IMPROVING LPC SPECTRAL ENVELOPE EXTRACTION OF VOICED SPEECH BY TRUE-ENVELOPE ESTIMATION

Fernando Villavicencio, Axel Röbel and Xavier Rodet

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

ABSTRACT

In this work we address the problem of all pole spectral envelope estimation for speech signals. The currently widely used all pole spectral envelope model suffers from well-known systematic errors and more severely from model order mismatch. We will propose a procedure to first establish a band limited interpolation of the observed spectrum using a recently rediscovered *true envelope* estimator and then using the band limited envelope to derive an all pole envelope model named TE-LPC . The band-limited envelope that is used to derive the all pole envelope model reduces the problem of the unknown all pole model order.

For the experimental investigation we propose a new perceptually motivated residual spectral peak flatness measure. The experimental results demonstrate that the proposed method significantly increases the spectral flatness for the perceptually especially important low order harmonics of voiced utterances.

1. INTRODUCTION

Several speaker modification applications as [1] are based on the extraction of the spectral envelope information from the speech signals giving special attention to voiced speech segments. This task can be achieved using various techniques offering different advantages and limitations.

Linear Prediction Coding (LPC) [2] is a well-known technique used for parametric representation of the spectrum in speech signals. Its main advantage comes from the all-pole characteristics of a simplified vocal tract model and the straightforward analogy of a source filter model with the speech production system. An alternative representation of the LPC polynomials, the Line Spectrum Frequencies (LSF) [3] offers advantageous interpolation properties.

We denote the term "spectral envelop" as a smooth function passing through the prominent peaks of the spectrum. For voiced speech, these peaks are principally related to the pitch harmonics, and therefore, the spectral envelope should be a transfer function that, if inverted, renders the sequence of spectral peaks as flat as possible. In this context, despite its advantages, the spectral envelope obtained by LPC suffers from a number of drawbacks. The first one is a model mismatch problem, related to the fact that the correct order of the all-pole model used in LPC for matching the signal spectrum is unknown, and usually it can hardly be obtained. Moreover, even if the model order were known, the spectral estimation contains systematic errors as a consequence of the aliasing that is taking place in the spectral domain due to the fact that the harmonic spectrum is sub-sampling the spectral envelope. These problems are especially manifested in voiced and high-pitched signals.

In our investigation we found that when using the standard LPC model for envelope estimation the perceptively particularly important information of the spectral envelope around the first few harmonics will be specially affected. These partials are important because they are individually resolved in the auditory system. Information that can be expected to be contained in the spectral envelope of these partials is the perceived speech quality (as voice "pressure" [4]), and formant dynamics.

Our proposition to improve the spectral envelope estimation is based on the *true envelope* estimator [5], for which an efficient real time implementation has recently been proposed [6]. This estimator is an iterative cepstral based technique that allows efficient estimation of the spectral envelope without the shortcomings of the discrete cepstrum [7], [8]. The resulting estimation can be interpreted as a band limited interpolation of the observed sub-sampled spectral envelope.

Using the *true envelope* estimation as input for the all pole model we follow the basic idea of [9] that is to keep the advantages offered by LPC speech processing but use interpolated envelope information with the aim of reduce the impact of the order mismatch described above, because the access order can not be used to flatten parts of the spectrum that are not related to spectral peaks. Results show a better performance of the proposed TE-LPC in a perceptually motivated spectral-peaks flatness measure, which is especially apparent in the low-frequency region.

The article is organized as follows. In section 2 we describe some deficiencies of the LPC modeling. The mentioned *true envelope* estimation is briefly described in section 3. The proposed modification in the LPC modeling is presented in section 4. A new spectral peak related spectral flatness measure and a comparison between conventional LPC and TE-ELPC performed on different speakers is found in section 5. The work ends with the results and conclusion in section 6.

2. LIMITATIONS OF LPC FITTING THE SPECTRAL ENVELOPE OF VOICED SPEECH

2.1. Autocorrelation matching

The use of LPC modeling implies some drawbacks fitting the spectral envelope. As we noted before, it suffers from systematic errors modeling the speech spectra specially manifested in harmonic segments (voiced speech) and being increased for high-pitched signals: the peaks of the spectral envelope estimated by LPC are highly biased towards the pitch harmonics. In part, the problem is due to the fact that the autocorrelation function computed from a sampled spectrum is an aliased version of the original continuous case [10] . As consequence, LPC matches the autocorrelation of an all-pole model with a distorted autocorrelation of the original signal. High-pitched signals contain fewer harmonics in the spectral representation, and therefore, the aliasing effect grows, leading to increased modeling errors. Using the discrete all pole (*DAP*) model [10] solves the problem if the transfer function is an all-pole filter and the order of the model is known. Unfortunately, the iterative optimization of the DAP model is computationally very demanding and its assumptions are generally not fulfilled.

