ROBUST FULL-BAND ADAPTIVE SINUSOIDAL ANALYSIS AND SYNTHESIS OF SPEECH

George P. Kafentzis1,2, Olivier Rosec2, Yannis Stylianou3

1Orange Labs, TECH/ACTS/MAS, Lannion, France
2Voxxygen S.A., Pole Phoenix, Pleumeur-Bodou, France
3Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece

gorge.kafentzis@orange.com, olivier.rosec@voxygen.fr, yannis@csd.uoc.gr

ABSTRACT

Recent advances in speech analysis have shown that voiced speech can be very well represented using quasi-harmonic frequency tracks and local parameter adaptivity to the underlying signal. In this paper, we revisit the quasi-harmonicity approach through the extended adaptive quasi-harmonic frequency tracks, and show that the application of a continuous f0 estimation method plus an adaptivity scheme can yield high resolution quasi-harmonic analysis and perceptually indistinguishable resynthesized speech. This method assumes an initial harmonic model which successively converges to quasi-harmonicity. Formal listening tests showed that eaQHM is robust against f0 estimation artefacts and can provide a higher quality in resynthesizing speech, compared to a recently developed model, called the adaptive Harmonic Model (aHM), and the standard Sinusoidal Model (SM).

Index Terms— Extended adaptive quasi-harmonic model, Speech modelling, Speech analysis, Sinusoidal modelling

1. INTRODUCTION

Sinusoidal analysis of speech has been in timestep for the last twenty years and have been proved to work well in many applications such as speech coding [1, 2, 3], speech analysis and synthesis [4, 5, 6, 7, 8], speech enhancement [9, 10, 11, 12], and speech modifications and transformations [5, 7, 13, 14, 15].

In that context, many different approaches have been suggested over the last thirty years, in order to provide high-quality, artefree, flexible and compact representations of the speech signal. After the milestone work of McAulay and Quatieri on the Sinusoidal Model (SM) [4], where speech is represented as a sum of stationary sinusoids on a frame-by-frame manner, people in the speech community have intensively worked on improving models that can represent speech more accurately than in SM, thus attaining high levels of flexibility and naturalness. Hybrid approaches have become a mainstream in speech representation due to the convenience in handling different types of speech components [5, 16, 17, 18, 19, 20]. Some of the most prominent representatives of these efforts that employ a sinusoidal component include the following: Stylianou [17] suggested to decompose speech into a deterministic and a stochastic component, with the former modelling the quasi-periodic phenomena of speech using harmonically related sinusoids, and the latter modelling its non-periodic characteristics, such as friction noise, using modulated Gaussian noise. It should be noted that voiced speech is considered to have both components, which are separated by a so-called maximum voiced frequency. Serra [5] suggested a similar model where the sinusoids are no longer constrained to be harmonic. George and Smith [7] presented a speech analysis/synthesis system based on the combination of an overlap-add sinusoidal model with an analysis-by-synthesis technique to determine the model parameters.

Agiomyrgiannakis and Rosec [20] discussed the use of a harmonic plus noise representation to model the residual of an LF-based analysis.

More recently, Pantazis [21] showed that by projecting the analyzing signal on a set of time-varying exponential basis functions inside the analysis window and by using a frequency correction mechanism on the frequency tracks, a high quality, quasi-harmonic representation of voiced speech can be obtained [22]. This model is termed as the adaptive Quasi-Harmonic Model - aQHM and it has been applied on a hybrid speech analysis-synthesis system, which is dubbed the adaptive Quasi-Harmonic plus Noise Model - aQHNM [8]. Kafentzis et al [23] showed that including amplitude adaptation can yield higher reconstruction rates for voiced speech, thus obtaining the extended adaptive Quasi-Harmonic Model - eaQHM. This adaptive scheme inspired Degottex and Stylianou [24, 25] to suggest the full-band adaptive Harmonic Model - aHM, which uses the frequency correction mechanism of aQHM to iteratively refine the fundamental frequency by a dedicated algorithm called Adaptive Iterative Refinement - AIR, and finally represents speech as a sum of harmonics up to the Nyquist frequency. Listening tests have shown that AIR-aHM provide almost perfect perceptual quality, provided that the estimated f0 is free of artefacts. Since all these models exploit the local adaptivity of the model on the analyzed signal, they are jointly called the adaptive Sinusoidal Models - aSMs.

