A Time Regularization Technique for Discrete Spectral Envelopes Through Frequency Derivative

Gilles Degottex

***Abstract*—In most applications of sinusoidal models for speech signal, an amplitude spectral envelope is necessary. This envelope is not only assumed to fit the vocal tract filter response as accurately as possible, but it should also exhibit slow varying shapes across time. Indeed, time irregularities can generate artifacts in signal manipulations or increase improperly the features variance used in statistical models. In this letter, a simple technique is suggested to improve this time regularity. Considering that time regularity is characterized by slowly varying spectral shapes among successive frames, the basic idea is to smooth the frequency derivative of the envelope instead of its absolute value. Even though, this idea could be applied to different envelope models, the present letter describes its application to the simple linear interpolation envelope. Using real speech signals, the evaluation shows that the time irregularity can be drastically reduced. Additional experiments using synthetic signals also show that the accuracy of the original envelope is not degraded by the process.**

***Index Terms*—Amplitude spectral envelope, time regularity, sinusoidal model, speech modeling.**

I. Introduction

Sinusoidal and harmonic models represent the speech signal by means of frequency, amplitude and phase parameters estimated in analysis frames at regular time instants (e.g. each 2.5ms) [1], [2]. A broad range of applications make use of these models [3]–[7]. Building an amplitude spectral envelope based on a discrete set of sinusoidal amplitude parameters [8]–[11] is a necessary step for many applications [4], [7], [12]. A spectral envelope needs the following properties: i) It should correspond, accurately enough, to the frequency response of the vocal tract filter (VTF). Even though this criterion is difficult to assess because the VTF's ground truth is unknown, improved naturalness in voice processing techniques can be expected with more accurate envelopes. ii) The regularity of the spectral envelope across time is also necessary. Indeed, when manipulating the speech signal, fast varying erratic shapes in the estimated spectral envelope would result in perceived artifacts. Time regularity can also be assumed to be important for statistical models (e.g. in HMM-based speech synthesis [4], [13]). An excessive time variation of the envelope parameters would increase unnecessarily the variance of the statistical models. iii) Finally, the sinusoidal amplitude parameters should be preserved in the estimated envelope. This property might not be critical for recognition tasks. However, it is critical for synthesis where the sinusoidal amplitudes have to be preserved in a basic analysis and re-synthesis process. In this letter, we mainly address property (ii), namely the time regularization (or inter-frame regularity).

In the state of the art, mainly three models are used for discrete envelopes: Linear Interpolation (LI) [8], [12], Discrete Cepstral Envelope (DCE) [9], [14], [15] and Discrete All-Pole (DAP) [11]. Estimation methods of the model parameters can be classified into Single Frame Analysis (SFA) [8], [9], [11], where the envelope of each frame is estimated independently from the other frames, and Multi Frame Analysis (MFA) [12], [15], where multiple frames are gathered for improving the estimation of the envelope at one instant. Shiga et al. [15], [16] extended the Least Squares (LS) solution of the DCE-SFA [9] to multiple, but non-successive, frames. Thus, a question remains about adapting the DCE-MFA to successive frames to address the problem of time regularization. Using 2D processing techniques, Wang et al. [12] also suggested an MFA approach to improve estimates of formants in high pitched voices. Surprisingly, the results in [12] show that a simple linear interpolation of all the amplitude parameters gathered from a set of frames (namely LI-MFA) provides almost always the best results. Therefore, it seems that time regularization of discrete spectral envelope using MFA is an open issue. A simplistic solution would be to low-pass lifter the cepstrum of envelope estimates across time, thus, reducing the quick and potentially erroneous variations (i.e. improving prop. ii). However, this would automatically over-smooth the formants' shape and reduce the accuracy of the envelope (i.e. degrade prop. i). Moreover, the sinusoidal amplitude parameters would not be preserved (i.e. degrade prop. iii).