2.2. Model mismatch and error criteria

An all-pole model is a simplified version of the acoustic model of the speech production system. However, it is not always well adapted to fit the speech spectra. Due to the form of the excitation pulse and for nasal sounds, due to the coupling between vocal and nasal cavities, the envelope of the speech spectrum will contain zeros that cannot be modeled by an all-pole transfer function. A possible solution to overcome this problem is to significantly increase the model order. Even if the observed spectral envelope could be correctly modeled by an all-pole filer the model order remains unknown.

Finally, the criteria of minimal residual energy used for LPC modeling is equivalent to the maximization of a spectral flatness measure of the complete residual spectrum [11]. For harmonic spectra, however, a spectral peaks flatness measure as for example the one used in the DAP model [10] is much more appropriate. Using the classical spectral flatness measure it is possible to keep some envelope features remaining in the residual spectrum. For the non-harmonic case, there are no more "valleys" formed by the harmonic peaks, and therefore, a spectral flatness measure that takes the whole spectrum into account is much more sensible. In that context, we can say that LPC performs better spectral fitting in unvoiced speech.

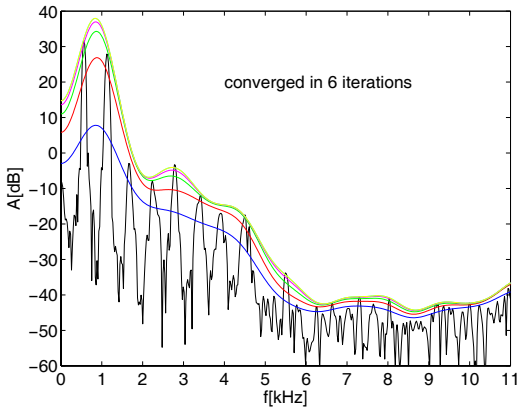

Fig. 1. True envelope estimator iterations

3. SPECTRAL ENVELOPE ESTIMATION BY TRUE ENVELOPE

The *true envelope* estimator has been proposed originally in [5]. Recently the iterative procedure has been significantly improved such that the computational costs are in the similar to the costs of the Levinson recursion such that real time processing can be achieved [6].

True envelope estimation is based on cepstral smoothing of the amplitude spectrum. Let $X(k)$ the K-point DFT of the signal frame $x(n)$ and $C_i(k)$ the cepstral representation of the smoothed spectral envelope at iteration i. The algorithm then iteratively updates the smoothing input spectrum $A_i(k)$ with the maximum of the original spectrum and the current cepstral representation

$$A_i(k) = \max(\log(|X(k)|), C_{i-1}(k)) \quad (1)$$

and applies the cepstral smoothing to $A_i(k)$ to obtain $C_i(k)$. The procedure is initialized setting $A_0(k) = \log(|X(k)|)$ and starting the cepstral smoothing to obtain $C_0(k)$.

As shown in Fig. 1 the estimated envelope will steadily grow. The algorithms stops if for all k the relation $A_i(k) < C_i(k) + \theta$ is true with θ being a user supplied threshold. For the current experiments $\theta = 2dB$ has been used. Given the fact that the cepstral order is limited the *true envelope* estimator performs a band limited interpolation of the prominent spectral peaks. The peaks that will be considered as prominent depend on the cepstral order. If a spectral envelope for an harmonic spectrum with fundamental frequency f_0 and sample rate F_s is required it is easy to show that the optimal order

is $\frac{F_s}{2f_0}$. Because the cepstral order can be changed independently for each frame the method allows to optimally interpolate the observed spectral peaks. Explicite peak selection that is necessary for the DAP estimator as well as for the discrete cepstre is not required.

4. TRUE ENVELOPE LPC MODELING

The nature of the drawbacks presented in the spectral modeling of voiced speech by the conventional LPC technique shows that LPC performance is limited by the method itself and also depends on the local characteristics of the signal.

