# TIME-SCALE MODIFICATIONS BASED ON A FULL-BAND ADAPTIVE HARMONIC MODEL

*George P. Kafentzis[1,3], Gilles Degottex[2,3], Olivier Rosec[4], and Yannis Stylianou[2,3]*

[1]Orange Labs, TECH/ACTS/MAS, Lannion, France
[2]Institute of Computer Science, Foundation for Research and Technology Hellas, Greece
[3]Multimedia Informatics Lab, Computer Science Department, University of Crete, Greece
[4]Voxygen S.A., Pole Phoenix, Pleumeur-Bodou, France

kafentz@csd.uoc.gr, degottex@csd.uoc.gr, olivier.rosec@voxygen.fr, styliano@ics.forth.gr

## ABSTRACT

In this paper, a simple method for time-scale modifications of speech based on a recently suggested model for AM-FM decomposition of speech signals, is presented. This model is referred to as the adaptive Harmonic Model (aHM). A full-band speech analysis/synthesis system based on the aHM representation is built, without the necessity of seperating a deterministic and/or a stochastic component from the speech signal. The aHM models speech as a sum of harmonically related sinusoids that can adapt to the local characteristics of the signal and provide accurate instantaneous amplitude, frequency, and phase trajectories. Because of the high quality representation and reconstruction of speech, aHM can provide high quality time-scale modifications. Informal listenings show that the synthetic time-scaled waveforms are natural and free of some common artifacts encountered in other state-of-the-art models.

***Index Terms***— Time-scale modifications, Adaptive quasi-harmonic model, Speech modeling, Harmonic model, Speech analysis, Adaptive Harmonic model

## 1. INTRODUCTION

In a great variety of speech applications, prosodic (i.e. time and pitch scale) modifications are required. From film industry, entertainment, and communications, to text-to-speech synthesis and pathological voice restoration, prosodic modifications have received increasing attention and have been thoroughly studied by the speech processing comminuty.

As a result, a number of time-scaling techniques have been proposed in speech literature, based on the corresponding analysis/synthesis models. These typically belong to two, different but not distinct, classes: parametric and non-parametric approaches. The latter include frequency domain and time domain PSOLA [1] and its variants, such as WSOLA [2] and MBR-PSOLA [3], and the phase vocoder-based techniques [4] [5]. Parametric techniques include narrowband models, such as the Sinusoidal Model (SM) developed by McAulay and Quatieri [6], and the Harmonic + Noise Model (HNM) [7] of Stylianou, and wideband models, which typically include the LF-ARX based source-filter methods of Agiomyrgiannakis and Rosec [8], the STRAIGHT method [9] of Kawahara, the GSS [10] of Cabral et al, and the SVLN [11] of Degottex. All these approaches provide high quality prosodic modifications. Among them, hybrid representations such as in [7] are considered well suited for prosodic modifications, since a well-manipulated separation of speech into a deterministic and a stochastic component leads to a better manipulation of the components and that aids to an enhanced quality of speech synthesis.

However, all these models share a common assumption; that speech is locally stationary, a fact that is not valid, since speech signals exhibit local nonstationarities, within an analysis window, both in amplitude and in phase. Attempts to solve this issue have been proposed, such as the use of small analysis windows [7] or a linear evolution of fundamental frequency [12]. To this direction, Quasi-Harmonic Model (QHM) evinced the ability to correct, in the least squares sence, contingent frequency estimation errors [13]. This way, amplitude and phase estimation bias due to frequency mismatches are alleviated. Even so, local nonstationarity is only partially addressed in QHM. It was shown in [14] that an adaptive sinusoidal model, called adaptive Quasi-Harmonic Model (aQHM) is able to efficiently tackle local phase nonstationarity, and in [15], an extension to include amplitude nonstationarity is proposed, referred to as the extended aQHM (eaQHM). Together, we propose to be referred to as the *adaptive Sinusoidal Models (aSMs)*, and all are achieved by estimating the frequency (and the amplitude, for eaQHM) trajectories of the deterministic part and then re-estimating the parameters using a new set of time-varying frequency (and amplitude, for eaQHM) basis functions. Thus, the model adapts to the local characteristics of the analyzed speech signal and a more accurate representation is attained. Using the adaptive sinusoidal models, an approach similar to that in [7] is followed in [16] for the analysis; a decomposition of speech into two bands is performed: a lower band, which represents the deterministic part and is modeled as a sum of quasi-harmonically related sinusoids using aQHM, and an upper band, which represents the stochastic part and is modeled by time and frequency modulated gaussian noise. This decomposition results in an adaptive Quasi-Harmonic + Noise (aQHNM) analysis and synthesis system.

