EXTENDING EFFICIENT SPECTRAL ENVELOPE MODELING TO MEL-FREQUENCY BASED REPRESENTATION

Fernando Villavicencio, Axel Röbel and Xavier Rodet

IRCAM-CNRS-STMS
Analysis-Synthesis team
Place Igor-Stravinsky 75004 Paris France
{villavicencio,roebel,rodet}@ircam.fr

ABSTRACT

In this work we consider the problem of spectral envelope estimation using spectra with perceptually warped frequency axis. The goal of this work is the reduction of the order of the spectral envelope model which will facilitate the use of these envelopes for training of voice conversion systems. We adapt the True-Envelope estimator to Mel-frequency representations and adapt a recently proposed cepstral model order selection criterion taking into account the distortion of the frequency axis. We evaluate the modified order selection procedure using a perceptual framework for the evaluation of envelope estimation errors. The experimental evaluation carried out with real speech confirms our modifications. The results demonstrate that the mel frequency based true envelope estimator achieves superior envelope estimation with significantly reduced model order.

Index Terms— speech synthesis, speech analysis, envelope detection, speech codecs, speech enhancement.

1. INTRODUCTION

The estimation of the spectral envelope represents one of the main tasks performed in several speech and audio processing applications. Defined as the smooth function passing through the prominent peaks of the spectrum, the spectral envelope is generally considered as one of the determining factors for the timbre of a sound. Accordingly, the quality of systems performing voice or timbre modification depends significantly on the quality of the envelope estimate.

The selection of the proper filter model (AR, MA, ARMA) and the corresponding model order are considered the main issues if aiming to perform efficient estimation, since they both are generally unknown. Autoregressive and cepstrum based methods are commonly used as envelope models. The main problems observed in such methods are the fact that they do not modelize the expected envelope, and that the model order is rarely adapted to the characteristics of the signal. On the other hand, efficient spectral envelope estimation can be achieved by means of the True-Envelope estimator (TE) [1]. Following [2], an optimal order selection is derived in terms of the F_0 and the sampling rate. However, for high-quality signals such selection could lead to high order values.

Human perception is frequently considered to reduce the dimensionality of parametric representations of the spectral information. It is well known that not all the information of the observed spectrum is equally important since a logarithmic scaling of the frequency axis is closer to the resolution used by the ear to perceive sound. We found that some envelope models considering such frequency scaling as the well-known Mel-Frequency Cesptral Coefficients (MFCC) and Warped Linear Prediction (WLPC) [3] are not well adapted to achieve signal modification since MFCC represents mainly the frequency-bands energy shape, while the envelopes deduced by WLPC include the harmonic structure. However, we argue that the application of frequency scaling in the TE method should lead us to achieve efficient modelisation of perceptually scaled spectra.

In this article we extend the estimation performed by TE and its autoregressive version (TELPC) to Mel-frequency based representations in order to reduce the model complexity when applying the order selection described in [2]. In this way we aim to obtain a reduced parametric model minimizing the perceptual effect of the missed envelope details. Also, we propose an adaptation of the order selection procedure by taking into account the frequency-axis warping. Finally, instead of typical spectral distortion measures, we propose the use of perceptive criteria inspired by the PESQ standard [4] to quantify the modeling error and evaluate the order selection adaptation.

The work is organized as follows. The efficient True-Envelope estimator and the order selection are described in section two. Mel-frequency scaling applied to True-Envelope and its autoregressive version are presented in section three and four. The methodology used for a perceptual evaluation of spectral envelope estimations is addressed in section five. An experimental evaluation of the proposed methods including the well-known technique based on linear prediction is found in section six. We finish the work in section seven with some conclusions.

2. IMPROVED CEPSTRUM-BASED SPECTRAL ENVELOPE MODELING

There are various techniques for cepstrum-based envelope estimation. In [1] an efficient spectral envelope estimator called True-Envelope (TE) was presented. This iterative technique allows efficient estimation of the spectral envelope [5] without the shortcomings of the discrete cepstrum [6]. The resulting estimation can be interpreted as a band limited interpolation of the major spectral peaks.