By time regularization of the envelope, we want actually to preserve the envelope's shapes, but not necessarily its absolute value. Thus, the frequency derivative of the spectral envelope might be the main component to regularize. In this letter, the suggested idea is to smooth, across time, the frequency derivative of a log envelope estimate. The smoothed derivative is then re-integrated across frequency to retrieve the improved envelope. Because the frequency derivative discards the log of the gain of the envelope, the suggested improved envelope is no more aligned on the given amplitude parameters. Thus, for each frame, the improved envelope has to be re-aligned. Note that the gain of the envelope is mainly driven by the glottal source energy. Thus, the suggested two-step approach (i.e. time regularization, then re-alignment) also allows to process separately two independent components, namely, the VTF response and the amplitude modulations of the glottal source energy. In this letter, we aim at demonstrating the success of the suggested approach by its application to the simple linear interpolation. In future works, this approach could also be applied to the cepstral and all-pole models.

The next section describes the suggested method of the derivative of the Linear Interpolation (dLI-MFA). Sec.III evaluates and compares the dLI-MFA to state-of-the-art methods.

Funded by the Swiss National Science Foundation (grant PBSKP2_140021) and the Foundation for Research and Technology Hellas (FORTH).

G. Degottex is with the University of Crete and FORTH, Heraklion, Greece, e-mail: degottex@csd.uoc.gr.

II. Time regularization by frequency derivative

This section basically describes the dLI-MFA method through application of the idea of frequency derivative into the LI model. The solution presented below makes use of harmonic peaks h defined by their frequency $f_{n,h}$ and their log amplitude $a_{n,h}$, which are estimated in frames n at regular time instants t_n (each 2.5ms), up to Nyquist frequency $f_s/2$ [17]. More technical details are provided in Sec. III.

The LI-SFA envelope simply consists in linking $a_{n,h}$ using straight lines across frequency [8]. The suggested dLI-MFA works in four passes. The first pass consists in computing the frequency derivative of the LI-SFA solution $dL_n[k]$ for each frame n, on a given number of frequency bins (e.g. $k = 0, ..., K$ with $K = 1024$), which is a discrete sampling of the continuous step function:

$$dL_n(f) = \sum_{h=0}^{H_n-1} (a_{n,h+1} - a_{n,h}) \cdot \chi_{[f_{n,h}, f_{n,h+1}]}(f) \qquad (1)$$

where $f = f_s \cdot k/K$, H_n is the number of harmonics at frame n up to Nyquist frequency and $\chi_A(f)$ is the indicator function of the frequency interval A. The second pass consists in a low-pass filtering of $dL_n[k]$ across time independently for each bin k using a Hamming window of 30ms (i.e. 30/2.5+1=13 frames; an odd number of frames being convenient). This provides a smoothed and regularized frequency derivative envelope $dR_n[k]$. During the third pass, the regularized log amplitude envelope $R_n[k]$ is retrieved at each frame through the cumulative sum of $dR_n[k]$ across frequency, which compensates for the derivative of the first pass. The upper plot of Fig. 1 shows an example of the derivatives $dL_n[k]$ and $dR_n[k]$. Finally, the frequency derivative removes the constant term of the log amplitude $a_{n,h}$, namely the logarithm of the gain of the estimated filter response. Thus, the resulting $R_n[k]$ is not aligned on $a_{n,h}$, even though its shape corresponds to that sampled by $a_{n,h}$. To solve this issue, for each frame, we re-align the mean of $R_n[k]$ on the mean of $a_{n,h}$ during the fourth pass. Because preserving the sinusoidal amplitudes is critical in terms of perception of the low frequencies, we suggest to consider only the first 4kHz in this fourth and last pass.

