Analysis and Synthesis of Speech using an Adaptive Full-band Harmonic Model

Gilles Degottex and Yannis Stylianou

***Abstract*—Voice models often use frequency limits to split the speech spectrum into two or more voiced/unvoiced frequency bands. However, from the voice production, the amplitude spectrum of the voiced source decreases smoothly without any abrupt frequency limit. Accordingly, multiband models struggle to estimate these limits and, as a consequence, artifacts can degrade the perceived quality. Using a linear frequency basis adapted to the non-stationarities of the speech signal, the Fan Chirp Transformation (FChT) have demonstrated harmonicity at frequencies higher than usually observed from the DFT which motivates a full-band modeling. The previously proposed Adaptive Quasi-Harmonic model (aQHM) offers even more flexibility than the FChT by using a non-linear frequency basis. In the current paper, exploiting the properties of aQHM, we describe a full-band Adaptive Harmonic Model (aHM) along with detailed descriptions of its corresponding algorithms for the estimation of harmonics up to the Nyquist frequency. Formal listening tests show that the speech reconstructed using aHM is nearly indistinguishable from the original speech. Experiments with synthetic signals also show that the proposed aHM globally outperforms previous sinusoidal and harmonic models in terms of precision in estimating the sinusoidal parameters. As a perspective, such a precision is interesting for building higher level models upon the sinusoidal parameters, like spectral envelopes for speech synthesis.**

***Index Terms*—Voice model, sinusoidal model, harmonic model, non-stationary**

I. Introduction

Sinusoidal and harmonic models aim to represent the speech signal with a set of parameters such as frequencies, amplitudes and phases [1], [2]. These models have been widely used in speech coding and synthesis [3], speech enhancement [4], for hearing aids [5] and voice transformation [6]. Additionally, the parameters can be used to build higher-level representations like spectral envelopes [7], [8], [9] or to estimate glottal source characteristics [10]. However, for this purpose, the accuracy and precision of the model parameters are key issues. A representation that reproduces sounds perceived as being of sufficient quality is another key issue.

Sinusoidal and harmonic models are mainly designed for representing the periodic (or deterministic) part of speech. In order to model the non-deterministic part of speech, these models often employ a random component [2]. Alternatively, the voiced speech spectrum can be represented using multiple bands, with some bands representing the deterministic part and others the non-deterministic part of speech using noise components [11], [12]. Simpler models have also been suggested in which the spectrum is split into two bands separated by the so-called maximum voiced frequency [13], [6]. The lower and higher bands are used for the deterministic and the non-deterministic components, respectively. For all multiband models, the reliable estimation of the voicing frequency limits are critical to avoid artifacts and provide a sufficient perceived quality of the synthesized sound. The need of frequency limits is however questionable for the following reasons. The voiced source is made of glottal pulses that are basically wideband signals whose amplitude spectrum is known to decrease smoothly [14], [15]. Thus, from the point of view of the voice production, there is no reason to abruptly low-pass the deterministic component of the voice. Additionally, the following observation support the presence of harmonic and deterministic content higher than usually observed with the DFT. In voiced segments, the speech signal is usually assumed to be stationary in a small analysis window (≈ 3 pitch periods). At low frequencies, this hypothesis is fairly acceptable because the variations of the fundamental frequency, f_0, of the glottal source are negligible compared to the stationary basis of the Discrete Fourier Transform (DFT). However, the variations of f_0 are proportional to 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 frequency bands.

In order to alleviate this problem of modeling sinusoidal non-stationarities, the Fan Chirp Transform (FChT) has been suggested that uses a chirp related frequency basis (i.e. linear frequency trajectories) adapted to the input signal [16]. Figure 1 shows the spectrograms of a short segment of voiced speech obtained by the DFT (left) and FChT (right) Although the low-frequencies in the DFT-based spectrogram seem to have a regular structure, this is not true for the frequencies around 3000 Hz where the frequency content is blurred. On the other hand, using the FChT, a regularity in the frequency content can be observed across all of the frequencies. This observation suggests that current voice models often underestimate the voicing frequency and a harmonic representation could be appropriate for both low and high frequencies.

Following the arguments above, we seek to use a full-band harmonics only representation of the speech spectrum. However, instead of relying on a chirp frequency basis, as in the FChT that limits the frequency tracks to linear time evolution, we suggest relying on a more flexible frequency model. For sinusoidal analysis, the Adaptive Quasi-Harmonic Model (aQHM) has been already suggested in which the

G. Degottex and Y. Stylianou are with the University of Crete, Computer Science Department and FORTH, Institute of Computer Science, Heraklion, Greece. (e-mail: degottex@csd.uoc.gr; yannis@csd.uoc.gr).

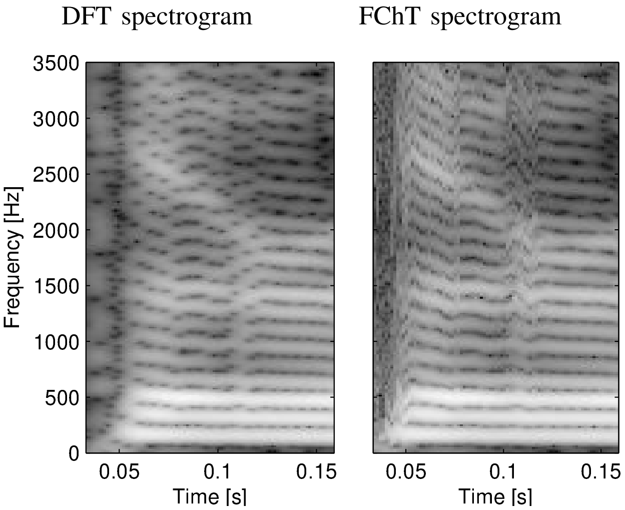

Fig. 1. Time-frequency segments of spectrograms using DFT and FChT. The FChT clearly reveals a harmonic structure in high frequencies which does not appear using the DFT.

frequency basis uses adaptive quasi-harmonic tracks based on an f_0 trajectory measured from the observed signal [17], [18]. As well as in FChT, this adaptivity allows for a non-stationary representation of the frequency components. However, it is not limited to linear trajectories, as is the case for FChT. By means of interpolation of anchor points, the adapted basis can follow any non-linear variations in the frequency modulations of the underlying signal. However, the estimation of aQHM parameters up to the Nyquist frequency is not straightforward. Indeed, aQHM assumes that the initial frequency tracks are in a restricted interval around the actual values [19], [17]. Thus, any potential error in the f_0 trajectory is multiplied by the harmonic number. For example, an error in the initial estimation of f_0 of only 1Hz at 100Hz results in an error of 50Hz at the 50^{th} harmonic. This error is half of f_0, thus placing the 50^{th} harmonic exactly halfway between two harmonics and outside of any reasonable interval for possible correction of frequency, as is necessary in the aQHM scheme. Furthermore, this generates a *frequency matching* problem, i.e. an ambiguity in terms of the connection between frequency components from neighboring frames. A correct frequency matching is, however, quite vital in order to preserve the quality of the reconstructed signal, especially when this is applied during the analysis stage as for aQHM. Consequently, from a point of view of either analysis or synthesis, an accurate f_0 estimate is critical in order to localize harmonic content in the high frequencies of the speech spectrum.

In [20], we recently suggested revisiting the simple harmonic model using adaptivity and full-band representation. This model was referred to as the Adaptive Harmonic Model, aHM. Additionally, regarding the potential error in f_0 leading to wrong localization of sinusoidal components as discussed above, an iterative algorithm had been also proposed in [20], called Adaptive Iterative Refinement (AIR), to allow a robust estimation of harmonic components up to the Nyquist frequency. In the current paper, we detail the technical description in order to facilitate the reproduction of the results and we present the results of a new comprehensive evaluation which should help to better understand both the model aHM and the algorithm AIR. Indeed, in the current paper, we estimate the accuracy and precision of the model parameters using synthetic signals in order to assess the advantage of these parameters before building higher-level models (e.g. spectral envelope). Then, we discuss also The Signal-to-Reconstruction Error Ratio (SRER) for both voiced and unvoiced segments. Finally, we present the results of two new listening tests in order to widen the number of existing state-of-the-art methods for comparison.