2.1. Order selection

A major advantage of the cepstral envelope estimation techniques is that a reasonable estimate of the optimal cepstral order can be provided. If the observed signal has the fundamental frequency F_0, the harmonic excitation spectrum samples the resonator filter with

This work was supported by the National Council for Science and Technology of Mexico, grant no. 143775

a sampling rate given by F_0. Therefore, one may deduce that the information from the original filter that exceeds the related Nyquist bandlimit in the cepstral domain is lost. Accordingly, a theoritical proper bound for the cepstral order can be obtained [2] in the form

$$\hat{O} = \frac{FS}{2F_0} = \alpha\frac{FS}{F_0}, \quad \alpha_c = 0.5 \tag{1}$$

While the real optimal order, that is the order that provides an envelope estimate with minimum error, depends on the specific properties of the envelope spectrum, the order selection according to (1) is reasonable for a wide range of situations and the resulting error has been found through experimentation to be rather close to the optimal order determined by exhaustive search.

3. EFFICIENT FREQUENCY-WARPED ENVELOPE ESTIMATION

3.1. Mel-scaled True-envelope estimation

The Mel scale is considered as an accurate representation of the frequential precision of human perception. The use of this scale has been found to have an advantage, since it allows a compression of the spectral information minimizing the perceptual effect of missed information. Accordingly, in our proposition the spectrum used by the True-Envelope estimator algorithm will be warped using a Mel-based function in order to obtain a perceptual resolution on the estimated envelope. We will refer to the resulting method as the Mel True-Envelope estimator.

3.2. Adaptation of the order selection to a warped representation

Following the ideas of order selection for cepstral models outlined above, an efficient order selection would depend on frequency due to the frequency-dependant warping factor. If the spectral envelope is represented in terms of a single cepstral model, however, the order has to be fixed. Moreover, the important question is which frequency band should be selected to guide the order selection process so that perceptually the envelope obtained in the mel frequency representation is as close as possible to the real envelope. Based on perceptual cues, we argue that the fact that we are mainly interested in a precise estimation at low frequencies give us a reason to propose an adaptation of the order selection based on the lowest Δ_{F0}. This value is found at the position of the first F_0 partial after frequency warping since we apply the Mel-scaling in a way that the axis limits are preserved ($0 \leq F_{melHz} \leq FS/2$). Accordingly, the adapation can be simplified to apply the inverse of the corresponding mel scaling factor at $F = F_0$ to the α parameter in (1). Therefore, the resulting α value should be found around 0.15 for $F_0 < 1KHz$ since the bark scaling is linear until this value. Clearly, this adaptation leads to reduced order values.

Note that due to the frequency-dependant compression of the frequency-axis, theoretically, the local order will increase with the frequency. Accordingly, the optimal order will generally be greater than the one selected using $\alpha = 0.15$ in (1). The impact of the different frequency bands for the order selection will depend on the energy and the perceptual sensitivity in the related bands. The evaluation of this phenomena is one of the motivations of the experimentation described in section 6.

4. IMPROVING ALL-POLE FREQUENCY-WARPED ENVELOPE MODELING

Spectral estimation based on frequency-warping is commonly done by replacing the unit delay elements of a conventional linear filter structure by first-order allpass filters. In [3], this technique, also known as Warped Linear Predictive Coding (WLPC), is described in detail and compared with the conventional LPC technique. As for the proposed MTELPC method, the aim of this technique is to perform an AR based estimation of a frequency warped version of a spectrum. The problem we found in WLPC is related to the fact that the extracted envelopes do not match the definition of spectral envelope since the most of the harmonic peaks at low frequencies are included in the estimation.

This characteristic could appear advantageous for inharmonic excitation signals or applications looking to keep the F_0 information (i.e. speech coding). However, for timbre modificiation it is essential to separate the contributions of the source and filter components.

