PITCH MODIFICATIONS OF SPEECH BASED ON AN ADAPTIVE HARMONIC MODEL

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

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

kafentz@csd.uoc.gr, degottex@csd.uoc.gr, olivier.rosec@voxygen.fr, yannis@csd.uoc.gr

ABSTRACT

In this paper, a simple method for pitch-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). The aHM models speech as a sum of harmonically related sinusoids that can adapt to the local characteristics of the signal. It was shown that this model provides high quality reconstruction of speech and thus, it can also provide high quality pitch-scale modifications. For the latter, the amplitude envelope is estimated using the Discrete All-Pole (DAP) method, and the phase envelope estimation is performed by utilizing the concept of relative phase. Formal listening tests on a database of several languages show that the synthetic pitch-scaled waveforms are natural and free of some common artefacts encountered in other state-of-the-art models, such as HNM and STRAIGHT.

Index Terms— Pitch modification, Speech analysis, Adaptive quasi-harmonic model, Adaptive harmonic model

1. INTRODUCTION

Pitch modification is defined as the change of the fundamental frequency of speech while preserving the short time envelope characteristics and the duration of a speech signal. Applications of pitch modifications range from entertainment, communications, and film industry, and extend up to text-to-speech synthesis, pathological voice restoration, and high-end hearing aids.

As a result, several pitch-scaling techniques have been proposed in literature. Typically, they belong to two, different but not distinct, categories: parametric and non-parametric techniques. The latter include frequency domain and time domain PSOLA [1] and MBR-PSOLA [2], and the phase vocoder-based techniques [3, 4]. The former include narrowband models, such as the Sinusoidal Model (SM) [5], the Harmonic + Noise Model (HNM) [6], and wideband models, which typically include the LF-ARX based source-filter methods [7, 8], the STRAIGHT method [9], the GSS [10], and the SVLN [11] methods. All these approaches provide high quality prosodic modifications. Among them, hybrid representations such as in [6, 8] are considered well suited for prosodic modifications, since a well-estimated 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 resynthesized speech.

Recent advances in signal modelling revealed that a family of *adaptive Sinusoidal Models (aSMs)* [12, 13] are able to efficiently tackle local amplitude and phase non-stationarities. The adaptation of these models is achieved by estimating the amplitude and frequency trajectories of the signal using Least-Squares (LS) and then re-estimating the parameters using a new set of amplitude and frequency varying basis functions. Thus, a more accurate representation of the analyzed speech signal is attained. Using this adaptive scheme, it was shown in [14] that such an approach can be used in a full-band harmonic analysis/synthesis system, thus providing synthetic speech that is perceptually indistinguishable from the original waveform. This model is called the *adaptive Harmonic Model (aHM)*, and it uses a similar analysis strategy as in aSMs, but reduces to strict harmonicity in the final representation of the signal, as it will be shown in Section 2. This model has been successfully applied in time-scaling of speech [15].

In pitch-scaling applications, the estimation of a new set of amplitude, frequency, and phase values is necessary due to pitch shifting. These values can be obtained by estimating the so-called *amplitude and phase envelopes* in the spectral domain. Spectral estimation is a field of study that has received increased attention because of the variety of its applications (voice conversion [16], word recognition [17], speech recognition [18], speaker verification [19], speaker identification [20], to name a few), and many algorithms are available to achieve it in a robust manner, such as cepstrum-based techniques [21, 22], AR models [23, 24, 25], and multi-frame analysis [26, 27]. In this work, the Discrete All-Pole (DAP) method is used [23]. For the phase envelope, the estimated principal values of the phase should be unwrapped in time and frequency domain. To this purpose, a simple approach similar to the one in [28] is suggested, which involves the computation and the interpolation of the *relative phase*. The method is described in detail in Section 3.

