ROBUST FULL-BAND ADAPTIVE SINUSOIDAL ANALYSIS AND SYNTHESIS OF SPEECH

George P. Kafentzis[1,3], Olivier Rosec[2], Yannis Stylianou[3]

[1]Orange Labs, TECH/ACTS/MAS, Lannion, France
[2]Voxygen S.A., Pole Phoenix, Pleumeur-Bodou, France
[3]Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
george.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 Model - eaQHM*, and we show that the application of a continuous f_0 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 f_0 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 classic Sinusoidal Model (SM).

Index Terms— Extended adaptive quasi-harmonic model, Speech modelling, Speech analysis, Adaptive Harmonic model, f_0 estimation

1. INTRODUCTION

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

In that context, many different approaches have been suggested over the last thirty years, in order to provide high-quality, artefact-free, flexible and compact representations of the speech signal. After the milestone work of McAulay and Quatieri on the Sinusoidal Model (SM) [3], where speech is represented as a sum of 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 [4, 13, 14, 15, 16, 17]. The most prominent representatives of these efforts that employ a sinusoidal component include the following: Stylianou [14] 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*. Other similar approaches include the work of Serra [4], where the sinusoids are no longer constrained to be harmonic, Levine [16], where multiresolutional sinusoidal modelling is employed for general audio processing, and Agiomyrgiannakis [17], who discusses the use of a harmonic plus noise representation to model the residual of an LF-based analysis.

More recently, Pantazis et al [18] 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 speech can be obtained [19]. 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* [6]. Kafentzis et al showed that including amplitude adaptation can yield higher reconstruction rates for *voiced* speech, thus obtaining the *extended adaptive Quasi-Harmonic Model - eaQHM* [20]. This adaptive scheme inspired Degottex et al [21, 22] 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 f_0 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 using such a model: first, as it is described in [22], 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 [20], the eaQHM is shown to provide highly accurate reconstruction of voiced speech, higher than the aQHM. In addition, Kafentzis et al [23] proposed the use of quasi-harmonics and 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 although the overall perceptual quality of AIR-aHM is high, it is sensitive to the f_0 estimation, as it is the case for most harmonic models.

In this paper, we extend the work presented in [20] by taking into account the latest developments in aSM and aHM and suggest a full-band, free of voicing decision, analysis-synthesis of speech based on eaQHM. The proposed system is shown to be robust in f_0 artefacts, by testing its performance using two well-known pitch estimators, called *SWIPE* [24] and *YIN* [25]. The eaQHM system assumes an initial harmonic frequency structure that successively 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 Sinusoidal Model, whereas it is superior to the recently developed AIR-aHM, especially in certain parts of speech such as unvoiced and transients.

The rest of this paper is organized as follows. Section 2 describes the eaQHM analysis and synthesis framework. Section 3 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

$$d(t) = \sum_{k=-K}^{K} A_k(t) e^{j\phi_k(t)} \tag{1}$$

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

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

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 $w(t)$ 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 modelled using a simple Harmonic Model as:

$$d(t) = \Big(\sum_{k=-L}^{L} a_k e^{j2\pi \hat{f}_k t} \Big) w(t) \tag{3}$$

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 [14]. As opposed to [6], 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 $\hat{f}_k$ values over successive analysis time instants t_i, thus obtaining

$$\hat{d}(t) = \sum_{k=-L}^{L} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{4}$$

where

$$\hat{A}_k(t) = |a_k(t)| \tag{5}$$

and

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} k\hat{f}_0(u) du, \qquad \hat{\phi}_k(t_i) = \angle a_k(t_i) \tag{6}$$

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 [20], by using the parameters a_k and b_k of the Quasi-Harmonic Model (QHM) [26]. This yields the eaQHM model:

$$d(t) = \left(\sum_{k=-L}^{L} \Big(a_k + t b_k\Big) \Big(\hat{A}_k(t) e^{j\hat{\phi}_k(t)}\Big) \right) w(t) \tag{7}$$

with

$$\hat{A}_k(t) = \frac{\hat{A}_k(t + t_i)}{\hat{A}_k(t_i)} \tag{8}$$

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 [20]. 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 [26], 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

$$\hat{\eta}_k = \frac{f_s}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2} \tag{9}$$

