A Full-Band Adaptive Harmonic Representation of Speech

Gilles Degottex[1] *and Yannis Stylianou*

University of Crete, Computer Science Dep. and FORTH, Inst. of Computer Science
Vasilika Vouton, 71110 Heraklion, Greece

degottex@csd.uoc.gr, yannis@csd.uoc.gr

Abstract

In this paper we present a full-band Adaptive Harmonic Model (aHM) that is able to accurately reconstruct stationary and non stationary parts of speech. The model does not require any voiced/unvoiced decision, neither an accurate estimation of the pitch contour. Its robustness is based on the previously suggested adaptive Quasi-Harmonic model (aQHM), which provides a mechanism for frequency correction and adaptivity of its basis functions to the characteristics of the input signal. The suggested method overcomes limitations of the initial method based on aQHM in detecting frequency tracks over time, especially at mid and high frequencies, by employing a bandlimited iterative procedure for the re-estimation of the fundamental frequency. Listening tests show that reconstructed speech using aHM is mainly indistinguishable from the original signal, outperforming standard sinusoidal models (SM) and the aQHM-based method, while it uses less parameters for the reconstruction than SM.

Index Terms: Sinusoidal model, quasi-harmonic model, non-stationary basis, speech analysis.

1. Introduction

Voice models are challenging for all speech technologies. They allow to represent the perceptual properties of speech into a set of meaningful parameters which can be interpreted (e.g. speech recognition), transformed or modeled statistically (e.g. voice conversion, speech synthesis). The robustness of the estimation and the completeness of their parameters in terms of ability to fully represent the perceived elements of the voice are among the key properties these technologies need. Currently, three main directions exist, the wide-band models like STRAIGHT [1] and WBVPM [2], those using glottal models [3, 4], and the sinusoidal models [5, 6]. Actually, these points of view do not exclude each other and a sinusoidal or harmonic representation of the signal is often a recurrent procedure for either estimating spectral envelopes [2] or estimating glottal parameters [7].

Regarding the robustness of estimation, the Frequency Modulation (FM) of the fundamental frequency f_0 in an analysis window is still a major source of errors. Indeed, a window of multiple pitch periods being necessary for frequency analysis, the signal is usually assumed to be stationary and thus, f_0 is assumed to be constant inside this window. However, the f_0 variations are far from negligible. Consequently, the harmonic structure in a spectrogram is blurred due to the mismatch between the Fourier basis having constant frequencies and the modulated harmonic structure of the speech signal [8, 9]. Accordingly, the Fan-Chirp Transform (FChT) has been proposed by adapting the Fourier basis using harmonic related chirps [8]. By visual inspection of the latter, it is worth noting that what may seem to be noise using the DFT spectrogram can look like de-interlaced deterministic components using the FChT spectrogram. The actual maximum voiced frequency may be therefore higher than usually found in standard DFT analysis. The estimation of deterministic components at high frequencies using frequency demodulation is thus interesting for sinusoidal models.

To address this problem, we will use an adaptation of the frequency basis of a sinusoidal model which has been already suggested for quasi-harmonicity, namely the adaptive Quasi-Harmonic Model (aQHM) [6]. This model uses a frequency basis built by interpolation of anchor frequencies. Thanks to some properties of QHM, it is possible to correct those key values in case of frequency mismatch [10]. The initial method for analysis/synthesis of speech proposed in [11] (termed adaptive Quasi-Harmonic + Noise Model (aQHNM) thereafter) assumes that, prior to any correction, the initial components of the basis built from an f_0 curve are in a reasonable interval around the actual ones. However, any potential error of the f_0 curve is multiplied by the harmonic number. For example, an error of only 2 Hz at 100 Hz results in an error of 100 Hz at the 50^{th} harmonic, which is one harmonic above the actual frequency and obviously outside of any reasonable interval. Assuming the initial guess close enough to the actual values might be sufficient by considering deterministic components only at low frequencies and then model the upper frequencies with noise as proposed in aQHNM. However, if we want a full-band model of the speech signal in order to reveal the harmonics also at high frequencies according to the phenomenon observed with the FChT, another method is necessary. In this paper we propose a new algorithm referred to as *Adaptive Iterative Refinement* (AIR) which starts with the lowest frequencies, where the error is assumed to be reasonable, and iteratively increases the number of harmonics. Additionally, we will show that the quasi-harmonicity can be used for frequency correction and removed in the final representation of the signal. The whole proposed method will be therefore referred as aHM-AIR in order to distinguish it from the model aHM which could be used in many other ways.