Although hybrid approaches have been successfully applied in prosody modifications, they do have some disadvantages - the first one is that the separation of the deterministic and the stochastic part can be problematic. 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 so-called noise part can be adequately modelled using a variety of techniques, such as modulated noise, but still does not attain the quality of the original waveform. So, a simple, robust, full-band representation would be preferable. Based on the aHM representation, a simple and flexible technique for pitch-scale modifications is presented in this paper. The aHM provide high resolution parameter trajectories which can be simply shifted in frequency, using appropriate envelope estimations for amplitudes and phases. The pitch-scale modified signal can be synthesized in a manner similar to the non-modified signal, as it will be shown in Section 3. The pitch-scaled signal sounds free of artefacts, such as "metallic" quality, chorusing, or musical noise. Based on formal listening tests it is shown that despite the simplicity of the model, its performance is comparable to certain state-of-the-art, but far more complex, meth-

G. Degottex was funded by the Swiss National Science Foundation (SNSF) (grants PBSKP2_134325, PBSKP2_140021)

ods (STRAIGHT, HNM). It should be noted that a recent work on aHM-based prosody modifications has been presented in [29]. However, in that work, the aHM-analysis part has been changed from its original form to include an intermediate signal, onto which the pitch modification is applied, leading to a two-step analysis procedure. Thus, the pitch shifting method used is more complex than the one presented in this paper, where we show that there is no need of an intermediate representation. Finally, the source-filter based aHM in [29] has been evaluated only for small pitch up-shifting factors ($+20\%$).

The rest of the paper is organized as follows. In Section 2, a review of the analysis and synthesis steps of aHM is presented. Section 3 provides the pitch-scale modification scheme and an example for the model in hand, along with the amplitude and phase envelope estimation methods that are used. Section 4 presents a formal evaluation of the proposed methods and discusses the results of the comparison with another well-known harmonic model, the HNM [6], and the widely used STRAIGHT method [9]. Finally, Section 5 concludes the paper.

2. THE aHM ANALYSIS/SYNTHESIS SYSTEM

In this section, a brief review of the adaptive Harmonic Model (aHM) is presented [14], 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 $f_0(t)$:

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

where f_s is the sampling frequency.

2.2. Analysis

In the analysis step, parametrizing the speech signal at each analysis time instant t_a^i is the first consideration. Initially, a sequence of the analysis time instants are created in the voiced parts of speech using the estimated $f_0(t)$ curve, so as to have one analysis time instant per pitch period. In unvoiced segments, the estimated $f_0(t)$ is not meaningful but it can be used to generate the corresponding analysis time instants. 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 curve $\phi_0(t)$ 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 curve of Eq.(2) is assumed to be known beforehand and can have a small deviation from its actual value, i.e.

$$\eta_0(t_a^i) = f_0(t_a^i) - \hat{f}_0(t_a^i) \tag{3}$$

where η_0 is called the *frequency mismatch*, f_0 is the actual fundamental frequency at an analysis time instant t_a^i, and $\hat{f}_0$ is an estimate of the latter. According to the adaptive scheme in [12], the amplitude $a_k(t)$ and fundamental frequency $f_0(t)$ curves are obtained by a linear and spline interpolation, respectively, of their estimated values, a_k^i 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* [12] 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 Eq. (2), a_k and b_k are the complex amplitude and the complex slope of the model, respectively, and K is the number of components. It has been shown in [30] that a_k and b_k obtained via a Least Squares minimization, can be used to provide an estimate, $\hat{\eta}_0$, of the frequency mismatch in Eq. (3). Thus, for the k^{th} component of the model, the latter can be computed as:

$$\hat{\eta}_k = \frac{f_s}{2\pi} \frac{\Re\{a_k\}\Im\{b_k\} - \Im\{a_k\}\Re\{b_k\}}{|a_k|^2} \tag{5}$$

Using this estimate, the fundamental frequency values f_0^i can be updated iteratively. However, as it is shown in [12], the frequency mismatch estimation is useful only if it is smaller than the main lobe of the analysis window. An iterative algorithm to update the frequencies has been proposed in [14].

2.4. Synthesis

In the synthesis step, each harmonic is generated in separate, one after the other, without using any window. Each harmonic component is synthesized by its parameters, namely its amplitudes $|a_k^i|$, its phases $\angle a_k^i$, and its fundamental frequency f_0^i. At 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. Obviously, the instantaneous phase $\angle a_k^i$ cannot be directly interpolated across time to obtain $a_k(t)$ because of its rotation due to the time advance between analysis time instants. To solve this, 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*:

$$\angle \tilde{a}_k^i = \angle a_k^i - k\phi_0(t_a^i) \tag{6}$$

Moreover, by assuming that the shape of the signal is changing smoothly, the RP values $\angle \tilde{a}_k^i$ are assumed to change smoothly as well, from one analysis time instant to the following one. Then, the RPs can be interpolated via splines in time domain to obtain its continuous counterpart, $\angle \tilde{a}_k(t)$. Finally, the instantaneous phase tracks, $k\phi_0(t)$, are obtained using Eq. (2), and they are added back to the continuous RP, $\angle \tilde{a}_k(t)$.

