ON THE RECOVERY OF TIME-VARYING SPECTRAL ENVELOPE INFORMATION FROM AQHM-DERIVED SPECTRA

Yannis Agiomyrgiannakis and Yannis Stylianou

Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
{agios, styliano}@ics.forth.gr

ABSTRACT

Spectral envelopes of speech signals are typically obtained by making stationarity assumptions about the signal which are not always valid. The Adaptive Quasi-Harmonic Model (AQHM), a non-stationary signal model, is capable of capturing the time-varying quasi-harmonics in voiced speech. This paper suggests the use of AQHM in a multi-layer scheme which results in a high-resolution time-frequency representation of speech. This representation is then used for the recovery of the evolving spectral envelope and thus, a time-frequency spectral envelope estimation algorithm is introduced related to the Papoulis-Gerchberg algorithm for data extrapolation. Results on voiced speech sounds show that the estimated spectral envelopes are smoother than those estimated by state-of-the-art spectral envelope estimators, while maintaining the important spectral details of the speech spectrum.

Index Terms— Spectral analysis, non-stationary analysis, speech synthesis, spectral envelope, true envelope, Papoulis-Gerchberg, voice modification, voice transformation.

1. INTRODUCTION

The estimation of the spectral envelope of voiced speech signals is a non-trivial task because a continuous frequency response is to be estimated from a limited number of samples located at the peaks of time-varying quasi-harmonic sinusoidal components. Most of the methods used for the estimation of the sinusoidal components rely on frame-level stationarity assumptions which are not consistent with the time-varying nature of voiced speech signals. To counter this, non-stationary linear AM-FM models like the Adaptive Quasi-Harmonic Model (AQHM) have been proposed [1, 2].

Quasi-Harmonic Model (QHM) describes the speech waveform as a sum of windowed exponentials and a set of functions that corresponds to their spectral derivative [3]. Based on this decomposition, a relatively simple iterative frequency estimator for the sinusoidal components has been devised and demonstrated to reach the Cramer-Rao lower bound on multi-component test signals in about 3 iterations [4]. An extension of QHM to locally non-stationary signals was recently presented in [1] resulting in a high-resolution and high-quality signal model which is referred to as the Adaptive Quasi-Harmonic Model (AQHM) [2, 1]. AQHM is merely a signal analysis tool aiming at providing continuous sinusoidal tracks. An application of AQHM to speech signals was presented in [2], where the speech signal is decomposed into a deterministic and a non-deterministic part. AQHM is used to capture the deterministic part, while the non-deterministic part is modelled as time-modulated colored noise.

This work is about recovering the spectral envelope of speech signals from time-varying instantaneous spectra obtained via AQHM decompositions. In the first part of this paper, we show that AQHM can be used in a multi-pass scheme that is able to model the whole speech signal, including transients, plosives and unvoiced parts, with a high mean utterance-level SNR of about 36 dB. Then, we use the derived *instantaneous* line spectra as a time-frequency representation in order to recover the underlying spectral envelope in voiced regions. The second part of this paper focuses on the recovery of the evolving spectral envelope by reconstructing what effectively is a 2-D spectral surface. The aim is to devise a high-quality accurate reconstruction that preserves the fine details of the spectrum while evolving smoothly over time. Such properties are desirable for high-quality analysis/synthesis, modification and transformation of speech signals [5, 6, 2].

In the single dimension case, the spectral envelope estimation problem is essentially equivalent to the recovery of a periodic signal from a few sparse samples. Traditionally, this is made by introducing some sort of prior knowledge regarding the underlying spectral envelope. For example, Discrete-All-Pole modelling describes the spectral envelope in terms of an auto-regressive linear system [7]. Another well-known high-quality spectral envelope estimator, STRAIGHT [6], uses spline interpolation to estimate the spectrum between the sinusoidal spectral peaks. Discrete Cepstrum Coefficient-based estimation [8] assume that the log-spectral envelope is bandlimited and uses regularization to penalize rapid variations. Bandlimiting assumptions are also used in the "True-Envelope" (TE) estimator, which, however uses a different way to find the "optimal" spectral envelope [9]. The TE estimator shares with STRAIGHT the advantage of avoiding an explicit peak-picking of the quasi-harmonic sinusoidal components. The algorithm is related to iterative bandlimited extrapolation algorithms and the Papoulis-Gerchberg algorithm [10, 11].

