Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation

Axel Roebel, Xavier Rodet

EFFICIENT SPECTRAL ENVELOPE ESTIMATION AND ITS APPLICATION TO PITCH SHIFTING AND ENVELOPE PRESERVATION

A. Röbel

Analysis/Synthesis Group
IRCAM, Paris, France
Axel.Roebel@ircam.fr

X. Rodet

Analysis/Synthesis Group
IRCAM, Paris, France
Xavier.Rodet@ircam.fr

ABSTRACT

In this article the estimation of the spectral envelope of sound signals is addressed. The intended application for the developed algorithm is pitch shifting with preservation of the spectral envelope in the phase vocoder. As a first step the different existing envelope estimation algorithms are investigated and their specific properties discussed. As the most promising algorithm the cepstrum based iterative *true envelope* estimator is selected. By means of controlled sub-sampling of the log amplitude spectrum and by means of a simple step size control for the iterative algorithm the run time of the algorithm can be decreased by a factor of 2.5-11. As a remedy for the ringing effects in the the spectral envelope that are due to the rectangular filter used for spectral smoothing we propose the use of a Hamming window as smoothing filter. The resulting implementation of the algorithm has slightly increased computational complexity compared to the standard LPC algorithm but offers significantly improved control over the envelope characteristics. The application of the true envelope estimator in a pitch shifting application is investigated. The main problems for pitch shifting with envelope preservation in a phase vocoder are identified and a simple yet efficient remedy is proposed.

1. INTRODUCTION

Pitch shifting is a signal transformation that is often used and many algorithms are available to achieve it. In the following article we address signal transposition based on a frequency domain representation notably the phase vocoder [1, 2, 3]. There exist two approachers to achieve pitch-shifting in the phase vocoder [4], both of which share the problem that they do not allow to change pitch without affecting timbre. If naturalness of the modified signals is desired the timbre modification poses a significant problem. A straightforward means to prevent timbre modification for monophonic signals consists of pre-warping the spectral envelope of each signal frame in the spectral domain before the transposition is actually performed [1]. To be able to perform the pre-warping a robust estimator of the spectral envelope is needed. To be able to achieve real time transposition [5] the estimation should be efficient and result in reasonable estimates for signals containing sinusoidal and noise signal components.

Note, however, that even the preservation of the spectral envelope is not necessarily assuring a naturally sounding transformed signal. A simple comparison of spectra of instrument samples reveals that the spectral envelope of sounds played with the same instrument changing only the pitch do in most cases not preserve the same spectral envelope. Using straightforward spectral comparison it can be easily found that for the same instrument and changing pitch, formants may change their relative level and may even appear or disappear completely. For speech signals it is well known that the first formant moves with the fundamental frequency. Given the rather small bandwidth that the lower formants usually have, this makes sense from a energy-saving point of view. The spectral deviations, however, depend on the instrument that is played and the playing style, and therefore, the optimal warping function would require a model of the envelope evolution, a topic that is beyond the following investigation.

The subject of the following investigation is first, the investigation into the properties of the different options that are available for the estimation of the spectral envelope. The comparison of the different estimators, i.e. all-pole models and cepstral models, shows that the cepstral based *true envelope* estimator has especially favorable properties, besides the fact that its estimation is computationally rather demanding. Despite of its favorable estimation properties the *true envelope* estimator has not received much attention outside of Japan, because the original publication has been in Japanese [6]. Therefore, the basic principle of the algorithm will first be explained, before an optimized implementation is proposed, that reduces the required computation to a minimum. The new implementation features the advantageous properties of the discrete cepstrum [7] without its drawbacks [8], and, with reduced computational complexity. The algorithm can be run in real time even in parallel with the signal transposition algorithm, i.e. the phase vocoder. The advantage of the cepstral representation is the fact that the optimal envelope model order can be derived from the fundamental frequency of the sound signal, which makes it simple to achieve semi automatic order adaptation. Considering the application to signal transposition we investigate into the problem that the envelope properties are ill conditioned for frequencies below the fundamental frequency. A revised pre-warping function is proposed that may achieve significantly improved perceived sound quality and may be used to influence the position of the first formant of the pitch shifted signal. Comparing the transposition results obtained with the true envelope and a standard LPC envelope estimation we find that for high pitched signals the artifacts related to the fact that the all-pole model may represent individual peaks are significantly reduced.