However, such an approach has some disadvantages - the first one is that the seperation of the determenistic and the stochastic part can be tricky. The so-called *transient* areas of speech need special treatment and their inclusion or exclusion (in whole or part of them) in the noise part can significantly degrade the resulting transformed signal. Moreover, the noise part can be adequately modeled using a variety of techniques, such as modulated noise, but still does not attain the quality of the original waveform. So, a simple, full-band representation would be preferable, and it was shown in [17] that such an approach can be used, thus providing synthetic speech that is perceptually indistinguishable from the original waveform. Ths model is called *adaptive Harmonic Model (aHM)*, and it uses a similar analysis strategy as in aQHM, but reduces to strict harmonicity

in the final representation of the signal, as it will be shown in Section 2.

Based on that representation, a simple and flexible technique for time-scale modifications is presented in this paper. The aHM provide high resolution parameter trajectories which can be simply stretched or compressed in time, without a seperate manipulation of noise parts of speech. The time-scale modified signal can be synthesized in a manner similar to the non-modified signal, as it will be shown in Section 3. The time-scaled signal sounds free of artifacts, such as "metallic" quality, chorusing, or musical noise, typically encountered in other state-of-the-art modification algorithms. Although the model is simple, its performance is superior to certain state-of-the-art methods (HNM, WSOLA), for moderate time-scaling factors (0.5 to 2.5).

The rest of the paper is organized as follows. In Section 2 we will review the analysis and synthesis steps of aHM. Section 3 provides the time-scale modification scheme for the model in hand. Section 4 demonstrates an example of application and Section 5 discusses the results of the comparison with the well-known Harmonic Plus Noise model (HNM) [7], and with a non-parametric approach, called WSOLA [18]. Finally, Section 6 concludes the paper.

## 2. DESCRIPTION OF AHM-BASED ANALYSIS/SYNTHESIS SYSTEM

In this section, a brief review of the adaptive Harmonic Model (aHM) is presented [17], along with a short description of the analysis and synthesis schemes.

### 2.1. The adaptive Harmonic Model - aHM

The adaptive Harmonic Model can be mathematically described as:

$$s(t) = \sum_{k=-K}^{K} a_k(t) e^{jk\phi_0(t)} \tag{1}$$

where $a_k(t)$ is a complex function that copes with the amplitude and the instantaneous phase of the $k^{th}$ harmonic component, while $K$ is the number of the components, and $\phi_0(t)$ is a real function defined as the integral of the fundamental frequency $f_0(t)$:

$$\phi_0(t) = \int_0^t 2\pi f_0(u) du \tag{2}$$

### 2.2. Analysis

In the analysis step, a parametrization of the speech signal at each analysis time instant $t_a^i$ is undertaken. At first, a sequence of the analysis time instants are created in the voiced parts of speech using the provided $f_0(t)$ track, such we have one analysis time instant per pitch period. In unvoiced segments, even though the estimated $f_0(t)$ is meaningless, it can be used to generate the corresponding analysis time instants. Moreover, if the distance between $t_a^i$ and $t_a^{i+1}$ is short enough, aHM can model the amplitude variations of the unvoiced signal (like in plosives). Thus, the upper limit of the size of the analysis window is 20ms and the lower limit comes from the provided $f_0(t)$ track, and is therefore set to 50Hz. Around each analysis time instant $t_a^i$, a Blackman window with a length of 3 local pitch periods is applied to the speech signal. The phase track $\phi_0(t)$ is then computed by means of spline interpolation of $f_0^i$ and using the integration formula in Eq.(2).

