SIMPLE MULTI FRAME ANALYSIS METHODS FOR ESTIMATION OF AMPLITUDE SPECTRAL ENVELOPE ESTIMATION IN SINGING VOICE

Gilles Degottex, Luc Ardaillon and Axel Roebel

IRCAM - UMR STMS IRCAM-CNRS-UPMC
Paris, France

ABSTRACT

In the state of the art, a single frame of DFT transform is commonly used as a basis for building amplitude spectral envelopes. Multiple Frame Analysis (MFA) has already been suggested for envelope estimation, but often with excessive complexity. In this paper, two MFA-based methods are presented: one simplifying an existing Least Square (LS) solution, and another one based on a simple linear interpolation. In the context of singing voice we study sustained segments with vibrato, because these ones are obviously critical for singing voice synthesis. They also provide a convenient context to study, prior to extension of this work in more general contexts. Numerical and perceptual experiments show clear improvements of the two methods described compared to the state of the art and encourage further studies in this research direction.

Index Terms— Multi frame analysis, spectral envelope, singing voice, voice analysis and modeling, voice synthesis.

1. INTRODUCTION

The estimation of the amplitude spectral envelope of voiced segments is a key aspect in singing voice synthesis for a good reconstruction of the formant effects and the overall timbre of the voice. The context of singing voice allows us to focus in this paper on the very specific context of vibrato segments, whose results could then benefit to voice processing in general. Note that for the following the glottal pulse shape does not need to be described independently [1, 2] so that we can include it into the Vocal Tract Filter (VTF) response and use the terms VTF and spectral envelope equivalently. Independently of the synthesis system used (e.g. parametric or concatenative synthesis [3, 4]), the envelope estimate mainly needs to minimize the cepstral error regarding ground truth. Additionally, it is also important that the difference between estimates of different VTF is reproduced, i.e. the *global* variance is preserved. Indeed, estimation techniques tend to underestimate this variance by averaging and, thus, *flattening* the VTF shapes. Even though this phenomenon is well known in statistical modeling [5], we will show that this phenomenon already exists during envelope estimation.

Supported by the ChaNTeR ANR project (ANR-13-CORD-0011).

Most estimation methods of spectral envelopes use a single frame of frequency analysis (e.g. using the DFT of a short time window) [6, 7, 8, 9, 10]. However, within a single frame, the sampling of the VTF by the harmonic structure of the voice source provides only a very limited set of sampling points, i.e. the integer multiples of the fundamental frequency f_0. To address this issue, the Multi-Frame Analysis (MFA) has been already suggested [11, 12, 13]. It consists in gathering the information of different frames for improving the estimation of a common envelope. However, we will show that the first method [11], the Discrete Cepstral Envelope using MFA (DCE-MFA), can be largely simplified. Additionally, the second one [12] seems to show no drastic improvement compared to a simple MFA-based Linear Interpolation (Linear-MFA). Finally, the accuracy of the last one [13] is not better than that of the DCE-MFA. Therefore, we conclude that room for improvement exists and we take the opportunity of the context of singing voice to focus on a very well defined case study, namely sustained segments of vibrato where the VTF is assumed stationary.

In this paper, we suggest two simple solutions that show very encouraging results using numerical and perceptual experiments. The first solution is a Simplified DCE-MFA (SDCE-MFA). We demonstrate that this method does not need the frame alignment that seems necessary for all MFA-based methods. The second solution is a low-pass liftering of the Linear-MFA (Linear-MFA-LIFT). As simple as it looks like this method provides very similar results to the SDCE-MFA and is computationally very efficient. Using Single Frame Analysis (SFA), spectral envelopes are known to have stability issues [9]. This issue is often addressed by using a regularization technique [9]. However, this approach risks also to flatten the envelope's shape and reduce the global variance. On the contrary, using MFA, the neighbor frames add constraints and useful information that solve the stability issue and improves the accuracy of the envelope estimate. This explains theoretically why the two simple solutions presented in this paper provide good improvements compared to SFA methods.