The basic idea of the algorithm AIR is the following. It starts by first modeling the lowest harmonics, where errors in the f_0 measurements can easily be corrected by the correction mechanism of QHM [19]. Next, the harmonic order of the model is iteratively increased by a continuous refinement of the f_0 trajectory. Consequently, the quasi-harmonicity is still used as a tool to estimate the adaptivity even though the quasi-harmonicity is not kept at the final speech representation of the suggested model. Strict harmonicity is thus used as a constraint in aHM in order to avoid ambiguities during frequency matching. Compared to other approaches for speech modeling (e.g. mixed excitation models, multi-band models, HNM [2]), aHM does not use a random component in voiced segments. Moreover, since aHM covers the whole spectrum and its frequency basis is not constrained to linear trajectories, it might also represent unvoiced segments properly. Thus, aHM can be used, and is used in this work, for the entire speech signal, whether or not the analyzed segment is voiced. Consequently, the suggested analysis/synthesis procedure does not need any detection of voiced/unvoiced transitions.

In the following paper, the description of aHM-AIR is split into two parts: Section II first describes the mathematical background and Section III then provides all of the technical details. The evaluation follows in Section IV with the necessary discussions and conclusions at the end of this document.

II. Theoretical background

Given the speech waveform $s(t)$, we first assume that its fundamental frequency curve $f_0(t)$ is known a priori, though we consider that there is a potential error on this curve. Then, in a single window of 3 pitch periods, we suggest using the following adaptive Harmonic Model (aHM) to represent the speech signal:

$$s(t) = 2\Re\Big(\sum_{k=1}^{K} a_k(t) \cdot e^{jk\phi_0(t)}\Big) \tag{1}$$

where $a_k(t)$ is a complex function of time representing both the amplitude and the instantaneous phase of the k^{th} harmonic and $\phi_0(t)$ is a real function defined by the integral of $f_0(t)$:

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

where the time reference $t=0$ is the center of the window, and f_s denotes the sampling frequency. According to the adaptive scheme proposed in [17], $a_k(t)$ and $f_0(t)$ are obtained by interpolating values a_k^i and f_0^i at specific instants t_i, termed anchor values in the following. The suggested method therefore provides estimates of these anchor parameters, which are assumed to be sufficient for the complete representation of the speech signal. The number of anchors has to be properly chosen. Indeed, too many anchors may overfit the signal and represent variations which are not related to a deterministic component in voiced segments. Such a behavior is meaningless for statistical modeling and may cause the voice characteristics to be difficult to control in voice transformation. On the other hand, underfitting the signal using too small a number of anchors must also be avoided. Assuming that, for speech, the frequency modulation is related to a change of pulse duration and not to any modulation inside a single pulse, one anchor per period should be sufficient. Even though the position of the anchors could play a role in the quality of the resynthesized sound, addressing the optimization of the anchor positions would overcharge this presentation and we expect this subject will be therefore addressed in future works. At the moment, we consider a pitch synchronous analysis in which the distance between anchors respects an input f_0 curve.

To estimate the aHM parameters in a robust way with the presence of potential fundamental frequency estimation errors, we will use the frequency correction mechanism of the adaptive Quasi-Harmonic Model (aQHM) [19]. This model is similar to the previous one:

$$x(t) = \sum_{k=1}^{K} (a_k + t b_k) \cdot e^{jk\phi_0(t)} \tag{3}$$

where $x(t)$ is the modeled analytic signal of the speech waveform, $\phi_0(t)$ is still defined by equation (2) and a_k, b_k are complex values that are constant in the window (in contrast to $a_k(t)$). To estimate the parameters a_k, b_k, we minimize the following squared error by discrete sampling between the windowed speech segment $s[n]$ and its model $x[n]$ (eq. 3):

$$\epsilon = \sum_{n=0}^{N-1} (s[n] - x[n])^2 \tag{4}$$

where N is the number of samples in the analysis window. The solution of this minimization problem can be found in [21, p.12]. As it has been shown in [19], a_k, b_k can be used to estimate the frequency correction of stationary components. Specifically, for each frequency component, a correction term can be computed as:

$$df_k = \frac{f_s}{2\pi} \cdot \frac{\Re(a_k)\Im(b_k) - \Im(a_k)\Re(b_k)}{|a_k|^2} \tag{5}$$

where $\Re(.)$ and $\Im(.)$ denote the real and imaginary parts, respectively. Using this correction, each anchor frequency f_0^i can be iteratively refined. The initial guess, however, has to be in a reasonable interval around the actual frequency, and the bandwidth of the main lobe of the analysis window can be used to define this interval [19]. The basic idea of the proposed iterative algorithm is the following. For a single analysis window, we first assume that the initial predicted frequencies $f_k = k \cdot f_0$ for a small number of harmonics, K, (e.g. 4) are close enough to the actual frequencies of the signal. This means that we assume the initial pitch estimate is free of octave errors. Then, estimating the parameters of equation (3), the correction term related to the fundamental frequency f_{corr} can be estimated as the mean of the correction terms df_k relative to f_0:

$$f_{corr} = \frac{1}{K} \sum_{k=1}^{K} df_k / k \tag{6}$$

The number of harmonics K can then be updated, taking into account this fundamental correction f_{corr}. Indeed, if $|f_{corr}|$ is low, the current set of K harmonics converges to the actual values. We can therefore assume that a few harmonics above K are now in a reasonable interval around their actual frequencies and K can thus be increased. To control the number of new harmonics added at each iteration, we propose linking K to f_{corr} in the following way. We first assume that the f_0 error remaining to be corrected is smaller or equal to f_{corr}. Therefore, the highest predicted harmonic frequency inside an interval of size $2N_w$ around the actual frequency is:

$$K = \lfloor N_w / |f_{corr}| \rfloor \tag{7}$$

According to [17], equation (5) holds only if the frequency to be corrected lies in a reasonable interval around the actual frequency. According to experiments, the size of this interval is about $B_w/3$ where B_w is the bandwidth of the squared window's main lobe [17]. Additionally, the highest frequency of the new set of harmonics has to be closer to its actual frequency than one of its neighboring frequencies (which are located $0.5 \cdot f_0$ around the actual frequency). Consequently, we chose N_w as the minimum between $B_w/3$ and $0.5 \cdot f_0$. Using (7), the initial number of harmonic K can also be chosen based on an assumed initial fundamental error (e.g. 20 Hz). Using the mechanism of frequency correction of aQHM, $|f_{corr}|$ will be reduced progressively along the iterations and K will thus be increased up to the Nyquist frequency.

III. Method

In this section, we describe the whole analysis/synthesis procedure. Compared to [20], we globally detail the description and comment on the stopping criterion as describe below. Note that the Matlab code of both analysis and synthesis is available on the following web-page:
http://gillesdegottex.eu/ExDegottexG2013jahmair

A. Analysis

Analysis consists of the parametrization of the speech signal at an analysis instant t_i based on (1). A sequence of instants is thus first created using the provided $f_0(t)$ curve, for example: $t_{i+1} = f_0(t_i)^{-1} + t_i$ and $t_0 = 0$. In unvoiced segments, even though the estimated $f_0(t)$ does not have a particular meaning, it is used nevertheless to generate analysis instants. In some speech segments, like in plosives, the time amplitude envelope can vary quickly. A minimum density of anchors has thus to be ensured in order to model properly this time variation. In the current implementation, we limit the distance between two anchors to a maximum of 20ms and we thus clip the provided $f_0(t)$ curve to a minimum value of 50Hz. Around each anchor time t_i, a Blackman window 3 local pitch periods long is applied to the speech signal. $\phi_0(t)$ is then computed by means of numerical integration and linear interpolation of f_0^i (eq. 2). The parameters a_k^i, b_k^i of the ith frame are computed using the LS solution of (4), as well as the frequency correction df_k and the fundamental correction f_{corr} (eq. 6). The number of harmonics for that frame, K^i is then updated using (7). Finally, the process is repeated for all frames until the Nyquist frequency is reached for all the frames. Algorithm 1 summarizes the analysis procedure.

