the "enhanced super resolution pitch determinator" (eSRPD) [3] methods. Because these prior methods do not explicitly process a spectral representation or decomposition of the input audio signal, they are not considered to be in the same class as the MP Generalized Harmonic Indicator (GHI). However, the SRPD and the eSRPD do provide a baseline for comparisons when assessing the performance of an MP based GHI and will therefore be referred to in the context of performance. Another method worth mentioning is an autocorrelation method described in [4]. In this paper the author emphasizes the comparison of the gross error, without comparing the deviation mean or the standard deviation, such as in [3] and used here.

The purpose of the GHI is to determine, assess and track the fundamental and harmonic frequencies of consecutive time segments of a signal.

Pre Processing

As a pre-processing step to the GHI process, the signal to be analyzed is first divided into consecutive overlapping or non-overlapping frames. Frame lengths and overlap percentages are typically chosen to be consistent with the stationary properties of the signal to be analyzed. In particular, multiple periods should be present in the segment, but the number of periods should not be arbitrarily large otherwise the fundamental and harmonic values may deviate excessively. Also, choosing too many periods can cause the computational complexity of super-resolution techniques to become prohibitive. Therefore, for this experiment it was determined that a frame size for a male speaker is near optimal at 25.6 ms, with a 50% overlap, corresponding to a 12.8 ms time steps for the beginning of each frame. For a female speaker the frame size is optimal at 12.8 ms, with a time step of 6.4 ms. Since a male speaker has a lower pitch period than a female speaker, it was necessary to use a larger frame size to capture the whole pitch period.

For each segment, a second pre-processing step is the calculation of the super-resolution representation of the segment, as provided by signal decompositions as in the MP technique [5]. As stated in the previous sections the MP technique is particularly effective at determining the frequency content of the signal, and includes frequency, decay rates, initial phases, and initial amplitudes in the decomposition. For each frame, 28 poles in the forward mode were determined to work well.

In a third and final pre-processing step, available decay and initial amplitude values are used to prune the original list of frequencies that the super-resolution process provides from the segment being decomposed. Frequencies that are too close to each other within the frequency resolution of the technique are eliminated. Likewise, frequency values that are not tone-like due to non-trivial decay (or growth) values are also eliminated. Therefore, poles with an absolute value of the decay coefficient less than .01 were also eliminated. Any zero valued frequencies that may result are also eliminated. The final pruning is the elimination of frequency values associated with trivial initial amplitudes relative to the number of bits of precision in the representation of the digitized signal. Therefore, poles with an amplitude less than $1/2^{16}$ were eliminated. The result is a list of frequency values $\vec{L}$, which serves as input to the GHI process. Reference is made to Figure 1 for a description of the GHI process which consists of the sequence of steps that follow.

Voiced/ Unvoiced Detection

In speech analysis, the voiced/unvoiced decision is often performed in conjunction with pitch analysis. The linking of the voiced/unvoiced decision to pitch analysis not only results in unnecessary complexity, but makes it difficult to classify short speech segments which are less than a few pitch periods in duration [6]. Therefore, it is better to classify a speech frame as voiced speech, or unvoiced speech, separately from estimating the pitch.

A voiced frame consists of a fundamental frequency with several related harmonics. An unvoiced frame does not have this property. A voiced/unvoiced detector determines if a speech frame is voiced or unvoiced. A voiced/unvoiced detector can rely on a few parameters to accomplish its goal. These parameters are energy, zero crossing, or prediction

gain. Relying on one parameter limits the robustness of the voicing detector. Whereas, increasing the number of parameters will increases its reliability [7].

In this experiment, the goal is to use a similar voiced detector that was used in [3]. This would make the comparison between the two pitch estimation algorithms fair. In [3] the author uses an energy based detector that uses a threshold. The singular values from the SVD are related to signal energy; therefore, they can be used by a voiced detector. Also, since the SVD is already used in the calculations of the MP algorithm, there is little or no processing overhead to use it as a voiced detector. To determine a voiced frame from an unvoiced frame, a threshold is needed. The maximum singular value of the frame is compared to the threshold. The frame is classified as an unvoiced frame if the threshold is larger than the maximum value. Otherwise, it would be classified as a voiced frame. This threshold is varied until the voice in error is the same as in [3]. This provides for the comparison of the pitch results in a fair way.