Since we have no knowledge of the true underlying spectral envelope it is not possible to objectively evaluate a spectral estimator nor its assumptions. In our case, the implicit peak-picking properties of the TE estimator make it a suitable candidate for estimating a spectral envelope from the AQHM-derived instantaneous line spectra. It exhibits, however, frame-to-frame fluctuations that seem to be related to the nature of the estimator and not to the underlying spectral process. These fluctuations are more evident in higher-frequencies and in spectral valleys. Therefore, we present a 2-D spectral surface estimator that does not suffer from these variations, seems to

This work was supported by LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623

preserve fine-details of the spectrum and allows a tradeoff between consistency and smoothness.

2. SPECTRAL ANALYSIS USING AQHM

2.1. Quasi-Harmonic Model

Sinusoidal models describe a single speech frame $h_l(t+t_l)$ centered at t_l using a stationary basis of windowed complex exponentials:

$$h_l(t+t_l) = \sum_{k=-K_l}^{K_l} a_k^l \left(w(t) e^{2\pi j f_k^l t} \right), \quad (1)$$

where f_k^l are the analysis frequencies of l-th frame, a_k^l are the complex amplitudes, which for real signals have a conjugate symmetry $a_k^l = (a_{-k}^l)^*$, and $w(t)$ is the analysis window which is zero outside a symmetric interval $[-T, T]$. The Quasi-Harmonic Model augments the harmonic model by a set of functions $tw(t)e^{2\pi j f_k^l t}$:

$$h(t+t_l) = \sum_{k=-K_l}^{K_l} a_k^l \left(w(t) e^{2\pi j f_k^l t} \right) + \sum_{k=-K_l}^{K_l} b_k^l \left(tw(t) e^{2\pi j f_k^l t} \right) \quad (2)$$

where b_k^l are the "complex-slope" parameters. The linearity of the model allows the estimation of a_k^l and b_k^l using typical least-squares methods [5] pg. 76-83.

An analysis of the properties of QHM reveals that it is able to resolve errors in frequency estimation. In fact, it is possible to analytically compute the frequency mismatch ρ_k^l between the analysis frequency and the actual frequency of the sinusoid [1]:

$$\rho_k^l = \frac{a_k^{l,R} b_k^{l,I} - a_k^{l,I} b_k^{l,R}}{|a_k^l|^2} \quad (3)$$

where x^R and x^I denotes the real and imaginary parts of x, respectively. Therefore, it is possible to use ρ_k^l to obtain an improved estimation of the analysis frequency

$$f_k^{l(new)} = f_k^l + \frac{1}{2\pi} \rho_k^l. \quad (4)$$

Hence, better estimates of the analysis frequencies can be obtained iteratively, resulting in more accurate amplitude and phase estimation. In practice, 3 iterations are enough to get a good estimation, which is shown to be very close to the Cramer-Rao lower bound [4].

2.2. Adaptive Quasi-Harmonic Model

Speech signals exhibit a local non-stationary behavior that QHM cannot resolve. In order to tackle local non-stationarity it has been suggested in [1] to use time-varying basis functions $e^{j\tilde{\phi}_k(t)}$ instead of the stationary complex exponentials $e^{2\pi j f_k^l t}$:

$$s(t+t_l) = \sum_{k=-K_l}^{K_l} a_k^l \left(w(t) e^{j\tilde{\phi}_k^l(t)} \right) + \sum_{k=-K_l}^{K_l} b_k^l \left(tw(t) e^{j\tilde{\phi}_k^l(t)} \right) \quad (5)$$

with

$$\tilde{\phi}_k^l(t) = \int_{t_l}^{t_l+t} f_k(u)du, \; t \in [-T, T] \quad (6)$$

where $f_k(t)$ is the frequency trajectory of the k-th component.

Comparing (5) to (2), we see that in (2) the analysis frequency of the k-th component, f_k^l, is considered to be constant during the analysis frame. To the contrary, the frequency of the k-th component in (5) is time-varying and the phase functions defining the basis for decomposing the signal is obtained by integration of the previously estimated frequency tracks $f_k(t)$. By considering the model in (5), a non-parametric evolution over time of the frequency of the k-th component is obtained. Therefore, the model suggested in (5) is more appropriate for modeling non-stationary signals as compared to the model suggested in (2).