Algorithm 1 Adaptive Iterative Refinement for aHM

```
Create a sequence of analysis instants $t_i$ according to $f_0(t)$
Initiate each $f_0^i = f_0(t_i)$
Initiate each $K^i$ using $f_{corr} = 20Hz$ and (7)
while $\exists i$ such as $f_0^i \cdot K^i < f_s/2$ do
  Compute $\phi_0(t)$ using (2) and interpolation of $f_0^i$
  for each anchor $c$ do
    Create a segment of 3 periods around $t_c$ using $f_0^c$
    Compute LS solution $(a_k^c, b_k^c)$ of (eq. 4)
    Compute $df_k$ (eq. 5) and $f_{corr} = \text{mean}(df_k/k)$
    Correct $f_0'^c = f_0^c + f_{corr}$
    if $f_0^c \cdot K^c < f_s/2$ then
      Update $K^c = \lfloor 0.5 \cdot N_w/|f_{corr}| \rfloor$
    end if
  end for
  Set $f_0^i = f_0'^i \ \forall i$
end while
```

In the above algorithm (AIR), the following three points also have to be considered:

1) Consistency of the correction terms: A df_k term whose harmonic lies in a frequency band made of noise cannot be interpreted as frequency correction. It is therefore necessary to check the consistency of the df_k values and ignore those which may degrade the f_0 curve instead of refine it. Any df_k term which does not satisfy the following three tests is discarded from the computation of f_{corr} (eq. 6): One, $|df_k|$ has to be smaller than $f_0/2$, otherwise two components may be close to each other turning the LS solution unstable. Two, $kf_0 + df_k$ has to be higher than 50Hz, this limit is assumed to be a minimum for f_0. Three, according to [17], df_k has to be smaller than $B_w/3$ where B_w is the main lobe's bandwidth of the squared window. Finally, the median value is also used to compute the fundamental correction term in (6) in order to avoid remaining outliers in the distribution of corrections.

2) Stopping criterion: Even though Algorithm 1 stops when the model is full-band, extra iterations may still improve the representation of the signal. In the current implementation, the iterations stop when the following two convergence criteria are met: i) the correction at the highest harmonic level $K \cdot |f_{corr}|$, has to be smaller than 10% of f_0 to ensure that the modifications of the frequency grid are negligible and ii) the maximum improvement of Signal to Reconstruction Error Ratio (SRER) for all of the frames is smaller than 0.1 dB.

3) Final iteration: Finally, Algorithm 1 provides parameters of aQHM and not aHM, the former having bigger flexibility than the latter because of the quasi-harmonicity in aQHM. Consequently, in order to ensure the consistency between the analysis and the synthesis models, the aHM model is used in the last iteration.

B. Synthesis

The synthesis procedure generates each harmonic successively (1) for the whole signal, without the use of any synthesis window [20]. Below, we describe the way to generate each harmonic from its estimated parameters, namely its amplitudes $|a_k^i|$, its phases $\angle a_k^i$ and the fundamental frequency f_0^i.

First, the instantaneous amplitude $|a_k(t)|$ is simply obtained by means of linear interpolation across time of the anchor amplitudes $|a_k^i|$ using a logarithmic scale. The instantaneous phase $\angle a_k^i$ cannot be interpolated like the amplitudes because of the linear phase term related to the time advance between each anchor instant. Consequently, we suggest first removing this time advance using the integral of $f_0(t)$ (eq. 2 with $t = 0$ at the start of the signal and $f_0(t)$ being obtained by linear interpolation of f_0^i):

$$\angle \tilde{a}_k^i = \angle a_k^i - k\phi_0(t_i) \tag{8}$$

With this preprocessing, the phase values change smoothly from one anchor to the next if the shape of the signal is also changing smoothly. $\angle \tilde{a}_k^i$ can then be interpolated to obtain its continuous counterpart $\angle \tilde{a}_k(t)$. In order to avoid phase jumps in the interpolation (e.g. between $-\pi$ and π), real and imaginary parts of $e^{j \cdot \angle \tilde{a}_k^i}$ are interpolated independently and the interpolated values are recovered through the arctangent function. Additionally, a spline or cubic interpolation is necessary so that the time-derivative of $\angle \tilde{a}_k^i$, i.e. the frequency, is still continuous. Finally, $\phi_0(t)$ is obtained using (2) ($t = 0$ being the start of the signal) and $a_k(t)$ is $|a_k(t)| \cdot e^{j \angle \tilde{a}_k(t)}$. All harmonics are finally summed as in (1) while discarding time segments of harmonics whose frequency are above the Nyquist frequency.

IV. Evaluation

In this section, we evaluate the suggested method aHM-AIR by making comparisons with state-of-the-art methods. Before evaluating the whole analysis/synthesis procedure, some comments are first given about the number of parameters used in aHM and the other methods. Then, we evaluate the relevance, i.e. accuracy and precision, of the model parameters estimated during the analysis step using synthetic signals. Since we designed the estimation algorithm to be robust against an error in the estimated fundamental frequency f_0, the accuracy and precision of the methods are evaluated as a function of this error. Then, the reconstruction of the speech signal is evaluated using the Signal to Reconstruction Error Ratio (SRER) as measurement. Using a set of 24 recordings (12 languages with both male and female voices), we show the SRER distributions of the compared methods for both voiced and unvoiced segments. Finally, the results of two formal listening tests are presented which evaluate the perceived subjective quality.

For the comparisons, the three following models are used:

SM The Sinusoidal Model [1] represents sinusoidal components which are estimated through peak picking on the DFT spectrum. For this method, the length of the window plays a crucial role in both parameter estimation and perceived quality of the resynthesis. Although the initial version of SM uses fixed length analysis windows, improved accuracy and precision are obtained if the analysis window is adaptive to f_0. Moreover, as the window length is not the subject of the following evaluations, its influence on the results has to be minimized. Therefore, the window length for SM is adapted to the input f_0 (similarly to the other methods). Also, in order to evaluate the accuracy and precision of the parameter estimation, it is necessary to associate a harmonic number to each peak observed on the amplitude spectrum. For this reason, the original version is slightly modified by using the closest peaks to the integer multiples of the input f_0. The DFT bins of each peak are then interpolated using a parabola in order to retrieve amplitude and frequency parameters and the phase parameter is obtained by linear interpolation of the phase spectrum [22].

HM The Harmonic Model [2] represents harmonically related sinusoids which are estimated through the LS solution of equation (3) using stationary frequency components (i.e. $\phi_0(t) = 2\pi t f_0/f_s$) and without quasi-harmonicity (i.e. $b_k = 0 \; \forall k$).

aQHM The adaptive Quasi-Harmonic Model [17] is based on (3), but with frequency components not restricted to being harmonically related (i.e. $\phi_0(t)$ becomes $\phi_k(t)$ and $f_0(t)$ becomes $f_k(t)$ in (2)). Note that it is not the mixed model proposed in [18] which uses a random component modulated in both time and frequency. Since all other methods are full-band representations, we preferred to make aQHM also full-band in order to obtain comparisons more straightforward, especially for the listening tests. This model comes with an iterative estimation algorithm which starts directly with a full-band representation unlike Algorithm 1 [17]. In the following evaluations, 6 iteration steps have been used.