The dLI-MFA method has also some practical advantages. First, besides the 4kHz used for the final alignment, this method is basically non-parametric. The number of frames and their time distance will obviously affect the results, as for any MFA-based method. However, no order is necessary (as in DCE or DAP models) or any regularization parameter (as in most DCE solutions [9]). Moreover, this method is straightforward to implement and use low computational resources since each pass needs only linear time operations compared to the DCE solutions which use least squares solution. Bottom plot of Fig. 1 illustrates an example of dLI-MFA estimate. It is interesting to note that, in the lowest frequencies, the envelope is similar to a straight linear interpolation. The harmonics do not span enough the frequency response of the filter to reconstruct a smooth envelope. However, in higher frequencies (e.g. 4kHz), the harmonic number multiplies the f0 variations and the higher harmonics better span the filter response. Thus, the envelope's shape is clearly smoother.

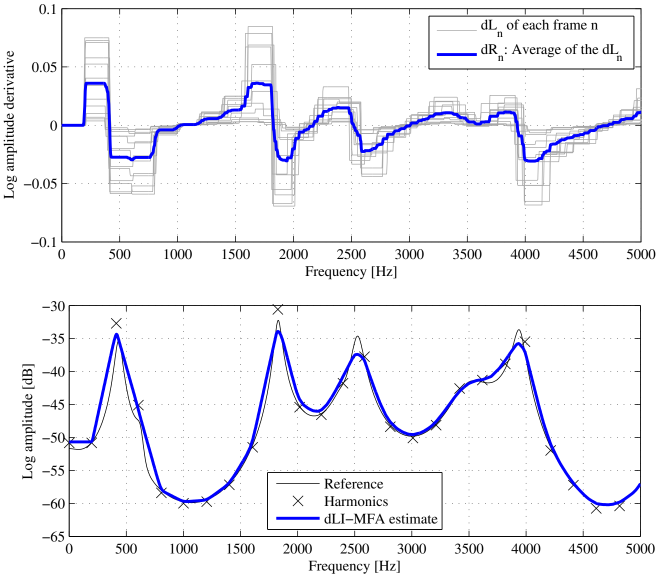


Fig. 1. Example of dLI-MFA estimate: Upper plot shows the construction of the frequency derivative of the averaged envelope. Bottom plot shows the final estimated envelope computed from a synthetic signal with a known reference frequency response.

III. Evaluation

This section presents comparisons of dLI-MFA with state-of-the-art methods. The used data is first described, followed by the means to assess the three properties.

A. Real and synthetic data

The first evaluation data set is made of 1000 real speech samples of 200ms extracted from random voiced segments of random utterances of the TIMIT database [18] (16kHz sampling). Since the ground thruth of the VTF (the reference) of real signals is unknown, one cannot assess the methods accuracy (prop. (i)) from this first evaluation set. Therefore, we built a second evaluation set with known VTF references. The $f_0(t)$ and energy curves are first extracted from each of the voiced segments above using the STRAIGHT vocoder [19] and 25ms Hamming windows, respectively. Then, a synthetic source is generated for each sample using Dirac impulse trains, obeying the extracted f_0 and energy curves. Finally, each synthetic source is convolved by a fixed VTF generated using a digital acoustic model [20]. A different set of uniform random articulatory parameters was used for each synthetic sample. These articulatory parameters include: the jaw, the tongue's position and shape, the tongue's tip position, the lip's height and protrusion, the larynx height and the velum opening (thus, producing also nasalized sounds) [20]. The length of the vocal tract was also set randomly and uniformly between 13 and 18cm. In conclusion, these synthetic signals offer a mean to assess envelope estimation errors while knowing the reference (necessary for the *relative cepstral error* measure below), which is unknown for real signals. Using the f_0 and energy curves of the real signals makes them as realistic as possible. The real signals are used in a complementary way in this evaluation procedure, by strengthening the evaluation validity given by the synthetic signals.

B. Error measurements for the properties assessment

1) Relative cepstral error: If the cepstrum c_k^* of a known reference is available, we assess the accuracy and precision of the envelope estimate for each speech sample (prop. (i)) using the relative cepstral error measure:

$$\epsilon_k = \frac{1}{N} \sum\nolimits_{n=1}^{N} \left| \frac{c^*_{n,k} - c_{n,k}}{c^*_k} \right| \tag{2}$$

where N is the number of frames in the speech sample.