In our implementation, the time-varying frequency tracks $f_k(t)$ are initialized as integer multiples of the fundamental frequency track obtained from a pitch detector. The adjacent frequencies ..., f_k^{l-2}, f_k^{l-1}, f_k^l, f_k^{l+1}, f_k^{l+2}, ... are interpolated using spline interpolation [1]. A detailed description on the construction of the time-varying phase functions $\tilde{\phi}_k^l(t)$, phase-coherence and the treatment of special cases that arise when the number of sinusoids varies from frame to frame is beyond the scope of this paper and can be found in [1]. Finally, an iterative refinement algorithm is used to optimize the instantaneous phase and the instantaneous amplitude tracks of the AM-FM sinusoidal components [1].

Using the estimated instantaneous amplitude and phase components, $\hat{a}_k(t)$, and $\hat{\phi}_k(t)$, respectively, the signal reconstruction is then simply provided as:

$$\hat{s}(t) = \sum_{k=1}^{K} \hat{a}_k(t) \cos\left(\hat{\phi}_k(t)\right). \quad (7)$$

Note that the signal model above describes a long segment of speech instead of a single frame. It consists of amplitude-modulated, frequency-modulated sinusoids with instantaneous frequencies

$$\hat{f}_k(t) = \frac{1}{2\pi} \frac{\partial \hat{\phi}_k(t)}{\partial t}. \quad (8)$$

2.3. Multi-layer Adaptive Quasi-Harmonic Model

AQHM decomposes the speech signal to AM-FM components according to equation (7). When applied to voiced speech, these components capture the time-varying quasi-harmonics and exhibit relatively smooth instantaneous frequency trajectories $\hat{f}_k(t)$. On the contrary, when applied to non-voiced speech, i.e. unvoiced parts and plosives, the derived AM-FM components show high variation on their instantaneous frequency and amplitude. Intuitively, AQHM components in non-voiced speech attempt to locate "optimal" frequency tracks that collectively minimize the mean-square-error. This states that AQHM can be used as a *full-signal model* and not only for voiced speech. Care -however- is needed on the interpretation of its parameters on non-voiced speech as the AM-FM components may not represent narrow-band processes and may be considerably correlated. In that aspect, we need to have further insight both analytically and experimentally on how AM-FM decompositions behave in such cases.

The latter behavior of AQHM, motivates a multi-layer AQHM decomposition. Let $s(t)$ be the original speech signal (a whole utterance) and $\hat{s}_1(t) = AQHM(s(t))$ be the AQHM reconstruction of $s(t)$. The residual $\epsilon_1(t) = s(t) - \hat{s}_1(t)$, can then be modeled by a second AQHM reconstruction $\hat{\epsilon}_1(t) = AQHM(\epsilon_1(t))$ and the residual of the residual $\epsilon_2(t) = \epsilon_1(t) - \hat{\epsilon}_1(t)$ by a third AQHM reconstruction $\hat{\epsilon}_2(t) = AQHM(\epsilon_2(t))$ and so forth. For the three layers discussed, the reconstruction of multi-layer AQHM is $\hat{s}_3(t) = \hat{s}_1(t) + \hat{\epsilon}_1(t) + \hat{\epsilon}_2(t)$. Excluding the first layer which corresponds to the AQHM decomposition, every additional layer captures AM-FM components that were missed by previous layers and successively minimizes the signal-wise energy of the residual. In voiced speech, the additional AQHM decompositions reveal weak components at spectral valleys that were buried in the side-lobes of the analysis window, residuals from salient harmonics that remained

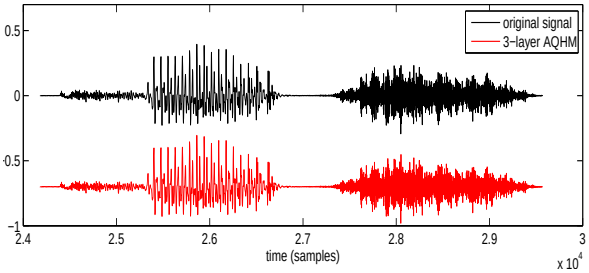

Fig. 1. Upper panel: Original speech signal. Lower panel: Copy-synthesis using a 3 layer aQHM

due to frequency mismatches as well as inter-harmonic noise structures. In non-voiced speech, the additional AQHM components follow trajectories that further minimize the residual error.