All methods use windows of 3 local pitch periods and the parameters are estimated each 5ms. In each analysis window, enough components are used to cover the full spectrum up to the Nyquist frequency. For resynthesis, the method described in section III-B is used for HM, aQHM and aHM. For SM, we used a standard overlap-add method since the frequency matching between components of neighboring frames is far from straightforward and not the subject of the following evaluations.

Each model uses an f_0 input. An algorithm is therefore necessary to estimate these values. For the sake of clarity in the following plots and discussions, a suffix such as STRAIGHT, YIN, AIR etc. is used to distinguish the algorithm used when this information need to be emphasized (e.g. HM-STRAIGHT, HM-AIR, aHM-AIR).

A. Number of parameters

Compared to aQHM, the number of parameter in aHM is reduced since only one parameter (i.e. f_0) is sufficient to reconstruct the full harmonic frequency grid. Conversely, aQHM uses a frequency parameter for each quasi-harmonic component. aHM is therefore similar to HM where $2 \cdot K + 1$ parameters are necessary per frame and aQHM is similar to SM where $3 \cdot K$ parameters are necessary.

B. Parameter estimation error

The purpose is here to evaluate the accuracy and precision of the estimated parameters in terms of a sinusoidal representation, compared to state-of-the-art methods. In the following, the estimated frequency, amplitude and phase values are compared to ground truth values of synthetic signals. To obtain a synthetic signal which is as close as possible to a natural speech signal, a Liljencrants-Fant glottal model [14] is used to synthesize the glottal source. To obtain a realistic vocal tract filter a digital simulator is used [23] that allows production of 13 different voiced phonemes. The synthetic signal is obtained as:

$$s(t) = 2\Re\Big(\sum_{k\in\mathbb{R}^+} G^{f_0(t)}(kf_0(t)) \cdot C(kf_0(t)) \cdot e^{jk\phi_0(t)}\Big) \qquad (9)$$

where $G^{f_0(t)}(f)$ is the spectrum of the Liljencrants-Fant model, $C(f)$ is the vocal tract filter representing a random phoneme among 13 covering the vocalic triangle, and $\phi_0(t)$ follows (2). The pulse shape of the glottal model is controlled by a random parameter $Rd \in [0.3; 2.7]$ [14] and its period is defined by $f_0(t)$, The definition of the synthetic fundamental frequency $f_0(t)$ will depend on the following evaluations.

1) Influence of additive f_0 error: Since AIR is designed to alleviate the consequences of potential errors in the f_0 curve, the following test evaluates the robustness of the different methods when errors in the initial $f_0(t)$ curve exist. In (9), the original $f_0(t)$ curve is first synthesized using 5 anchors per seconds with random values in $[80; 400]$ Hz. Then, a zero-mean Gaussian noise with various STandard-Deviation (STD) is added to this curve which is finally input to the methods. In the following plots, the estimation error of the sinusoidal parameters is plotted as a function of the STD of this additive f_0 error. A total of 320 test samples of 500ms duration each are generated using a sampling frequency of 44.1kHz. The samples are analyzed at regular intervals of 5ms and the differences between the estimated parameters using each method and the reference parameters, are computed. For each method, Fig. 3 shows the mean of the estimation error on the first three rows and the STD using a base-10 logarithmic scale on the last three rows. For all figures, we follow the same line style convention which is shown in Fig. 2. To avoid the influence of outliers in the computation of the mean and the standard-deviation, we computed these two values through the median and the interquartile range, respectively.

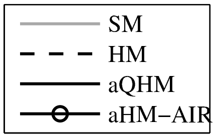

Fig. 2. Line styles for all the methods shown in Fig. 3 and Fig. 6

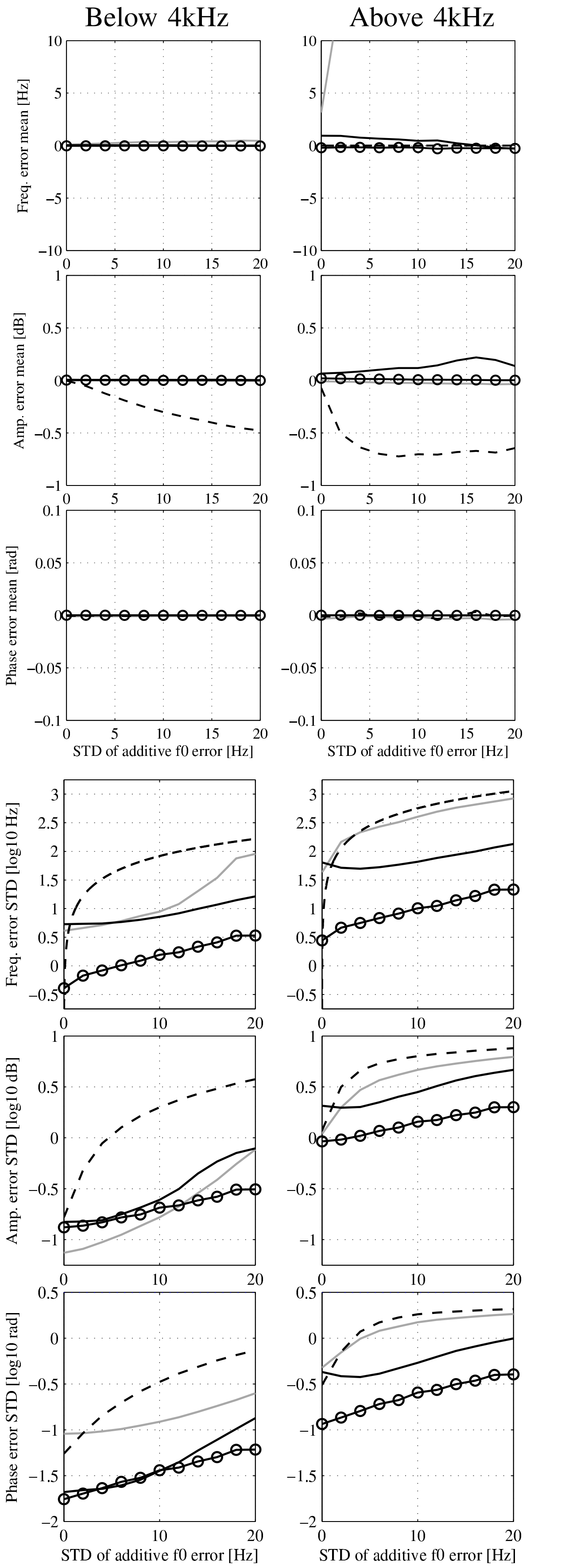

Fig. 3. Error of sinusoidal parameters with respect to a potential error on the f_0 curve provided to the analysis methods.

In the last three rows of Fig. 3, where the error STD is shown, the suggested method aHM-AIR always shows a smaller STD than the other methods except for the amplitude estimation under 4kHz. The estimation of the frequency grid is thus most precise when using aHM-AIR (4th line). The estimation of the phase is also more precise especially above 4kHz (last line, right column). Globally, the improvement provided by aHM-AIR compared to the other methods is most apparent when considering the upper band of the signal. The aHM-AIR method thus provides better parameter precision in the high frequencies. For the Harmonic Model (HM), the error increases quickly as the f_0 error increases because no correction method is used to reduce the influence of the f_0 errors. On the other hand, the SM method selects the observed peaks in the amplitude spectrum even though the input f_0 values can be erroneous. Also, aQHM/aHM-AIR both use an iterative method for the refinement of the input f_0. Concerning the precision of SM in the estimation of the amplitudes below 4kHz (5th row, left column), an explanation could be the following. The SM method always modifies the integer multiples of f_0 by means of quadratic interpolation in order to fit the maximum amplitude of a peak. Even though the frequency can be modified towards an erroneous value, this behavior ensures that the amplitude is always maximized. However, for aHM and aQHM, if the harmonic frequency, $k \cdot f_0$, is not properly aligned with the peak before the LS solution is computed and it slides down the main lobe of the window, the estimated amplitude can be substantially erroneous and consequently have higher variability than the maximized amplitude provided by SM.