Note, that the spectral flatness measure used as a performance measure in [3] is not well adapted to harmonic spectra since, as opposed to the one defined in [7], it takes into account the whole spectrum. Therefore, the improvements show by WLPC using this measure are mostly interesting for unvoiced speech.

On the other hand, once a proper order is selected to perform a MTE estimate, it can be used as shown in [7] to fit an all-pole model of the frequency warped representation. We denote this method as the Mel-based True-Envelope LPC (MTELPC). As stated in the first section, the estimation performed by True-Envelope follows the shape defined by the spectral peaks and not the peaks itself for either linear or warped frequency resolutions. Therefore, MTELPC is expected to match, in contract to WLPC, the desired shape wtihout extracting the harmonic structure of the spectrum.

Note that MTELPC performance will be limited firstly by the MTE estimate given as input of the autoregressive model and secondly by the order of the autoregressive model itself. To resolve the first point, we are especially interested in using an adequate order selection criteria, as the one we proposed in the previous section. Once the MTE is defined, we can deduce that the resulting all-pole envelope will become closer to the MTE estimate as the order of the AR system increases.

5. PERCEPTUAL EVALUATION OF SPECTRAL ENVELOPE ESTIMATION

A problem of the evaluation of spectral envelope models by means of simply comparing the differences of the transfer functions is that the perceived impact of the differences is not taken into account. Therefore, we propose to use a perceptual framework to evaluate the envelope estimations. In this way, considering the current state of the art of speech and audio quality evaluation, we followed some concepts found in the PESQ standard[4] to include percpetual criteria in our evaluation. Firstly, the use of a perceptually-based frequency resolution. Secondly, the application of the middle-ear filter which represents a frequency dependent attenuation of the spectral energy and the use of a spectral energy quantification based on the concept of loudness. Finally, the consideration of perceptual bands, as performed by the inner ear.

Note that when an envelope estimation evaluation is carried out using real speech signals, the reference envelope is not available since we have no acces to the underlying transfer function. Nevertheless, given the subsampled nature of the observed envelope, it seems reasonable to use the ideally band-limited interpolation as the

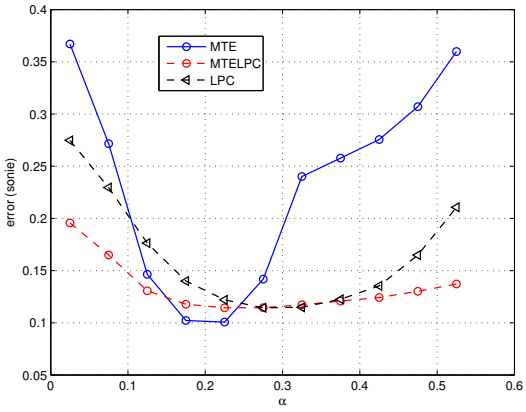

Fig. 1. Perceptual error as a function of the local α (male speaker). In all cases MTELPC uses $\alpha = 0.15$ for the MTE estimate.

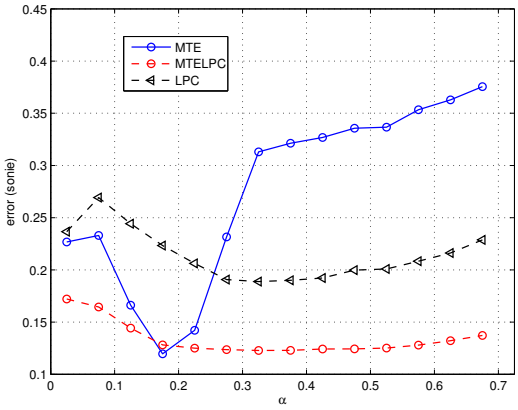

Fig. 2. Perceptual error as a function of the local α (female speaker). In all cases MTELPC uses $\alpha = 0.15$ for the MTE estimate.

reference envelope. As has been explained in [2], this interpolation is approximately performed on the linear frequency-axis by the TE estimator using the order selection according to (1). The resulting modeling error when comparing both reference and evaluation envelopes (obtained from the MTE, MTELPC and LPC methods) is also averaged at each perceptual band.