A simple experiment was made using a set of 20 different utterances from different speakers, 10 males and 10 females, randomly selected from TIMIT. A 3-layer AQHM decomposition was made and the utterance-level SNR was computed using each additional layer. The mean reconstruction SNR from the first layer was 25.77 dB, the second layer increased SNR to 32.55 dB, while the third layer increased it further up to 36.02 dB. AQHM analysis was made using a 2 ms analysis step, while we found that SNR improved when a small offset dt of 1 ms was introduced at the analysis time instants t_l according to formula: $t'_l = t_l + dt * mod(b-1, 2)$, where b is the index of the layer and $mod(.,.)$ the modulo-division operator. The offset is used to avoid computing the AQHM parameters of each level at the same analysis instants with the previous level.

An informal subjective listening test has shown that the reconstructed utterances are indistinguishable from the original signals. Finally, an example of a 3-layer signal reconstruction of a speech segment is depicted in Figure 2.3, where we can observe that the fine details of the unvoiced part are accurately described.

3. ESTIMATION OF SPECTRAL ENVELOPES

The instantaneous amplitudes from the multi-layer AQHM components can be used to construct a high-resolution time-frequency representation of the speech signal. The motivation for using the multi-layer AQHM for this purpose arises from 1) the fact that the AM-FM components track time-varying quasi-harmonics in voiced speech signals and, 2) the robustness of the multi-layer approach in revealing weak components in spectral valleys and higher frequencies. Such components, missed by the first layer will now be captured by the following layer.

The time-frequency representation is constructed by quantizing the instantaneous frequencies on an FFT frequency grid. For example, assuming an FFT of size $N = 2048$, a 1025-bin grid is used with center frequencies $\frac{k}{N}F_s, k = 0, \ldots, \frac{N}{2}$, where F_s is the sampling rate of the speech signal. The instantaneous frequency of each AQHM component is computed using formula (8) and a spectrogram-like representation is made by setting the instantaneous amplitudes on the corresponding frequency bins. When more than one instantaneous amplitudes are allocated to a single frequency bin, we keep the maximum. The resulting representation consists of time-varying frequency-quantized line spectra. The rest of this section is devoted on the estimation of a smoothly evolving spectral envelope that preserves the fine details of the spectrum.

3.1. 1-D "True-Envelope"

The "True Envelope" estimator is an iterative method that successively refines an estimation of the spectral envelope [12]. Let $\vec{x}$ be the measured log-spectrum vector and H_q be a "band-limiting" matrix constructed as follows $H_q = F^{-1} \cdot W_q \cdot F$, where F is the DFT matrix and W_q a diagonal matrix with zeros and ones on its diagonal such that the operation $\tilde{\vec{x}} = H_q \cdot \vec{x}$ corresponds to low-pass filtering $\vec{x}$. Therefore, since $\vec{c} = F \cdot \vec{x}$ is the real discrete cepstrum of $\vec{x}$, the matrix H_q is performing low-pass liftering on $\vec{x}$ and the diagonal of W_q defines the corresponding lifter. Let $W_{f,n}$ be a diagonal frequency selection matrix with ones around the location of spectral peaks and zeros elsewhere. Then, the "True-envelope" algorithm performs the following iterations until convergence:

1. $\vec{x}_0 = \vec{x}$
2. $W_{f,n} = diag\{1(\vec{x} > \vec{x}_n)\}$
3. $\vec{x}_n = W_{f,n} \cdot \vec{x} + (I - W_{f,n}) H_q \cdot \vec{x}_{n-1}$

where $1(\vec{c})$ is the element-wise unit operator that yields 1 when the condition $\vec{c}(.)$ is true and zero otherwise. Thus, the i-th diagonal element of $W_{f,n}$, $W_{f,n}(i,i)$ is equal to 1 when the i-th component of $\vec{x}$, $\vec{x}(i) > \vec{x}_n(i)$. After convergence in n' iterations, the "True-envelope" $\vec{e}$ is obtained as $\vec{e} = H_q \cdot \vec{x}_{n'}$. If $W_{f,n}$ is fixed for all iterations (i.e. to reflect the spectral peaks), the iterations above correspond to the well-known Papoulis-Gerchberg algorithm [10, 11].

The cepstral liftering matrix H_q must restrict the spectral envelope from adapting to individual harmonics. This can be achieved using an appropriate selection of the order of the discrete cepstrum $\vec{c}$ via the quefrency weighting matrix W_q. A good tradeoff between over-smoothing and over-fitting can be achieved by using the order $P_c = \frac{F_s}{2f_0}$, where F_s and f_0 are the sampling rate and the fundamental frequency of the speech signal, respectively [12].