The article is organized as follows. In section **2** we describe the main tools that are available today for spectral envelope estimation. We compare the different algorithms with respect to results and performance. In section **3** we describe the steps that are necessary to reduce the computational complexity of the algorithm. In 4 we describe the automatic selection of the appropriate cepstral smoothing order and in section **5** we evaluate the performance of the optimized algorithm and investigate into its application for en-

velope preservation.

2. ESTIMATING THE SPECTRAL ENVELOPE

The term spectral envelope denotes a smooth function that passes through the prominent spectral peaks. For an harmonic signal the prominent spectral peaks are generally the harmonics, however, if some of the harmonics are missing or weak (clarinet) the spectral envelop should not pass through these. From this explanation the main problem of the spectral envelope becomes apparent. There exists no technical or mathematical definition and what is desired depends to some extend on the signal

Most of the existing techniques are based on either linear prediction (LPC) [9] or the real cepstrum [10]. For an LPC model of order P_l the standard all-pole model estimation using the Levinson-Durbin recursion requires $O(P_l^2)$ floating point operations. For reasonable P_l (< 120) it is possible to run the LPC estimation in real time within the phase vocoder such that for each frame the spectral envelope can be pre-warped as required. For harmonic sounds the all-pole model suffers from systematic errors that have been addressed in [11]. For high pitched signals or for signals that contain individual, outstanding sinusoids , i.e. sinusoids with significantly increased amplitude, the all-pole model may start to represent those peaks and, if it is applied for envelope preservation, the sound quality of the transposition degenerates.

The second approach that is available for envelope estimation is cepstral smoothing. This method is based on a Fourier representation of the log amplitude spectrum of the signal. Similar to the all-pole model a number of proposals have been made to find the spectral envelope using the cepstrum. In the following article we will briefly discuss two of the proposed strategies, the discrete cepstrum [7, 8] and the true envelope [6].

The real cepstrum $C(l)$ of a signal is the Inverse Fourier transform of the log amplitude spectrum of the sound. If we define $X(k)$ to represent the K-point DFT of the signal frame $x(n)$ the real (discrete) cepstrum $C(l)$ of a signal frame is

$$C(l) = \sum_{k=0}^{K} \log(|X(k)|) e^{i\frac{2\pi kl}{K}}. \quad (1)$$

Because the spectral envelope is considered to be a smoothed version of the amplitude spectrum a simple means to obtain an estimate of the spectral envelope is to set all the high frequency elements in the cepstrum to 0. As is well known the resulting Fourier representation is the optimal approximation of the log amplitude spectrum, in the MSE sense, using as mathematical model a Fourier series with limited number of harmonics and fundamental period given by the sample rate. The number of harmonics used on the non negative frequency axis is called the order P_c of the cepstrum. Unfortunately, the filtered cepstrum will create an envelope following the mean of the spectrum and not the contour of the spectral peaks. The discrete cepstrum proposed in [7, 8] establishes a method to find the cepstrum parameters considering only the peaks of the signal amplitude spectrum, such that the mean spectrum is close to what is considered a spectral envelope. The problems are, that the the method requires a fundamental frequency analysis or another means to preselect the spectral peaks, that it is often ill-conditioned, and has computational complexity of $O(P_c^3)$.

There is, however, another procedure to cope with the fact that the filtered cepstrum represents the mean value. The method has been developed in [6] where it was introduced as a method that estimates the *true envelope*. The algorithm is iterative and a straightforward implementation requires rather extensive computation, especially if the FFT size is large. Let $V_i(k)$ be the cepstral representation of the spectral envelope at iteration i, that is the Fourier transform of the filtered cepstrum, and further initialize the iteration using $A_0(k) = \log(|X(k)|)$ and $V_0(k) = -\infty$. The algorithm then iteratively replaces the current target amplitude spectrum according to

