Efficient Periodicity Extraction Based on Sine-Wave Representation and its Application to Pitch Determination of Speech Signals

Dan Chazan, Meir Tzur (Zibulski)[†]*, Ron Hoory and Gilad Cohen*

IBM Research
MATAM Haifa 31905, ISRAEL
chazan@il.ibm.com

Abstract

This paper presents a novel low-complexity method for extracting periodicity of signals based on their sine-wave representation. In this representation, the signal is modeled as a finite sum of sine-waves, with time-varying amplitudes, phases and frequencies. We describe how one can modify the familiar spectral-comb analysis method to obtain a guaranteed and effective procedure to find the fundamental-frequency which gives the best harmonic approximation of the signal spectrum. The search is efficiently carried out in the frequency domain. The procedure obtains a successive refinement of possible pitch values which are consistent with an increasing number of sine wave components. Other pitch intervals are pruned at an early stage of the search. The advantage of this algorithm is its high accuracy achieved at a relatively low complexity. We also briefly describe one possible application in the area of pitch determination of speech signals.

1. Introduction

In many research fields, such as audio and biomedical signal processing, seismic event analysis and communication, one often requires a proper harmonic representation of a given signal. The analysis usually begins with determining the dominant periodicity, or the *"fundamental-frequency"* (also referred to as the *"pitch"*). For example, classical methods for determining pitch in speech signals can be found in [1, 2]. Weighted least-squares or other optimization methods are then used (if necessary) to find the amplitude and phase of each harmonic component, giving the best overall approximation of the signal [3].

"Spectral-comb" based analysis methods have successfully been used for fundamental-frequency extraction in the frequency domain [4], along side other popular methods such as the cepstrum or the time-domain correlation. A challenging research goal is the development of computationally efficient implementations of these and other algorithms, while maintaining or even improving their accuracy. In this paper the specific case where a signal is represented by a finite sum of sine-waves is considered, which is referred to as the *"sine-wave representation"*[3]. A spectral comb like criterion which tries to measure the fit of the harmonics of the pitch with the frequencies of the sine wave components is used. The procedure tries to maximize that criterion.

This paper is organized as follows. In section 2 we describe the spectral-comb analysis method. Section 3 focuses on signals with sine-wave representation, and describes how the search can be modified to make it more computationally efficient. In section 4 we briefly discuss the application for pitch determination of speech signals. Concluding remarks are given in section 5.

[†] Meir Tzur (Zibulski) is now with BIT Innovation Technologies Inc., P.O.B 431 Tirat HaCarmel, 39032 ISRAEL.

2. Fundamental-frequency determination using the spectral comb method

Let $x(t),\ \ t \in \mathbb{R}$ be a real continuous-time signal, with a Fourier transform $X(f) = \int_{-\infty}^{+\infty} x(t)e^{-j2\pi ft}dt,\ \ f \in \mathbb{R}$. We wish to determine the fundamental-frequency F_0, such that a close estimation $\hat{x}(t)$ of the signal $x(t)$ can be found where $\hat{x}(t)$ is the sum of sine-waves at the fundamental-frequency F_0 and all its harmonics $(2F_0, 3F_0, ...)$.

In the spectral-comb determination method, we maximize a correlation between the signal spectrum and a frequency-domain comb function. For each candidate fundamental-frequency f, the comb function $c_f(\nu)$ is defined such that it receives its maxima values at the arguments $\nu = f, 2f, 3f, \ldots$, corresponding to the candidate pitch harmonics. A *"utility function"* $U(f)$ is then defined as follows:

$$U(f) = \int_0^{+\infty} c_f(\nu)|X(\nu)|d\nu \ \ , \ \ c_f(\nu) \geq 0. \qquad (1)$$

The frequency F_0 which maximizes the above utility function is selected as the fundamental-frequency of the signal $x(t)$:

$$F_0 = \arg\max_f\{U(f)\}, \qquad (2)$$

where the search is limited to some region $f \in [F_{min}, F_{max}]$. This maximization indicates that the signal's energy is concentrated around the frequency F_0 and its harmonics, giving rise to a good harmonic approximation of $x(t)$. Note that a global maximum selection strategy as in (2) may result in a frequency F_0 which is an integer divider of the true fundamental-frequency (i.e., an integer multiply of the true period). This ambiguity can be resolved if local maxima points of the utility function are also examined.