2) Parameter estimation error corresponding to f_0 estimation methods: In practice, the additive f_0 error in the evaluations above is related to that of f_0 estimation methods. In the test below, using the same synthetic signals as in the previous test, we measured the parameter error with respect to three f_0 estimation methods which are well known state-of-the-art methods for speech signals: YIN [24], SWIPEP [25], STRAIGHT [26, p.9] (the f_0 method used in STRAIGHT) and the refined f_0 estimate obtained from aHM-AIR. The AIR algorithm needs an input f_0 estimate. The comparison with the other methods would not be fair if the most accurate and precise f_0 method was used for AIR since AIR would anyway refine and improve the results. Thus, in this test, we used the f_0 estimate of SWIPEP because its precision is between that of the two other state-of-the-art f_0 estimation methods, i.e. YIN and STRAIGHT (see Fig. 5). Additionally, according to our experiments, SWIPEP seems to be more robust to octave errors. Fig. 4 shows the results of this evaluation with the mean of the error to the left and the STD of the error to the right using a base-10 logarithmic scale. Fig. 4 shows only the results computed on the full-band of the signal. From our experiments, the results computed on the first 4kHz are very similar to these results and do not provide additional information. For the sake of comparison, Fig. 5 shows the f_0 error of each method. The plots related to the mean of the error (left column) show that the AIR algorithm provides the smallest bias for frequency and phase estimation and a slightly larger bias for amplitude estimation compared to STRAIGHT.

Also, the right column shows that the STD of the error is unequivocally smaller using AIR for all the parameters. In conclusion, the AIR method clearly provides the most robust parameters estimation. According to the right plot of Fig. 5, it is also worth noting the improvement of the f_0 estimation using AIR compared to the SWIPEP method.

3) Influence of the f_0 chirp rate: As argued in the introduction, variations in the f_0 curve can be significant within an analysis window. In this third test, we therefore evaluate the estimation error with respect to a constant variation of $f_0(t)$, that is, the f_0 chirp rate. In (9), the $f_0(t)$ curve is synthesized using: $f_0(t) = f_0(t_c) + c_r \cdot t$, where t_c is the time at the middle of the segment and c_r is the chirp rate. For each synthetic sample, the polarity of the rate is chosen randomly as is the value of $f_0(t_c)$ which is chosen in $[80; 400]$ Hz, so that the f_0 boundaries at the start and end of the segment lie in the same frequency interval. Due to this limitation at the boundaries, the $f_0(t_c)$ values are restricted by the duration of the segment. Accordingly, a duration of only 100ms has been used and 1000 samples have been generated with a sampling frequency of 44.1kHz. All of the methods use the same input f_0. Since the ideal f_0 values in synthesis would not be realistic, we blurred the f_0 in synthesis by adding a zero-mean Gaussian noise with standard-deviation 10^{-1}[Hz] and we provided these blurred f_0 values to all of the methods evaluated. The standard-deviation of the Gaussian noise is chosen according to the previous evaluation (see right plot of Fig. 5). Fig. 6 shows the mean of the estimation error in the first three rows and the STD using a base-10 logarithmic scale in the last three rows (the legend, see Fig. 2, is the same as in Fig. 3). First, when looking at the 1st and 4th rows, it is worth noting that the unbiased and precise frequency estimates corresponding to HM are only due to the input f_0 which has been fixed as described above. According to Fig. 5, this precision of estimation can only be reached using aHM-AIR. Then, similarly to Fig. 3, the suggested aHM-AIR method again shows good precision for frequency and phase estimation (4th and last rows) especially for significant high chirp rate and above 4kHz. Also, the SM method again shows a more precise estimate than the other methods concerning the amplitude parameter below 4kHz (5th row to the left).

C. SRER distributions

Comparing the two adaptive models, aHM is less flexible than aQHM, since the former imposes a harmonic relationship between the frequency components of the signal. On the other hand, the frequency components in aQHM are free to deviate from the harmonicity. To evaluate this difference globally, we measured the Signal to Reconstruction Error Ratio (SRER) between recorded utterances and their models. The set of recordings should cover the voice variability as much as possible. In order to maximize this coverage in the test, we used recordings made of 12 different languages, assuming that the different phonemes and the different origins of these languages provides a sufficient voice variability. Each language was represented by one example from one male voice and one example from one female voice, so 24 recordings were used in total. These utterances were approximately 3 seconds long with a sampling frequency between 16kHz and 44.1kHz. The

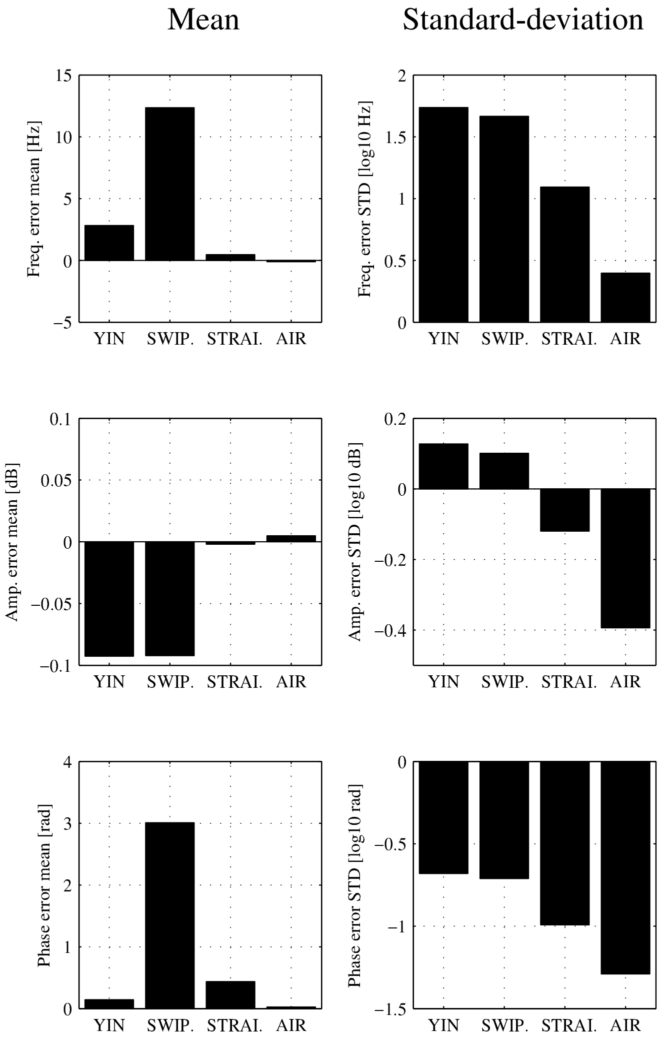

Fig. 4. Error of sinusoidal parameters according to state-of-the-art f_0 methods and the suggested f_0 refinement method AIR (computed on the full-band of the signal).

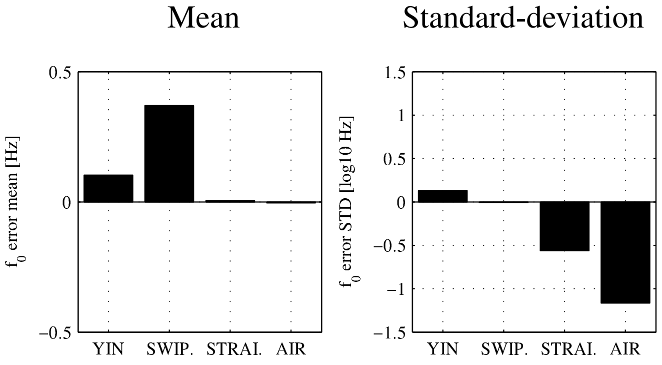

Fig. 5. f_0 error corresponding to f_0 estimation methods using the same synthetic signals as in Figure 4.