In the interpolation scheme of the adaptive basis, the regularity of the anchor values have to be properly chosen according to the analyzed signal. Too many anchors may overfit the signal and represent variations which are meaningless for statistical modeling or difficult to control in voice transformation. Underfitting the signal is obviously not desired either. For speech, since we consider that the FM is related to a change of pulse duration and not to any modulation inside a single pulse, one anchor per period should be necessary. However, the position

[1]This work was supported by the Swiss National Science Foundation (PBSKP2_134325) and LISTA project (E.U. FP7 FET-OPEN, grant agreement: 25623).

of the anchors can be critical in case of erratic modulation like in creaky voice. Addressing this subject in this single paper would overcharge this presentation and we therefore fixed experimentally the number of anchor to 4 per period in this study. According to informal listening tests, this leads to satisfactory reconstruction quality as it will be shown in the evaluation. Decreasing the time resolution while keeping the synthesis quality untouched will be the subject of a forthcoming publication.

Managing transients in speech model is always problematic since the detection of voiced/unvoiced transitions and the estimation of a maximum voiced frequency is a tricky task. A unified model covering both voiced and unvoiced segments is therefore also an interesting solution. Here, the solution of the harmonic model is computed using the Least Squares (LS) method. Since this model covers spectral content with regularly spaced components, the LS solution makes also sense in a random segment (e.g. fricative), especially the adaptive model thanks to the flexibility of its non-stationary basis. The second point we want to tackle in this paper is the ability of the harmonic model to also represent properly the noise segments of the speech signal.

The next section describes the mathematical background necessary to describe the method presented in the next section. The Evaluation section then compared the modeling error using Sinusoidal Model (SM), aQHNM and aHM-AIR and the results of a listening test are presented.

2. Theoretical background

First, we assume that a fundamental frequency curve $f_0(t)$ is known a priori while considering a potential error on this curve. Then, in a single window of 3 pitch periods, we represent the speech signal using the adaptive Harmonic Model (aHM):

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

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

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

According to the adaptive scheme, $a_k(t)$ and $f_0(t)$ are obtained by interpolation of values a_k^i and f_0^i at specific instants t_i which are termed anchor values in the following. The proposed method will therefore provide estimates of these anchor parameters.

In order to do so, we suggest to use the LS solution of the adaptive Quasi-Harmonic Model (aQHM) [6]. This model, similar to the previous one, is the following:

$$s(t) = \sum_{k=-K}^{K} (a_k + tb_k) \cdot e^{jk\phi_0(t)} \tag{3}$$

where $\phi_0(t)$ is still defined by equation (2) and b_k is also a complex value. As it has been shown in [10], a_k and b_k can be used to estimate the frequency mismatch of stationary components. Specifically, for each frequency component, a correction term can be computed as:

$$df_k = \frac{f_s}{2\pi} \cdot \frac{\Re(a_k)\Im(b_k) - \Im(a_k)\Re(b_k)}{|a_k|^2} \tag{4}$$

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 has to be however in a reasonable interval around the actual frequency. The bandwidth of the main lobe of the analysis window can be used for this purpose [10].

The basic idea of the proposed iterative algorithm is the following. In a single analysis window, we can first assume an arbitrary small number of harmonics K (e.g. 4) where the guessed harmonic frequencies are assumed to be close enough to the actual ones. By computing the LS solution for (3), the correction term related to the fundamental frequency f_{cor} can be then estimated:

$$f_{corr} = \frac{1}{K}\sum_{k=1}^{K} df_k/k \tag{5}$$

The signal being real, only positive harmonics are used. The number of harmonics K, can be then iteratively updated taking into account this fundamental correction f_{corr}. Indeed, if $|f_{corr}|$ is low, this indicates that the current set of K harmonics converged to the actual values such as K can be increased in order to add new harmonics whose guessed frequencies will be in a reasonable interval around the actual ones. On the contrary, K has to be kept to its current value until successive iterations reduce $|f_{cor}|$ sufficiently such as K can be increased. To control the number of new harmonics added at each iteration, we propose to link K to f_{corr}. Indeed, the highest predicted harmonic frequency lying inside an interval of size N_w is:

$$K = \lfloor 0.5 \cdot N_w/|f_{corr}| \rfloor \tag{6}$$

where we chose N_w as the minimum between the bandwidth of the window's main lobe and f_0 (because the latter is also the distance between the harmonics). Note that instead of fixing the initial K values, we can also chose an initial fundamental error (e.g. 20 Hz) from which the initial K can be deduced using (6). Thanks to the property of frequency refinement of aQHM, $|f_{corr}|$ will be reduced progressively and K increased up to the Nyquist limit.

3. Method

The speech model used for synthesis is therefore aHM (eq. 1). We will now describe the proposed method to estimate the parameters of this model, namely the Adaptive Iterative Refinement method (AIR) which uses aQHM (eq. 3) as an intermediate model for the estimation of the correction terms df_k.

3.1. Analysis

The procedure of the analysis step consists of a parametrization of the speech signal at each anchor i at their time instant t_i. A sequence of the anchor instants is thus first created using the provided $f_0(t)$ curve such as $t_{i+1} = \frac{1}{4}f_0(t_i)^{-1} + t_i$ and $t_0 = 0$. As discussed in the introduction, 4 anchors per period allows to represent f_0 variations which depend significantly on the anchor positions like in creaky voice. In unvoiced segments, even though the estimated $f_0(t)$ is meaningless, the latter can be used to generate anchor instants. Moreover, if the distance between two anchors is short enough, aHM can model the amplitude variations of the unvoiced signal (like in plosives). Here, 20ms is used and the lower limit of the provided $f_0(t)$ curve is therefore set to 50Hz. Around each anchor time t_i, a Blackman window of 3 local pitch periods long is applied to the speech signal. $\phi_0(t)$ is then computed by means of linear interpolation of f_0^i and using (2). Using the LS solution of eq. (3) in this window, a_k^i, b_k^i are computed as well as the frequency mismatch df_k and the fundamental correction f_{corr} (eq. 5). The number of harmonics K^i is then updated using eq. (6). 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 times 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 eq. (6)
while ∃i such as f_0^i K^i < f_s/2 do
    for each anchor c do
        Create a segment of 3 periods around t_c using f_0^c
        Compute φ_0(t) using eq. (2) and interp. of all f_0^i
        Compute LS solution (a_k^c, b_k^c) of aQHM (eq. 3)
        Compute df_k (eq. 4) and f_corr = mean(df_k/k)
        Correct f_0'^c = f_0^c + f_corr
        if f_0^c K^c < f_s/2 then
            Update K^c = ⌊0.5 · N_w/|f_corr|⌋
        end if
    end for
    Set f_0^i = f_0'^i  ∀i
end while
```

Although df_k can be interpreted as a frequency mismatch, it is possible that, for some k, this correction is irrelevant (e.g. the spectrum is made of noise only). Therefore, it is necessary to check the consistency of df_k values. In our current implementation, any k-value which doesn't satisfy the following two tests is set to zero: 1) $|df_k|$ has to be smaller than $f_0/2$. Otherwise two components may be close to each other turning the LS solution unstable. 2) $kf_0 + df_k$ has to be higher than some minimal value, here 50Hz. The median value is also used to compute the mean in eq. (5) in order to avoid remaining outliers in the distribution of mismatches.

At the output of Algorithm 1, the estimated amplitude and phase values correspond to the aQHM model and not aHM which is used for synthesis. Therefore, after convergence of the algorithm, the aHM model is used in an extra iteration step to ensure the consistency between the models used in the analysis and the synthesis.

3.2. Synthesis

The synthesis procedure generates harmonic after harmonic (eq. 1) without the use of any window. We describe below the mean to generate each sinusoidal harmonic from its estimated parameters, namely its amplitudes $|a_k^i|$, its phases $\angle a_k^i$ and f_0^i.