Sec. 2 describes the two MFA methods mentioned above and Sec. 3 will present experimental results comparing these methods to state-of-the-art SFA methods.

2. MULTI-FRAME ANALYSIS (MFA) METHODS

2.1. Simplified Discrete Cepstral Envelope for MFA (SDCE-MFA)

This first method is based on the works of Shiga et al.[11], which is an MFA version of the LS cepstral solution [8, 9]. The cepstral model of the log amplitude spectral envelope is:

$$E(f) = c_0 + 2\sum_{n=1}^{P} c_n \cos(n2\pi f/f_s) \tag{1}$$

where c_n are the cepstral coefficients, P the cepstral order. Given a set of harmonic parameters (pairs of frequency and amplitude), the problem is to estimate c_n, such as $E(f)$ is smooth and it passes as close as possible to the harmonics [8, 9, 10, 14]. The MFA solution in [11] minimizes the error:

$$e = \sum_{k=1}^{K} \|\boldsymbol{W}_k \cdot (\boldsymbol{a}_k - d_k\boldsymbol{u}_k - \boldsymbol{B}_k\boldsymbol{c})\| \tag{2}$$

where k and K are the frame index and number of frames considered, $\boldsymbol{a}_k$ contains the components' log amplitudes, d_k is a scalar correcting all the amplitudes belonging to frame k, $\boldsymbol{u}_k = [1, \ldots, 1]^T$, $\boldsymbol{c}$ contains the cepstral coefficients, $\boldsymbol{W}_k$ is a weighting matrix emphasizing the importance of the low frequencies, and for each frame $\boldsymbol{B}$ is:

$$\boldsymbol{B} = \frac{1}{H}\begin{bmatrix} 1 & 2\cos(1\omega_1) & 2\cos(2\omega_1) & \cdots & 2\cos(P\omega_1) \\ 1 & 2\cos(1\omega_2) & 2\cos(2\omega_2) & \cdots & 2\cos(P\omega_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos(1\omega_h) & 2\cos(2\omega_h) & \cdots & 2\cos(P\omega_h) \end{bmatrix} \tag{3}$$

with $\omega_h = 2\pi h f_0(t)/f_s$ and H the number of harmonics in the frame. In this work, $\boldsymbol{W}_k$ is a diagonal matrix whose terms corresponds to a Gaussian window centered at DC and 3kHz of standard-deviation [11]. Compared to [9] and [14], (2) has a weighting term, but no regularization term. We continued to use (2) in this work, since the motivation for a regularization term drops when using MFA, as argued in the introduction.

The voice source is likely to have an amplitude modulation component (e.g. tremolo in singing voice). This problem concerns all MFA-based approaches, as it is necessary to compensate for this modulation, i.e. align the frames in amplitude prior to the computation of the envelope. In [11], d_k and $\boldsymbol{c}$ are estimated jointly for this purpose in an iterative way. Here below, we show that these amplitude corrections can actually be dropped, thus leading to a simpler solution. In [11](eq.9), the estimate of $\boldsymbol{c}$ is given by:

$$\Big(\sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{B}_k\Big)\boldsymbol{c} = \sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k(\boldsymbol{a}_k - d_k\boldsymbol{u}_k) \tag{4}$$

whose inverse of the left-hand side can be distributed:

$$\begin{aligned}\boldsymbol{c} \;=\; & \sum_{k=1}^{K}\Big(\sum_{l=1}^{K} \boldsymbol{B}_l^T\boldsymbol{W}_l\boldsymbol{B}_l\Big)^{-1} \cdot \Big(\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{a}_k\Big) \\ & - \sum_{k=1}^{K} d_k \cdot \Big(\sum_{l=1}^{K} \boldsymbol{B}_l^T\boldsymbol{W}_l\boldsymbol{B}_l\Big)^{-1} \cdot \Big(\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{u}_k\Big)\end{aligned} \tag{5}$$