A simple method for defining the comb function $c_f(\nu)$ for each candidate fundamental-frequency f, is by scaling a fixed unit-period comb function $c(\nu)$ in the following way: $c_f(\nu) = c(\nu/f)$. A suitable comb function has the following properties:

1. $c(\nu + 1) = c(\nu) \quad \forall\, \nu$;
2. $0 \leq c(\nu) \leq 1 \quad \forall\, \nu$;
3. $c(0) = 1$;
4. $c(\nu) = c(-\nu) \quad \forall\, \nu$;
5. $c(\nu) = 0,\ r_1 \leq |\nu| \leq 0.5$, where $r_1 < 0.5$ is a positive parameter;

6. the function is continuous and non-increasing in the region $[0, r_1]$.

Figure 1 shows a single period of a comb function $c(\nu)$ which fulfills the conditions above. In this example, a piece-wise linear trapezoid function was chosen, characterized by two parameters r_1 and r_2 ($0 \leq r_2 < r_1 < 0.5$) . The benefits of using such a piece-wise linear function will be clear in the next section.

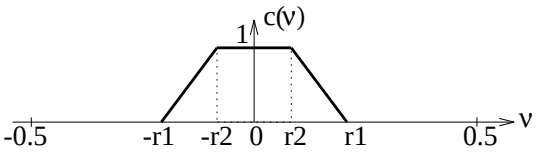

Figure 1: A unit-period trapezoid comb function, $c(\nu + 1) = c(\nu)\ \forall\, \nu$.

Some signals may have a very poor harmonic representation. This can be tested by comparing the maximum utility function value $U(F_0)$ to some predefined fixed threshold U_{THR} (this requires a prior normalization of the signal). If this value is lower than the threshold, the signal may be marked as not having a proper harmonic representation.

In the general case where no assumptions can be made on the analyzed signal $x(t)$, the search for the best fundamental frequency F_0 begins with a selection of an analysis frequency resolution $\triangle f$. Expression (1) is then evaluated for each of the frequency candidates on the discrete grid $f \in \{F_{min}, F_{min} + \triangle f, F_{min} + 2\triangle f, \ldots, F_{max}\}$, to find the frequency that maximizes the expression. Such an exhaustive search is very time consuming, and usually not feasible when the desired accuracy is high (i.e., small $\triangle f$).

It is important to notice that since the peaks of the comb function $c(\nu/f)$ become narrower as the fundamental-frequency candidate f decreases, the utility function $U(f)$ may have a very "chaotic" (noise-like) nature for low frequencies f. A finite resolution gird search may therefore completely miss a local maximum of the utility function in this region, which can result in the selection of a wrong fundamental-frequency.

3. Fundamental-frequency determination based on sine-wave representation

Assuming x(t) can be approximated by a finite sum of sine-waves, we present an efficient method for performing the fundamental frequency search without missing any local maxima of the utility function. Let the approximation of x(t) be of the form: $x(t) = \sum_{i=1}^{N} a_i \sin(2\pi f_i t + \phi_i)$, where $\{a_i, f_i, \phi_i\}_{i=1}^{N}$ are the N sine-wave amplitudes (positive and real), frequencies (in Hz) and phase offsets (in radians), respectively. No a-priori assumptions are made on the set $\{f_i\}$. The Fourier transform amplitude of this signal is $|X(f)| = \sum_{i=1}^{N} a_i \delta(f - f_i)$ for $f \geq 0$, and therefore the utility function (1) can be written as:

$$U(f) = \sum_{i=1}^{N} a_i c_f(f_i) = \sum_{i=1}^{N} a_i c(f_i/f)$$

We wish to efficiently find the maximum of this function at a given search range $f \in [F_{min}, F_{max}]$. We will assume the following:

1. Without loss of generality, we assume that the amplitudes are normalized to a unit sum, and are sorted in decreasing order, i.e., $a_1 \geq a_2 \geq \ldots \geq a_{N-1} \geq a_N$, $\sum_{i=1}^{N} a_i = 1$.
2. The comb function $c(\cdot)$ is selected to be the trapezoid function plotted in Figure 1, or a similar piece-wise linear function having the six properties specified in section 2.
3. A threshold $0 \leq U_{THR} < 1$ is chosen a-priori, such that candidate fundamental frequencies with utility function values below this threshold are discarded, since the corresponding harmonic structure does not match the signal spectrum well.

