A COMPUTATIONALLY EFFICIENT REFINEMENT OF THE FUNDAMENTAL FREQUENCY ESTIMATE FOR THE ADAPTIVE HARMONIC MODEL

Veronica Morfi, Gilles Degottex, Athanasios Mouchtaris

Multimedia Informatics Lab, Computer Science Department, University of Crete
Institute of Computer Science, Foundation for Research and Technology Hellas
Heraklion, Greece
morfi@csd.uoc.gr, degottex@csd.uoc.gr, mouchtar@ics.forth.gr

ABSTRACT

The full-band Adaptive Harmonic Model (aHM) can be used by the Adaptive Iterative Refinement (AIR) algorithm to accurately model the perceived characteristics of a speech recording. However, the Least Squares (LS) solution used in the current aHM-AIR makes the f_0 refinement in AIR time consuming, limiting the use of this algorithm for large databases. In this paper, a Peak Picking (PP) approach is suggested as a substitution to the LS solution. In order to integrate the adaptivity scheme of aHM in the PP approach, an adaptive Discrete Fourier Transform (aDFT) is also suggested in this paper, whose frequency basis can fully follow the frequency variations of the f_0 curve. Evaluations have shown an average time reduction of 5.5 times compared to the LS solution approach, while the quality of the re-synthesis is preserved compared to the original aHM-AIR.

Index Terms— Fundamental frequency, speech analysis/synthesis, peak picking, Harmonic Models

1. INTRODUCTION

Harmonic Models (HM) are used for speech coding [1], concatenative speech synthesis [2], speech modeling [3], voice transformation [4]. After the analysis step, these models provide a set of sinusoidal parameters, such as frequencies, amplitudes and phases, which can later be used to build higher-level representations (e.g. spectral envelopes). For synthesis purpose, the perceived quality is crucial. Additionally, current speech synthesis technologies often need to process large recording databases. Computationally efficient algorithms are, thus, preferred.

In most analysis approaches, the fundamental frequency f_0 is assumed to be constant, in an analysis window, whereas f_0 actually varies throughout the speech signal. This mismatch between the frequency basis of the model (e.g. DFT or HM) having constant frequencies and the modulated harmonic structure of the speech signal causes the harmonic structure in a spectrogram to be blurred, and can make the harmonic structure difficult to track.

To alleviate this issue, the Adaptive Quasi-Harmonic Model (aQHM) has been suggested [5] in which the frequency basis is adapted to the f_0 curve estimated from the speech signal. Thus, the adapted frequency basis can follow any non-linear variations of the frequency basis of the underlying signal. However, a proper estimation can be obtained only if the input components of the frequency basis built from the f_0 curve are in a reasonable interval around the actual values of the speech signal [5]. Any error on the f_0 curve being multiplied by the harmonic number, the tracking of the harmonic structure in mid and high frequencies can be easily compromised.

In [6], the Adaptive Iterative Refinement (AIR) was proposed to address this problem by refining the f_0 curve, leading to accurate parameter estimates of the aHM model, a simpler version of aQHM. The AIR algorithm begins with the lowest frequencies, where the f_0 error is assumed to be small, and iteratively increases the number of harmonics up to the Nyquist frequency by successive refinement of the f_0 curve at each iteration step. In order to compute the sinusoidal parameters of the harmonic model, the Least Squares (LS) solution was used. However, even though aHM-AIR allows a robust estimation of the harmonic components through the refinement of the f_0 curve, the computational load of the LS solution does not allow processing of large databases in a convenient time duration.

This paper addresses this problem of computational efficiency by replacing the LS solution of the aHM-AIR method with a Peak Picking (PP) approach [7]. In order to integrate the adaptivity scheme of aHM to the PP approach, the Adaptive Discrete Fourier Transform (aDFT) is also suggested in this paper. In contrast to the standard DFT, the frequency

This work is co-funded in part by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program ”Education and Lifelong Learning” of the National Strategic Reference Framework-(NSRF), Research Funding Program: THALES, Project ”MUSINET”.

