REAL TIME SIGNAL TRANSPOSITION WITH ENVELOPE PRESERVATION IN THE PHASE VOCODER

A. Röbel
IRCAM
1, place Igor Stravinsky, Paris

X. Rodet
IRCAM
1, place Igor Stravinsky, Paris

ABSTRACT

The following article presents a new real time implementation of an iterative cepstrum based spectral envelope estimation technique that was originally published under the name *true envelope*. Because the original algorithm is hardly known outside Japan we will first describe the algorithm and compare it to the standard techniques, i.e. LPC and discrete cepstrum. The estimation properties are compared and it is shown that the true envelope estimator achieves convincing envelope estimations even for problematic, high pitch signals. The algorithm is analyzed with the objective to find an efficient implementation that sufficiently reduces the computational complexity such that the algorithm can be used in real time within the phase vocoder. The implementation that is presented reduces the run time required by the algorithm depending on the cepstral order on the estimation parameters by a factor of 2 to 9 such that real time processing becomes feasible.

1. INTRODUCTION

Use of the digital phase vocoder technique [3] for signal transformation has a long history. If recent improvements are respected [4, 11] the phase vocoder can be considered an efficient technique that achieves high quality time-scale modifications in real time [1]. The application of the phase vocoder is not limited to time-scale modifications, it can be used for pitch-shifting as well. There exist two approaches to achieve pitch-shifting in the phase vocoder [8], 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 consists of pre-warping the spectral envelope of each signal frame in the spectral domain before the transposition is actually performed [3]. To be able to perform the pre-warping a robust estimator of the spectral envelope is needed. Due to real time constraints the estimation should be efficient and result in reasonable estimates for signals containing sinusoidal and noise signal components.

In the following we will present a optimized implementation of the *true envelope* estimator. The *true envelope* estimator, despite its very robust estimation results, has not received much attention outside Japan, because the original publication has been in Japanese [7]. Before discussing the new implementation we will briefly describe the problems that are related to the estimation of the spectral envelope and will give a short overview over the estimation techniques that are currently available. We will summarize the computational complexity and demonstrate the favorable properties of the *true envelope* estimator. An optimized implementation is proposed that reduces the required computation to a minimum. The implementation features the advantageous properties of the discrete cepstrum [6] without its drawbacks [2], and with reduced computational complexity. In comparison with all-pole modeling the algorithm requires an increase in run time of less than 60% compared to the Levinson-Durbin recursion. Accordingly, the algorithm can be run in real time and in parallel with the phase vocoder.

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 and in section **4** we compare the results obtained with the optimized version with the results obtained with the original algorithm and discuss further improvements.

2. ESTIMATING THE SPECTRAL ENVELOPE

The estimation of the spectral envelope is not a straightforward operation. Numerous investigations have been conducted to find suitable methods. Most of the techniques are based on either linear prediction (LPC) [9] or the real cepstrum [10]. For a LPC model 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, especially if they are high pitched, the all-pole model suffers from systematic errors and sound quality degenerates. The discrete all-pole model proposed in [5] resolves part of the problem, however, requires a nonlinear adaptive optimization such that it can not be used for real time applications.

The alternative approach is based on cepstral smoothing which relies 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 [6, 2] and the true envelope [7].

The real cepstrum $C(r)$ 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 $C(r)$ of a sig-

nal frame is

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

Because the spectral envelope is considered to be a smooth contour that connects the spectral peaks a simple means to obtain an estimate of the spectral envelope is to set all the high frequency elements in the cepstrum to 0. The number of bins 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 as desired the contour of the spectral peaks. The discrete cepstrum proposed in [6, 2] 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 method requires a fundamental frequency analysis, 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 [7] where it was described as a method that estimates the *true envelope*. The algorithm is iterative and a straightforward implementation is computationally expensive. Let $V_i(k)$ be the cepstral representation of the smoothed 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 target amplitude 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 in the original spectrum will be filled by the mean spectrum and the estimated envelope will steadily grow until all the peaks are covered. As stopping criterion of the procedure the condition a parameter Δ is required that defines the maximum excess hat a peak of the observed spectrum is allowed to have above the spectral envelope (in the following experiments $\Delta = 2dB$ is used). The graceful handling of the signal spectrum according to the requested cepstral order is a very nice feature of the algorithm avoiding all the problems observed due to ill-conditioned setup of the discrete cepstrum [2]. The drawback of the procedure is 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, the algorithm scales with $O(K\log(K))$.