2) Relative cepstral time irregularity: To assess the time irregularity of the envelope estimate along a voiced segment (prop. (ii)), we approximate a variance measure of the cepstral coefficients across time. Since we want to use this irregularity measure with real signals, thus, without knowing the reference, this variance is measured through the difference of the successive cepstral values, as shown in (3). We also normalize the difference by the cepstrum in order to reveal the importance of the irregularity, as in a relative error:

$$\rho_k = \begin{cases} \frac{1}{N-1} \sum_{n=1}^{N-1} \left| \frac{c_{n,k} - c_{n-1,k}}{c_{n,k}} \right|, & \text{if } |c_{n,k}| > 0 \; \forall n \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

3) Harmonic error: To assess how well the estimated envelope preserves the harmonic log amplitudes $a_{n,h}$ in each sample (prop. (iii)), we suggest to use the mean absolute error:

$$\eta_h = \frac{1}{N} \sum\nolimits_{n=1}^{N} |a_{n,h} - R(f_{n,h})| \tag{4}$$

Finally, for reason of clarity in the results presented below, the above measurements are interpolated on continuous scales in log10 ms or Hz, e.g. $\eta_h \Rightarrow \eta(f)$.

C. Compared methods

The dLI-MFA method is compared to the following state-of-the-art methods. First, the traditional LI [8] is used and its MFA version (LI-MFA) (see examples in Fig. 2), which was used for comparison in [12]. In LI-MFA, a small set of successive frames is pre-aligned using the energy of the first 4kHz. Then, all the harmonic peaks of these frames are interpolated as if they were from a single frame. Fig. 2 shows erratic shapes in the LI-MFA solution. Even though this appears to be an issue in the frequency envelope, this problem vanishes when looking at the lower cepstral coefficients (shown later in Sec. III-D). The DCE-SFA [9] method is also used (with $\lambda = 0.035$ [9]). On a linear frequency scale, the cepstral order is given by its optimum [21]: $o = \lfloor 0.5 \cdot f_s / f_0 \rfloor$, which corresponds to the “Nyquist quefrency” of the harmonic sampling. Since a Mel frequency scale is often used in DCE estimates in order to minimize the harmonic error in the low frequencies [4], we also compare with this variant (DCEmel-SFA). Finally, the MFA version of the DCE (DCE-MFA) has already been suggested for non-successive frames [15], [16]. However, for successive frames, the solution described in [15] is not stable because it does not make use of a regularization term, as described in [9]. Thus, for our experiments, we added the regularization term to the DCE-MFA solution (with same order and λ as in the DCE-SFA).

There are obviously numerous other methods (e.g. DAP [11], splines-based [10]). For reason of space, we selected the methods that compare the best to dLI-MFA. Envelope estimates computed without sinusoidal parameters should be also considered in a more comprehensive study [19], [22]–[24].