Besides the fact that the autocorrelation function used in LPC suffers some aliasing, it does not represent the desired spectral information to be modeled since we are interested in fitting the spectral envelope as close as possible and not the original spectra. Accordingly we follow the proposition of [9] to first interpolate the spectrum using optimal band limited interpolation and then impose an high order all-pole model such that the LSF representation of the spectral envelope is still achievable. Besides this advantage we conjecture that the high order LPC model will obtain a better representation of the spectral narrow formants which are generally to broad after the band limited interpolation.

Denoting the K-point DFT of the speech segment $x(n)$ as $X(w_k)$, and the *true envelope* estimator as the operator $\mathrm{E_T}$, we can summarize the steps for the proposed modification as follows

$$\mathrm{E_T}\colon\ X(\omega_k) \rightarrow E_T(\omega_k) \tag{2}$$

We obtain the autocorrelation function RE(i) of the estimated spectral envelope by IDFT.

$$R_{E_T}(i) = \frac{1}{K}\sum_{k=1}^{K} E_T(\omega_k)^2 e^{j\omega_k i} \tag{3}$$

Now, the new TE-LPC filter coefficients are computed resolving the well-known Yule-Walker equations system with the modified autocorrelation information

$$\begin{bmatrix} R_{E_T}(0) & R_{E_T}(-1) & \dots & R_{E_T}(-p) \\ R_{E_T}(1) & R_{E_T}(0) & \dots & R_{E_T}(-p+1) \\ \vdots & \vdots & \ddots & \vdots \\ R_{E_T}(p) & R_{E_T}(-1) & \dots & R_{E_T}(0) \end{bmatrix} \begin{bmatrix} 1 \\ \hat{a}_1 \\ \vdots \\ \hat{a}_p \end{bmatrix} = \begin{bmatrix} \hat{e}_p \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{4}$$

We remark that after the modification of the autocorrelation function, the minimal residual energy characteristic of LPC is no longer valid. Related to the original signal, the resulting predictor is not optimal in the sense of the MSE criteria but it is supposed to fit closer the spectral envelope. A comparison between LPC and TE-LPC is shown in Fig. 2.

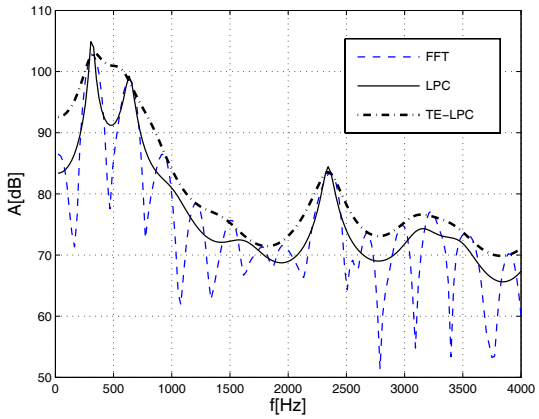

Fig. 2. Example of LPC and TE-LPC spectral fitting (model order=50)

5. EVALUATION

A widely used measure for the performance of an LPC model is based on measuring the whitening property of the inverse filter. The Spectral Flatness Measure, SFM, is defined as the ratio between geometric and arithmetic averages of a power spectrum according to

$$\mathrm{SFM}_{db}(S(\omega_i)) = -20\log_{10}\frac{G_m(S(\omega_i))}{A_m(S(\omega_i))} \tag{5}$$

Where

$$G_m(S(\omega_i)) = \sqrt[1/N]{\prod_{i=1}^{N}|S(\omega_i)|^2} \tag{6}$$

$$A_m(S(\omega_i)) = \frac{1}{N}\sum_{i=1}^{N}|S(\omega_i)|^2 \tag{7}$$

Decibel values are often used so in the case of a constant spectrum SFM=0 dB. Denoting $X(\omega_k)$ and $\hat{A}(\omega_k)$ as the K-point DFT of voiced speech and their correspondent all-pole model, we may express the residual SFM as follows

$$\mathrm{SFM}(S_r(\omega_k)) = \mathrm{SFM}(X(\omega_k)/\hat{A}(\omega_k)) \tag{8}$$

However, as we mentioned before, for voiced speech we are interested in the spectral flatness of spectral peaks of the residual spectra. To stay as close as possible to the original spectral flatness measure we are not using the discrete Itakura-Saito measure proposed for DAP but we are going to use a discrete version of the SFM which takes into account the harmonic peaks only. This spectral-peaks flatness measure, SPFM, is defined according to

$$\mathrm{SPFM}_{db}(S_r(\hat{\omega}_i)) = -20\log_{10}\frac{G_m(S_r(\hat{\omega}_i))}{A_m(S_r(\hat{\omega}_i))} \tag{9}$$