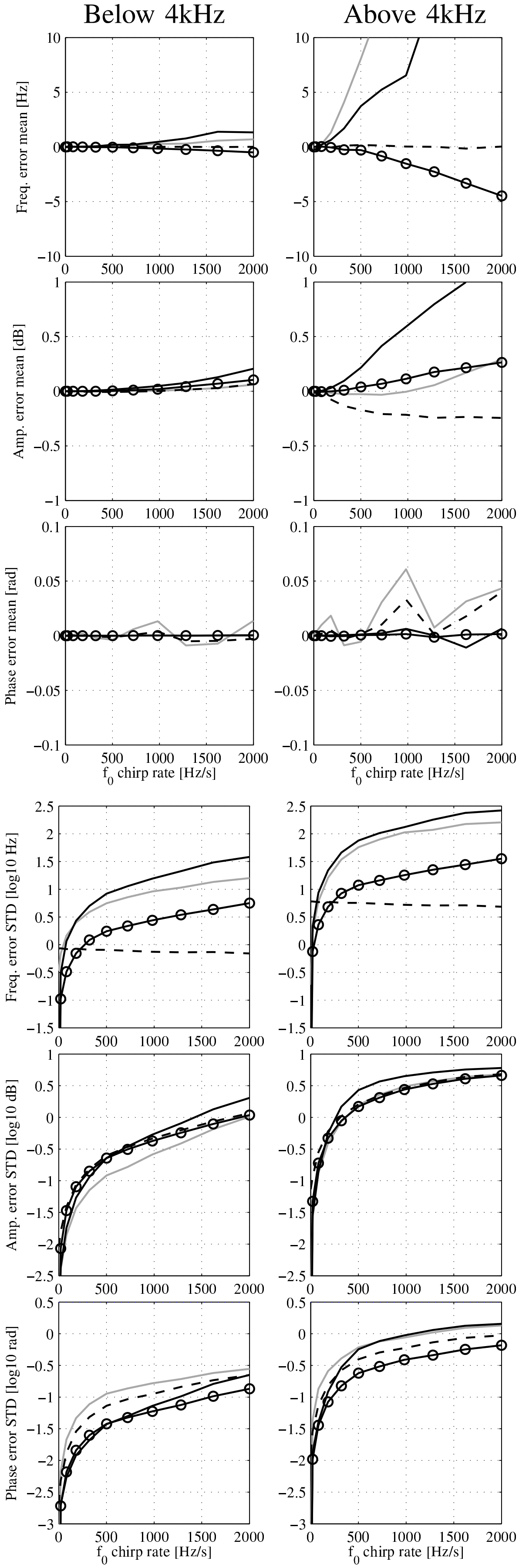

Fig. 6. Sinusoidal parameter error with respect to a chirp rate of f_0.

samples can be found on the following web-page with their corresponding resynthesis using the four analysis/synthesis methods: http://gillesdegottex.eu/ExDegottexG2013jahmair

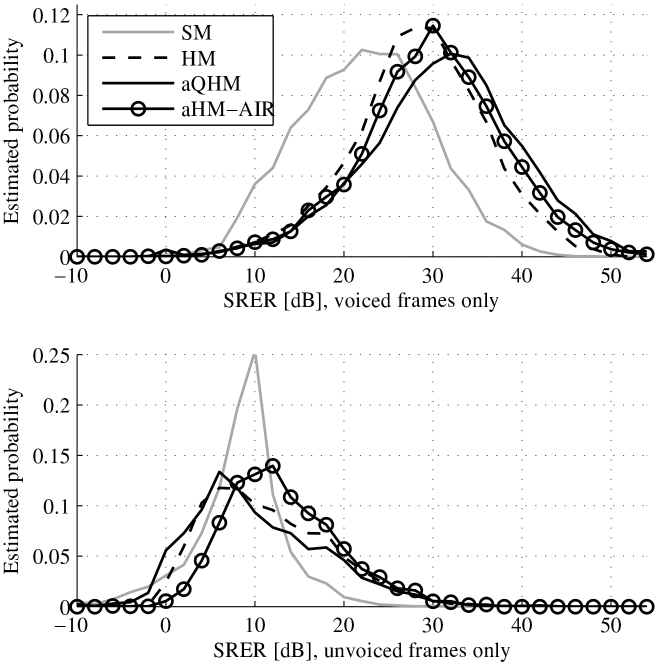

Fig. 7. Estimation of the full-band SRER distributions for voiced and unvoiced frames on top and bottom plots respectively.

In order to minimize the influence of the input f_0 curve on the results, the refined f_0 values given by the output of the AIR algorithm are used for all methods since this method provides the best frequency estimation results according to Fig. 5. Similarly to section IV-B2, SWIPEP has also been used to provide the initial f_0 estimate to the AIR algorithm. Fig. 7 shows the distribution of SRER for each method using a sliding window of 10ms with 50% overlap. The SRER was computed using the full-band of the recordings and its distribution over the voiced and unvoiced segments is shown on the top and bottom plot, respectively. The 24 sentences were sufficient to obtain more than 8000 values for each distribution. Globally, the three models HM, aQHM and aHM have very similar distributions compared to the SM model. For the voiced frames, the mean of these distributions are clearly higher than that of SM. The mean corresponding to aQHM is more than 10dB above SM which is in accordance with the results given in [17]. On the one hand, the three models HM, aQHM and aHM use the LS solution, which explicitly minimizes the reconstruction error during the parameters estimation. On the other hand, in the SM method, it is only assumed that estimating sinusoidal parameters by peak picking provides a set of sinusoids which properly represent the signal. The observed difference between the harmonic models and SM in Fig. 7 thus makes sense. Finally, the aQHM model has a slightly better SRER compared to aHM. One can also expect this result since aQHM is more flexible than aHM, thanks to the quasi-harmonicity. Concerning the unvoiced frames, the average SRER is obviously lower for all of the methods since the limited number of sinusoids of the models cannot properly cover the noise that fills the whole spectrum. The aHM model also provides a better fitting of the noise than the HM model because of its adaptivity. However, as for voiced segments, we could expect that aQHM provides a better SRER than aQHM, which is not the case in Fig. 7. An explanation might be that the corrections terms df_k (in eq. 5) are meaningless for noise and lead to misplaced quasi-harmonics. On the other hand, the strict harmonicity of aHM ensures, at least, that the full-band is regularly sampled.

D. Perceived subjective quality of the models

In this part of the evaluations, the perceived quality of the reconstructed signals using the four models was evaluated subjectively using listening tests. According to Fig. 5, aHM-AIR provides the most precise f_0 estimate. Thus, all of the compared methods in this test use the f_0 estimate of aHM-AIR, as in the SRER evaluation step. This minimizes the influence of f_0 errors in this 1st listening test. The influence of f_0 errors on the perceived quality has been evaluated in another test whose results are presented in the next section. All listening tests have been carried out using a web interface. Compared to tests carried out locally in a laboratory, we believe that for the addressed subject, there is a variability in the listening conditions of web-based tests which is more realistic than that of a specific room prepared especially for experiments. One can also note that it improves the objectivity of the results by using listeners who are not related to the author's work [27]. The listeners were first asked to listen to one original recording among the 24 utterances used for the SRER measurements. Then, they had to rate the impairment of five sounds: four of them were the synthesized made with SM, HM, aQHM and aHM, while the fifth sound was the original recording, which was added to the comparison set in order to check the consistency of the answers. The test duration needed to be moderate in order to keep the listeners focused. An exhaustive evaluation of the 24×5 sounds was therefore not possible. In this test, each listener was asked to grade only 2 languages randomly selected from the set of the 12 languages. Since each language was represented by one male and one female voice, each listener evaluated the resynthesis of 4 recordings. According to the recommendation ITU-R BS[28], we used the following grading scale of impairment: (5)Imperceptible, (4)Perceptible but not annoying, (3)Slightly annoying, (2)Annoying, (1)Very annoying. In order to optimize the listening conditions, we kept only the answers of listeners who used headphones or earphones. Additionally, answers from listeners who did not rate the original recordings systematically between 4 and 5 were discarded considering that the instructions were not understood or the listener was not focused enough. 48 people answered the test and the answers of 44 listeners were kept. Since the sounds to evaluate were selected randomly, the number of occurrences of each sound was not uniform (even though it tends to be when the number of listeners increases). In order to remove any possible bias, the mean and confidence intervals of the results were therefore normalized according to the number of occurrence of each sound. Figure 8 shows the results of this listening test. Firstly, the SM method has been clearly graded lower than the other methods. By the authors, significant artifacts appear in the high frequencies of the resynthesis using this method. Then, globally, the three other methods provide very similar results. These three methods use a harmonic or quasi-harmonic frequency grid which ensures minimal continuity of the sinusoidal components. Conversely, in SM, a component can disappear from one frame to the next which generate a persistent artifact mainly in the high frequencies. This lack of continuity can partially explain the substantial difference between SM and the other models. Note that by replacing the overlap-add method used for SM by a birth-and-death technique [1], we noticed the same artifacts.