### 2.3. Adaptive Iterative Refinement - AIR

The fundamental frequency track of Eq.(2) is assumed to be known beforehand and can have a potential error, i.e.

$$\eta_0 = f_0 - \hat{f}_0 \tag{3}$$

that is called *frequency mismatch*, where $f_0$ is the actual fundamental frequency at a certain time instant and $\hat{f}_0$ is an estimate of the latter. Following the adaptive scheme presented in [14], the amplitude $a_k(t)$ and fundamental frequency $f_0(t)$ values are obtained by a linear interpolation, respectively, of their values, $a_i^k$ and $f_0^i$, at the analysis time instants, $t_a^i$. In order to have an estimate of these values, the *adaptive Quasi-Harmonic Model - aQHM* is used, that is given by the following equation:

$$s(t) = \sum_{k=-K}^{K} (a_k + t b_k) e^{jk\phi_0(t)} \tag{4}$$

where $\phi_0(t)$ is the same as in 2, $a_k$ and $b_k$ are the complex amplitude and the complex slope of the model, respectively, and $K$ is again the number of the components. It has been shown in [13] that $a_k$ and $b_k$, that are obtained via a Least Squares minimization, can be used to provide an estimate, $\hat{\eta}$, for the frequency mismatch of Eq.(3). Thus, for the $k^{th}$ component in general, this can be computed as:

$$\hat{\eta}_k = \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2} \tag{5}$$

where $a_k^R, b_k^R$ and $a_k^I, b_k^I$ are the real and imaginary parts, respectively, of the complex amplitude and the complex slope of the model. Using this estimate, the fundamental frequency values $f_0^i$ can be updated in an iterative manner. However, as it is shown in [14], this term cannot be larger than the main lobe of the analysis window.

In [17], an iterative algorithm has been proposed to update the frequencies. Its main idea is discussed here. In a single analysis window, an arbitrary small number of harmonics $K$ (e.g. 4)can be assumed. These harmonics are considered not to vary too much from their actual values, i.e. the mismatch $\eta$ is small. By computing the LS solution for Eq.(4), the correction term, $\eta_0$, related to the fundamental frequency $f_0$ can be then estimated by the following equation:

$$\hat{\eta_0} = \frac{1}{K} \sum_{k=1}^{K} \frac{\hat{\eta}_k}{k} \tag{6}$$

This estimaton can be furtherly used to update the number of harmonics, $K$. If $\hat{\eta_0}$ is small, this means that the current set of harmonics have converged very close enough to their actual values. Then, $K$ can be further increased to add new harmonics in a new set of harmonics. If $\hat{\eta_0}$ is large, then the current set of harmonics have not converged to their actual values and further iterations are necessary to successively reduce $\hat{\eta_0}$. The number of harmonics that are added in each iteration are given by the following equation:

$$K = \left\lfloor \frac{\frac{1}{2} N_w}{|\hat{\eta_0}|} \right\rfloor \tag{7}$$

where $N_w$ is given by $N_w = min\{B_w, f_0\}$, where $B_w$ is the bandwidth of the main lobe of the analysis window. Using the LS solution of Eq.(4), the local parameters $a_k^i, b_k^i$ are computed, along with the $k^{th}$ frequency mismatch, $\hat{\eta}_k$, and the fundamental frequency correction, $\hat{\eta}_0$. The number of harmonics, $K^i$, is then updated using Eq.(7). As a last step, the process is repeated for all frames until the

Nyquist frequency is reached for all frames. This approach is termed as the *Adaptive Iterative Refinement - AIR* and a pseudocode for it is given in [17].
It should be noted that the estimated amplitude and phase values that are obtained at the analysis step correspond to the aQHM model and not aHM which is used for synthesis. Therefore, the aHM model is used in a last iteration step to ensure the consistency between the models used in the analysis and the synthesis.