First, the instantaneous amplitude $|a_k(t)|$ is simply obtained by means of linear interpolation across time of the anchors' amplitudes $|a_k^i|$ on a logarithmic scale. The instantaneous phase $\angle a_k^i$ cannot be interpolated directly across time to obtain $\angle a_k(t)$ because of its rotation due to the time advance between anchor instant. Consequently, we propose to first remove this effect using the integral of $f_0(t)$ from the start of the signal (eq. 2), $f_0(t)$ being obtained by linear interpolation of f_0^i:

$$\angle \tilde{a}_k^i = \angle a_k^i - k\phi_0(t_i) \quad (7)$$

Thanks to 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 therefore be interpolated to obtain its continuous counterpart $\angle \tilde{a}_k(t)$. Note that to avoid phase jumps in the interpolation (e.g. between $-\pi$ and π), real and imaginary parts are interpolated independently. Additionally, a spline or cubic interpolation is necessary such as its time-derivative, the frequency, is still continuous. To conclude, in eq. (1), $\phi_0(t)$ is obtained using eq. (2), the start of the signal being the time reference, and $a_k(t)$ is $|a_k(t)| \cdot e^{j\angle \tilde{a}_k(t)}$ whose terms are described here-above.

All along the iterative process, and since the harmonic numbers K^i increase independently from one anchor to the other, there are often missing components in the interpolations of amplitude and instantaneous phase. If a component is missing its amplitude is set to -300 dB; and $\angle \tilde{a}_k(t)$ is set to zero.

4. Evaluation

Compared to aQHNM, aHM-AIR uses a pure harmonic frequency grid. Additionally, the noise component in aHM-AIR is modeled the same way as the deterministic components without using spectrally shaped noise. It is therefore important to show that at least the reconstruction quality is kept. In the following tests, we used 12 utterances of various databases (6 female and 6 male utterances in 6 different languages, between 2s and 5s length, with sampling frequency between 16kHz and 48kHz). The samples can be found on the following web-page with their corresponding re-synthesis:
`http://gillesdegottex.eu/ExDegottex2012a`
Three methods are compared for each test: the proposed aHM-AIR, aQHNM [11] and a standard stationary sinusoidal model (SM) [5]. The initial method uses a maximum of 60 components, a maximum voiced frequency of 5.5kHz and 3 adaptation steps. The spectral envelope of the noise component is modeled using an LPC of order 18 and its time domain envelope is represented using 4 sinusoidal components. The SM method used a window of 4 periods of the median period $1/\hat{f}_0$ obtained through the initial $f_0(t)$ shared between all the methods and enough components to cover the full spectrum using the same $\hat{f}_0$ (i.e. $K = 0.5f_s/\hat{f}_0$). Finally, the step size has been set to 2.5ms for all of the methods. Since the sampling of the parameters of aHM-AIR depend initially on the fundamental frequency curve, these parameters were therefore resampled uniformly for the comparison.

4.1. SNR distributions

We first computed the distribution of the Signal to Noise Ratio (SNR) of the resynthesized signals for each method (Fig. 1). SNR are computed on voiced segments only since the methods are hardly comparable for unvoiced segments. A sliding window of 10ms with 50% overlap is used. Only the first 4kHz were taken into account because aQHNM uses a fixed number of harmonics which never reaches the Nyquist frequency. The 12 sentences were sufficient to obtain more than 4000 values for each distribution. The mean corresponding to aHM-AIR and aQHNM is higher than 10dB above SM. This corresponds to the results given in [6]. The distribution of aHM-AIR is slightly below and has more spread than the one of aQHNM. This means that aQHNM fit slightly better the harmonic content in the first 4kHz of the voiced signal. However, the listening test below shows that this reduction is not important considering the full-band of the speech signal.

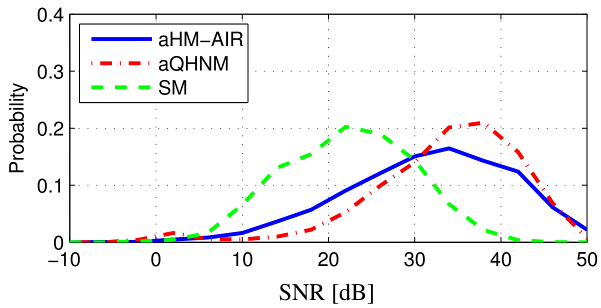

Figure 1: SNR distributions.

4.2. Listening test