To evaluate the real world performance of the algorithms we have selected two examples of singing voice sounds. The model orders have been selected such that the acoustic result obtained when used to pre-warp the spectral envelope during transposition achieves subjectively the highest quality for the whole sound. Due to the high frequency resolution of a spectral representation using an AR model the LPC order has to be limited to lower orders to prevent representation of partials in the spectral envelope. For the cepstral models there exists a simple relation between the largest distance δ_F between two neighboring spectral peaks that need to be interpolated (F0 for harmonic sounds) and the optimal cepstral order P_c. Given

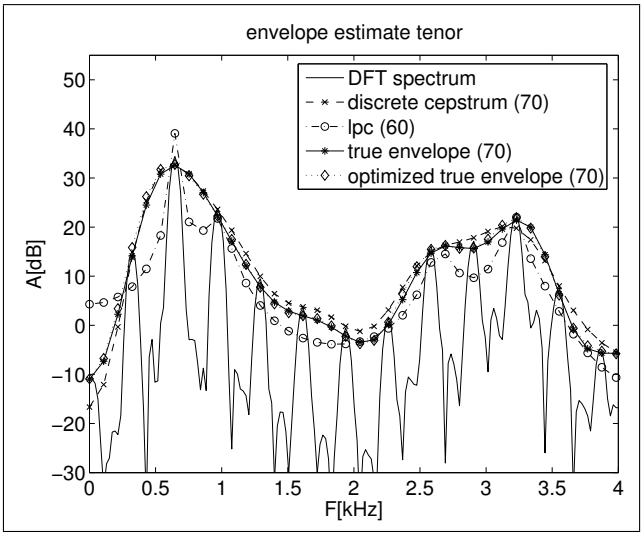

Figure 1. The signal spectrum of a tenor singing voice segment with the envelope estimates obtained with standard LPC $P_l = 60$, the discrete cepstrum and two versions of the true envelope estimator (standard and optimized). All cepstral models use $P_c = 70$, which is derived from the maximum F0 of the sound according to eq. (3).

δ_F and the sample rate R the order P_c has to be selected according to

$$P_c \leq \frac{R}{2\delta_F} \quad (3)$$

to prevent the sinusoidal peaks to be individually resolved in the envelope.

The first example is a tenor singer displayed in fig. 1. The sound segment used for fig. 1 has a rather 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 even over-estimates their amplitude, while at the same time it under-estimates the envelope for low amplitude peaks. 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 passes exactly through the selected peaks deviating only if the cepstral order or the regularization requires it. The true envelope follows the same contour, however, due to the selected Δ its level is slightly lower. The difference is hardly perceptually relevant.

As a second sound example we consider a soprano singer. The signal 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. The discrete cepstrum and the true envelope perform similar, however, in this case the bad constraint of the discrete cepstrum at frequency zero is becoming a problem.

The run time of all the algorithms are compared in the left three data columns of 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 suf-

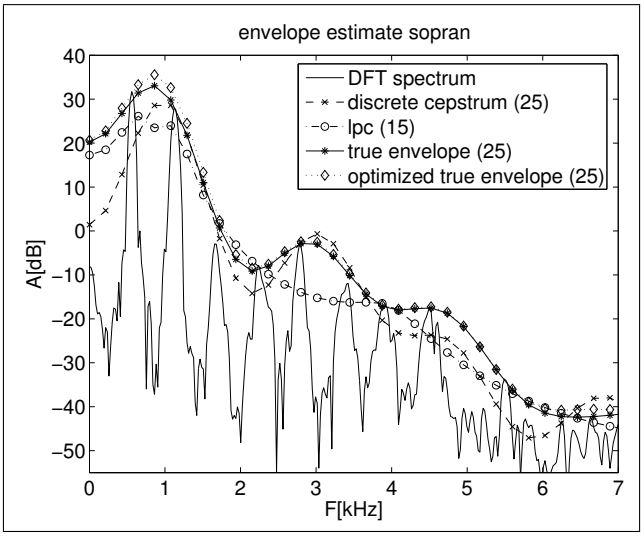

Figure 2. Estimating the spectral envelope for a sopran singer. The spectrum is shown with envelope estimates obtained with standard LPC $P_l = 15$, the discrete cepstrum and the true envelope estimator original and new optimized version (all $P_c = 25$).

sound	LPC	dis. cep	true env	true env opt.
tenor	1.6s	4.58s	7s	2.7s
sopran	0.70s	1.30s	8.21s	0.86s

Table 1. Calculation times for the spectral estimates in for the whole tenor and soprano sound signals. The sound examples last 6s and 3s, respectively. The calculation has been performed using a 2.4GHz Pentium 4 processor.

ficiently 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 to converge. For this a higher model order appears to be advantageous, because the difference between the initial cepstrum and the final envelope will be smaller. Considering the true envelope algorithm we recognize that there exists a lot of wasted computation due to the large sized cepstrum. By means of carefully adapting the algorithm a significant reduction of the run time appears to be possible.

3. REDUCING THE COMPUTATIONAL COMPLEXITY OF THE TRUE ENVELOPE

There exist two steps that are proposed to reduce the run time of the true envelope estimator. First, irrelevant information is removed by means of sub-sampling the log amplitude spectrum. Second, the knowledge of the cepstral order is used to further reduce the size of the cepstrum such that the required transformations become cheaper.

Given the fact that it is hardly possible to perceive the form of the spectral envelope with a precision well below the fundamental frequency the over sampled representation of the spectrum that is used for most applications is not required for the envelope estimation. A single bin per sinusoid is sufficient to represent the amplitude spectrum.