### 2.4. Synthesis

In the synthesis step, each harmonic is generated in seperate, one after the other, without using any window. Each harmonic component is synthesized by its parameters, namely its amplitudes $|a_k^i|$, phases $\angle a_k^i$, and fundamental frequency $f_0^i$. First, the instantaneous amplitude, $|a_k(t)|$, of the $k^{th}$ harmonic is simply obtained by linearly interpolating the estimated $|a_k^i|$ on the analysis time instants $t_a^i$, on a logarithmic scale. The instantaneous phase $\angle a_k^i$ cannot be interpolated directly across time to obtain $a_k(t)$ because of its rotation due to the time advance between analysis time instants. Therefore, it is proposed to remove this effect using the integral of $f_0(t)$ from the start of the signal, and obtain the *relative phase - RP*:

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

Thus, by assuming that the shape of the signal is changing smoothly, the phase values change also smoothly from one analysis time instant to the other. Then, the RP $\angle \tilde{a}_k^i$ can be interpolated to obtain its continuous counterpart, $\angle \tilde{a}_k(t)$. Additionally, a spline or cubic interpolation is necessary such as its time derivative, the frequency, is still continuous. All along the iterative process, and since the harmonic numbers $K^i$ increase independently from one analysis time instant to the other, there are often missing components in the interpolations of amplitude and instantaneous phase. If this is the case, then the amplitude of the missing component is set to $-300$ dB and the corresponding phase $\angle \tilde{a}_k(t)$ is set to zero.

## 3. TIME-SCALE MODIFICATION SCHEME

The purpose of time-scale modification is to maintain the perceptual quality of the original speech signal while changing the apparent rate of articulation. On the contrary to most parametric state-of-the-art systems, there is *no* seperate modification of the deterministic and the stochastic part, since the aHM approach is full-band.

The pitch contour (and thus the harmonics) should be stretched or compressed in time, and the formant structure should be changed at a slower or faster rate than the rate of the input speech, but otherwise not modified. For an arbitrary time-scale modification, the time $t$ in the original signal is mapped to a time $t'$ in the modified signal. For that, a mapping function reffered to as *the time-scale warping function* is defined:

$$D(t) = \int_0^t \beta(\tau) d\tau \tag{9}$$

where $\beta(\tau) > 0$ is the time-varying time-scaling rate. When $\beta(\tau) > 1$, then the articulation rate is slowed down, whereas the opposite happens when $\beta(\tau) < 1$. Note that for a fixed $\beta(\tau) = \beta$, then the time-scale warping function is reduced to a linear function of time, i.e. $D(t) = \beta t$.
In the adaptive sinusoidal model context, the paramaters should be transformed in the way described next. Let us remind that in an analysis window centered at $t_k^i$, the instantaneous components $\{a_k^i, f_0^i\}$, are known. From these, we can compute their continuous counterparts, which are the instantaneous amplitudes $A_k(t) = |a_k(t)|$ and frequencies $f_0(t)$, obtained by interpolating $a_k^i$ and $f_0^i$, respectively. Then, the time-scaled waveform, $s_{TS}(t)$, for a constant time-scale factor is given by:

$$\hat{s}_{TS}(t') = \sum_{k=-K}^{K} \hat{A}'_k(t') e^{j\hat{\phi}'_k(t')} = \sum_{k=-K}^{K} \hat{A}_k(\beta^{-1}t) e^{j\hat{\phi}_k(\beta^{-1}t)} \tag{10}$$

Thus, time scaling requires the following steps to be performed:

1. The instantaneous amplitudes are time-scaled:

$$A'_k(t') = A_k(D^{-1}(t)) \tag{11}$$

2. The instantaneous frequencies in the modified signal at time $t'$ correspond to the instantaneous frequency in the original signal at time $D^{-1}(t')$:

$$kf'_0(t') = kf_0(D^{-1}(t')) \tag{12}$$

where $D^{-1}(t)$ is the inverse time-scale warping function. For the interpolation of the phase, the relative phase, RP, is first computed by extracting the integral of the frequency from the phase information at analysis time instant $t_a^i$, as in Eq.(8). Then, the RP values are interpolated, thus obtaining $\angle \tilde{a}_k(t')$, and the integrated time-scaled frequency is added back to the interpolated RP values:

$$\hat{\phi}'_k(t) = \angle \tilde{a}_k(t') + \int_0^{t'} 2\pi k \hat{f}'_0(u) du \tag{13}$$