3. PITCH-SCALE MODIFICATION SCHEME

The purpose of pitch-scale modification is to change the pitch contour of the original speech signal while maintaining the apparent rate of articulation. The pitch contour (and thus the harmonics) should be shifted in frequency, and the formant structure should *not* be changed at a different rate than the rate of the input speech. For that, pitch scaling requires the estimation of amplitudes and phases in the new, shifted harmonic frequencies. In this work, the Discrete All-Pole method [23] is used. For the phase, there are also several approaches for its estimation. In this work, a method is used that utilizes the concept of *relative phase*, as it will be discussed in this section.

For an arbitrary pitch-scale modification, the input $f_0(t)$ contour is mapped to a different one, $f_0'(t) = \rho(t) f_0(t)$ in the modified signal, where $\rho(t)$ is the pitch-scale factor function. When $\rho(t) > 1$, then the pitch increases, whereas the opposite happens when $\rho(t) <$

1. Note that for a fixed $\rho(t) = \rho$, the pitch modification is constant throughout the whole waveform.

In the adaptive Harmonic model context, the parameters should be transformed in the way described next. Let us remind that in an analysis window centered at t_a^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 pitch-scaled waveform, $s_{PS}(t)$, for a constant pitch-scale factor is given by:

$$\hat{s}_{PS}(t) = \sum_{k=-K}^{K} A'_k(t) e^{j\phi'_k(t)} \tag{7}$$

where $A'_k(t)$ and $\phi'_k(t)$ are computed using the following way:

1. In order to compute $\phi'_k(t)$, it is first necessary to compute the pitch-scaled frequencies. Thus the new frequencies are given by:
$$kf_0(t) \leftarrow \rho k f_0(t) \tag{8}$$
2. The instantaneous amplitudes at analysis time instants t_a^i, $A'_k(t_a^i)$, are computed from sampling the spectral envelope at the corresponding frequencies $\rho k f_0$:
$$A'_k(t_a^i) = DAP(t_a^i, \rho k f_0) \tag{9}$$
where $DAP(t_a^i, f)$ is the Discrete All-Pole-based envelope constructed around time instant t_a^i. Then, the k^{th} instantaneous amplitude is linearly interpolated across successive time instants.
3. Then, the instantaneous phase should be re-computed. For this, the RP is first computed by extracting the integral of the initial fundamental frequency from the phase information at analysis time instant t_a^i, as in Eq. (6). Then, the RP values are interpolated, thus obtaining $\angle\tilde{a}_k(t)$, and finally, the integrated pitch-scaled frequency is added back to the interpolated RP values:
$$\hat{\phi}'_k(t) = \angle\tilde{a}_k(t) + \frac{2\pi}{f_s}\int_0^t \rho k f_0(u)du \tag{10}$$

The details for amplitude and phase estimation are following next.

3.1. Amplitude Estimation

Amplitude estimation is performed via an all-pole technique, called the Discrete All-Pole method (DAP). This method utilizes a discrete version of the Itakura-Saito (IS) distortion measure as its error criterion, instead of a time-domain criterion that most of other all-pole models use. The IS error measure is given by

$$E_{IS} = \frac{1}{N}\sum_{m=1}^{N} \frac{X(\omega_m)}{\hat{X}(\omega_m)} - \log\frac{X(\omega_m)}{\hat{X}(\omega_m)} - 1 \tag{11}$$

where $X(\omega_m)$ is the given discrete spectrum defined at N frequency points, and $\hat{X}(\omega_m)$ is the all-pole model spectrum evaluated at the frequencies $\omega_m \in [0, f_s/2]$, where f_s is the sampling frequency. This method manages to overcome the well-known limitations of linear prediction [31] and produces better fitting of spectra that are represented with a small set of discrete values, such as in the case for harmonic models.