Although hybrid models have been proved to provide flexibility in manipulation and resynthesis of speech, in this paper a full band quasi-harmonic analysis of speech is described, using the eaQHM. There are several reasons for suggesting such an approach: first, as it is described in [25], a maximum voiced frequency is not necessary from a speech production point of view in the analysis of voiced speech, thus giving rise to a full-band model for voiced speech. Moreover, in [23], the eaQHM is shown to provide highly accurate reconstruction of voiced speech, higher than the aQHM. In addition, Kafentzis et al [26] proposed the use of quasi-harmonics and a local adaptivity to accurately represent voiced and voiceless consonants. Also, the perceptual quality of consonants in AIR-aHM is high, thus showing that local adaptivity and harmonicity can perceptually represent all parts of speech. However, it should be noted that AIR-aHM is sensitive to the f0 estimation, as it is the case for most harmonic models.

In this paper, we extend the work presented in [23] by taking into account the latest developments in aSM and aHM and suggest a high-quality, full-band, and free of voicing decision analysis-synthesis system of speech based on the eaQHM. The proposed system is shown to be robust in f0 artefacts, by testing its perfor-
mance using two well-known pitch estimators, called SWIPE [27] and YIN [28]. The eaQHM system assumes an initial harmonic frequency structure that subsequently converges in quasi-harmonicity, thus allowing frequencies to deviate from their harmonic grid by applying the frequency correction mechanism of eaQHM. Formal listening tests and objective measures on the resynthesized speech are utilized, and show that eaQHM outperforms by far the standard SM, whereas it is superior to the recently developed AIR-aHM, and compares it to the competition. Section 4 discusses the eaQHM analysis and synthesis framework. Section 3 describes the eaQHM analysis and synthesis framework. Section 4 presents the framework for objective and subjective evaluation of eaQHM and compares it to the competition. Section 4 discusses the results and Section 5 concludes the paper.

2. DESCRIPTION OF eaQHM-BASED ANALYSIS/SYNTHESIS SYSTEM

The full-band signal is described as an AM-FM decomposition

\[ d(t) = \sum_{k=-K}^{K} A_k(t) e^{i\phi_k(t)} \]  

where \( A_k(t) \) is the instantaneous amplitude and \( \phi_k(t) \) is the instantaneous phase of the \( k^{th} \) component, respectively. The instantaneous phase term is given by

\[ \phi_k(t) = \phi_k(t_i) + \int_{t_i}^{t} \frac{2\pi}{f_s} f_k(u) du \]  

where \( \phi_k(t_i) \) is the instantaneous phase value at the analysis time instant \( t_i \), \( f_s \) is the sampling frequency, and \( f_k(t) \) is the instantaneous frequency of the \( k^{th} \) component.

2.1. Analysis

Having an initial and continuous \( f_0 \) estimation for all frames (usually separated as voiced and unvoiced), noted by \( \hat{f}_0 \), the next step is to assume a full-band harmonicity to obtain a first estimate of the instantaneous amplitudes of all the harmonics. Using a Blackman analysis window \( u(t) \) centered at \( t_i \) and with support in \( [t_i-T, t_i+T] \), where \( 2T \) is of 3 local pitch periods length, a frame of the analyzed speech is initially modeled using a simple Harmonic Model as:

\[ d(t) = \left( \sum_{k=-L}^{L} a_k e^{i\hat{\phi}_k t} \right) w(t) \]