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 power of 2 above half the size of the analysis window. The initial sub-sampled spectrum used for the first iteration of the algorithm is created by means of the maximum filter

$$S_0(m) = \max_{k=m(u-0.5)}^{m(u+0.5)-1}(\log(X(k))), \text{where} \quad u = \frac{K}{M} \quad (4)$$

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.

The spectral envelope is available only after sub-sampling due to the spectral peaks of the (harmonic or in-harmonic) sound source. According to the previous discussion the cepstrum size L is selected to be the power of two larger or equal to $8P_c$. Concerning the sub-band filter we propose the use of the maximum filter eq. (4) instead of the mean filter used in [12]. For the proposed sub-sampling

of the observed spectrum the maximum filter introduces an error that is related to an envelope displacement of $\frac{\delta_F}{4}$ which, compared to the possible spectral resolution determined by the harmonic peaks, appears negligible for quasi harmonic sounds.

To proceed with the initialization of the algorithm we further sub-sample the initial log amplitude spectrum

$$A_0(l) = \max_{m=lh}^{l(h+1)-1}(S_0(m)), \text{where} \quad h = \frac{M}{L} \qquad (5)$$

and start the algorithm at iteration $i = 1$. Note, that the two maximum filters eq. (4) and eq. (5) can be combined.

In the following we will investigate into the speed of convergence of the iterative algorithm. Assume first a target spectrum A_0 that can be represented by the cepstral coefficient without error. Here the algorithm will converge in a single step. If the target spectrum is less smooth the convergence will require more iterations, because the out of band energy creates local amplitude peaks above the cepstrally smoothed envelope. If all the cepstral coefficients would be phase aligned, as for example in the case of an ideal harmonic impulse spectrum with arbitrary spectral envelope the optimal step size could be calculated form the total sum of amplitudes of the cepstrum divided by the sum of amplitudes within the selected quefrency band. In the general case, however, the different cepstral basis functions will not be phase aligned such that the peak level above the cepstrally smoothed envelope will be much smaller as indicated by the sum of amplitudes. A more defensive assumption is that the amplitudes add according to their energies. We now distinguish the cepstral coefficients C_i' obtained by computing the cepstral coefficients of the target envelope $A_i(k)$ and the coefficients that are output as result of the i-th iteration C_i. If we denote the in-band energy of the change of the cepstral coefficients, which is the energy of the difference between C_i and C_{i-1} confined to the region below the cepstral order P_c as E_I and the out-of-band energy as E_O we propose to speed up convergence by amplifying the change of the cepstral coefficients for all but the first iteration as follows

$$\lambda = \sqrt{\frac{E_I + E_O}{E_I}} \qquad (6)$$

$$C_i(r) = \lambda(C_i(r)' - C_{i-1}(r)) + C_{i-1}(r). \qquad (7)$$

Here r is the quefrency bin. With the updated C_i the new envelope model V_i and according to (eq. (2)) the next target amplitude A_i are derived. Iteration stops if $max(A_i(k) - V_i(k)) < \Delta$. After convergence the first level of interpolation is done in the spectral domain using h shifted IFFT operations of size L while for the next level linear interpolation can be used.

4. EVALUATION AND CONCLUSION

Having explained the new implementation of the true envelope estimator we will briefly study the results of the optimized estimator in the figures fig. 1 and fig. 2. The visual inspection shows that the differences to the original implementation are very small. We conclude that the strong sub-sampling that is performed for the low order model of the soprano singer does not have any negative impact on the results. With respect to the decrease in computational complexity the results in the last column in table (1) demonstrate that with the new implementation the computational requirements are close to the ones of the LPC model requiring less than 60% increase in run time. The effect of the step size control depends on the spectral characteristics of the log amplitude spectrum. It accounts for a small part of the improvement with up to 20% run time reduction.

Concerning further improvements that are currently under investigation we mention the use of a smooth window function (Hanning or Hamming) in stead of a rectangular window to when creating the cepstral smoothing.

5. REFERENCES

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

[2] Cappé O. and Moulines E. "Regularization techniques for discrete cepstrum estimation". *IEEE Signal Processing Letters*, 3(4):100–102, 1996.

[3] Dolson M. "The phase vocoder: A tutorial". *Computer Music Journal*, 10(4):14–27, 1986.

[4] Dolson M. and Laroche J. "Improved phase vocoder time-scale modification of audio". *IEEE Transactions on Speech and Audio Processing*, 7(3):323–332, 1999.

[5] El-Jaroudi A. and Makhoul J. "Discrete all-pole modeling". *IEEE Transactions on Signal Processing*, 39(2):411–423, 1991.

[6] Galas T. and Rodet X. "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 (ICMC)*, pp. 82–84. 1990.

[7] Imai S. and Abe Y. "Spectral envelope extraction by improved cepstral method". *Electron. and Commun. in Japan*, 62-A(4):10–17, 1979. In Japanese.

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

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

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

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

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