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

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_i) + \frac{2\pi}{f_s} \int_{t_i}^{t} \hat{f}_k(u) du \tag{10}$$

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 $\hat{\eta}_k$ terms. By iteratively adding the $\hat{\eta}_k$ 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 [6, 19], where the frequency correction estimation $\hat{\eta}_k$ on each adaptation should be less than $f_0/2$, 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

$$|\hat{\eta}_k(t_i)| \leq \frac{\hat{f}_0(t_i)}{m+1} \tag{11}$$

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

$$SRER = 20\log_{10}\frac{std(d(t))}{std(d(t)-\hat{d}(t))} \tag{12}$$

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 continuation over frame boundaries as described in [19]. Finally, the speech signal can be approximated by its time-varying components using:

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

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

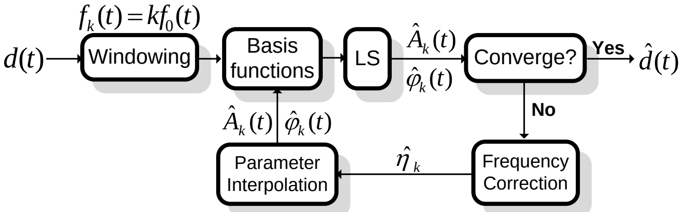

Fig. 1. *Block diagram of the eaQHM system.*

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 first one, called *SWIPE*, has been introduced in [24], and a description of the second one, called *YIN*, can be found in [25].

In objective evaluation, the SRER is computed for the whole waveform, serving as an estimate of the total residual energy "missed" by each model. The higher the SRER value, the more information is captured by the model used.

In subjective evaluation, a formal listening test has been conducted in order to measure perceptual quality. In these experiments, a database of 32 speech utterances was used, including 16 male and 16 female speakers from 16 different languages: Greek, French, English, Spanish, Finnish, Chinese, Portuguese, Basque, Japanese, Italian, German, Korean, Russian, Arabic, Indonesian, and Turkish. All waveforms were sampled at 16 kHz.

The parameters for the models were the following: for SWIPE and YIN pitch estimators, the pitch was estimated every 1 ms and its fundamental frequency estimation limits were $[70, 220]$ Hz and $[120, 350]$ Hz for males and females, respectively. A median smoothing was performed after estimation to suppress outlier estimates. 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 $2K + 1$ parameters per synthesis frame are used in *all* models (A_k, ϕ_k), 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.

SRER Performance				
Model	Speakers			
	SWIPE		YIN	
	Males	Females	Males	Females
SM	18.6(1.90)	18.6(3.64)	14.3(2.20)	16.2(3.28)
aHM	23.9(2.66)	18.9(3.27)	23.9(2.61)	19.9(3.05)
eaQHM	34.5(2.39)	30.9(3.00)	34.4(2.45)	30.7(3.19)

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.*

Figure 2 shows the first 16 frequency tracks in the analysis step for an utterance produced by Greek male speaker, the local SRER for a sliding window of 30 ms, and the corresponding speech waveform. It should be noted that the overall SRER for eaQHM is 34.67 dB whereas for the aHM is 25.60 dB for this sample, which contains both voiced and unvoiced areas. In this figure, it is obvious that in AIR-aHM all components are purely harmonic, and any slight fluctuation of the f_0 propagates in the higher harmonics. In eaQHM however, the upper frequency components deviate from the multiples of the f_0 and their structure seems smoother. Based on the lower panel (time-varying SRER), it seems that the representation suggested by eaQHM (middle panel) is more accurate compared to that one obtained by aHM (upper panel). Also, it should be mentioned that in our experiments, no manual refinement of the estimated f_0 is performed.

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 [27], with each scale being (1) "Very bad", (2) "Bad". (3) "Good", (4) "Very good", (5) "Perfect". The results from 18 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 4 were familiar with signal processing and listening tests.