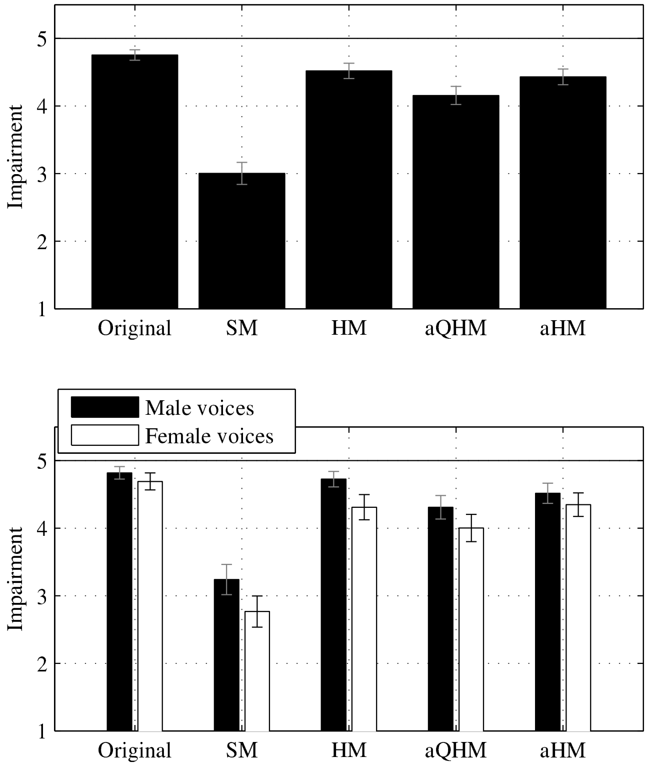

Fig. 8. Impairment evaluation of the resynthesis quality by 44 listeners using 24 utterances of 12 different languages, with the 95% confidence intervals. Global results above and gender-specific results below. The used f_0 values are those provided by the aHM-AIR method.

The slight downward trend of the aQHM method compared to aHM and HM can be explained by some musical sounds which can be sparsely perceived along the resynthesis. Having the frequency components completely independent, as in aQHM, may provide better flexibility, though it also adds a risk that components leave the frequency band in which they are supposed to be. Conversely, the strict harmonicity may oversimplify the representation, even though it offers a global constraint stabilizing the resynthesis. This argument of stability has already been discussed for the SRER distributions in Fig. 7 where one can see that the SRER of aQHM is lower than that of aHM in unvoiced segments. Even though the SRER of aQHM is higher than that of aHM in voiced segments, the SRER difference around 10dB in unvoiced segments is easier to perceive than that around 30dB in voiced segments. The global difference can therefore explain the slight downward trend seen in the listening test. Finally, the results specific to gender show that the resynthesis of the male voices made by the HM method are clearly indistinguishable from their original recordings.

E. Influence of the f_0 estimate on the subjective quality

The f_0 curve obviously has an influence on the resynthesis quality. Consequently, we carried out a second listening test for this purpose by comparing state-of-the-art methods used in section IV-B2. Unlike the test in section IV-D, the goal here is to evaluate the f_0 methods and not the models. Since HM provides the best quality according to Fig. 8, we chose this model for this new test. The same evaluation scheme as in the previous listening test was used. Listeners were asked to evaluate the impairment of sound files compared to an original recording using a web interface. The same 24 utterances previously described were used. In order to reduce the number of sounds to evaluate and thus encourage participation in the test, the YIN method was not used since it provides the least precise f_0 estimates according to Fig. 5. SWIPEP, STRAIGHT and AIR were therefore compared, while the original recordings were again used for verification purposes. Similarly to the previous tests, SWIPEP was used to provide the input f_0 of the AIR algorithm. 52 people answered the test and 46 answered the test by rating the original recordings systematically between 4 and 5. Note that the listening test described here and that from above have been carried out independently and the listeners were not necessarily the same. Globally, HM-AIR and HM using the f_0 given by STRAIGHT (HM-STRAIGHT) provide a quasi-perfect reconstruction unlike HM-SWIPEP. Comparing HM-STRAIGHT and HM-AIR, HM-AIR shows only a slight positive trend. By informally listening to the resynthesis, HM-AIR has indeed slightly less localized artifacts.

F. Discussions

According to the evaluation of the parameter estimation error (Figures 3, 4, 6), even though the simple peak picking method provides a more precise estimation of the amplitude than the suggested aHM-AIR below 4kHz, the latter method offers more precise estimates of the frequency and phase values. It is worth noting that in order to build higher level models upon the sinusoidal parameters (e.g. spectral envelopes), the accuracy and precision of the amplitude parameters is necessary in addition to its location in the time-frequency space. From this point of view, we show in this paper that the aHM-AIR method clearly improves the localization of the frequency tracks compared to state-of-the-art methods. The comparison between the f_0 estimation methods in Figure 4 also shows that the f_0 estimate plays a crucial role in the reliability of the sinusoidal parameters estimates. According to this evaluation, we also show in this paper that AIR clearly improves the f_0 estimates used for sinusoidal parameter estimation.

Concerning the listening tests, a ceiling effect can be observed. The simple harmonic model, when f_0 estimation is precise enough, provides already almost perfect reconstruction quality in terms of perception (Figures 8 and 9). Improvements are therefore difficult to obtain. Based on the results specific to the gender, it is however interesting to note that it is not possible to state if a difference exist between the male voices resynthesized using HM-AIR and their original recordings. This observation is even further supported by the fact that the two tests have been carried out independently. Comparing HM-SWIPEP and HM-AIR in Figure 9, the AIR method clearly improves the quality of the resynthesis based on HM-SWIPEP by refining the f_0 values provided by SWIPEP. Finally, trends can also be observed that indicate aHM-AIR slightly improves the perceived quality compared to aQHM. By informally listening to the resynthesis, aQHM has indeed more artifacts which are mainly localized in time rather than persistent along the sound. We also observed the same difference between HM-AIR and HM-STRAIGHT in the second test (Fig. 9), with HM-AIR having less artifacts than HM-STRAIGHT.