The second term of (5) containing the cepstral coefficients of the constant spectrum $\boldsymbol{u}_k$, we can write:

$$\begin{aligned}\boldsymbol{c} \;=\; & \sum_{k=1}^{K}\Big(\sum_{l=1}^{K} \boldsymbol{B}_l^T\boldsymbol{W}_l\boldsymbol{B}_l\Big)^{-1} \cdot \Big(\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{a}_k\Big) \\ & - \sum_{k=1}^{K} d_k \cdot [1, 0, \ldots, 0]^T\end{aligned} \tag{6}$$

Therefore, the c_n with $n > 0$ are not influenced by d_k. The estimation of the shape of the envelope is independent of the frames alignment. As a conclusion, the SDCE-MFA suggested in this paper consists in computing (6), disregarding the frame alignment. If an amplitude alignment is necessary for the application, the final envelope can be eventually re-aligned on the harmonic amplitudes of the central frame of the K contiguous frames.

2.2. Linear interpolation for MFA with cepstral Liftering (Linear-MFA-LIFT)

The Linear-MFA has been suggested and used for comparison in the works of Wang et al.[12], which is a MFA version of the SFA linear interpolation [15]. In this method, the K successive frames are first pre-aligned using the energy of the first 4kHz. Then, all the harmonic peaks of these frames are interpolated as if they were from the same frame. Erratic shapes appear on the estimate because of the noise in the sinusoidal parameters (See [13](Fig.2)). This problem makes the Linear-MFA practically unusable without further processing. In [12], they applied an AR model on top of the Linear-MFA, which is then used for estimation of formants' position. Because we chose to focus on cepstral models in this paper, we suggest to filter the erratic shapes by low-pass liftering of the envelope, thus, leading to a new method called Linear-MFA-LIFT. One can note the simplicity of the implementation of this method. Also, conversely to the SDCE-MFA that involves the computation of an LS solution, the Linear-MFA-LIFT needs only the frame alignment, a linear interpolation and a cepstral liftering, which is far more efficient computationally.

2.3. Order selection

Assuming a strict harmonic grid $f_h = h \cdot f_0$, the SFA envelope estimation is similar to the traditional signal reconstruction based on uniform sampling. Thus, according to the Nyquist theorem, the maximum cepstral order is:

$$P^* = \Big\lfloor \frac{0.5 \cdot f_s}{f_0} \Big\rfloor \tag{7}$$

which is called *usual order* in the following [16]. Using the SDCE-MFA, the LS problem become overdetermined when using (7). As a consequence, the order can be increased, to some extent, as it will be shown in Sec. 3.1.1.

The choice of the cepstral order is far from straightforward for the Linear-MFA-LIFT. Therefore, in the following, for SDCE-MFA and Linear-MFA-LIFT, the order will be expressed in terms of the Usual Order Factor (UOF) which is a multiplicative factor of the usual order (7).

3. EVALUATION

In this evaluation, the sinusoidal parameters (frequency and amplitude) are extracted by peak picking [17] using a 4096 bins DFT and a Blackman window of 3 periods duration, estimated each 5ms. Because there is no such peak at DC and Nyquist frequencies, artificial components are added at these frequencies using first and last harmonic's amplitude, respectively. Additionally, for the MFA methods compared below, a 400ms window was used ($\approx$ 2 vibrato periods for a 5Hz vibrato). For the sake of the comparison, we compared with two state-of-the-art SFA methods: i) The "True-Envelope" (TE) [18, 10, 19], which is also based on cepstral representation and does not use sinusoidal parameters. ii) The traditional Discrete Cepstral Envelope (DCE), which makes use of a regularization term because the LS problem is not overdetermined in the SFA case [9] (using reg. coef. λ =0.035).