$$A_i(k) = \max(A_{i-1}(k), V_{i-1}(k)) \quad (2)$$

and iteratively applies the cepstral filter to the updated target spectrum A_i. With this procedure the valleys between the peaks of the target spectrum will be filled by the cepstral filter and the estimated envelope will steadily grow until all the peaks are covered. As stopping criterion of the procedure a parameter Δ is used that defines the maximum excess that a peak of the observed spectrum is allowed to have above the spectral envelope (in the following experiments $\Delta = 2dB$ is used). The graceful interpolation of the signal spectrum according to the requested cepstral order is a nice feature of the algorithm avoiding all the problems related to the ill-conditioned setup of the discrete cepstrum [8]. Nevertheless, the method suffers from two drawbacks. The first is due to the fact that the complexity is not related to P_c but to the DFT size K. Because the algorithm basically calculates a sequence of DFTs of order K, it scales with $O(K \log(K)$.

The second problem is due to the fact that, as is commonly done, the window applied to the spectrum is a rectangular window. The rectangular windowing will smooth the log amplitude spectrum by means of convolution with a periodic sinc function. This smoothing kernel creates oscillations whenever the log amplitude spectrum changes rapidly. This is the well known *Gibbs Phenomenon*. Artificial oscillations in the spectral envelope are rather annoying, especially for applications that try to detect the formant structure from the envelope. They can be reduced significantly if a Hamming window (centered at frequency zero) is used as the smoothing filter in the cepstral domain. Selecting the window approximately 1.66 as large as the rectangular window creates a smoothing kernel with approximately the same central lobe, but nearly no side lobes. The resulting spectral envelopes are much smoother which facilitates their use for further processing.

To evaluate the real world performance of the algorithms we have selected two examples of singing voice sounds. The first is a tenor singer displayed in fig. 1. The model orders have been selected such that the acoustic result obtained when used to prewarp the spectral envelope during transposition achieves subjectively the highest quality for the whole sound. The sound segment that is displayed in fig. 1 has relatively high pitch such that the model problems will be apparent. With respect to the envelope obtained by means of the LPC it is clearly visible that the all-pole model obtained by means of the levinson algorithm favors the representation of large amplitude peaks and may over-estimate the amplitude. Accordingly, as soon as transposition is not an integer factor the transformed sound contains whistling artifacts. The discrete cepstrum is obviously doing a very good job for the present signal. The envelope closely follows the selected peaks deviating only if the cepstral order or the regularization require it. The true envelope describes a similar contour, however, it follows the peak contour more faithfully passing generally ΔdB below the peaks. The difference, however, is small and hardly perceptually relevant.

As a second sound example we consider a soprano singer. The

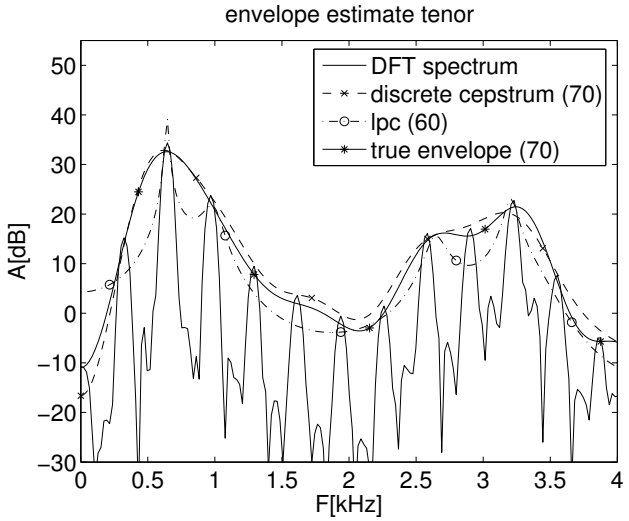

Figure 1: *The signal spectrum of a tenor singing voice segment with the envelope estimates obtained with standard LPC $P_l = 60$, the discrete cepstrum and two versions of the true envelope estimator (standard and optimized). All cepstral models use $P_c = 70$.*