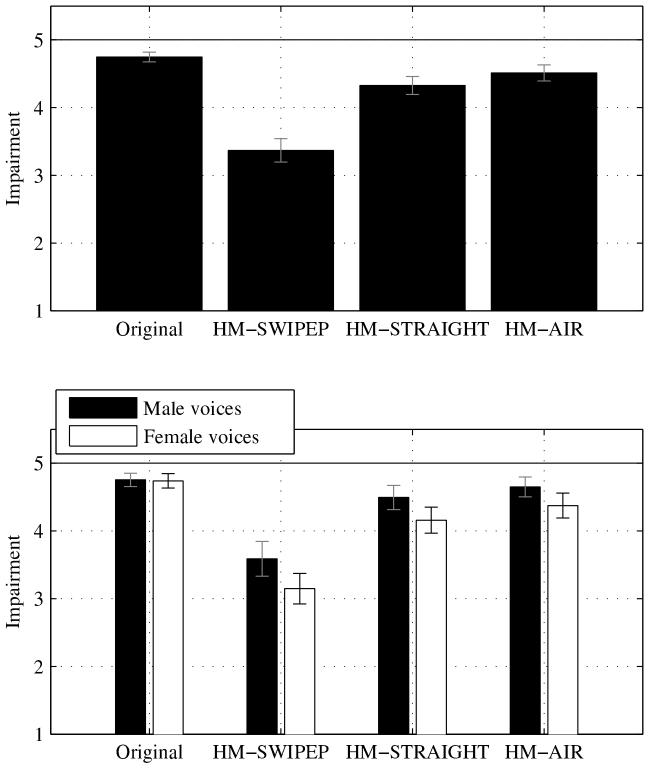

Fig. 9. Impairment evaluation of the resynthesis quality by 46 listeners using the same 24 utterances of 12 different languages, with the 95% confidence intervals.

V. Conclusions

Arguing that the need of frequency limits is questionable to model the speech spectrum and inspired by the observation of the FChT, we assumed that the speech spectrum could be modeled using a full-band harmonic model. Taking advantage of the non-stationary frequency basis of the Adaptive Quasi-Harmonic Model (aQHM), which adapt its frequency basis to the time variations of the frequency components, we suggested in a previous publication a full-band Adaptive Harmonic Model (aHM) for both voiced and unvoiced segments of the speech signal. We had also suggested a new algorithm, the Adaptive Iterative Refinement (AIR), to deal with the localization of the high frequency harmonics up to the Nyquist frequency. In this paper, we provided a new comprehensive evaluation of aHM-AIR. First, using synthetic signals, we evaluated the accuracy and precision of the parameter estimation of aHM-AIR and other state-of-the-art methods and we carried out listening tests to assess the perceived quality provided by the suggested analysis/synthesis procedure compared to other methods. These listening tests clearly show that a full-band Harmonic Model (HM) is sufficient to reproduce a quasi-perfect quality when the fundamental frequency curve has the necessary precision. Compared to aQHM, aHM globally provides the same high quality, with the benefit of a slight positive trend, without using quasi-harmonicity, and thus reducing the number of parameters. Compared to HM, the aHM model does not improve the perceived quality. However, as shown in the evaluation with synthetic signals, the algorithm, AIR, allows for precise estimation of the sinusoidal parameters which is important to build higher-level representation like spectral envelopes.

VI. Acknowledgments

This work was supported by the Swiss National Science Foundation (grants PBSKP2_134325,PBSKP2_140021) and LISTA project (E.U. FP7 FET-OPEN, grant agreement: 25623). The authors would like to thank Elizabeth Godoy for her valued help in the writing and also the numerous listeners who took the time to answer the listening tests.

References

[1] R. McAulay and T. Quatieri, "Speech analysis/Synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[2] Y. Stylianou, *Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, TelecomParis, France, 1996.

[3] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[4] J. Jensen and J.H.L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 731–740, 2001.

[5] Yi Hu and P. C. Loizou, "On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants," *Journal of Acoustic Society of America*, vol. 127, no. 1, pp. 427–434, 2010.

[6] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal-tract filter estimate for voice transformation and synthesis," *Speech Communication*, Accepted, in publishing process, DOI:10.1016/j.specom.2012.08.010 2012.

[7] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[8] M. Campedel-Oudot, O. Cappe, and E. Moulines, "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 469–481, 2001.

[9] Jordi Bonada, *Voice Processing and Synthesis by Performance Sampling and Spectral Models*, Ph.D. thesis, Universitat Pompeu Fabra, Spain, 2008.

[10] G. Degottex, A. Roebel, and X. Rodet, "Phase Minimization for Glottal Model Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1080–1090, 2011.

[11] D. W. Griffin, "Multi-Band Excitation Vocoder," Tech. Rep. RLE Technical Report No. 524, Massachusetts Institute of Technology, 1987.

[12] X. Serra, *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*, Ph.D. thesis, Stanford University, 1989.

[13] S.-J. Kim and M. Hahn, "Two-Band Excitation for HMM-Based Speech Synthesis," *IEICE - Transactions on Information and Systems*, vol. E90-D, no. 1, pp. 378–381, 2007.

[14] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis.," *STL-QPSR*, vol. 36, no. 2-3, pp. 119–156, 1995.

[15] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[16] M. Kepesi and L. Weruaga, "Adaptive Chirp-based time-frequency analysis of speech signals," *Speech communication*, vol. 48, no. 5, pp. 474–492, 2006.

[17] Y. Pantazis, O. Rosec, and Y. Stylianou, "Adaptive AM-FM Signal Decomposition With Application to Speech Analysis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 2, pp. 290–300, 2010.

[18] 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 International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010.

[19] Y. Pantazis, O. Rosec, and Y. Stylianou, "Iterative Estimation of Sinusoidal Signal Parameters," *Signal Processing Letters, IEEE*, vol. 17, no. 5, pp. 461–464, may 2010.

[20] G. Degottex and Y. Stylianou, "A Full-Band Adaptive Harmonic Representation of Speech," in *Proc. Interspeech*. ISCA, September 2012, p. N/A.

[21] Yannis Pantazis, *Decomposition of AM-FM Signals with Applications in Speech Processing*, Ph.D. thesis, University of Crete, 2010.

[22] J. O. III Smith, *Spectral Audio Signal Processing*, W3K Publishing, 2011.

[23] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3-4, pp. 199–229, 1982.

[24] A. de Cheveigne and H. Kawahara, "YIN, A fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[25] Arturo Camacho, *SWIPE: A sawtooth waveform inspired pitch estimator for speech and music*, Ph.D. thesis, University of Florida, USA, 2007.

[26] H. Kawahara, I Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[27] H. Honing and O. Ladinig, "The Potential of the Internet for Music Perception Research: A Comment on Lab-Based Versus Web-Based Studies," *Empirical Musicology Review*, vol. 3, no. 1, pp. 4–7, 2008.

[28] The ITU Radiocommunication Assembly, "ITU-R BS.1284-1: EN-General methods for the subjective assessment of sound quality," Tech. Rep., ITU, 2003.

Gilles Degottex received the Diploma degree in computer science from University of Neuchâtel (UniNE), Switzerland. After a one-year specialization at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, he obtained his Ph.D. degree in 2010 at the Institut de Recherche et Coordination Acoustique/Musique (Ircam), Université Pierre et Marie Curie (UPMC), Paris, France. He is currently holding a postdoctoral position regarding voice modeling for voice transformation at University of Crete, Heraklion, Greece. His research interests include glottal source features, wide and narrow band voice models for voice transformation, speech synthesis and speech enhancement.

Yannis Stylianou (M'95, SM'12) is Professor at University of Crete and Associated Researcher at FORTH. During 2011-2012 he was visiting Professor at University of the Basque Country, Bilbao, Spain. He received the Diploma of Electrical Engineering from the National Technical University of Athens and the M.Sc. and Ph.D. degrees in Signal Processing from the Ecole National Superieure des Telecommunications (ENST), Paris, France in 1992 and 1996, respectively. Then, he was with AT&T Labs Research, NJ, USA, as a Senior Technical Staff Member. In 2001 he joined Bell-Labs Lucent Technologies, NJ, USA. He is on the Board of the International Speech Communication Association (ISCA), on the Editorial Board of the Digital Signal Processing Journal of Elsevier, of Journal of Electrical and Computer Engineering, Hindawi JECE, Associate Editor of the EURASIP Journal on Speech, Audio, and Music Processing and of the EURASIP Research Letters in Signal Processing. He was member of the IEEE Speech and Language Technical Committee (2007-2010) and Associate Editor for the IEEE Signal Processing Letters (2000-2002)