GHI Algorithm

The GHI process is shown in Figure 1. After the MP algorithm generates a set of frequencies for a given frame, preprocessing eliminates some frequencies as previously described. The result is a list of frequency values $\vec{L}$, which serves as input to the GHI process. The n elements of the nx1 list vector $\vec{L}$ are ordered in the Frequency Sorter, for example in ascending order, to form the ordered frequency list vector, $\vec{F}$. The nx1 vector $\vec{F}$ is then input to the Column Duplicator, which forms the nxn matrix $\mathbf{F}$ by replicating $\vec{F}$ for each column of $\mathbf{F}$. Thus $\mathbf{F} = \vec{F}\vec{1}^T$, where $\vec{1}^T$ is a 1xn dimension row vector, the elements of which are all 1. The frequency matrix $\mathbf{F}$ is then input to the Candidate Generator, where the nxn matrix of candidate fundamentals, $\mathbf{D}$ is formed as $\mathbf{D} = \mathbf{F} - \mathbf{F}^T$. When ascending ordering is used for $\vec{F}$, the matrix $\mathbf{D}$ can be represented as the sum of an upper triangular matrix and a lower triangular matrix,

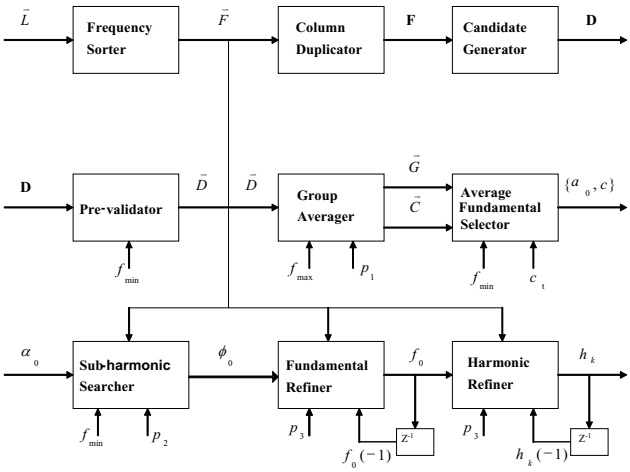

Figure 1: The Generalized Harmonicity Indicator.

and will have diagonal elements that are each zero. Thus the elements below the diagonal for the described ascending ordering will be the frequency differences which can be used to determine the fundamental and harmonics in subsequent steps.

The matrix $\mathbf{D}$ is input to the Pre-validator which forms a vector $\vec{D}$ whose elements are chosen from the positive elements of $\mathbf{D}$ that are greater than some minimum value, $f_{\min} > 0$. The elements of the mx1 vector $\vec{D}$ are arranged in ascending order and will result in $m \le 0.5n^2 - 0.5n$. The pre-validated candidate fundamental list, $\vec{D}$, is then input to the Group Averager, which produces both a vector of averaged groupings of fundamentals, $\vec{G}$, and an associated count vector, $\vec{C}$. To generate the groupings, $\vec{G}$, group boundaries are formed by inspecting the elements of the candidate fundamental list, $\vec{D}$. Starting with the second element of $\vec{D}$, a difference is formed between each current element and the previous element in the vector. If this difference is less than a fraction p_1 times the current element, then the element is grouped with the prior element. Otherwise, a new group is started with the current element. The parameter p_1 is typically chosen to be 0.1 (10 percent). Because elements are in ascending order, each group represents a distinct positive change in candidate fundamentals. For each defined group, the number of elements in each group are used as the elements of the count vector, $\vec{C}$. Using these counts, groups of candidate fundamentals are averaged to form the corresponding elements of the vector $\vec{G}$. Averages greater than the parameter $f_{\max}$ are not allowed, and likewise the corresponding elements of the count vector $\vec{C}$ are eliminated. The group average vector, $\vec{G}$, and the count vector, $\vec{C}$, are both input to the Average Fundamental Selector. If after such processing there are no elements in $\vec{G}$, then it is arbitrarily assigned a single element equal to $f_{\max}$, and the count vector $\vec{C}$ is assigned a corresponding single element equal to a count threshold, c_t. For example, the count threshold for the speech pitch estimation application was set to 3. From the group average vector, $\vec{G}$, a subset of elements is chosen which correspond to the largest elements of the count vector, $\vec{C}$, greater than or equal to the count threshold c_t. For the speech pitch estimation example application, the elements corresponding to the 3 largest counts are used. The initial fundamental estimate, α_0, is chosen as the minimum of the group averages from the subset. The count, c, is chosen as the largest count. Thus the Average Fundamental Selector is biased away from simply using the largest group average. This results in an enhanced selection process that allows for the possibility that a valid fundamental is not the one associated with the largest count.