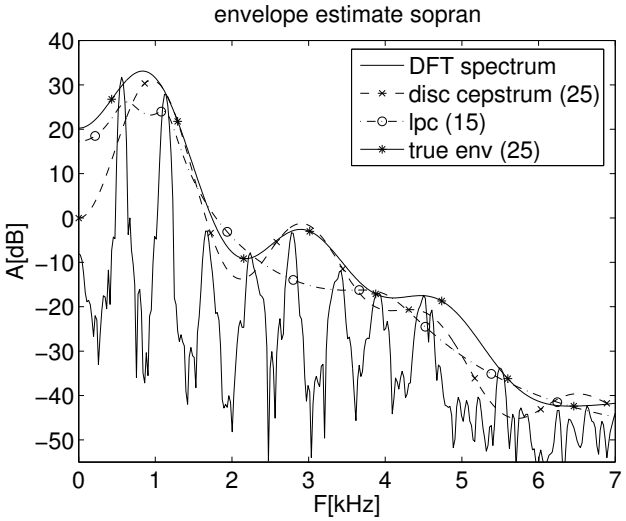

Figure 2: *Estimating the spectral envelope for a soprano singer. The spectrum is shown with envelope estimates obtained with standard LPC $P_l = 15$, the discrete cepstrum and the true envelope estimator original and new optimized version (all $P_c = 25$).*

signals spectrum together with the envelope estimates are shown in fig. 2. Note, that again the LPC model is mainly influenced by the major spectral peaks and misses completely the formant that exists around 3kHz. For this example the discrete cepstrum performs not as well. The problem arises due to the fact that the discrete cepstrum requires special treatment of the region below the fundamental to prevent the discrete cepstrum from creating a high level bump around frequency 0Hz. This is the problem of ill-conditioning that has been extensively discussed in the literature. In contrast to this the true envelope performs much better. It fills the valley at frequency zero according to its cepstral order and the result is intuitively correct. While the true envelope is not optimal in making up the envelope around DC (see the discussion in section **5.1**), it nevertheless will never create a local maximum around frequency zero.

The run time of all the algorithms used to estimate the data in fig. 1 and fig. 2 are compared in table (1). The run time for the discrete cepstrum strongly increases with the model order which is related to the computational complexity of matrix inversion. The run time for the LPC is smaller and is sufficiently fast for real time processing. The computational complexity of the true envelope estimator appears to be inversely related to the model order. This is due to the fact that the main change in complexity comes from the number of iterations that are required until convergence. For this a higher model order appears to be advantageous, because the difference between the initial cepstrum and the final envelope will be smaller. A similar argument may explain that the Hamming window smoothed cepstrum, that has been used for run time comparison in table (1) and that uses a larger number of coefficients, requires slightly less computation.

sound\method	LPC	dis. cep	TE	TE Hamming
tenor	1.6s	4.6s	7.s	5.45s
sopran	0.7s	1.3s	8.21s	6.67s

Table 1: *Calculation times for the spectral estimates for the whole tenor and soprano sound signals. The sound examples last 6s and 3s, respectively. The cepstral order is as specified in fig. 1 and fig. 2. For the Hamming windowed cepstrum the order has been selected to $P_C = 116$.*

3. REDUCING THE COMPUTATIONAL COMPLEXITY

Considering the true envelope algorithm we recognize that there exists a lot of wasted computation due to the large cepstrum size used. By means of carefully adapting the algorithm a significant reduction of the run time appears to be possible.

There exist two steps that are proposed to reduce the run time of the true envelope estimator. First the highly irrelevant information that is contained in the spectrum due to its oversampling is removed by means of sub-sampling of the log amplitude spectrum. Further sub-sampling can be performed if the cepstral model order is taken int account. In a second step the step size of the iterative algorithm can be controlled to achieve slightly faster convergence. These two steps together significantly reduce the computational complexity of the iterative algorithm, such that the true envelope estimator becomes much cheaper than the discrete cepstrum, and only slightly more expensive than the LPC.