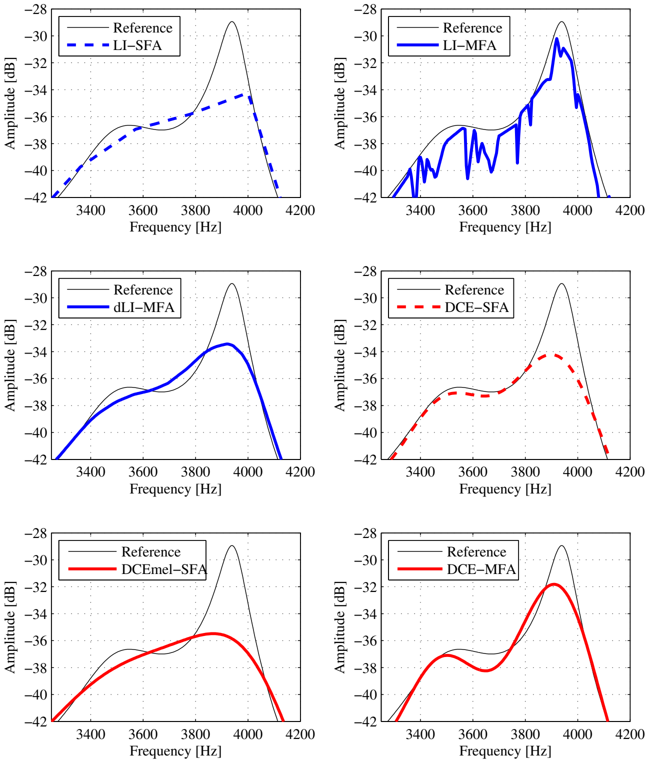
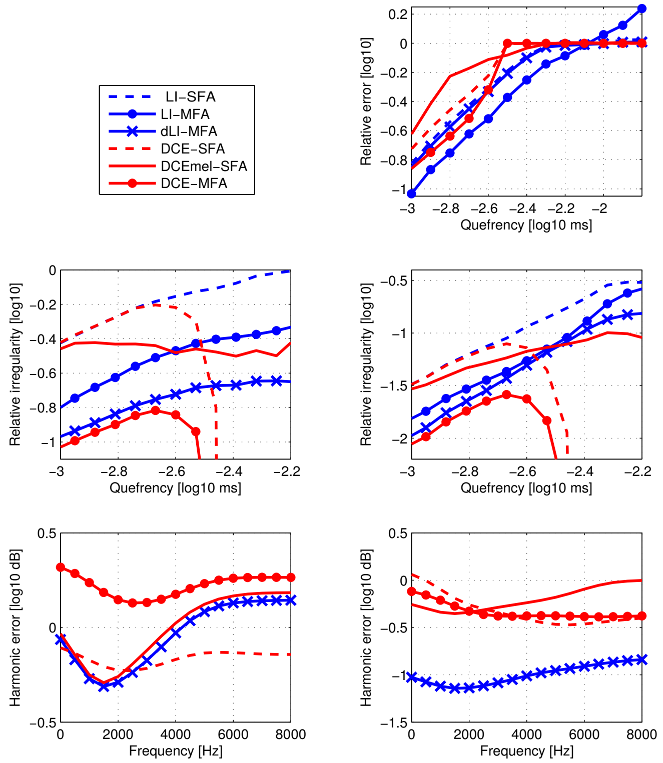


Fig. 2. Examples of estimate of formants’ shape using the methods compared and a synthetic signal with a known reference frequency response.

All the methods compared in this work make use of a discrete set of spectral peaks to obtain the sinusoidal amplitude parameters. These peaks are picked from a 4096 bins DFT [1], using Blackman windows of 3 periods duration, estimated each 2.5ms. The amplitude of each peak is fit by quadratic regression of the closest peak to $h \cdot f_0$ in the DFT, as described in [8]. The zero log amplitude at DC frequency, as well as the peaks close to Nyquist frequency, are often responsible of technical issues in all the methods compared. To minimize this problem, artificial harmonics were added at DC and up to Nyquist, using first and last available amplitudes, respectively. For the MFA-based methods, a window duration of 30ms was used, i.e. 30/2.5+1=13 successive frames. The impact of the window duration is studied in Sec. III-E.