G. Degottex was funded by the Swiss National Science Foundation (SNSF) (grants PBSKP2_134325, PBSKP2_140021) and the Foundation for Research and Technology Hellas (FORTH).

basis of the aDFT is fully adapted to the input f_0 curve of the signal. Therefore, while keeping the basis structure of the aHM-AIR, this paper describes how to replace the LS solution by the PP approach based on the aDFT. In the evaluation, the Fan Chirp Transform (FChT), which uses a linear chirp related frequency basis adapted to f_0, will also be compared to the aDFT.

In the rest of this paper the aDFT is first presented (Sec. 2), followed by the method section (Sec. 3), where the modifications of the AIR algorithm are explained. In the evaluation section (Sec. 4), the LS solution will be compared to the suggested PP approach, using aDFT or FChT, in terms of SRER. Assessment of the perceived quality of the reconstructed signal will also be carried out using PESQ [8].

2. ADAPTIVE DISCRETE FOURIER TRANSFORM (ADFT)

The theoretical novelty of the suggested approach lies in the Adaptive Discrete Fourier Transform (aDFT). In order to emphasize the importance of adaptivity for the AIR algorithm and describe the aDFT, the difference between DFT and FChT is first described in this section.

The DFT of a windowed signal $x(n)$ is defined as

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j2\pi kn/N} \tag{1}$$

where N refers to the length of the windowed signal $x(n)$ and the DFT length, $k = 0, .., N-1$. In the Discrete Fourier Transform (DFT), the frequency basis is constant inside the analysis window which implies the assumption of stationarity of the analyzed signal. For speech signals this assumption is valid only when the variations of the fundamental frequency, f_0, are negligible compared to the stationary basis of the DFT. However, the variations of the harmonics are proportional to those of f_0 and the harmonic number. The non-stationarity of the voiced signal is therefore highly increased as frequencies increase, making the validity of the stationarity hypothesis questionable for mid and high frequencies of speech signals. To alleviate this issue, the Fan Chirp Transform (FChT) has been suggested [9]. This method uses a chirp related frequency basis (i.e. linear frequency trajectories) whose slope is adjusted to the average slope of the f_0 curve in the analysis window. The FChT of a signal $x(n)$ is defined as

$$X(k, a) = \sum_{n=0}^{N-1} x(n) \xi^*(n, k, a) \tag{2}$$

where N stands for the length of the windowed signal $x(n)$ and the FChT length, $k = 0, .., N-1$, * denotes the complex conjugate and $\xi(n, k, a)$ is the basis of the FChT:

$$\xi(n, k, a) = \sqrt{|\phi_a'(n)|} e^{-j2\pi k \phi_a(n)}, \tag{3}$$

where $\phi_a(n)$ rules the time dependence of the frequency basis exponent

$$\phi_a(n) = \Big(1 + \frac{1}{2} a \Big(n - \frac{N}{2}\Big)\Big)\Big(n + \frac{N}{2}\Big) - \frac{N}{2} \tag{4}$$

whose first frequency basis component is

$$\phi_a'(n) = (1 + an) \tag{5}$$

where the parameter a is the chirp rate, the f_0 slope. Using the FChT, a regularity in the frequency content can be observed in mid and high frequencies [9]. Even though the FChT basis better fits the frequency modulations of the speech signal than the DFT, the frequency basis is constrained to linear trajectories.

In order to better follow the non-linear f_0 variations the Adaptive Discrete Fourier Transform (aDFT), which is based on the adaptivity scheme of aQHM [5], is suggested in this paper. The aDFT, is similar to DFT but uses a frequency basis that follows the variations of the f_0 curve. In this paper, we define the aDFT of a signal $x(n)$ as

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-jk\phi_0(n)} \tag{6}$$

where $k = 0, .., N-1$, N is the aDFT length and $\phi_0(n)$ is the fundamental phase of the frequency basis, which is a real function defined by the integral of the fundamental frequency $f_0(t)$.

$$\phi_0(t) = \frac{2\pi}{f_s} \int_0^t f_0(\tau)\, d\tau \tag{7}$$

According to the adaptivity scheme, $f_0(t)$ is obtained by linear interpolation of $f_0^{t_i}$ values at analysis instants t_i, where the time zero corresponds to the window center. More details can be found in the following sections.