3.1. Reducing size of the cepstrum

Given the fact that it is hardly possible to perceive the form of the spectral envelope with a precision given by the width of the spectral peaks it appears reasonable to down-sample the log amplitude spectrum such that a frequency bin has approximately the width of a sinusoidal peak. While the exact position of the amplitude samples is perceptually irrelevant, and therefore, can be safely quantized, the amplitude values of the spectral peaks should be treated more carefully. It is clear that it is the maximum value

of a peak that contains accessible information about the value of the spectral envelope. Therefore, the sub-sampling operation that is used to remove irrelevant spectral information replaces the original log-amplitude spectrum of size K by means of a sub-sampled spectrum $S(m)$ having size M being the largest power of 2 below the size of the analysis window. To keep the relevant information contained in the maxima of the spectral peaks sub-sampling is done using a maximum filter as follows

$$S(m) = \max_{k=m(u-0.5)}^{m(u+0.5)-1}(\log(X(k))), \text{where} \quad r = \frac{K}{M}. \qquad (3)$$

Because the removed information is irrelevant from a perceptive point of view a simple linear interpolation can be used to up-sample the spectral envelope that will be obtained from the down-sampled representation to the initial DFT size.

To further reduce the computation the order P_c of the cepstrum is considered. As has been shown in [12] a sub-band DFT can be used to reduce the amount of computation if the signal under investigation is band limited. They show that the standard K-point DFT can be decomposed into two K/2-point DFT, one operating on the lower half band and the other on the upper half band. The sub-band DFTs operate on two low resp. high pass filtered and sub-sampled signals. The filtering and sub-sampling introduces linear distortion and aliasing, however, in combining the two sub-band DFTs the errors will cancel and the complete K-point DFT is obtained. While the whole process is computationally more demanding than the direct evaluation of the K-point DFT, savings can be achieved if one of the sub-band DFT can be neglected because the energy in the band is low. The whole process can be iterated to further reduce computational complexity and effective bandwidth of the DFT. For the true envelope estimation we are seeking to calculate the low frequency bins of a DFT. Therefore, we will consider the result of the DFT of the lowest frequency band, only.

Due to the use of only the lower sub-band DFT to construct the cepstral coefficients related to the complete K-point DFT a systematic error is introduced. One part of the error is due to the pass-band attenuation of the sub-band filter, the other part due to the weak stop-band rejection of the sub-band filter which leads to aliasing errors. From the detailed investigation presented in [12] it can be concluded that for sub-sampling 8 times above the required band limit of the DFT the aliasing error is in the order of 20dB below the maximum amplitude in the aliased band. Note however, that for the iterative procedure considered here the out of band energy and the aliasing due to limited stop-band rejection will reduce with the ongoing iteration. While initially the out of band energy may be significant this energy and the resulting aliasing error will diminish with the envelope approaching the final position. The algorithm stops if the complete spectrum is covered by the envelope in which case the out of band energy will be zero and no aliasing takes place any more.

According to the previous discussion the cepstrum size L is selected to be $8P_c$ such that even for the first iteration only a small amount of aliasing will take place. Moreover, instead of the mean filter used in [12] prior to sub-sampling, we again use a maximum filter to best preserve the relevant information

$$A_0(l) = \max_{m=lh}^{l(h+1)-1}(S(m)), \text{where} \quad h = \frac{M}{L}. \qquad (4)$$

Note, that the two maximum filters eq. (3) and eq. (4) can be combined into a single operation. With the selected L the placement error is smaller than an 8-th part of the period of the cepstral component with highest frequency. Under the assumption that the amplitude of the cepstral coefficients decrease with their frequency this offset appears to be acceptable. This is even more true for the cepstral smoothing using a Hamming window, because in this case the highest cepstral coefficients are used only with a very weak weight. For interpolation from this stage of sub-sampling, two options are available. We may either perform another linear interpolation together with the up-sampling from the previous stage, or, to obtain a smooth contour, a band limited interpolation may be performed in the spectral domain. In the present investigation the second option has been selected.

3.2. Controling the step size of the iterative algorithm

In the original version of the iterative algorithm the cepstral coefficients of the next iteration are created by simply filtering the cepstrum. The speed of the adaptation depends on the match between the part of the cepstrum that will be retained and the energy distribution of the spectrum to be enveloped. If the log amplitude spectrum has a form such that its spectrum is completely contained in the part of the cepstrum below the cepstral order the algorithm will converge in a single step. If, however, the log amplitude spectrum contains impulses (partials) the energy in the cepstrum will have a large band width and the part of the energy that is used to adapt the envelope will only be the small portion of the energy that falls into the kept band. Generally speaking the more the changes that are required to update the envelope are related to energy in the band above the cepstral order, the smaller will be the correlation with the basic functions in the lower cepstral band, and, the longer the algorithm will need to converge.