For

$$S_r(\hat{\omega}_i) = S_r(\omega_i), \quad \forall \quad \omega_k \approx (hF_0), \quad hF_0 < FS \quad (10)$$

For the following comparison, we define the *low band* as the low-pass spectral band containing the first few harmonic peaks (4 for this test), and the *high band* for the rest of harmonic or quasi-harmonic peaks (defined as the spectral maxima closer to expected harmonic frequencies). This choice has a psycho-acoustic basis: in the auditory human model, the first harmonics are resolved in individual perceptual bands, in contrast to the rest, for which an average of some peaks takes effect in each band. For completeness, even if we are mainly interested in the low band case, we will evaluate the SPFM also in the high band and the whole spectrum.

We evaluate the SPFM on normalized voiced speech residuals computed from conventional LPC and the proposed TE-LPC on low-medium pitched signals (male speakers) and high pitched signals (female speakers).

6. RESULTS AND DISCUSSION

Table 1 contains the resume of the results obtained with a speech corpus composed of 100 phonetically-balanced utterances of 4 different speakers (3 males, 1 female), the f_0 averages over all the sentences for each speaker are also shown. As performance measure we used the reduction percentage of the SPFM performed by TE-LPC compared to LPC. The evaluation was done with 2 different model orders.

	Low band (0-4$\mathbf{f_0}$)	**High band (4$\mathbf{f_0} - \mathbf{F_S}$)**	**Whole spectrum**	**Unvoiced speech**
order=56	SPFM reduction percentage [%]			
female	60	6	13	3
male 1	41	1	2	3
male 2	36	1	2	5
male 3	28	1	2	5
order=80	-	-	-	-
female	60	20	25	16
male 1	45	3	5	4
male 2	39	1	2	3
male 3	42	0	2	2

Table 1. LPC and TE-LPC performance comparison for different speakers with $\bar{f}_0$ = 242Hz, 153Hz, 131Hz and 129Hz respectively.

As expected, the results show that TE-LPC performs better SPFM maximization in all the cases we measured. While the improvements are rather small if measured over the whole spectrum or the high frequency band, they are significant in the perceptually especially important low frequency band. The improvement is bigger for high-pitched signals and is not very sensitive to the model order for the selected order values. Improvements found for the unvoiced cases could be due to voiced and mixed parts remaining in the unvoiced segments.

In conclusion, the proposed TE-LPC method performs better spectral envelope extraction of voiced speech in the perceptually important low-band region, allowing to improve applications where spectral envelope extraction is involved. The authors are currently investigating into objective performance comparison between the DAP and TE-LPC method.

7. REFERENCES

[1] A. Kain and M. Macon, "Spectral Voice Conversion for Text-to-Speech Synthesis," in *Proc. of the ICASSP'98*, USA, 1998, pp. 285–288.

[2] J. Makhoul, "Linear Prediction: A tutorial review," *Proc of the IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

[3] T. Bäckström W.B. Kleijn and P. Alku, "On Line Spectral Frequencies," *IEEE Signal Processing Letters*, vol. 10, no. 3, pp. 75–77, 2003.

[4] B. Doval N. Henrich and C. D'Alessandro, "Glottal Open Quotient Estimation using Linear Prediction," in *Proc. of the International Workshop on Models and Analysis of Voiced Emissions for Biomedical Applications*, Italy, 1999.

[5] S. Imai and Y. Abe, "Spectral Envelope Extraction by Improved Cepstral Method," *Electron. and Comm. (in Japan)*, vol. 62, no. 4, pp. 10–17, 1979.

[6] A. Röbel and X. Rodet, "Efficiente Spectral Envelope Estimation and its Application to Pitch Shifting and Envelope Preservation," in *Proc. of the DAF'x'05*, Spain, 2005.

[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] H. Fujisaki H. Hermansky and Y. Sato, "Spectral Envelope Sampling and Interpolation in Linear Predictive Analysis of Speech," in *Proc. of the ICASSP'84*, 1984, pp. 2.2.1–2.2.4.

[10] A. El-Jaroudi and J. Makhoul, "Discrete All-Pole Modeling," *IEEE Transactions on Signal Processing*, vol. 39, no. 2, pp. 411–423, 1991.

[11] S. Kay, *Modern Spectral Estimation: Theory and Application*, Prentice-Hall Signal Processing Series, USA, 1988.