6. EVALUATION ON REAL SPEECH

6.1. Procedure

Firstly, we were interested in evaluating the α parameter for the MTE order selection. Once a proper value was defined, it was used to perform MTELPC estimates in order to carry out a comparison of the estimation performance among LPC, MTE, and MTELPC using the perceptive model previously outlined. We used real speech signals from a male and a female speaker so that we could also evaluate the order selection related to the mean F_0 of the speaker. For the three

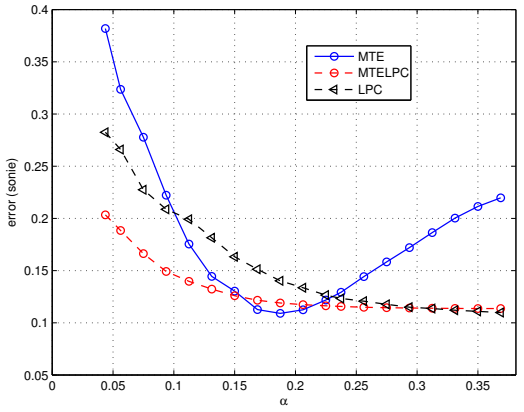

Fig. 3. Perceptual error as a function of the average α (male speaker). In all cases MTELPC uses $\alpha = 0.15$ for the MTE estimate.

methods, we used several model orders within the range $\hat{O} = [7, 59]$. This range, and the one defined by the F_0 values, covers approximately α within the range $[0, 0.7]$. On the other hand, a discrete grid was defined to quantify the resulting α value in function of the F_0 value and the analysis order at each frame. Linear TE estimation with order selection is also performed since it will represent our reference envelope. Ten-thousand frames (75 percent of them considered as voiced) were analyzed for each speaker, corresponding to approximately 40 short phrases. The mean F_0 values for the male and female speaker are close to $150Hz$ and $240Hz$, respectively.

6.2. Results

The resulting average error produced by the MTE estimates showed a parabolic form and its minimum was found to be, as expected, slightly higher than the proposed value (α=0.15), as shown in fig. 1 for the male speaker. This performance is especially clear in fig. 2, which corresponds to the female speaker case and then the mean F_0 is significantly higher.

We computed the MTELPC estimates fixing the resulting α value to 0.15 since we are specifically interested in evaluation the effect of the AR model order when using already optimized MTE estimation. The results are already included in fig. 1 and fig. 2. As we stated before, it showed an improved performance for increasing α, which can be explained by the increasing capacity of the AR model to fit the underlying MTE estimate. We remark, however, that the best result of MTELPC can not outperform the one of MTE.

Clearly, LPC was found to perform worse than the MTE based methodss. As expected, this effect increases since the aliasing of the autocorrelation function used by LPC grows for increased F_0. [8].

In order to measure the effect of the voicing nature we carried out the evaluation without any voicing discrimination criterion on the frames. In theory, this is benefical to LPC since it represents the best method for envelope estimation of inharmonic spectra. Also, it clearly reduces the cases where the order selection has a theoretical basis. Despite this fact,the improvements shown by LPC were not significant, affirming the proposed α value as a proper choice.

We also considered the case when the model order remains unchanged since many applications are generally restricted to this con-

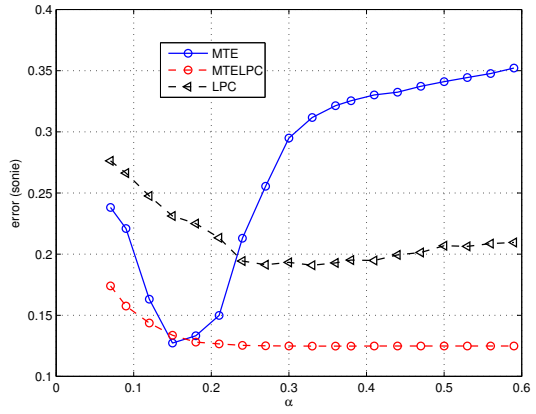

Fig. 4. Perceptual error as a function of the average α (female speaker). In all cases MTELPC uses $\alpha = 0.15$ for the MTE estimate.