3. PROPOSED METHOD

In this section the suggested method to estimate the parameters of aHM will be described, namely the Adaptive Iterative Refinement (AIR) [6] method which uses the Peak Picking (PP) approach [7] on the Adaptive Discrete Fourier Transform (aDFT). In order to use the PP approach for extracting harmonic peaks, a method based on [10] was used. The main structure of the original AIR algorithm has been described in detail in [6]. The modifications brought by the new algorithm are explained below.

A parametrization of the speech signal at time instants t_i takes place during analysis. A sequence of anchor instants t_i is first created using a rough estimate of the f_0 curve, with distance of one period between each instant t_i. A Blackman window of 3 local pitch periods is then applied to the speech signal around each t_i. The aDFT length (N) is defined as twice the window's length. This means that voices with high

f_0 (e.g. female voices) will need a shorter aDFT length than voices with lower pitch (e.g. male voices).

The basic idea of the AIR algorithm is to begin with a low harmonic level ($K_i = 8$) for each instant i. For each iteration step, the corrected $f_0'^{t_i}$ is estimated from the PP-aDFT at each time instant t_i. After each iteration, the $f_0(t)$ curve is replaced by the new $f_0'(t)$ curve. In our implementation each corrected $f_0'^{t_i}$ is estimated using the median of the harmonic peaks divided by their harmonic number (i.e. $f_k^{t_i}/k$). Before the next iteration begins, K_i is updated, as in the original AIR algorithm [6]. Eventually, this process is repeated for all frames until the Nyquist frequency is reached for all of them. Algorithm 1 describes this specific analysis procedure.

Algorithm 1 AIR for aHM using Peak Picking

```
Create a sequence of times $t_i$ according to $f_0(t)$
Initiate each for $f_0^{t_i}$ = $f_0(t_i)$
Initiate each $K_i$ = 8
while ∃ i such as $f_0^{t_i} K_i < f_s/2$
    for each instant $t_c$
        Create a segment of 3 periods around $t_c$ using $f_0(t_c)$
        Compute the aDFT of the segment
        Pick the harmonic peaks $f_k$ up to $K_c$ in the aDFT
        Correct $f_0'^{t_c}$ = median($f_k/k$)
        Compute $f_{corr} = f_0'^{t_c}$ - $f_0^{t_c}$
        if $f_0^{t_c} K_c < f_s$
            Update $K_c = \lfloor 0.5 N_w / |f_{corr}| \rfloor$
        end if
    end for
    Set $f_0^{t_i} = f_0'^{t_i}$
end while
```

Considering that the reason for replacing the LS solution with aDFT and PP was to improve the speed of the aHM-AIR method, a few more techniques are suggested in this direction. These techniques are explained in the following subsections.

3.1. Reduction of Computational Load

The number of harmonics (K_i) at the first iteration starts from a low value ($K_i = 8$). Since only the first 8 harmonics will be used in the PP, there is only a need to compute the aDFT bins containing these harmonics. Thus, aDFT only computes up to the current harmonic level K_i instead of the whole aDFT length (N), hence, avoiding computation of bins above the current harmonic level.

The second improvement regarding the method's complexity was based on the observation that the f_0 values computed in each iteration eventually converged for each window. Thus, the frequency basis is almost the same for the low frequencies as the harmonic levels, K_i, used in its computation increase. The aDFT in low frequencies is, thus, very similar between iterations and the correction of the frequency basis becomes more and more negligible for low frequencies. Hence, it can be assumed that below a certain extent of f_0 correction, the peaks estimated during the previous iteration would remain almost the same, and they can be kept the same for the following iterations. In order to implement this in the proposed method, a threshold, B_i in the frequency bins of the aDFT, below which the peaks are kept the same, needs to be set:

$$B_i = \frac{\text{tol} \cdot f_0^{t_i} \cdot N_i}{f_{corr}^{t_i} \cdot f_s} \tag{8}$$

where $f_0^{t_i}$ is the frequency of the time instant t_i, N_i is the aDFT length for the current frame, $f_{corr}^{t_i}$ is the correction of $f_0^{t_i}$ computed in the previous iteration and f_s is the signal's sampling frequency. A tolerance factor of 10% of the f_0 (i.e. $\text{tol} = 0.1 \cdot f_0^{t_i}$) was chosen, which provided an important reduction of the computational time without altering drastically the results.