The scalar value initial fundamental, α_0, and the associated count, c, are input to the Sub-harmonic Searcher. The Sub-harmonic Searcher forms the nx1 sub-harmonic candidate vector as $\vec{S} = \vec{F} - 0.5\alpha_0\vec{1}$ and uses this vector to determine whether or not α_0 should be reduced by a factor of 0.5. Reduction is performed if $0.5\alpha_0$ is greater than $f_{\min}$ while at the same time, the minimum of absolute values of the

elements of $\vec{S}$ is less than $0.5p_2\alpha_0$. Here, p_2 is a fractional parameter that restricts the search space. A typical value for this parameter is 0.1 (10 percent). The resulting output of the Sub-harmonic searcher is designated as ϕ_0, and represents the fundamental estimate prior to optional refinement processes.

The pre-refined fundamental estimate, ϕ_0, is input to the Fundamental Refiner. A pair of nx1 error vectors are formed as $\vec{E}_{-1} = \vec{F} - f_0(-1)\cdot\vec{1}$ and $\vec{E} = \vec{F} - \phi_0\vec{1}$. Here, $f_0(-1)$ is the refined fundamental estimate from the previous signal segment, and $\vec{F}$ is the ordered list vector from the output of the Frequency Sorter. Thus the z^{-1} block represents a unit segment delay. A scalar, $x = p_3 f_0(-1)$, is also calculated and is used to restrain the refinement process. Typical values for the fractional parameter p_3 is also 0.1 (10 percent). A comparison is made to determine if the minimum of the absolute values of the elements of $\vec{E}$ is less than the minimum of the absolute values of the elements of $\vec{E}_{-1}$ and is also less than *x*. If so, f_0 is the element of $\vec{F}$ associated with the minimum of the absolute values of the elements of $\vec{E}$. If both of these conditions are not met, then $f_0 = \phi_0$ (no refinement is made).

The output of the Fundamental Refiner, f_0, is input to the final optional step, the Harmonic Refiner. This step is identical in form to the Fundamental Refiner, and is repeated for all harmonic frequencies of interest. For example a harmonic is formed as the product $\phi_k = kf_0$, where the integer *k* is greater than 1. A pair of nx1 error vectors are formed as $\vec{E}_{-1} = \vec{F} - h_k(-1)\cdot\vec{1}$ and $\vec{E} = \vec{F} - \phi_k\vec{1}$. Here, $h_k(-1)$ is the refined harmonic estimate from the previous signal segment, and $\vec{F}$ is the ordered list vector from the output of the Frequency Sorter. A scalar, $x = p_3 h_h(-1)$, is also calculated and is used to restrain the refinement process. Typical values for the fractional parameter p_3 is also 0.1 (10 percent). A comparison is made to determine if the minimum of the absolute values of the elements of $\vec{E}$ is less than the minimum of the absolute values of the elements of $\vec{E}_{-1}$ and is also less than *x*. If so, h_k is the element of $\vec{F}$ associated with the minimum of the absolute values of the elements of $\vec{E}$. If both of these conditions are not met, then $h_k = \phi_k$ (no refinement is made).

Results

Shown in Table 1 are the performances results for the MP GHI process for the application of speech pitch estimation which in the present context refers to fundamental frequency estimation. The top half of Table 1 refers to results from male speech and the bottom half refers to female speech. The speech database that was used, is described in [3], and was downloaded from the author's website [8]. This database consists of a female and male speaker each speaking 50 English sentences sampled at 20 kHz. This database includes the recording of laryngeal frequency for each speech file in the database, which acts as the ground truth for fundamental estimation. This laryngeal data was created by placing a sensor on the subject's throat, while the speech data was collected. As previously described, a special property of speech is the fact that each segment of an utterance can be classified as either voiced or unvoiced. As implied, the voiced segments of the speech are segments that contain fundamental and harmonic frequency content, whereas unvoiced segments are either silence or fricatives and plosives. These latter segments contain either weak or no fundamentals and harmonics.

The ground truth given in this database consisted of the voiced frames time locations and their respective fundamental frequency. In this experiment, the start of each frame was slightly different than the ground truth time marks. Therefore, there was a need to adjust the ground truth to correspond to the beginning of each frame. The ground truth data fundamental frequency was interpolated for each sample. Once the data was interpolated, an interpolated fundamental frequency was determined by locating the beginning time of each frame. This allowed for one to adjust the frame size of the experiment and determine which frame size was optimal.