## 4. EXAMPLE OF APPLICATION

Time scale modification for a factor of $1.5$ is applied on a speech signal sampled at 16 kHz. Figure 1 shows the original speech signal (upper panel) and the time-scaled signal for a factor of $1.5$ (lower panel). It can be seen that the time-stretched signal preserves the shape of the original waveform, in all of its parts (fricative, transient, and voiced). Moreover, the first $50$ harmonic frequency trajectories $k\hat{f}_0(t)$ and $k\hat{f}'_0(t')$ are depicted in Figure 2 in the upper and lower panel, respectively, for the same pair of signals.

## 5. DISCUSSION AND RESULTS

It has already been shown in [17] that aHM-based synthetic speech outperforms state-of-the-art methods, such as the Sinusoidal Model (SM) [6] and the adaptive Quasi-Harmonic+Noise Model (aQHNM) [14]. In this case, informal listenings have been conducted to examine the performance of our modification scheme and two well-known state-of-the-art approaches, a parametric and a non-parametric one: the HNM approach [7] and the WSOLA technique [2]. The time-scale modification factors were selected to be $0.5, 0.8, 1.2, 1.5, 2.0$, and $2.5$, which are typical values for moderate speech prosodic modifications. A database consisting of $15$ male and $15$ female clear speech recordings was selected, with speakers of different languages (Greek, German, Italian, Japanese, French, and American).
The sampling frequency of all waveforms is 16 kHz. For the HNM, the maximum voiced frequency is fixed to $5500$ Hz, and the analysis is pitch synchronous. The analysis window size is two local pitch periods. The order of the AR filter for the noise part is set to

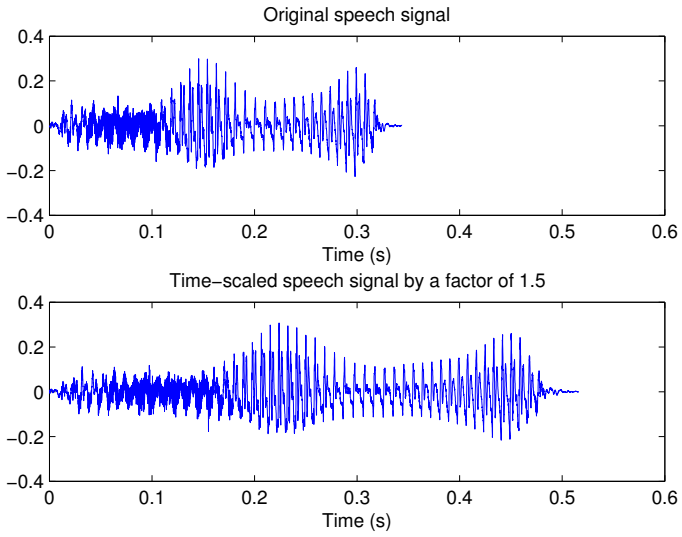


**Fig. 1**. *Original signal (upper panel) and time-scaled signal (lower panel) for a factor of* 1.5.

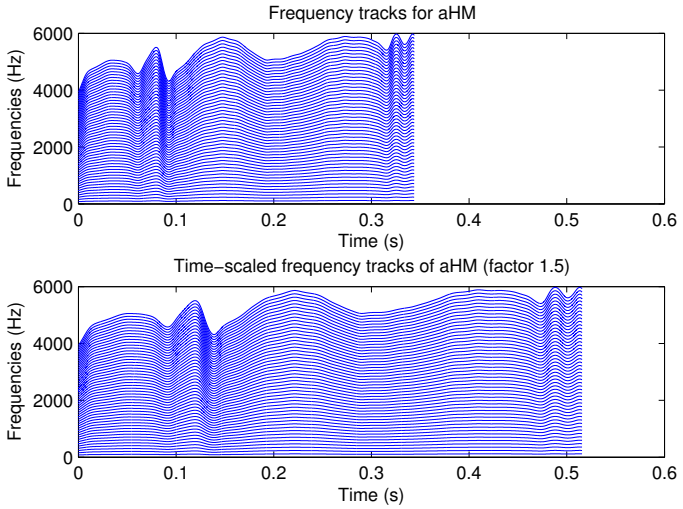


**Fig. 2**. *First* 50 *frequency trajectories for the original signal (upper panel) and time-scaled frequency trajectories (lower panel) for a factor of* 1.5.