The DAP method works iteratively to solve a nonlinear set of equations, in order to converge to a global minimum. The order, P, of the method does not differ from the empirical choice that is employed in most all-pole methods, that is

$$P = \frac{f_s}{1000} + 2 \tag{12}$$

where f_s is in Hertz. More details on DAP can be found in [23].

3.2. Phase Estimation

For the phase, the following simple method is selected, that utilizes the relative phase described in [14, 15]. Specifically, the linear phase term is sought to be removed during the resampling process in pitch shifting. Related work on linear phase removal has been suggested in other speech processing applications, such as concatenative speech synthesis [28], speech transformations [32], and speaker verification [33]. Let us consider a sinusoid

$$x_0(t) = \cos\left(2\pi\int_0^t f_0(u)du + \theta_0\right) \tag{13}$$

which we will consider as the reference sinusoid, and another one,

$$x_k(t) = \cos\left(2\pi k\int_0^t f_0(u)du + \theta_k\right), \quad k \in Z^+ \tag{14}$$

The instantaneous phases of the two sinusoids are

$$\phi_0(t) = 2\pi\int_0^t f_0(u)du + \theta_0, \quad \phi_k(t) = 2\pi k\int_0^t f_0(u)du + \theta_k \tag{15}$$

respectively. Let us consider that $\theta_0 = 0$, meaning that the time origin is set as the point where $\phi_0(0) = 0$. If we choose any analysis time instant t_a^i, the instantaneous phases become

$$\phi_0(t_a^i) = 2\pi\int_0^{t_a^i} f_0(u)du \quad , \quad \phi_k(t_a^i) = 2\pi k\int_0^{t_a^i} f_0(u)du + \theta_k \tag{16}$$

respectively. By changing variables, we get

$$\theta_k = \phi_k(t_a^i) - k\phi_0(t_a^i) \tag{17}$$

which is the so-called *relative phase*. These values allow the reconstruction of the shape of the signal, using the reference phase $\phi_0(t_a^i)$ in a synchronous reconstruction. For the purposes of pitch-scale modification, the f_0 track can be changed without any recomputation of the phase, because if the RPs are kept constant, the waveform will stretch or shrink accordingly without any other change. Our approach on the phase estimation is to interpolate the RPs in the frequency domain in order to obtain the RPs of the modified frequencies. Having these, the synthesis is quite simple and follows the same approach as in the synthesis scheme without modifications.

Pitch scale modifications for a factor of 2 and 0.5 are applied on a speech signal sampled at 16 kHz. Figure 1 shows the original speech signal (upper panel), the pitch-scaled signal for a factor of 2 (middle panel), and the pitch-scaled signal for a factor of 0.5 (lower panel).

3.3. Other types of frequency modifications

It is well known that formant manipulation along with pitch scaling can alter the perceived age, gender, and size of the speaker [34]. Since DAP method offers formant-shape spectral envelope estimation (unlike other methods, such as Discrete Cepstrum or linear-spline interpolations), it is straightforward to modify formants through the vocal tract length (VTL). Scaling or shifting the formant structure results in transformations like *child voice* or *big man voice*. For example, speakers can be modified to sound like their "child " counterpart by scaling formants and significantly raising the pitch.