3.1. Search method

We define the contribution of a single sine-wave component (a_i, f_i) to the utility function as:

$$U_i(f) = a_i c(f_i/f)\ ,\quad i = 1, 2, .., N$$

then $U(f) = \sum_{i=1}^{N} U_i(f)$. Since $c(\cdot)$ is a piece-wise linear function, the values of $c(\cdot)$ at the first-derivative discontinuity points define all function values. These points will be referred to as *"break-points"*. Note that $U_i(f)$ is not a piece-wise linear function. We however approximate $U_i(f)$ to be linear at all regions which are not originally constant, and mark this approximation by $\hat{U}_i(f)$. The resulting approximated utility function is also similarly defined as $\hat{U}(f) = \sum_{i=1}^{N} \hat{U}_i(f)$. This was found to be a good approximation resulting in the reduced complexity search algorithm described below. An example is illustrated in Figure 2.

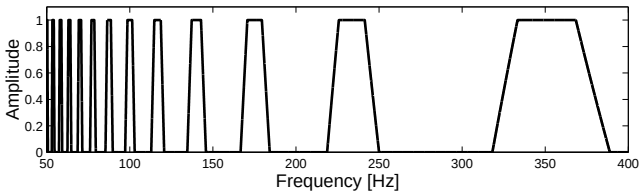

Figure 2: An example of a contribution $U_i(f)$ to the utility function based on a sine-wave component with $f_i = 700$ Hz (assuming $a_i = 1,\ r_1 = 0.2,\ r_2 = 0.1$). The region 50 – 400 Hz is plotted. The function is very close to a piece-wise linear function, with peaks in the vicinity of $f_i/2 = 350$ Hz, $f_i/3 = 233$ Hz, etc.

It can be seen that $U_i(f)$ receives its maxima values at the vicinity of the frequencies f_i/m, $m = 1, 2, 3, \ldots$, all with an equal weight of a_i. These are the fundamental-frequency candidates which produce a harmonic at f_i. By summing such weighted contributions from each sine-wave component of the signal, high utility function values will be observed for all common denominators of the sine-wave frequencies. The above approach is somewhat similar to a one-dimensional view of the familiar Hough transform used in image processing [5]. It should be noted that other weighting methods can be used, such as the log or square of the amplitudes or any other function of them, as long as the weights are normalized to a unit sum.

To construct the utility function, we now have to successively sum up the piece-wise linear functions $\hat{U}_i(f), i =$

$1, 2, \ldots, N$ one after the other, to get $\hat{U}(f)$. This is done efficiently since only a set of break-points needs to be updated and their partial sum values recalculated and stored.

The existence of a predefined threshold U_{THR} can also be used to speed up the search. To show this, we define the sum of the remaining sine-wave amplitudes at iteration i as:

$$R_i = \sum_{k=i+1}^{N} a_k.$$

and we also define the *"partial utility function"*, as the utility function computed for all previous sine-wave components:

$$PU_i(f) = \sum_{k=1}^{i} \hat{U}_k(f). \tag{3}$$

We then have for each i,

$$\begin{aligned} \hat{U}(f) &= PU_i(f) + \sum_{k=i+1}^{N} \hat{U}_k(f) \\ &\leq PU_i(f) + \sum_{k=i+1}^{N} a_i = PU_i(f) + R_i \quad \forall\, f\ . \end{aligned}$$

This means that there exist an easy-to-calculate upper bound on the utility function for each partial sum. Therefore, at iteration i, all frequency intervals where $PU_i(f) + R_i$ is less than the threshold U_{THR} can be deleted from the valid frequency-search region.

The following steps summarize the algorithm for efficiently constructing $\hat{U}(f)$:

1. Initialization: Set $i = 1$, $PU_0(f) = 0$, and set the initial search region to be $[F_{\min}, F_{\max}]$.
2. Calculate the values of the function $\hat{U}_i(f)$ at all its break-points within the current set of valid search intervals. Also, calculate the values of $\hat{U}_i(f)$ at the break-points of current partial utility function $PU_{i-1}(f)$.
3. Calculate the values of the partial utility function $PU_{i-1}(f)$ at the break-points of the function $\hat{U}_i(f)$.
4. Calculate, at all break-points, a new partial utility function $PU_i(f)$ by adding $\hat{U}_i(f)$ to the current partial utility function $PU_{i-1}(f)$.
5. Update the set of valid search intervals by deleting intervals where $PU_i(f) + R_i$ is less than the threshold U_{THR}.
6. End if the last sine-wave component was processed (i.e. $i = N$) or if the valid search intervals is an empty set. Otherwise, increase i and return to step 2.