It can be further expected that for later iterations the differences between the smoothed envelope and the original spectrum, eq. (2), will be more and more restricted to small regions. This means the differences will become smaller but also approach the form of impulses. Accordingly the rate of convergence of the algorithm will slow down. In the following the smoothed cepstral vector at iteration i is denoted as C_i and the cepstrum obtained from the current target vector according to eq. (2) will be denoted as C_i'. Moreover, we distinguish in band energy E_I of the cepstral change, that is the energy of the difference between C_i' and C_{i-1} confined to the region below the cepstral order P_c and the energy of the cepstral change outside of this band, which is denoted as E_O. Using these entities it is easy to show that for spectra with a single or a sequence of harmonic impulses the instantaneous convergence can be obtained if the change of the cepstral coefficients from the last step is multiplied by $\lambda = \frac{E_I+E_O}{E_I}$. For the general case the optimal step size control has not been found, however, our experiments showed that also for noisy spectra the use of the energy ratio leads to better performance than the more defensive strategy to use its square root as proposed in [13]. To update the cepstral coefficients we use now

$$C_i(r) = \lambda W(r)(C_i(r)' - C_{i-1}(r)) + C_{i-1}(r). \qquad (5)$$

The function $W(k)$ represents the window function, which here is either rectangular or Hamming, and which is centered at the origin of the cepstrum. The energy relation between in-band and out-of-band energy controls the step size of the algorithm. The important point why the windowing operation needs to be integrated into the coefficient update is that a repeated windowing of the cepstral coefficients would systematically increase the smoothing of the envelope even if no out-of-band energy is present. For all but

the rectangular window repeated windowing is not a transparent operation and needs to be avoided. This is achieved here because the window acts only ones on every change of the coefficients such that the estimated envelope will not be systematically smoothed.

3.3. Summary of the revised algorithm

The algorithm starts with the input of a amplitude spectrum of size K. Based on the selected cepstral order P_c a sub-sampled initial log amplitude spectrum $A_0(l)$ is generated according to eq. (3) and eq. (4). The last stage cepstral spectrum $C_{-1}(r)$ is initialized to 0. The first iteration of cepstral smoothing uses eq. (5) with step size fixed to $\lambda = 1$. The cepstral coefficients are then transformed back into log amplitude domain to get $V_0(l)$ which is compared by means of eq. (2) with the original input spectrum $A_0(l)$ to obtain the modified target spectrum for the next generation $A_1(l)$. From this a new cepstrum C_1' is calculated and the algorithm continues using a step size λ according to eq. (5). The iteration finishes if the initial log spectrum $A_0(l)$ is nowhere larger than the smoothed spectrum $V_i(l)$ by more than ΔdB. If required $V_i(l)$ is then up-sampled by $\frac{M}{L}$ in the spectral domain by means of multiplying the current set of cepstral coefficients with an appropriate phase vector and applying repeatedly a complex Fourier transform. The resulting spectrum is up-sampled again by means of linear interpolation to obtain the final spectral envelope. Note, that for the calculation of the real cepstrum the input and the output vector are both real and symmetric, such that the order of FFT and IFFT can be chosen as convenient. To be able to apply the frequency domain interpolation we have chosen to use a real valued FFT transform to obtain the cepstrum and to use the IFFT function to transform the cepstral coefficients that become complex after frequency domain interpolation back into the log amplitude domain.

4. CEPSTRAL ORDER SELECTION

One of the main advantages of the cepstrum based envelope estimation is the fact that the required cepstral order can be directly related to the frequency resolution that the algorithm should obtain. For an harmonic input spectrum the envelope estimator should not resolve the individual harmonics such that the main lobe of the smoothing kernel should have a width of about 2 times the fundamental frequency. For non harmonic spectra the width of the main lobe of the smoothing kernel should be adapted to the largest distance between spectral peaks that should be connected.