3.1. Numerical evaluation using synthetic signals

Since the ground truth of the VTF is unknown in voice recordings, we first evaluate the methods numerically using synthetic signals with known reference VTFs. 1000 samples are generated using the following procedure. First, for each sample, an $f_0(t)$ curve is generated whose average frequency is a random value in $[80, 800]$Hz, the vibrato has a random extent in $[0, 150]$cents and a random frequency in $[4, 6]$Hz. A Dirac impulse train is then generated accordingly, which is then amplitude modulated in order to reproduce a tremor of standard-deviation 0.5dB (using a low-pass filtered Gaussian noise with 5Hz cutoff). Finally, each synthetic source is convolved by a stationary VTF generated using a digital acoustic synthesizer [20]. A random set of articulatory parameters was used for each of the 1000 synthetic samples.

To assess the methods, we used the measurements below.

Absolute cepstral error:

$$\epsilon_{n,i} = \frac{1}{M}\sum_{m=1}^{M}\left|c^*_{n,i} - c_{m,n,i}\right| \qquad (8)$$

where $c^*_{n,i}$ is the reference of the sample i, M is the number of frames in the sample and $c_{m,n,i}$ is the nth coefficient of the frame m. $\epsilon_{n,i}$ is then averaged over i for the figures below.

Cepstral Variance:

The following ratio assesses the capacity of a method to reproduce the *global* variance of the reference VTFs:

$$\bar{\sigma}_n = \frac{\text{std}_i(\bar{c}_{n,i})}{\text{std}_i(\bar{c}^*_{n,i})} \quad \text{where} \quad \bar{c}_{n,i} = \frac{1}{M}\sum_{m=1}^{M} c_{m,n,i} \qquad (9)$$

where $\bar{c}_{n,i}$ is the average cepstrum over the M frames of each sound sample i and $\text{std}_i(.)$ is the standard-deviation over i. If $\bar{\sigma}_n < 1$ (negative log value in Fig. 1), the variance of the estimates is smaller than it should be. This corresponds to an averaging and *flattening* effect of the estimates. If $\bar{\sigma}_n > 1$ (positive log value), the envelopes are more different than they should be and unnatural resonances are expected in synthesis.

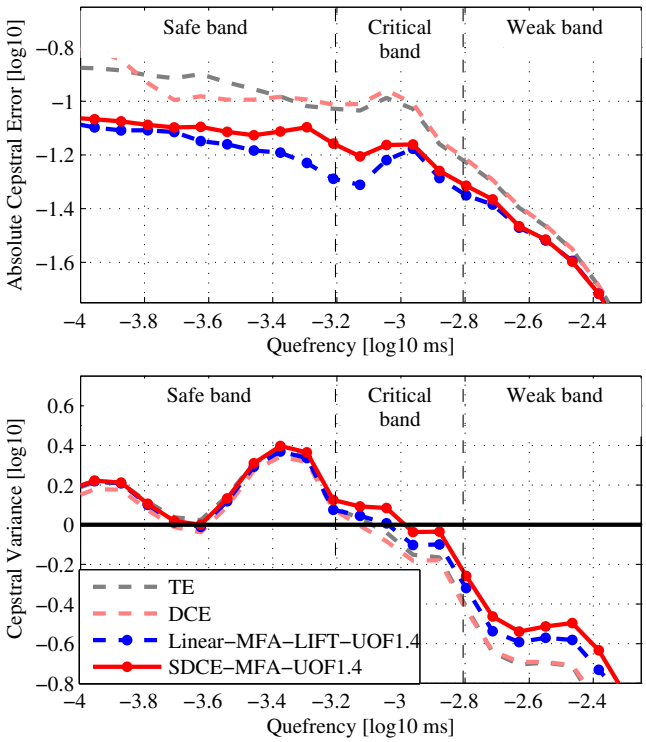

Fig. 1. Numerical evaluation of envelope estimation methods.