The cepstral liftering matrix H_q is using a rectangular quefrency window in W_q. The rectangular quefrency windowing corresponds to convolution with a periodic sinc function in the log-spectrum domain and introduces oscillations whenever the log-spectrum is changing rapidly, a behavior known as *Gibbs phenomenon*. It can be reduced however if at the last iteration we apply a Hamming window approximately 1.66 times the size of the rectangular window during the reconstruction of $\vec{e}$ [12].

The main advantage of TE estimator is that it avoids explicit peak-picking and depends only on the estimated fundamental frequency, while it is also relatively robust to pitch errors.

3.2. 2-D "True-Envelope"

The 1-D "True-Envelope" exhibits frame-to-frame variations that are more related to the estimation process rather than to the "true" underlying spectral envelope. For the purpose of high-quality voice modification and voice transformation it is important to reduce these variations as much as possible. This is not a trivial task because any operation that effectively smooths the evolution of the spectral envelope may also remove fine details of the spectrum. If the estimation of the spectral envelope is seen as the recovery of a 2-D spectral surface from a few spectrogram samples (in our case, from the AQHM-derived spectrogram-like representation), then we can apply additional constraints, similar to those that enabled the estimation of 1-D spectral envelopes, in order to obtain 2-D estimators that satisfy both smoothness and spectral-fitting. The "True Envelope" algorithm is well placed for such a task because it fits the locally salient spectral peaks while avoiding explicit peak picking. Additionally, locally salient spectral peaks provide accurate information regarding the underlying spectral envelope. In order to address these considerations, we propose a 2-D "True envelope" algorithm that uses neighboring frame information during the estimation of a spectral envelope.

The algorithm is conceptually the same with its 1-D counterpart. Let l_c be the index of the current frame and X an $(\frac{N}{2} + 1)$-by-

$(2K+1)$ matrix that contains a patch of the log-spectrogram. The columns of X contain log-spectra from frames $l_c-K,...,l_c+K$, while the lines of X correspond to the $\frac{N}{2}+1$ frequency bins of an N-length FFT. Let $Lifter2D(\cdot)$ be a two dimensional low-pass liftering operator and $max(.,.)$ an element-wise maximum function. Then, the iterations of the 2-D true envelope algorithm are:

1. $X_0 = X$
2. $X_n = Lifter2D(max(X, X_{n-1}))$

After convergence, we obtain the 1-D spectral envelope of the l_c-th frame from the $(K+1)$-th column of X_n. This procedure is repeated for every frame. Again, as in Section 3.1, at the last iteration we perform the lowpass liftering along the frequency dimension using a Hamming window in order to avoid the *Gibbs phenomenon*.

The $Lifter2D(\cdot)$ operator corresponds to an 2-D FFT-based low-pass filtering that can be efficiently implemented using 1-D FFT. In fact, since log-spectra are real with an even symmetry, fast pruned DCT algorithms can be used for the implementation. The liftering operation requires two parameters, the order of the cepstrum P_c along the vertical (frequency) dimension and the order P_a of the filter along the horizontal (time) dimension. P_c is computed as in Section 3.1, while in our experimental evaluations we set $P_a = K+1$.

An proper objective evaluation of a spectral envelope estimator is not possible due to the lack of a reference. A subjective evaluation using analysis/synthesis and modifications of speech signals is likely to reveal more information but it is also biased to the specifics of the modification algorithm. It is possible -however- to evaluate certain properties of the evolving spectral envelopes like "consistency" [6] and smoothness of the spectral envelope evolution. Consistency measures how well the estimator fits the quasi-harmonics.

To evaluate consistency, an experiment was made using the test-set described in Section 2.3. Voiced segments of 50-150 ms were extracted from these utterances and three algorithms were tested on them; 1-D TE, 2-D TE, and STRAIGHT. As a reference we used the quasi-harmonics of the first 2 kHz, as estimated by the AQHM decomposition. We visually confirmed that the AQHM components were actually capturing the quasi-harmonics. The evaluation was made using the spectral distortion criterion computed at quasi-harmonic frequencies. The 1-D TE yielded a very low mean error of 0.78 dB (CI=0.01) while the 2-D TE showed a mean error of 1.44 dB (CI=0.01). The evaluation of smoothness was made by measuring the modulation spectra of the evolving spectral envelope surface computed at FFT-bin frequencies. In other words, we extract the trajectory of the spectral envelope for each frequency of the FFT-bin and compute its frequency content using a sliding window of length 16 (16 frames, 32 ms). Then we average the modulation spectra in eight 1-kHz bands. Figure 3.2 shows the average modulation spectra of two bands, the 2-3 kHz and the 4-5 kHz, respectively. We can observe that 1) the 1-D TE is more or less as smooth as STRAIGHT and, 2) the 2-D TE is considerably smoother than both 1-D TE and STRAIGHT. A similar behavior is observed on the other bands as well, with TE 2-D being increasingly smoother than its counterparts as the frequency increases. It should also be noted that all three estimators provide relatively smooth trajectories in the first 1 kHz band.