[1] http://www.csd.uoc.gr/~kafentz/listest/pmwiki.php?n=Main.EAQHM-LT

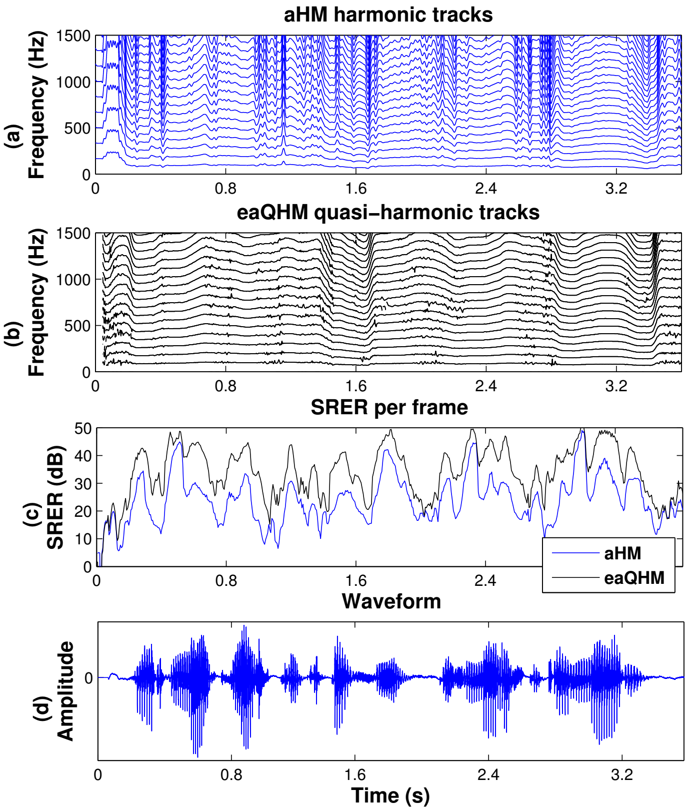

Fig. 2. *Analysis data of a Greek male speaker for both adaptive models: (a) aHM tracks, (b) eaQHM tracks, (c) Local SRER for both models over time, (d) Speech waveform.*

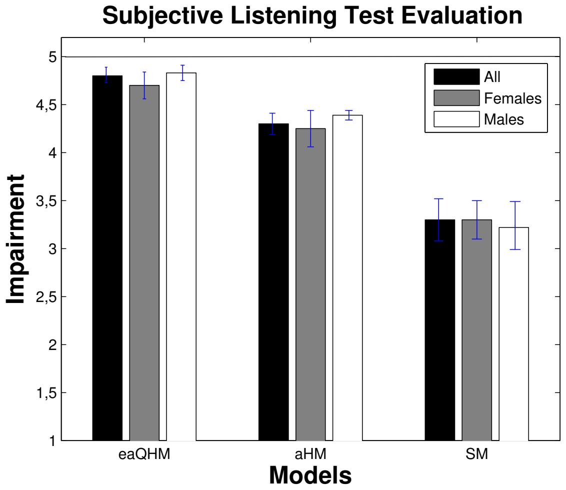

Fig. 3. *Impairment evaluation of the resynthesis quality between the original recording and the reconstructions with all three models, with the* 95% *confidence intervals.*

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.

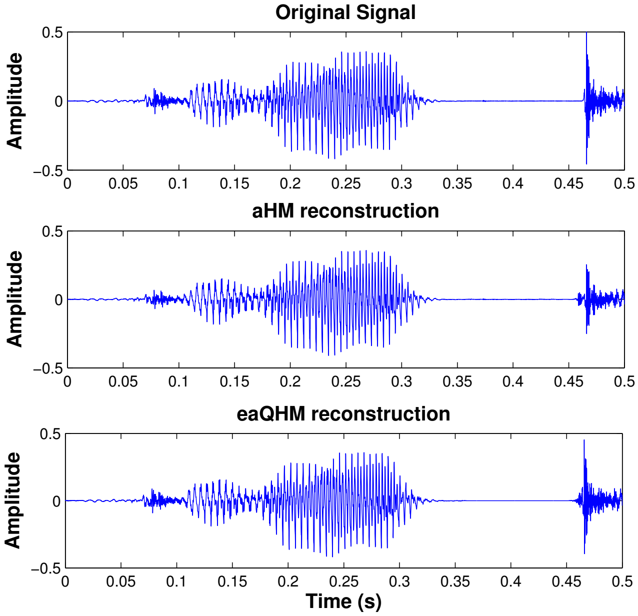

Fig. 4. *Speech utterance (/krɔk^{h}ɛ/) in Korean language by a female subject. Upper panel: Original signal, Middle panel: aHM reconstruction, Lower panel: eaQHM reconstruction.*

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 a 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

[1] R. J. Mcaulay and T. F. Quatieri. Low-rate speech coding based on the sinusoidal model. In S. Furui and M. M. Sondhi, editors, *Advances in Speech Signal Processing*. Marcel Dekker Inc., New York, 1992.