For Fig. 1, the UOF has been set to 1.4, which is justified in Sec. 3.1.1. For the sake of the discussion, we split the cepstrum into 3 bands. We call *safe band* the cepstral coefficients below the lowest usual order because they are independent of f_0 (-3.20[log10 s] when using a max f_0 of 800Hz). We call *critical band* the cepstral coefficients that are influenced by f_0 and whose magnitude is above a noticeable amplitude difference of 1dB [21, 22] (in average -2.81[log10 s], when computed on the 1000 synthetic VTFs used in these experiments). Finally, we call *weak band* the quefrency band above the noticeable limit, because we assume their magnitudes are close to negligible. From Fig. 1, we observe the following. The absolute error shows that, in the safe band, both MFA methods divide the error by a factor slightly less than two compared to the SFA methods. Also, the lack of improvement in the weak band is not problematic since this band has almost negligible impact on the perception. In the safe band the presence of errors explains the increase of cepstral variance above 0 [log10]. However, in the critical and weak bands, the cepstral variance is basically lower than 0 for most methods. This lack of cepstral variance demonstrates an averaging effect of the envelope estimates. Nevertheless, MFA methods exhibits a higher variance than the SFA methods.

3.1.1. *Measurements with respect to UOF*

We study here the influence of the Usual Order Factor (UOF) on the critical band. We do not consider the *safe band* since it is independent of f_0, and thus independent of UOF. The *weak band* is also excluded because it would bias the results with perceptually irrelevant information. In Fig. 2, we can see

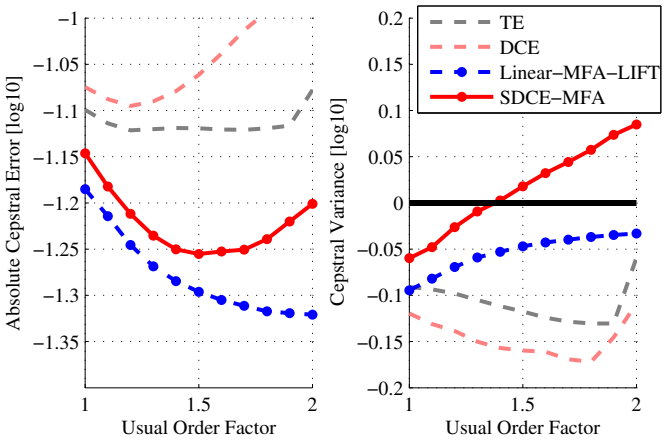

Fig. 2. Measures with respect to Usual Order Factor (UOF).

that the absolute error of Linear-MFA-LIFT decreases when UOF increases, and, the SDCE-MFA exhibits a local minimum around 1.5. The cepstral variance helps to better understand this behavior. When the order is low, the cepstral variance shows that the error is due to a flattening of the envelope estimates. After UOF=1.4, the cepstral variance continues to increase and add an extra error to the estimates. On the contrary, the error of the Linear-MFA-LIFT continues to decrease because its cepstal variance is bounded below 0 [log10]. On the down-side, its estimates are always a bit flattened. By using UOF≈1.4, the SDCE-MFA is actually the only method which is able to reconstruct the global variance that is observed in the reference, thus, justifying the value used for Fig. 1. For the sake of the comparison, this same value is used for the Linear-MFA-LIFT and for the following experiment.

3.2. Listening tests about pitch scaling

In this section, we present the results of a proof-of-concept experiment in the context of pitch scaling. The purpose of our research project ChaNTeR is to synthesize singing voice in French. Thus, we used recordings of approximately 2s of the 15 French vowels, containing natural vibrato, of the two singers recorded for the project, one male and one female. Note, that two singers are a rather small sample, so that this experiment can only serve as a first indication of the potential of the method. The 15 recordings of each singer are pitch scaled downwards and upwards, using 0.75 and 1.25 scaling factors, respectively. To keep the duration of the test manageable by the listeners, we compared only 3 methods: SDCE-MFA, Linear-MFA-LIFT and TE, since the numerical results of the DCE are very close to those of the TE. Also, each listener evaluates only 4 different vowels, one per singer for both scaling direction, i.e. 12 comparisons pairs. For each listener, the 4 vowels are taken randomly among the 15. Using a web-based interface, the listeners gave their pairwise preferences following the procedure of a standard Comparison Mean Opinion Score (CMOS) [23], based on *the clarity of the pronunciation of the phoneme*. To reinforce the results against the pitch scaling techniques, we present here the results of 2 independent listening tests using two techniques: i) A harmonic synthesis [24, 25, 26], whose implementation comes from COVAREP [27](v1.3.2). ii) A phase vocoder [28, 29, 19] using shape preservation [30]. [1]