For a rectangular cepstral window the limit of the cepstral order that prevents the envelope to adapt to individual harmonics is given by

$$P_c \leq \frac{R}{2\delta_F}. \tag{6}$$

Here R is the sample rate of the signal and δ_F is the frequency peak spread limit, which is generally the maximum fundamental frequency to be treated. For a hamming window the main lobe is approximately $1\frac{2}{3}$ times as wide as for the rectangular window, such that the limits for the order are simply $1\frac{2}{3}$ times larger.

5. EVALUATION AND APPLICATION TO PITCH SHIFTING

To evaluate the new implementation of the true envelope estimator we first compare with the results of the spectral envelopes estimated with the original algorithm that have shown in fig. 1 and fig. 2. Despite the strong sub-sampling that is applied especially for the soprano example the deviations between the envelope is only marginal. The two envelopes with and without sub-sampling are shown for the soprano spectrum in fig. 3. The deviation due to the sub-sampling is as expected much less than a 4-th part of the fundamental frequency. Due to space constraints the figure for the comparison of the results for the tenor signal is not displayed. Due to the low pitch and higher order of the cesptral filter the variations are considerably smaller and the envelope deviation of the optimized version is less than 1dB.

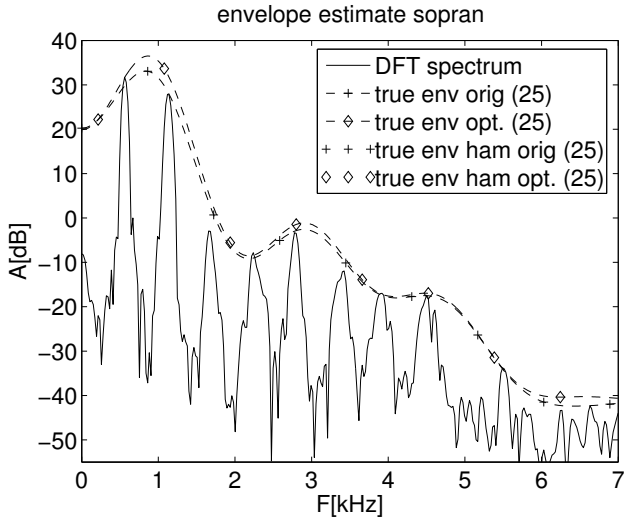

Figure 3: *Comparison of the estimated spectral envelope for the soprano singer using the original and the optimized true envelope estimator and $P_c = 25$. The variation in the true envelope due to the severe sub-sampling is less than 3dB.*

sound\method	orig.	sub	sub/step	sub/step (Ham)
tenor	7s	3.24s	2.2s	2.2s
soprano	8.21s	0.88	0.72	0.86s

Table 2: *Calculation times for the spectral estimates for the tenor and soprano sound signals using the optimized true envelope estimator with and without step size control. The cepstral order is the same as in table (1).*

The run time that is required for the estimating the spectral envelope of the test signals is listed in table (2). Using only the sub-sampling the run time is reduced by a factor of about 2 for the tenor signal and by a factor of 9 for the soprano signal. The additional step size control further reduces the run time for the tenor to achieve a total reduction of about 3.2 (hamming windowed 2.5) and for the soprano signal to 11.4 (hamming windowed 7.8) Compared to the LPC model the increase in run time is below 40%.

5.1. Application to pitch shifting

The true envelope estimator allows a significantly improved envelope estimation especially for high pitched sounds. The question arises, whether this improved estimate is perceptible when being used for spectral envelope preservation in combination with

a pitch shifting transformation. To investigate this question the true envelope estimator has been integrated into a phase vocoder application that allows to preserve the spectral envelope by means of pre-warping the envelope of the STFT frames before the pitch shifting takes place. The pre-warping is done by means of simple amplitude multiplication using the pre-warping factor $P(k)$ that is designed to compensate the transposition of the envelope that will be obtained in conjunction of the pitch shift. The pre-warping multiplier for pitch shifting with a transposition factor f and a spectral envelope $A(k)$ is

$$P(k) = \frac{A(k \cdot f)}{A(k)}. \tag{7}$$

Compared to the previous version that worked with an LPC based envelope estimator the artifacts are considerably reduced because the true envelope estimator will never adapt to individual harmonics of the spectrum if the order is selected according to eq. (6).