3.2. Unvoiced Segments

In unvoiced segments, no harmonic structure exists, hence using a harmonic model in those parts is questionable. However, it has been shown that it is possible to use aHM for both voiced and unvoiced segments thus providing a uniform representation across time which does not need any voicing decision [6]. Using the suggested PP approach in unvoiced segments, substantial deviations from the input f_0 curve were often observed in the newly computed f_0 curve. This is obviously caused by the lack of harmonic structure in addition to the low harmonic level used during the first steps (e.g. $K_i = 8$), which prevent convergence of the f_0 values. In this paper, we suggest to discard any substantial f_0 deviations and force the harmonic level to increase, before the next iteration step. In the current implementation, a deviation threshold of 8% from f_0 is used to decide whether or not each f_0' will be discarded. In such a case, we suggest to force the harmonic level by the following way $K_i' = |f_0' - f_0| \cdot K_i$.

4. EVALUATION

For the following evaluations, three versions of aHM-AIR were compared, each one using a different method to compute the sinusoidal parameters of the harmonic model. These three methods, as previously mentioned, are: the LS solution; PP using FChT (PP-FChT); and PP using aDFT (PP-aDFT). These tests were applied on a small database of 32 utterances (16 female and 16 male in 16 different languages, between 2s and 4s length, with sampling frequency varying between 16kHz and 44kHz). We assume that the different phonemes and origins of these languages provide a sufficient voice variability for supporting the validity of the results. The samples can be found on the following web-page with their corresponding re-synthesis using the three methods:
`http://www.csd.uoc.gr/∼morfi/`

For FChT, the chirp-factor a for each time instant t_i, was estimated based on linear interpolation of the four neighbouring f_0 values, around t_i.

4.1. Computational Time

For each method, the running time has been measured for each recording and the time reduction ratios, with respect to the LS-based method, were averaged (Table 1). On average, by using PP-FChT, aHM-AIR becomes 5 times faster, while, with PP-aDFT, it becomes 5.5 times faster, as shown in Table 1. Amongst the used sentences, the maximum ratio of time improvement observed was 11.8 for the PP-aDFT over the LS solution.

Methods	Male Voices	Female Voices	All
LS/PP-FChT	4.5	5.6	5.0
LS/PP-aDFT	4.8	6.2	5.5

Table 1. Average Time Reduction Ratios

4.2. Signal-to-Reconstruction Error Ratio (SRER)

To evaluate the global reconstruction accuracy of the suggested methods, the segmental Signal-to-Reconstruction-Error Ratio (SRER) between the recorded utterances and their models was computed. A sliding window of 10ms with 50% overlap was used. In order to evaluate only the impact of the AIR algorithm, which basically refines the fundamental frequency, the LS solution was kept to estimate the final sinusoidal parameters used for synthesis, as previously used in [6]. The SRER was computed using the full-band of the recordings and its distribution of the voiced and unvoiced segments is shown on the top and bottom plot of Fig.1, respectively. The sole 32 sentences were sufficient to obtain more than 10000 values for each distribution.

It can be observed that the distributions of all three methods are very similar. Especially, the comparison between the LS solution and the aDFT methods shows that these two methods have very close results. A smaller SRER for the FChT method is noted in the voiced segments. This is caused by the fact that the frequency basis in FChT is constrained to linear trajectories and cannot fully adapt to the input f_0 curve, in contrast to the LS and the aDFT based solutions.

4.3. Objective Quality Assessment

It is expected that, since the results of SRER for all three methods are very similar, the re-synthesized signals would sound quite the same. In order to verify this, the PESQ method [8] is used to assess the perceived quality of the reconstructed signals compared to the originals. Table 2 presents the PESQ scores for the three methods using the same database as in the previous tests. Due to the fact that the sampling frequency for the signals in the database varied from 16kHz to 44kHz, a re-sampling of all signals to 16kHz was performed in order for the PESQ measurement to be used. The results show that there is no clear difference in quality, which can be confirmed by informal listening:
`http://www.csd.uoc.gr/~morfi/`