20. The parameters of aHM are the ones described in the previous section. For the WSOLA, an analysis window length of 15 ms is used. A tolerance variable $\Delta$ (a tolerance factor on the desired time-warping function to ensure signal continuity at segment joins) of 7 ms is selected, which according to [2], usually produces high-quality time-scaled speech.

In general, the participants acknowledged the proposed method natural and free of common artifacts, such as "metallic" quality, chorusing, or musical noise. Although the model is simple, it is shown to perform similarly or even better than the - more complex - HNM, for speech prosody modifications, especially in voiced parts of speech, where the well-known problem of *lack of presence* is addressed. Please note that HNM decomposes speech into a deterministic and a stochastic component, and although it shares the harmonicity assumption in its deterministic component, it handles differently its stochastic part (modulated noise). In our WSOLA samples, a step effect in the amplitude of the time-scaled speech was observed, that led to audible artifacts. No such artifacts were present in the aHM time-scaled samples. Finally, it should be noted that although WSOLA performs quite close to aHM and is much faster, it is completely inappropriate of providing higher level representations of speech (i.e. spectral envelopes).

## 6. CONCLUSIONS AND FUTURE WORK

A time-scale modification scheme based on the recently developed adaptive Harmonic Model (aHM) analysis/synthesis system is presented. The system utilizes a full-band representation of speech based on quasi-harmonic analysis and strict harmonic synthesis. The model itself results in very high reconstruction quality for both voiced and unvoiced parts [17]. This scheme provides flexibility in time scaling modifications, avoiding the separation of speech into determenistic and stochastic components. Listenings result in that time-scale modifications are of very good quality, compared to the HNM and WSOLA approach. Since pitch scaling is also of great importance in speech modifications, future work will focus on designing a scheme for pitch and frequency scaling, taking advantage of the high quality frequency estimation provided by the aHM.

## 7. REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, 1990.

[2] W. Verhelst and M. Roelands, "An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for High Quality Time-Scale Modification of Speech," *ICASSP*, pp. 554–557, 1993.

[3] T. Dutoit and H. Leich, "Improving the td-psola text-to-speech synthesizer with a specially designed mbe re-synthesis of the segments database," in *EUSIPCO*, 1992, pp. 343–347.

[4] J. Laroche and M. Dolson, "Imroved Phase Vocoder Time-Scale Modification of Audio," in *IEEE Trans. on Speech and Audio Processing*, vol. 7, May 1999, pp. 323–332.

[5] P. Depalle and G. Poirot, "SVP: A Modular System for Analysis, Processing and Synthesis of Sound Signals," *Proceeding of the 1991 International Computer Music Conference*, 1991.

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

[7] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, E.N.S.T - Paris, 1996.

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

[9] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. IEEE ICASSP*, Munich, Apr 1997, pp. 1303–1306.

[10] J. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Interspeech*, Brisbane, Australia, 2008, pp. 1829–1832.

[11] G. Degottex, "Glottal source and vocal-tract separation," Ph.D. dissertation, UPMC-Ircam, France, 2010.

[12] A. Robel, "Parameter Estimation for Linear AM-FM Sinusoids using Frequency Domain Demodulation," *Signal and Image Processing*, pp. 162–166, 2007.

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

[14] ——, "Adaptive AMFM signal decomposition with application to speech analysis," *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, pp. 290–300, 2011.

[15] 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.

[16] 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.

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

[18] M. Demol, W. Verhelst, K. Stuyve, and P. Verhoeve, "Efficeint non-uniform time-scaling of speech with wsola," *Int. Conf. on Speech and Computers (SPECOM)*, pp. 163–166, 2005.