To properly take into account the voiced/unvoiced classification process, Table 1 and 2 includes the percentage of voiced segments in error (voiced classified as unvoiced) and the percentage of unvoiced segments in error (unvoiced classified as voice). This is necessary for a fair comparison because misclassifying voiced segments can affect important performance metrics, such as the absolute deviation mean and population standard deviation (p.s.d.). For example, a higher voiced in error percentage will cause the mean and p.s.d metrics to improve (become lower) as a result of eliminating weak voiced portions of the signal in the metric calculations. Likewise, higher unvoiced in error percentages will cause the metrics to degrade (become higher) as a result of including unvoiced segments in the calculations. This can be seen in Table 2, which shows the results of the MP GHI when using different threshold for the SVD voiced/unvoiced detector. The absolute deviation mean (adm) per utterance is calculated as

$$adm_j = \frac{1}{N_j}\sum_{i=1}^{N_j}\left| f_{ij} - \hat{f}_{ij} \right|, \qquad (1)$$

where f_{ij}, and $\hat{f}_{ij}$ is the actual and estimated fundamental frequency respectively for sample i, and N_j is the number of samples in utterance j. The population absolute deviation mean is expressed as

$$adm = \frac{1}{K}\sum_{j=1}^{K} adm_j . \qquad (2)$$

Where K is the number of speech signals in the population. The standard deviation (sd_j) per utterance is expressed as

Method	Unvoiced in error (%)	Voiced in error (%)	Gross high errors (%)	Gross Low errors (%)	Absolute deviation (Hz) mean	Absolute deviation (Hz) p.s.d.
SRPD	4.05	15.78	0.62	2.01	1.78	2.46
eSRPD	4.63	12.07	0.90	0.56	1.40	1.74
MP GHI	2.88	12.08	0.96	1.07	1.67	2.16
SRPD	2.35	12.16	0.39	5.56	4.14	5.51
eSRPD	2.73	9.13	0.43	0.23	4.17	5.13
MP GHI	0.89	9.06	1.09	0.22	3.02	3.91

Table 1: Fundamental estimation evaluation for male speech (top) and female speech (bottom).

MP GHI SVD Threshold	Unvoiced in error (%)	Voiced in error (%)	Gross high errors (%)	Gross low errors (%)	Absolute deviation (Hz) mean	Absolute deviation (Hz) p.s.d.
1.0	12.32	2.05	1.51	2.36	2.09	2.84
1.5	8.77	4.10	1.44	1.93	2.01	2.74
2.0	6.10	6.40	1.33	1.60	1.91	2.59
2.5	4.15	9.14	1.00	1.34	1.80	2.41
2.75	3.64	10.55	0.97	1.26	1.74	2.31
3.0	2.92	11.80	0.94	1.08	1.68	2.17
3.03	2.88	12.08	0.96	1.07	1.67	2.16
3.1	2.74	12.65	0.98	0.92	1.65	2.11
3.2	2.57	13.2	0.98	0.94	1.64	2.09
1.0	2.50	4.24	2.02	0.59	3.51	4.91
1.5	1.57	5.79	1.43	0.47	3.32	4.55
2.0	1.11	7.6	1.28	0.29	3.14	4.17
2.25	0.96	8.65	1.12	0.25	3.05	3.97
2.33	0.89	9.06	1.09	0.22	3.02	3.91
2.38	0.84	9.31	1.08	0.23	3.00	3.87
2.5	0.75	9.88	1.03	0.16	2.94	3.70

Table 2: Fundamental estimation evaluation for male speech (top) and female speech (bottom) with varying thresholds for the SVD voiced unvoiced detector

Gender	Unvoiced in error (%)	Voiced in error (%)	Gross high errors (%)	Gross Low errors (%)	Absolute deviation (Hz) mean	Absolute deviation (Hz) p.s.d.
male	0	0	3.57	2.50	2.15	2.91
female	0	0	9.67	1.05	3.79	5.46

Table 3: MP GHI fundamental estimation evaluation during perfect detection.

$$sd_j = \frac{1}{N_j - 1} \sum_{i=1}^{N_j} \left(\left| f_{ij} - \hat{f}_{ij} \right| - adm_j \right)^2 , \quad (3)$$

and the p.s.d as

$$p.s.d = \frac{1}{K} \sum_{j=1}^{K} sd_j . \quad (4)$$

As seen in Table 1, the voiced in error is the same for the eSPRD and the MP GHI. The performance is commensurate with prior super-resolution techniques.