D. Results of error measurements

Fig. 3 shows the three error measurements computed from the real and synthetic signals and averaged among the 1000 samples. Note that LI-SFA and LI-MFA do not appear in the harmonic error plots because they perfectly preserve the harmonic amplitudes. Fig. 3 motivates the following observations. Firstly, the methods are ordered the same with respect to the relative irregularity between real and synthetic signals (besides a different scaling). This supports the experimental setup. Even though the harmonic errors exhibit more differences between the synthetic and real signals, these differences can be explained by the noise component which is present in the real signals, but not reproduced in the synthetic ones. Secondly, if we consider only the cepstral coefficients below -2 log10 ms of the relative error, LI-MFA exhibits a lower error than the other methods (which is shown even more clearly in Fig. 4). This simple result is encouraging for MFA approaches



Real signals Synthetic signals

Fig. 3. Comparison of the methods using the 3 error measurements (using base-10 log scale, the smaller the better), using the real and the synthetic speech signals, on the left and the right, respectively.

because this suggests that the average cepstral order of SFA-based methods (around -2.5 log10 ms) could be increased by exploiting successive frames. Thirdly, by inspection of the relative error using synthetic signals, the accuracy of dLI-MFA is similar that of its SFA version (i.e. LI-SFA). Thus, the suggested processing does not degrade the accuracy of the original linear interpolation. Moreover, its irregularity (second row) is more than 3 times smaller than that of LI-SFA for both synthetic and real signals. Finally, even tough the LI-SFA method has obviously no harmonic error at all, that of dLI-MFA is still fairly small (around 1 and 0.1dB, for real and synthetic signals, respectively). This shows that the suggested dLI-MFA clearly improves the time regularity of its SFA version, without degrading the accuracy of the envelope and with very limited consequences on the harmonics error, thus, supporting globally the suggested approach.

The DCE-MFA method also shows globally promising results. Its relative error and irregularity is smaller than its SFA version and dLI-MFA. Its drawback is mainly the harmonic error in real signals. However, considering the potential increase of the cepstral order and the possibility to use a non-linear frequency scale to better preserve the low harmonics, as in DCEmel-SFA, there is clearly room for improvement through development of better DCE-MFA solutions.

E. Impact of the window duration

Whereas a 30ms window (13 frames) was used in previous experiment for the MFA methods, this section presents results where the window duration varies. For reason of space, only the results using synthetic signals are reported here. A more comprehensive study could also detail this evaluation with the impact of the step size of the frames (here 2.5ms). The same experimental setup is used, as previously. Fig. 4 shows the average cepstral errors lower than -2.7 log10 ms and the average harmonic errors below 4kHz. Compared to Fig. 3, Fig. 4 better details the differences between the methods in terms of relative errors. One can see that dLI-MFA has a lower average relative error than all SFA-based methods. Moreover, the other MFA-based methods (DCE-MFA and LI-MFA) exhibit even lower relative errors. The relative error decreases constantly with the window duration for all MFA-based methods. The relative cepstral irregularity also decreases substantially up to 50ms window length, whereas the harmonic errors stays below 1dB. This suggests that, using MFA approaches, room for improvement exists for improving both accuracy and time regularity of envelope estimates while keeping satisfactory harmonic errors.

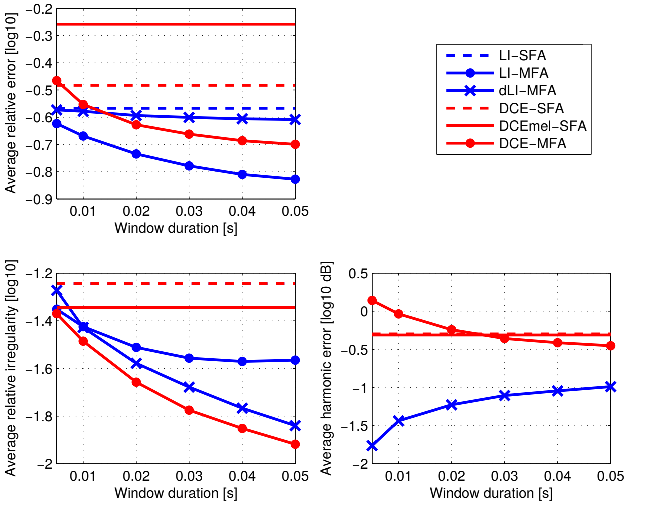


Fig. 4. Results with respect to the window duration, using synthetic signals.