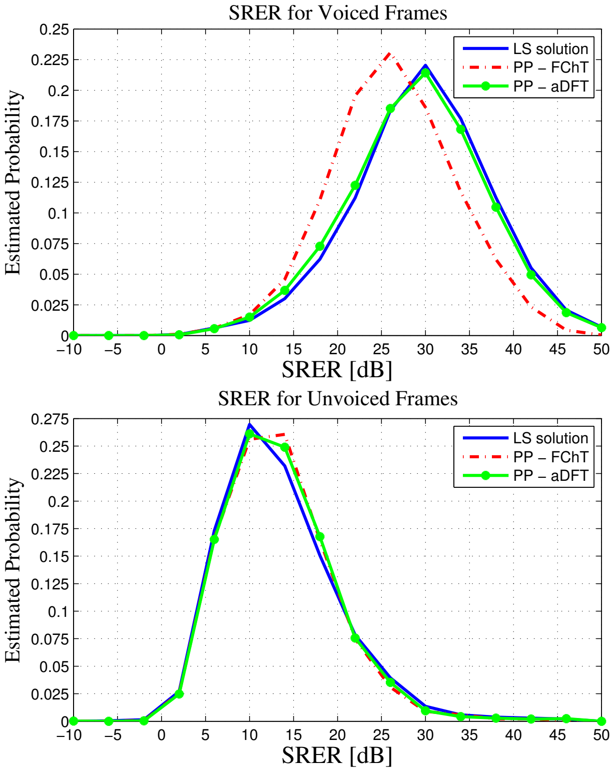

Fig. 1. Estimation of the full-band SRER distributions for voiced and unvoiced frames.

PESQ Ratings (up to 4.5)	
LS solution	4.13
PP - aDFT	4.12
PP - FChT	4.04

Table 2. PESQ scores assessing the overall quality of the three re-synthesized signals of the methods and the originals.

5. CONCLUSIONS

Taking advantage of the good perceived quality provided by aHM-AIR, a Peak Picking approach was suggested to replace the LS solution for the f_0 refinement, in order to reduce the computational time of the AIR algorithm. Two different transforms were used for Peak Picking, the existing FChT and a new aDFT, whose frequency basis fully adapts to the input f_0 curve and which is presented in this paper. Evaluations have shown that by performing this substitution, the computational load of the AIR algorithm decreases, in average, by a factor of 5.5. Moreover, according to objective assessment, this replacement did not degrade the perceived quality of the re-synthesized signal. Therefore, the speed and quality of the aHM-AIR method using Peak Picking make it ideal for processing large databases during a convenient time duration.

6. REFERENCES

[1] L. Almeida and J. Tribolet, “Harmonic coding: A low bit-rate, good-quality speech coding technique,” *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP ’82*, 1982.

[2] Y. Stylianou, “Applying the harmonic plus noise model in concatenative speech synthesis,” *IEEE Trans. on Speech and Audio Proc.*, vol. 9, pp. 21–29, 2001.

[3] ——, *Modeling Speech based on Harmonic Plus Noise Models*. Springer Berlin / Heidelberg, 2005, pp. 244–260.

[4] G. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, “Time-scale modifications based on a full-band adaptive harmonic model,” in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, August 2013.

[5] Y. Pantazis, G. Tzedakis, O. Rosec, and Y. Stylianou, “Analysis/Synthesis of Speech based on an Adaptive Quasi-Harmonic plus Noise Model,” in *Proc. IEEE ICASSP*, Dallas, Texas, USA, Mar 2010.

[6] G. Degottex and Y. Stylianou, “Analysis and synthesis of speech using an adaptive full-band harmonic model,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085–2095, 2013.

[7] R. J. McAulay and T. F. Quatieri, “Speech Analysis/Synthesis based on a Sinusoidal Representation,” *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 34, pp. 744–754, 1986.

[8] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP ’01). 2001 IEEE International Conference on*, vol. 2, 2001.

[9] M. Kepesi and L. Weruaga, “Adaptive chirp-based time-frequency analysis of speech,” *Speech Communication*, vol. 48, pp. 474–492, 2006.

[10] D. B. Paul, “The spectral envelope estimation vocoder,” *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 29, pp. 786–794, 1981.