where \( a_k \) is the complex amplitude of the \( k^{th} \) harmonic, \( \hat{f}_k = k \hat{f}_0 \) are the analysis frequencies, and \( L \) is the number of harmonics that span the whole spectrum up to Nyquist frequency. The estimation of the model parameters is obtained via Least Squares, as described in [17]. As opposed to [8], where the initial \( f_0 \) estimation is refined using an iterative QHM, in our work no \( f_0 \) refinement is necessary, thus reducing the overall complexity of the algorithm, and a simple amplitude estimation for each component is performed. As a final step, the overall signal can be synthesized by interpolating the \( |a_k| \) and \( \hat{\phi}_k \) values over successive analysis time instants \( t_i \), thus obtaining

\[ \hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{i\hat{\phi}_k(t)} \]

and

\[ \hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} k \hat{f}_0(u) du, \quad \hat{\phi}_k(t_i) = \angle a_k(t_i) \]

2.2. Adaptation

The above model is still harmonic and stationary within an analysis frame. Therefore, in order to converge to quasi-harmonicity and to confront the stationarity issue, the projection of the signal onto a set of time-varying basis functions is suggested in [23], by using the parameters \( a_k \) and \( b_k \) of the Quasi-Harmonic Model (QHM) [29]. This yields the eaQHM model:

\[ d(t) = \left( \sum_{k=-L}^{L} (a_k + tb_k) \left( \hat{A}_k(t) e^{i\hat{\phi}_k(t)} \right) \right) w(t) \]

where

\[ \hat{A}_k(t) = \frac{\hat{A}_k(t + t_i)}{A_k(t_i)} \]

and \( \hat{\phi}_k(t) \) as in Eq. (6). In this model, \( a_k, b_k \) are the complex amplitude and the complex slope of the \( k^{th} \) component, and \( \hat{A}_k(t), \hat{f}_k(t) \), \( \hat{\phi}_k(t) \) are estimates of the instantaneous amplitude, frequency, and phase of the \( k^{th} \) component, respectively, from the previous analysis step. The \( a_k, b_k \) parameters are obtained via Least Squares [23]. It is apparent that the basis functions where the signal is projected are time-varying. The adaptation is completed by using the frequency correction mechanism first introduced in [29], and states that an estimate of the mismatch between the actual \( k^{th} \) frequency and the estimated one, termed \( \eta_k = f_k - \hat{f}_k \), is given by

\[ \hat{\eta}_k = \frac{\hat{f}_k}{2\pi} \Re\{a_k\} \Im\{b_k\} - \Re\{a_k\} \Im\{b_k\} \]

Hence, at the first adaptation, for the analysis time instant \( t_i \), the instantaneous frequencies are \( \hat{f}_k(t_i) = k \hat{f}_0(t_i) + \eta_k(t_i) \) and the instantaneous phases become

\[ \hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} \hat{f}_k(u) du \]

Then, a Least Squares solution for the \( a_k, b_k \) using these refined frequencies (and phases) leads to a better estimate of the instantaneous amplitudes \( \hat{A}_k(t) = |a_k(t)| \) and the \( \eta_k \) terms. By iteratively adding the \( \eta_k \) term of the current adaptation on the \( k^{th} \) frequency track of the previous adaptation, the frequency tracks deviate from strict harmonicity and represent the underlying actual frequencies better. Additionally, and on the contrary to previous works [8, 22], where the frequency correction estimation \( \hat{\eta}_k \) on each adaptation should be less than \( \hat{f}_0/2 \), in our approach it is supposed that after each adaptation the estimated frequencies become more and more localized to the actual frequencies, so the frequency correction for a given analysis time instant \( t_i \) is constrained as in

\[ |\hat{\eta}_k(t_i)| \leq \frac{\hat{f}_0(t_i)}{m + 1} \]