A subjective test was conducted using a simple web interface. The same 12 samples were proposed to the listeners who were asked to grade the impairment of the re-synthesis after listening to the original recording according to the recommendation ITU-R BS[12]: (5)Imperceptible, (4)Perceptible but not annoying, (3)Slightly annoying, (2)Annoying, (1)Very annoying. The original recording was also added in the comparison in order to check the consistency of each answer. 22 persons answered the test and 20 have been kept because the corresponding listeners answered the test properly using headphones or earphones. Figure 2 shows the results of this listening test. Accordingly,

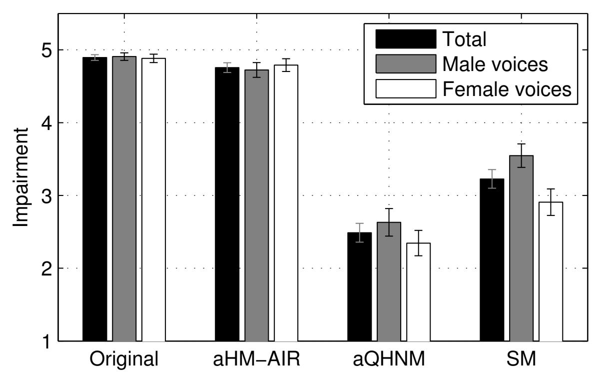

Figure 2: Impairment evaluation by 20 listeners with the 95% confidence interval.

the reconstruction provided by the proposed method is close to transparent and is clearly of better quality than aQHNM. This strong difference can be partly explained by: 1) The modeling of the noise in aQHM which can be simplistic; 2) Some erratic behavior of the frequency tracks in aQHM which are strongly constrained by the common f_0 curve. Using aHM-AIR, female and male voices seem to be also resynthesized with the same quality conversely to SM where clear differences appear between the two genders.

It is worth noting that the number of component is reduced in aHM-AIR compared to SM and aQHNM. Whereas SM and aQHNM represent the frequency grid using a frequency value for each component at each analysis instant, aHM-AIR represents the same structure using only the fundamental frequency. On the other hand, aQHNM uses only a limited number of components (here 60) and a reduced representation of mid and high frequencies with spectrally shaped noise. Specifically for the 12 sentences used above, the number of parameters between aQHNM and aHM-AIR is almost the same whereas the number of parameters in aHM-AIR represents only 65% of those of SM.

5. Conclusions

Taking advantage of an adaptive frequency basis for Harmonic Model (aHM), we proposed a full-band speech model representing both voiced and unvoiced segments up to the Nyquist frequency. According to this model, since components have to be estimated at high frequencies, we have also suggested a new algorithm for this purpose, the Adaptive Iterative Refinement (AIR). The full method is thus called aHM-AIR. Evaluations have shown that using aHM-AIR, the SNR of voiced segments is comparable to the method initially proposed for aQHM. Additionally, a listening test has shown than the quality provided by the proposed method is close to transparent. This listening test also shows that an adaptive and purely harmonic frequency grid can be used to properly represent high frequency content in both voiced and unvoiced segments. The homogeneity of this model across frequency and time is also interesting for spectral envelope estimation (for both amplitude and phase) which is a necessary step for potential applications like voice transformation and statistical speech modeling.

6. References

[1] 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.

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

[3] Y. Agiomyrgiannakis and O. Rosec, "Towards flexible speech coding for speech synthesis: an LF + modulated noise vocoder," in *Proc. Interspeech*, 2008, pp. 1849–1852.

[4] G. Degottex, A. Roebel, and X. Rodet, "Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5128–5131.

[5] 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.

[6] 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.

[7] 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.

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

[9] N. Malyska and T.F. Quatieri, "Spectral representations of non-modal phonation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 34–46, 2008.

[10] 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.

[11] Yannis Pantazis, Georgios Tzedakis, Olivier Rosec, and Yannis 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.

[12] The ITU Radiocommunication Assembly, "Itu-r bs.1284-1: En-general methods for the subjective assessment of sound quality," Tech. Rep., ITU, 2003.