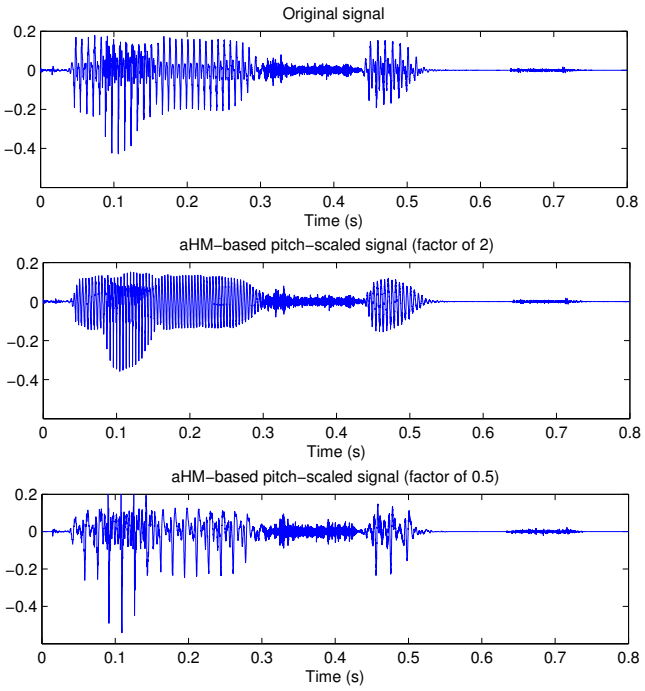

Fig. 1. *Upper panel: Original signal, Middle panel: pitch-scaled signal for a factor of 2, Lower Panel: pitch-scaled signal for a factor of 0.5.*

3.4. Joint time and pitch scaling

In [15], it was shown that aHM can successfully perform time scaling. A joint time and pitch scale modification scheme can be easily obtained since time scaling simply resamples the amplitudes, frequencies, and relative phases, no matter whether they came from the analysis part or the pitch-scaling algorithm described earlier. Thus, a very simple, flexible, and high-quality speech modification system based on the aHM can be built.

4. DISCUSSION AND EVALUATION

In general, first informal listenings acknowledged that common artefacts, such as "metallic" quality, chorusing, or musical noise do not appear in aHM more than they do in the state-of-the-art methods in hand. However, for large pitch scale factor and due to the harmonic nature of the representation, the spectral area between the distant successive harmonics manifests a sense of *tenseness* in voice. This is apparent especially in unvoiced parts, where the number of harmonics is not high enough to represent these parts well. It should be noted that both HNM and STRAIGHT use some kind of noise component, whereas aHM does not. HNM uses time and frequency modulated noise to represent unvoiced parts and high-frequency components of voiced parts, whereas STRAIGHT uses all-pass filters to compensate for the buzz timbre of minimum-phase vocal tract filter. To compensate the lack of "randomness" in pitch-shifting for aHM, harmonic information from the analysis part of the original signal is used as inter-harmonic content between the pitch-scaled harmonics that represent unvoiced speech. Specifically, the unvoiced parts are synthesized using the harmonics from the analysis of the original signal. In voiced parts, a random offset is added in each harmonic phase, sampled from the interval $[-\pi, \pi]$. The noise variance is inversely proportional to the harmonic number, that is, lower harmonics have less noise compared to higher, in a linear scale. The discrimination between voiced and unvoiced (V-UV) parts of speech has been obtained using the V-UV presented in [6], although any V-UV detector can be used.

Formal listening tests have been conducted to examine the performance of our modification scheme and two well-known state-of-the-art parametric approaches: the Harmonic + Noise Model [6], and the STRAIGHT [9] method. The pitch-scale modification factors were selected to be 0.5, 0.8, 1.2, 1.5, and 2.0, which are typical values for speech prosodic modifications. In this experiment, 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. For both genders of speakers, we have posed a minimum and maximum value for the pitch estimation: $f_{0(min,max)} = (120, 300)$ Hz for females, and $f_{0(min,max)} = (70, 200)$ Hz for males.

For the HNM, the maximum voiced frequency is fixed to 5500 Hz, and the analysis is pitch synchronous. The analysis window size is set to two local pitch periods. The order of the AR filter for the noise part is set to 20. For the STRAIGHT, default parameters were used, as they are provided by the on-line version of the code. The parameters of aHM are the ones described in the previous section. Part of the listening test is currently available on-line[1]. A number of 23 listeners participated in the test, 2 of them were familiar with signal processing techniques. The test is forced choice, thus it means the listener did not have the option of selecting *no preference*, and the preference results are presented in Table 1, where it is shown the percentage of listeners who preferred the first model over the second of each pair. The preference test shows that aHM clearly outperforms HNM and it is comparable with STRAIGHT for moderate pitch shifting factors.