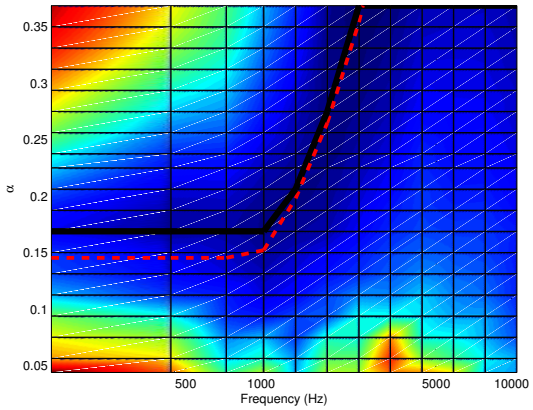

Fig. 5. Perceptual error (intensity) as a function of α and the frequency band (male speaker), MTE estimation. At each band, the solid line follows the resulting minimal error location while the dotted one represents the theoretical α value considering the local frequency warping.

dition. Principally, we aim to measure the optimality of the order selection in MTE when considering only the mean F_0. For MTELPC only the AR model order remains unchanged since, as the previous test, the selection criterion denoted by $\alpha = 0.15$ is already considered at each analysis frame to compute the MTE estimates used to fit the AR model. The results are shown in fig. 3 and fig. 4 and are found to be rather similar to those obtained using variable order. We can therefore state that average order selection should be preferred over any arbitrary criterion when the model order must remain unchanged.

Taking into account the modeling error per perceptual band, let us also clarify the effect of the local adaptation of α related to the frequency bands. We show in fig. 5 the resulting error-surface of the MTE modeling related to α within each perceptual band. Clearly, the α value corresponding to the minimal error at each perceptual band (solid line) follows the local scaling of δ_F (dotted line) after warping of the frequency axis. This view also lets us appreciate the fact that, compared to LPC, by using MTE based methods with order selection we are primarily maximizing the matching of the estimated envelope on the low-band region, which is, clearly, perceptually preferable.

7. CONCLUSIONS

In this work, we have presented an adaptation of the cepstrum based True envelope estimator for Mel-scale frequency axis warping. The previously known order selection criterion has been adapted taking into account the effective frequency warping. An evaluation framework based on perceptual criteria was proposed to estimate the perceptual impact when measuring envelope differences. The experimental comparison to linear predictive envelope estimation demonstrates that the Mel-frequency representation of the spectra allows us to achieve equivalent (for low pitch signals) or significantly improved (for high pitch signals) spectral envelope estimations with significantly reduced model orders. Further subjective tests will be carried out in the future to demonstrate the reliability of the perceptual evaluation framework.

8. REFERENCES

[1] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication (in Japanese)*, vol. 62, no. 4, pp. 10–17, 1979.

[2] A. Röbel, F. Villavicencio, and X. Rodet, "On cepstral and all-pole based spectral envelope modelling with unknown model order," *Pattern Recognition Letters*, vol. 28, no. 11, pp. 1343–1350, 2007.

[3] A. Härmä and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, July 2001.

[4] International Telecommunication Union, "Perceptual evaluation of speech quality (pesq)," Itu-t recommendation p.862, ITU-T.

[5] A. Röbel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proceedings of the International Conference on Digital Audio Effects, 2005. DAFx'05*, Spain, 2005.

[6] 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, 1990. ICMC'90.*, 1990, pp. 82–84.

[7] F. Villavicencio, A. Röbel, and X. Rodet, "Improving lpc spectral envelope extraction of voiced speech by true-envelope estimation," in *Proceedings of ICASSP '06.*, France, 2006.

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