where \( m \in \{1, \cdots, M\} \) is the current adaptation number and \( M \) is the maximum number of allowed adaptations (in our experiment, \( M = 6 \)). This way, any relatively large frequency correction value - which often leads to audible artefacts - that might be obtained in a higher adaptation step will be suppressed. Finally, this adaptation scheme continues until a convergence criterion is met, which is
related to the overall Signal-to-Reconstruction-Error Ratio (SRER), that is, when the SRER stops increasing after each adaptation, then the algorithm is considered to have converged. The SRER is defined as

\[ SRER = 20 \log_{10} \frac{std(d(t))}{std(d(t) - d(t))} \]  
(12)

where \(d(t)\) is the original waveform, \(\hat{d}(t)\) is the model representation, and \(std(\cdot)\) is the standard deviation.

2.3. Synthesis

In the synthesis stage, the \(k^{th}\) instantaneous amplitude track, \(\hat{A}_k(t)\), is computed via either linear or spline interpolation of the successive estimates from the last adaptation step. The \(k^{th}\) instantaneous frequency track, \(f_k(t)\), is also computed via spline interpolation. Also, it is worth noting that a frequency matching mechanism is trivial, since the analysis frequencies are integer multiples of a fundamental and the number of components is constant. As for the \(k^{th}\) instantaneous phase track, \(\hat{\phi}_k(t)\), the non-parametric approach based on the integration of instantaneous frequency is followed, as it is shown in the adaptation steps of the analysis. In addition, phase coherence over frame boundaries is an issue that needs to be addressed. Therefore, a constant term is added in order to guarantee phase continuity over frame boundaries as described in [22]. Finally, the speech signal can be approximated by its time-varying components using:

\[ \hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t)e^{j\hat{\phi}_k(t)} \]  
(13)

A block diagram of the algorithm is depicted in Figure 1.

![Block diagram of the eaQHM system.](image)

**Fig. 1. Block diagram of the eaQHM system.**

3. EVALUATION

In this section, objective and subjective measures of quality of the resulted synthetic speech from all different available models (SM, aHM, eaQHM) are presented. To show the robustness on pitch estimation differences, two well-known pitch estimators were used. The \(aHM\), \(eaQHM\) are presented. To show the robustness on pitch estimation limits were [70, 220] Hz and [120, 350] Hz for males and females, respectively. For AIR-\(f_0\), which was used in the aHM model only, the analysis window is of Blackman type and its length is 3 local pitch periods, whereas the step size is pitch period synchronous. For the model parameter estimation, the analysis window is of Blackman type for aHM, and Hamming type for eaQHM and SM. Their size is 3 times the local pitch period and the analysis step size was 2.5 ms, for all models. It should also be noted that 2\(K + 1\) parameters \((A_k, \phi_k)\) per synthesis frame are used in all models, where \(K\) is the number of sinusoids.

3.1. Objective Evaluation

In objective analysis, the Signal-to-Reconstruction-Error Ratio (SRER) is chosen to measure the accuracy of the numerical representation between the original and the synthesized speech. In Table 1, the mean and the standard deviation of the SRER for all utterances in our database are presented for both pitch estimators. It is clearly evident that quasi-harmonicity can capture more information of the underlying speech signal, with the same number of synthesis parameters. Figure 2 shows the first 16 frequency tracks in the analysis step for an utterance produced by Greek male speaker, the local pitch periods, whereas the step size is pitch period synchronous. For the model parameter estimation, the analysis window is of Blackman type for \(aHM\), and Hamming type for \(eaQHM\) and SM. Their size is 3 times the local pitch period and the analysis step size was 2.5 ms, for all models. It should also be noted that 2\(K + 1\) parameters \((A_k, \phi_k)\) per synthesis frame are used in all models, where \(K\) is the number of sinusoids.