We have observed a tradeoff between smoothness and consistency. At this point, it is hard to make a decision regarding the optimal tradeoff. This has to be made in the context of a high-quality voice modification/transformation system with a careful subjective evaluation. It is important, though, to state that in a practical application, we can easily combine the 1-D TE and the 2-D TE in a weighted average that uses the 1-D TE at the perceptually important

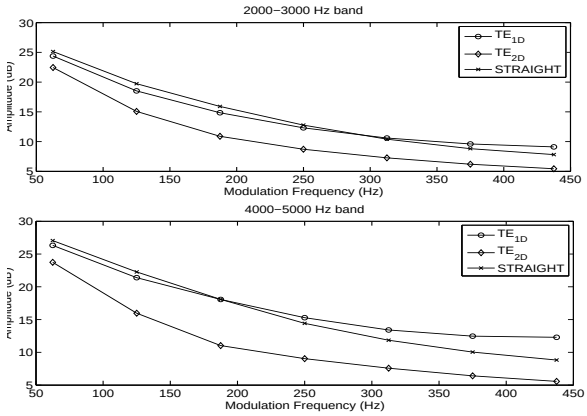

Fig. 2. Average Modulation Spectra

first 1.5 kHz and the 2-D TE for frequencies above 1.5 kHz, so that we can benefit both from the good fit of 1-D TE at lower frequencies and from the smoothness of 2-D TE at higher frequencies.

4. CONCLUSIONS

The proposed multi-layer AM-FM decomposition based on AQHM is capable of modelling voiced as well as non-voiced speech with a high average utterance SNR of 36 dB. The instantaneous amplitudes of this decomposition provide an accurate high-resolution estimation of the instantaneous amplitudes of the quasi-harmonics in voiced speech. An attempt to recover the underlying spectral envelope using the "True-Envelope" estimator shows undesirable fluctuations on evolution of recovered spectral envelopes. A new 2-D "True-Envelope" estimator that uses spectra from neighboring frames to provide smoother trajectories is proposed and evaluated. A combination between these two envelopes seems to be a good practical solution. However, an optimal tradeoff between smoothness and consistency is yet to be found.

5. REFERENCES

[1] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou, "AM-FM estimation for speech based on a time-varying sinusoidal model.," in *Interspeech*, Brighton, 2009.

[2] Yannis Pantazis, Georgios Tzedakis, Olivier Rosec, and Yannis Stylianou, "Analysis/Synthesis of Speech based on an Adaptive Quasi-Harmonic plus Noise Model," in *ICASSP*, Dallas, 2010.

[3] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou, "On the Properties of a Time-Varying Quasi-Harmonic Model," in *Interspeech*, Brisbane, 2008.

[4] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou, "On the robustness of the Quasi-Harmonic Model of Speech," in *ICASSP*, Dallas, 2010.

[5] Y. Stylianou, *Harmonic-plus-Noise Models for Speech, combined with Statistical methods for speech and speaker modification*, Ph.D. thesis, Ecole Nationale. Superieure des Telecommunications, Paris, France, 1996.

[6] Hideki Kawahara, "Straight, exploitation of the other aspect of vocoder: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, June 2006.

[7] Amro El-Jaroudi and John Makhoul, "Discrete All-Pole Modeling," *IEEE Trans. Signal Processing*, 1991.

[8] O. Cappe, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA, October 1995.

[9] A. Robel and X. Rodet, "Real time signal transposition with envelope," in *Proc. Int. Computer Music Conference (ICMC)*, 2005.

[10] Paulo Jorge S.G. Ferreira, "Interpolation and the Discrete Papoulis-Gerchberg algorithm," *IEEE Trans. Signal Processing*, vol. 42, no. 10, October 1994.

[11] Benjamin G. Salomon and Hanoch Ur, "Accelerated Iterative Band-Limited Extrapolation Algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 11, pp. 871, November 2004.

[12] A. Robel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. of the 8th Int. Conference on Digital Audio Effects (DAFx05)*, Madrid, Spain, September 20-22 2005.