Note that steps 2,3 and 5 may require simple linear interpolation to calculate the new utility function values between existing break-points. The above construction algorithm is computationally efficient because of the following reasons:

- The existence of a threshold U_{THR} is utilized during the partial summing, to reduce the valid frequency search regions and therefore speed-up the process. In some other traditional methods, a threshold is only utilized at the end to check the validity of the selected fundamental-frequency.
- The construction of the utility function is done through a small set of break-points, defining all function values. We do not use some arbitrary resolution $\triangle f$ (referring to section 2). This results in a more accurate search which is guaranteed to cover all possible local maxima points of the utility function.

The output of the above algorithm is a small set of frequency break-points and the values of the utility function at these frequencies. The frequency which receives the global maximum of the utility function can be selected as the fundamental-frequency of $x(t)$ according to (2). To avoid detection of multiple periods, an alternative selection process may be preferred. We extract the break-points which are local maxima of the utility function. These break-points are then examined and one of them is selected while giving preference to higher frequency break-points.

3.2. A simulated example

In Figure 3 the spectrum of an infinite duration signal composed of 8 sine-waves is shown. They are sorted and numbered according to decreasing amplitudes. To find the dominant fundamental-frequency, we selected a trapezoid comb function with parameters $r_1 = 0.2,\ r_2 = 0.1$ and a threshold $U_{THR} = 0.8$. Figures 4a-4c show the partial utility function $PU_i(f)$ of equation (3) for $i = 1, 3, 8$, in the predefined search region 100-350 Hz. The partial utility function break-points are plotted as dots connected by straight lines. The function is only calculated (and plotted) in the frequency regions where the upper bound exceeds the threshold. The highlighted portions of the horizontal frequency axis represent the updated search regions for the next partial summing. One can see that as more sine-waves components are used to construct the partial utility function, these regions quickly shrink in size and number. Figure 4c displays the final utility function, which includes only a small set of break-points. Clearly the emerging fundamental-frequency is in the close vicinity of 250 Hz or 125 Hz. To avoid double period detection, a local maximum break-point at 251 Hz is identified and selected as the fundamental frequency.

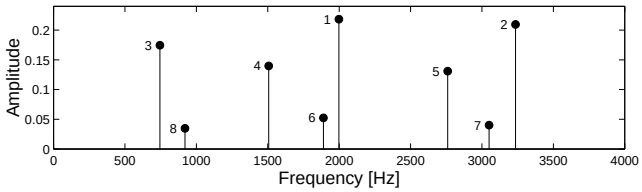

Figure 3: Spectrum of the input signal

4. Application to pitch determination of speech signals

The above method was used for the determination of pitch in an 8 Khz digitized speech signal. The Short Time Fourier Transform (STFT) of the signal was calculated using a 25 msec (200 sample) Hamming window and a 256 point Discrete Fourier Transform (DFT). To increase the frequency resolution, the DFT values were interpolated by a factor of two. A sine-wave representation of the signal was defined using a set $\{f_i, a_i\}$ of the local maxima of the DFT amplitude, as described in [3]. The pitch frequency was then calculated as described in section 3. The process repeated itself by advancing the analysis