Nevertheless, for high pitched sounds the transposed signals sound rather dull. This is especially true if the pitch shift lowers the pitch of the signal. Further inspection of the problem reveals the following issues. The spectral envelope below the fundamental partial will generally have a rather steep slope towards 0. Pitch shifting down will therefore attenuate the fundamental and create a less complete sound perception. Moreover, due to the fact that the cepstral order needs to be adjusted to fit the highest fundamental frequency that is present in the whole signal, the formants that may be observed for lower pitched signals will be smoothed such that the sound is perceived as dull. A third problem for down transposition is that the originally unvoiced high frequency parts will be amplified by the larger amplitude of the lower frequency envelope.

There are two possible improvements of the algorithm that may reduce the first two problems. The third problem is still under investigation. First, the the cepstral order should be adapted (using eq. (6)) to an estimate of the fundamental frequency of the signal. Doing this slightly increases the perceived quality for high pitched sounds. Considering the attenuation of the fundamental we propose a second adjustment that concerns the amplitude warping function $P(k)$ that will be applied. Instead of pre-warping the amplitude using a fixed pre-warping factor it appears more appropriate to not pre-warp the lower part of the envelope below and around of the fundamental frequency. The pre-warping should start only for the second or even higher partials. This modification has two benefits: the fundamental keeps its amplitude, and, if the first formant is located close to the fundamental it will move synchronously with the pitch. The modified pre-warping function becomes

$$P(k) = \frac{A(k(D(k) + (1 - D(k))f))}{A(k)}, \text{with} \tag{8}$$

$$D(k) = \frac{1}{1 + \exp((k - k_0)/T_k)} \tag{9}$$

The transition location between pitch shifted and fixed envelope is defined by means of the bin position of the fundamental k_0. The transition bandwidth is specified by means of T_k. For pitch shifting up T_k can be selected without restrictions, for pitch shifting down it should be adapted to be large enough such that the resulting index will be monotonically increasing.

6. CONCLUSIONS

In the present article the problem of the estimation and preservation of the spectral envelope has been addressed. From the different algorithms available for envelope estimation the true envelope estimator appears to be a very promising choice, despite the fact that its run time requirements are rather high. A new implementation of the true envelope estimator has been presented that reduces the run time according to the cepstral order on average by a factor of 2.5 to 11. As a result the algorithm can be used for real time applications. The scaling behavior is reversed from initially decreasing run time with increasing cepstral order into increasing run time with cepstral order. Despite the increased efficiency the estimation results are not significantly affected.

The performance of the algorithm for the envelope preservation has been studied. Two problems have been identified that can be remedied by means adaptation of the cepstral smoothing order and the envelope pre-warping function to the current pitch.

7. REFERENCES

[1] M. Dolson, "The phase vocoder: A tutorial," *Computer Music Journal*, vol. 10, no. 4, pp. 14–27, 1986.

[2] M. Dolson and J. Laroche, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 3, pp. 323–332, 1999.

[3] A. Röbel, "Transient detection and preservation in the phase vocoder," in *Proc. Int. Computer Music Conference (ICMC)*, 2003, pp. 247–250.

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

[5] N. Bogaards, A. Röbel, and X. Rodet, "Sound analysis and processing with audiosculpt 2," in *Proc. Int. Computer Music Conference (ICMC)*, 2004.

[6] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electron. and Commun. in Japan*, vol. 62-A, no. 4, pp. 10–17, 1979, in Japanese.

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

[8] O. Cappé and E. Moulines, "Regularization techniques for discrete cepstrum estimation," *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 100–102, 1996.

[9] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*, Springer Verlag, 1976.

[10] A. V. Oppenheim and R. W. Schafer, *Digital Signal Processing*, NJ: Prentice Hall, 1975.

[11] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[12] O. V. Shentov, A. N. Hossen, S. K. Mitra, and U. Heute, "Subband DFT - interpretation, accuracy, and computational complexity," in *Proc of 25-th Asilomar Conf. on Signals, Systems and Computers*, 1991, pp. 95–100.

[13] A. Röbel and X. Rodet, "Real time signal transposition with envelope preservation in the phase vocoder," in *Proc. Int. Computer Music Conference (ICMC)*, 2005, to appear.