The gross error, the metric used in [6], represents outliers of the estimated fundamental frequency. This metric does not measure the method's resolution. Gross high errors are the percentage of estimated fundamental frequencies that are 20% greater than the actual. Likewise, gross low errors

are the percentage of estimated fundamental frequencies that are 20% lower than the actual. As seen in Table 1, the MP GHI is commensurate with the eSRPD and SRPD methods in the gross error metric.

The last results that are worth examining are based on perfect detection results. Perfect detection is when the voice/unvoiced detector correctly classifies every frame. Unfortunately, [3] does not examine this test. As seen in Table 3, the results are very good. The male results are very impressive; whereas the female's gross high has increased significantly. This may be a concern, but without eSPRD results to compare with, this is left as an open question.

Conclusions

The MP GHI process is a novel approach to estimating, tracking and assessing the fundamental and harmonic frequencies of a speech signal. Pitch Estimation is a key step in accomplishing many speech processing applications, such as automatic speaker identification, speech data compression, pitch alteration, pitch prediction and natural sounding time compressions and expansions. The GHI could also be applied to other types of signals that are periodic in nature.

The GHI process is computationally efficient in that it consists of a small number of trivial matrix calculations and comparisons. In many signal processing applications, signal decomposition such as the MP technique may already be required. Therefore, the computational efficiency of the GHI process can be easily leveraged by these signal decomposition processes. Also, the GHI process can be implemented as a real-time process, in that an output fundamental and harmonic estimate can be generated for each signal segment, without the need to wait for future segments to be processed.

The GHI process is not confined to any particular super-resolution signal decomposition, but is particularly suited to the MP technique due to its ability to pre-condition the decomposition, based on decay or growth rates, frequencies, initial phases, and initial amplitudes. The GHI allows for super-resolution tracking of both the fundamental and harmonics. It does not require a fundamental component to actually be present in the original signal, since the fundamental candidates are generated based on the spacing between frequency components. A variety of outputs are provided including average fundamental, α_0, harmonic assessment count, c, refined fundamental and harmonic estimates, all of which can be more useful as a group as opposed to methods that simply yield the fundamental estimate itself. Tracking is enhanced as a result of incorporating the estimates of fundamental and harmonics from the previous signal segment. Finally, because the GHI process uses the super-resolution list, $\bar{F}$, for refinement, the output harmonic estimates, h_k, can be used to assess inharmonicity. Inharmonicity occurs when the harmonics are not exact integer multiples of the fundamental, and can be fairly common for example in musical instruments.

The GHI has many advantages as discussed. These advantages, in addition to the results that were observed, demonstrate that the MP GHI is a very attractive pitch estimator.

Reference

[1] Gold, N. Morgan, "Speech and Audio Signal Processing", John Wiley & Sons, Inc., 2000.

[2] Y. Medan, E. Yair, D. Chazan, "Super Resolution Pitch Determination of Speech Signals," IEEE Trans. On Signal Processing, ASSP-39(1):40-48, 1991.

[3] P. Bagshaw, S. Hiller, M. Jack, "Enhanced Pitch Tracking and the Processing of F0 Contours for Computer Aided Intonation Teaching," 3rd European Conference on Speech Communication and Technology, EUROSPEECH'93, Berlin, Germany, September 1993.

[4] de Cheveigné, Alain; Kawahara, Hideki "YIN, a fundamental frequency estimator for speech and music", Acoustical Society of America Journal, Volume 111, Issue 4, pp. 1917-1930 (2002).

[5] Haddad, D.M.;Sarkar, T.K.;Noga, A.J.; "Speech Compression Using the Matrix Pencil Technique"; IEEE 12th DSP Workshop; Sept. 2006;Page(s):218-221

[6] Bishnu S. Atal, Lawrence R. Rabiner; "A Pattern Recognition Approach to Voiced-Unvoiced-Silence Classification with Applications to Speech Recognition", IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume ASSP-24, No. 3, June 1976, pp. 201-212.

[7] Wai C. Chu, "Speech Coding Algorithms", New Jeresy: John Wiley & Sons, Inc., 2003

[8] http://www.cstr.ed.ac.uk/research/projects/fda/