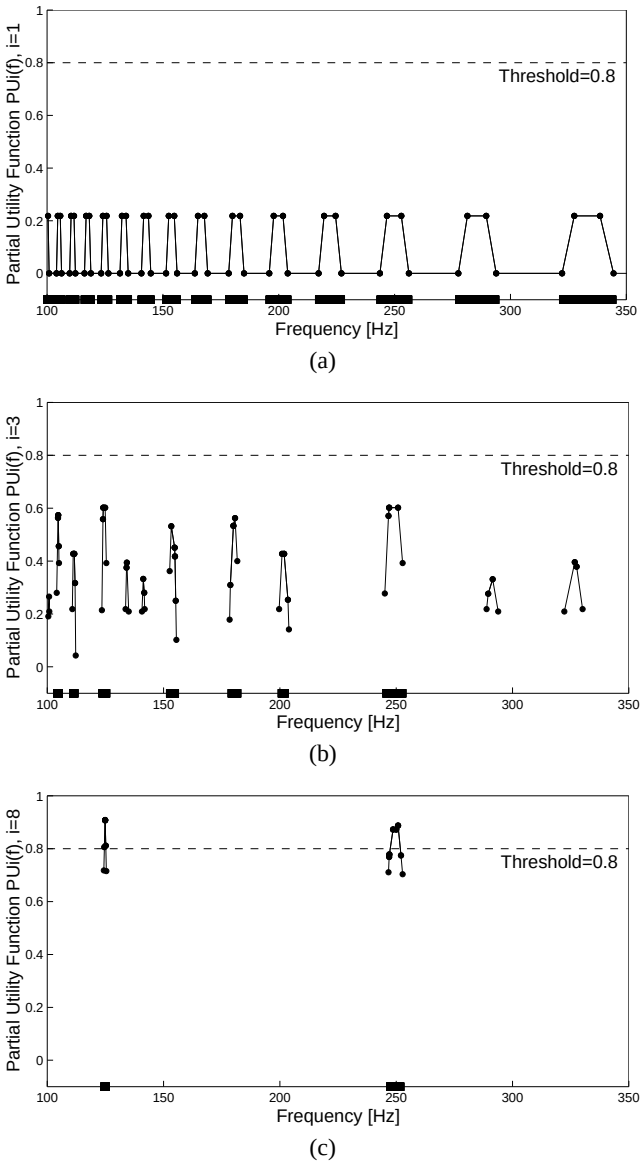

Figure 4: Partial utility functions for the spectrum in Figure 3. See text in section 3.2 for more details.

window 10 msec at a time to produce a pitch track. Since an analysis window of 25 msec is not long enough to find large pitch periods, a search scheme based on combining (in the DFT domain) two successive analysis windows was also utilized.

The selected threshold U_{THR} serves as a voiced/unvoiced classifier. It was found that in order to improve the classification, a threshold value selected in some range $[U_{THR}^{\min}, U_{THR}^{\max}]$ is preferred to a fixed threshold. The threshold value is recalculated at each frame, taking into account (among other things) the maximum value of the utility function at the previous frame (the "voicing degree").

Finally, as expected for voiced speech, we would like to give preference to the determination of a continuous pitch track. To do so, while we examine the remaining local maxima break points, we give additional weight to the selection of the local maximum which is in the close vicinity of the pitch frequency of the previous frame.

It was found that in the case of speech signals, the high order harmonics of the pitch have an important role in resolving pitch ambiguities, especially double or half pitch errors. The sine-wave representation therefore must cover spectral peaks extracted from the full signal bandwidth. In some of the other pitch determination algorithms, the signal is first low-passed in order to reduce the complexity of the search [1], which therefore restricts the accuracy of the solution. The pitch determination algorithm presented in this paper has the advantage of a full-bandwidth analysis while maintaining a low complexity search.

The pitch determination algorithm was successfully integrated into a low bit-rate speech coder, for usage in distributed speech recognition systems. Strict demands for a low complexity, highly accurate pitch determination in the encoder were met. For more details, refer to [6, 7].

5. Conclusions

This paper presented the spectral comb analysis method for extracting the fundamental-frequency, which requires an exhaustive search. It was shown that when a sine-wave representation of the signal is given, a computationally efficient search can be performed. The existence of a periodicity matching threshold, such as the voiced/unvoiced classifier in speech signals, was also used for a quick reduction of the frequency search intervals. This method can serve as the core of an efficient, real-time fundamental-frequency determination algorithm, where the parameters, thresholds and detailed search logic are fine-tuned according to the specific application.

6. References

[1] Wolfgang Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, 1983.

[2] L. R. Rabinter et. al., "A comparative performance study of several pitch detection algorithms," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, no. 5, pp. 399–418, October 1976.

[3] Robert J. McAulay and Thomas F. Quatiery, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds., chapter 4, pp. 121–170. Elsevier, 1995.

[4] Philippe Martin, "Comparison of pitch detection by cepstrum and spectral comb analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1982, pp. 180–183.

[5] R. C. Gonzalez and P. Wintz, *Digital Image Processing*, Addison-Wesley Publishing, 1987.

[6] Dan Chazan, Ron Hoory, Gilad Cohen, and Meir Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2000.

[7] Dan Chazan, Ron Hoory, Gilad Cohen, and Meir Zibulski, "Low bit rate speech compression for playback in speech recognition systems," in *European Signal Processing Conference (EUSIPCO)*, 2000.