IV. Conclusions

In this letter, a time regularization technique has been suggested, which exploits the frequency derivative of envelope estimates from a small set of successive frames. Even though this idea could be applied potentially to various envelope estimation methods (e.g. cepstral or all-pole models), this letter describes its application to the simple linear interpolation envelope. Experiments have been carried out to evaluate the relative error, the relative irregularity and the harmonic error of spectral envelopes. Experiments using synthetic signals have shown that the relative error of the envelope is not degraded using the suggested method compared to state-of-the-art envelope estimates, whereas the time regularity is clearly improved. The experiments show also promising results for the Discrete Cepstral Envelope (DCE) exploiting successive frames. Even though the suggested improved linear interpolation technique is straightforward to implement and has less parameters than the DCE, one may expect future developments of the frequency derivative approach using cepstral models and multiple frame analysis.

V. Acknowledgements

The author would like to thank Yannis Stylianou for the valued discussions about the discrete cepstral envelopes.

References

[1] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[2] Y. Stylianou, *Harmonic plus Noise Models for Speech combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. thesis, TelecomParis, France, 1996.

[3] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, 2001.

[4] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, 2014.

[5] J. Jensen and J.H.L. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 731–740, 2001.

[6] Yi Hu and P. C. Loizou, "On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants," *Journal of Acoustic Society of America*, vol. 127, no. 1, pp. 427–434, 2010.

[7] G. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Pitch modifications of speech based on an adaptive harmonic model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[8] D. B. Paul, "The spectral envelope estimation vocoder," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 29, no. 4, pp. 786–794, 1981.

[9] M. Campedel-Oudot, O. Cappe, and E. Moulines, "Estimation of the spectral envelope of voiced sounds using a penalized likelihood approach," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 469–481, 2001.

[10] J. Bonada, "High quality voice transformations based on modeling radiated voice pulses in frequency domain," in *Proc. Digital Audio Effects (DAFx)*, 2004.

[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] T.T. Wang and T.F. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 1, pp. 171–186, 2010.

[13] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234–1252, 2013.

[14] T. Galas and X. Rodet, "Generalized discrete cepstral analysis for deconvolution of source-filter system with discrete spectra," in *Applications of Signal Processing to Audio and Acoustics, 1991. Final Program and Paper Summaries., 1991 IEEE ASSP Workshop on*, 20-23 1991, pp. 0_71–0_72.

[15] Y. Shiga and S. King, "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," *EUROSPEECH*, 2003.

[16] Yoshinori Shiga, *Precise Estimation of Vocal Tract and Voice Source Characteristics*, Phd thesis, University of Edinburgh, 2005.

[17] G. Degottex and Y. Stylianou, "Analysis and synthesis of speech using an adaptive full-band harmonic model," *IEEE Trans. on Audio, Speech and Lang. Proc.*, vol. 21, no. 10, pp. 2085–2095, 2013.

[18] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathon G. Fiscus, David S. Pallett, and Nancy L. Dahlgren, "Timit acoustic-phonetic continuous speech corpus," Philadelphia: Linguistic Data Consortium, 1993, Web Download.

[19] H. Kawahara, I Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptative time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[20] S. Maeda, "A digital simulation method of the vocal-tract system," *Speech Communication*, vol. 1, no. 3-4, pp. 199–229, 1982.

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

[22] S. Imai and Y. Abe, "Spectral envelope extraction by improved cepstral method," *Electronics and Communication*, vol. 62-A, no. 4, pp. 10–17, 1979, in japanese.

[23] A. Roebel and X. Rodet, "Efficient spectral envelope estimation and its application to pitch shifting and envelope preservation," in *Proc. Digital Audio Effects (DAFx)*, 2005, pp. 30–35.

[24] Y. Agiomyrgiannakis and Y. Stylianou, "On the recovery of time-varying spectral envelope information from aqhm-derived spectra.," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 5400–5403.