Preference Test			
Factor	aHM-HNM	HNM-STRAIGHT	aHM-STRAIGHT
0.5	72%-28%	5%-95%	32%-68%
0.8	81%-19%	12%-88%	44%-56%
1.2	90%-10%	8%-92%	48%-52%
1.5	86%-14%	12%-88%	44%-56%
2.0	78%-22%	10%-90%	40%-60%

Table 1. *Preference test for pitch-shifting for all models and* 32 *speakers.*

5. CONCLUSIONS AND FUTURE WORK

A new and simple approach on pitch-scale modification based on the recently developed adaptive Harmonic Model (aHM) analysis/synthesis system is presented. The system utilize a full-band representation of speech based on quasi-harmonic analysis and strict harmonic synthesis. The proposed pitch scaling scheme provides flexibility and simplicity. Amplitude and phase envelopes are estimated using Discrete All-pole modelling and a simple approach which takes into account the relative phase between the harmonics, respectively. A noise-like effect is added via phase randomization of mid-to-higher frequencies to enhance naturalness. Formal listening tests show that pitch-scale modifications are of very good quality, compared to other state-of-the-art approaches, such as HNM and STRAIGHT. Future work will focus on handling the randomness of speech in a more concrete way and in dropping the voiced-unvoiced decision, thus leading to a simpler and more robust system.

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

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

[3] J. Laroche and M. Dolson, "New phase-vocoder techniques for pitch shifting, harmonizing and other exotic effects," *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pp. 91–94, 1999.

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

[5] 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.

[6] 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.

[7] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling." *ICASSP*, pp. 525–528, 2007.

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

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

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

[11] 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*, vol. 55, no. 2, pp. 278–294, 2013.

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

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

[14] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Transactions on Acoustics, Speech and Language Processing*, vol. 21, no. 10, pp. 2085–2095, 2013.

[15] G. P. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Time-scale Modifications based on a Full-Band Adaptive Harmonic Model," in *Proc. IEEE ICASSP*, Vancouver, CA, 2013.

[16] E. Godoy, O. Rosec, and T. Chonavel, "Voice conversion using dynamic frequency warping with amplitude scaling, for parallel or nonparallel corpora," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 20, no. 4, pp. 1313–1323, 2012.

[17] C. Magi, J. Pohjalainen, T. Backstrom, and P. Alku, "Stabilised Weighted Linear Prediction," *Speech Communication*, vol. 51, pp. 401–411, 2009.

[18] M. Wolfel and J. McDonough, "Minimum variance distortionless response spectral estimation," *Signal Processing Magazine, IEEE*, vol. 22, no. 5, pp. 117–126, 2005.

[19] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise degradation," *Proc. IEEE ICASSP*, pp. 4769–4772, 2012.

[20] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[21] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source filter systems with discrete spectra: Application to musical sound signals," *Proceedings of the International Computer Music Conference*, pp. 82–84, 1990.

[22] O. Cappe and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, pp. 100–102, 1996.

[23] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. on Signal Processing*, vol. 39, pp. 411–423, 1991.

[24] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis : a unified approach to speech spectral estimation," *In Int. Conf. Spoken Lang. Proc.*, pp. 1043–1045, 1994.

[25] J. Markel and A. Gray, *Linear prediction of speech*. Springer Verlag, 1976.

[26] Y. Shiga and S. King, "Estimating the spectral envelope of voiced speech using multi-frame analysis," *in Proc. EUROSPEECH2003*, pp. 1737–1740, 2003.

[27] T. Toda and K. Tokuda, "Statistical approach to vocal tract transfer function estimation based on factor analyzed trajectory hmm," *Proc. IEEE ICASSP*, pp. 3925–3928, 2008.

[28] Y. Stylianou, "Removing linear phase mismatches in concatenative speech synthesis," *IEEE Trans. on Speech and Audio Proc.*, vol. 9, pp. 232–239, 2001.

[29] J. Lee, F. K. Soong, and H.-G. Kang, "A source-filter based adaptive harmonic model and its application to speech prosody modification," *Interspeech*, 2013.

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

[31] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE*, vol. 63, pp. 561–580, 1975.

[32] R. D. Federico, "Waveform preserving time stretching and pitch shifting for sinusoidal models of sound," *COST-G6 Digital Audio Effects Workshop*, pp. 44–48, 1998.

[33] P. D. Leon, M. Pucher, J. Yamagishi, I. Hernáez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[34] D. R. Smith and R. D. Patterson, "The interaction of glottalpulse rate and vocal-tract length in judgements of speaker size, sex, and age," *J. Acoust. Soc. Am.*, vol. 118, no. 5, pp. 3177–3186, 2005.