<table>
<thead>
<tr>
<th>Model</th>
<th>SWIPE</th>
<th>YIN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Males</td>
<td>Females</td>
</tr>
<tr>
<td>SM</td>
<td>18.6(1.90)</td>
<td>18.6(3.64)</td>
</tr>
<tr>
<td>aHM</td>
<td>23.9(2.66)</td>
<td>18.9(3.27)</td>
</tr>
<tr>
<td>eaQHM</td>
<td>34.5(2.39)</td>
<td>30.9(3.00)</td>
</tr>
</tbody>
</table>

**Table 1. Signal to Reconstruction Error Ratio values (dB) for all models on a database of 32 utterances (16 of male speakers, 16 of female speakers) using SWIPE and YIN pitch estimators. Mean and Standard Deviation are given.**

3.2. Subjective Evaluation

For perceptual quality evaluation, a formal listening test was designed. A part of it is currently available on-line\(^1\). The listeners were asked to evaluate the perceptual quality of the resynthesized speech compared to the original one, for all different models. An 1 - 5 scale was used in the evaluation according to the recommendation ITU-R BS [30], with each scale being (1) “Very bad”, (2) “Bad”, (3) “Good”, (4) “Very good”, (5) “Perfect”. The results from 34 listeners are depicted in Fig. 3. In the same plot we show the 95% confidence interval. This shows that the obtained results are statistically significant. Please note that among these listeners, only 6 were familiar with signal processing and listening tests.

\(^1\)http://www.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.EAQHM-LT
Fig. 2. Analysis data of a Greek male speaker for both adaptive models: (a) aHM tracks, (b) eaQHM tracks, (c) Local SRER for both models over time, (d) Speech waveform.

Fig. 3. Impairment evaluation of the resynthesis quality between the original recording and the reconstructions with all three models, with the 95% confidence intervals.

4. DISCUSSION

According to the listeners, the overall quality of both adaptive models is much higher than the traditional Sinusoidal Model. Moreover, perceptual differences between the two adaptive models were easy to find, and it was clearly stated that these differences are mostly present in the unvoiced parts, and especially in transients and sharp onsets of voiceless stop sounds (for example, in an aspirated velar /k/ in the utterance of Figure 4 by a Korean female). Additionally, it is interesting that although AIR-aHM performs significantly lower in terms of reconstruction, this does not translate to a respective quality degradation, whereas in the SM, there is a substantial perceptual quality degradation, compared to the other two models. Finally, it is interesting that although the pitch estimators behave differently, both the adaptive models appear to be very stable in the reconstruction of output speech, as Table 1 shows.

Fig. 4. Speech utterance (/krʌkʰe/) in Korean language by a female subject. Upper panel: Original signal, Middle panel: aHM reconstruction, Lower panel: eaQHM reconstruction.

Regarding the complexity of the algorithms, on average it takes about 80 seconds for eaQHM and about 55 seconds for aHM to perform analysis and synthesis of a 4-seconds long speech utterance on an Intel Core i7 CPU with 6 GB of RAM using MATLAB programming environment. Most of the computational burden comes from the refinement of \( f_0 \) for AIR-aHM and from the successive adaptations for eaQHM until it converges. In our experiments, a mean number of 2.3 adaptations for eaQHM and a mean number of 14 iterative refinements of the \( f_0 \) for AIR-aHM were required in order for the models to converge.

5. CONCLUSIONS AND FUTURE WORK

In this paper, the extended adaptive Quasi-Harmonic Model - eaQHM analysis/synthesis system for speech is presented, and we showed that high resolution analysis and perceptually indistinguishable resynthesized speech is rendered. The system assumes an initial harmonic model which successively converges to quasi-harmonicity. Numerical evaluations showed that eaQHM can outperform all state-of-the-art systems, such as SM, and the recently proposed AIR-aHM, and it is insensitive to \( f_0 \) estimation errors, thanks to the iterative adaptation mechanism. From a perceptual point of view, listeners found differences between the adaptive Harmonic Model and the suggested model, which concludes that quasi-harmonicity plus adaptivity is adequate to overcome any \( f_0 \) estimation errors and provide transparent resynthesized speech. In the near future, the development of prosodic modifications will be the primary focus regarding this model.
6. REFERENCES