[2] Sassan Ahmadi and Andreas S. Spanias. Low bit-rate speech coding based on an improved sinusoidal model. *Speech Communication*, 34(4):369 – 390, 2001.

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

[4] X. Serra. *A System for Sound Analysis, Transformation, Synthsis based on a Determistic plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[5] J. Laroche Y. Stylianou and E. Moulines. HNM: A Simple, Effecient Harmonic plus Noise Model for Speech. In *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 169–172, New Paltz, NY, USA, Oct 1993.

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

[7] Michael E. Deisher and Andreas S. Spanias. Speech enhancement using state-based estimation and sinusoidal modeling. *The Journal of the Acoustical Society of America*, 102(2):1141–1148, 1997.

[8] J. Jensen and J.H.L. Hansen. Speech enhancement using a constrained iterative sinusoidal model. *Speech and Audio Processing, IEEE Transactions on*, 9(7):731–740, 2001.

[9] E. Zavarehei, S. Vaseghi, and Qin Yan. Noisy speech enhancement using harmonic-noise model and codebook-based postprocessing. *IEEE Trans. on Audio, Speech and Lang. Processing*, 15(4):1194–1203, 2007.

[10] Y. Stark and J. Tabrikian. MMSE-based speech enhancement using the harmonic model. In *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of*, pages 626–630, 2008.

[11] T.F. Quatieri and R.J. McAulay. Shape-Invariant Time-Scale and Pitch Modifications of Speech. *IEEE Trans. on Acoust., Speech and Signal Processing*, 40:497–510, 1992.

[12] J. Laroche Y. Stylianou and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.

[13] D. W. Griffin and J. S. Lim. Multiband Excitation Vocoder. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36(8):1223–1235, 1988.

[14] Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, E.N.S.T - Paris, 1996.

[15] A. J. Abrantes, J. S. Marques, and I. Trancoso. Hybrid sinusoidal modeling of speech without voicing decision. In *EUROSPEECH*. ISCA, 1991.

[16] S. Levine. *Audio Representations for Data Compression and Compressed Domain Processing*. PhD thesis, Stanford University, 1999.

[17] Y. Agiomyrgiannakis and O. Rosec. ARX-LF-based source-filter methods for voice modification and transformation. In *Proc. IEEE ICASSP*, Taipei, Taiwan, Apr 2009.

[18] Y. Pantazis. *Adaptive AMFM Signal Decomposition With Application to Speech Analysis*. PhD thesis, Computer Science Department, University of Crete, 2010.

[19] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AMFM signal decomposition with application to speech analysis. *IEEE Trans. on Audio, Speech, and Lang. Proc.*, 19:290–300, 2011.

[20] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou. An Extension of the Adaptive Quasi-Harmonic Model. In *Proc. IEEE ICASSP*, Kyoto, March 2012.

[21] G. Degottex and Y. Stylianou. A full-band adaptive harmonic representation of speech. In *Interspeech*, Portland, Oregon, U.S.A, 2012.

[22] G. Degottex and Y. Stylianou. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Trans. on Audio, Speech, and Language Processing*, 21(10):2085–2095, 2013.

[23] G. P. Kafentzis, O. Rosec, and Y. Stylianou. On the Modeling of Voiceless Stop Sounds of Speech using Adaptive Quasi-Harmonic Models. In *Interspeech*, Portland, Oregon, USA, September 2013.

[24] A. Camacho and J. G. Harris. A sawtooth waveform inspired pitch estimator for speech and music. *J. Acoust. Soc. Am.*, 124:1628–1652, 2008.

[25] A. de Cheveigne and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[26] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Interspeech*, Brisbane, Sep 2008.

[27] The ITU Radiocommunication Assembly. ITU-R BS.1284-1: EN-general methods for the subjective assessment of sound quality. Technical report, ITU, 2003.