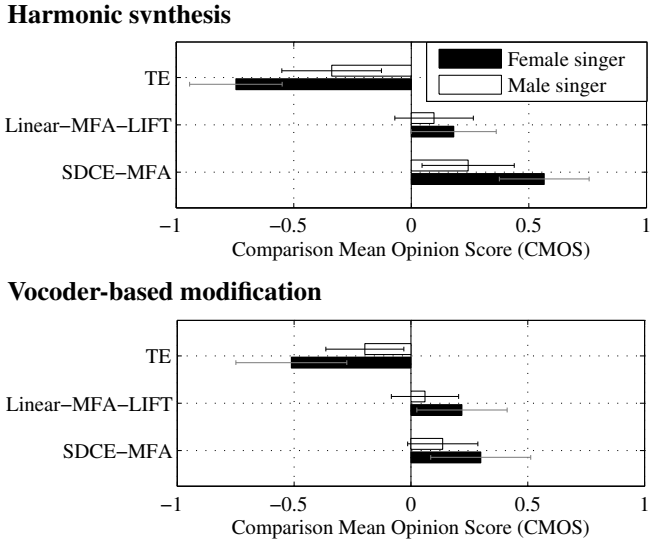

Fig. 3. Results of a listening tests about pitch scaling with the 95% confidence intervals.

31 and 33 listeners took the tests for the harmonic and the vocoder techniques, respectively (See Fig. 3). Based on these results we first conclude that, for both techniques, the MFA methods are clearly preferred compared to the TE envelope. One can also note that the very simple Linear-MFA-LIFT provides a very interesting improvement compared to the TE. The SDCE-MFA is only preferred to the Linear-MFA-LIFT for the female voice using the harmonic technique. Since the numerical experiment shows a smaller absolute error for the Linear-MFA-LIFT compared to the SDCE-MFA, one can assume that the preference shown for the SDCE-MFA in this test is mainly due to its capacity for reproducing the cepstral variance.

4. CONCLUSIONS

In this paper, we have described two MFA methods for amplitude spectral envelopes estimation that are based on previous works. We have shown that the existing LS solution DCE-MFA can be simplified to avoid the frame alignment coming with MFA approach. We also suggested to simply low-pass lifter the MFA-based linear interpolation. Numerical evaluation using synthetic signals have shown that the error is almost divided by two compared to state-of-the-art SFA methods. We have also shown that only the DCE-MFA is able to reproduce the *global* cepstral variance, thus avoiding a flattening effect of the estimates. Finally a listening test dedicated to the voices used in our research project suggests clear improvements in terms of pitch scaling quality.

[1] The samples generated can be found in the following web-page: `http://gillesdegottex.eu/Demos/DegottexG2016mfasings`

5. REFERENCES

[1] G. Degottex, P. Lanchantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Comm.*, vol. 55, no. 2, pp. 278–294, 2013.

[2] S. Huber and A. Röbel, "On the use of voice descriptors for glottal source shape parameter estimation," *Computer Speech and Language*, vol. 28, pp. 1170–1194, 2013.

[3] Kazuhiro Nakamura, Keiichiro Oura, Yoshihiko Nankaku, and Keiichi Tokuda, "Hmm-based singing voice synthesis and its application to japanese and english," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 265–269.

[4] L. Ardaillon, G. Degottex, and A. Roebel, "A multi-layer f0 model for singing voice synthesis using a b-spline representation with intuitive controls," in *Proc. Interspeech*, Dresden, Germany, September 2015.

[5] T. Toda and S. Young, "Trajectory training considering global variance for hmm-based speech synthesis," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4025–4028.

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

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

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

[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] 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.

[11] Y. Shiga and S. King, "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," *Proc. EUROSPEECH*, pp. 1737–1740, 2003.

[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] G. Degottex, "A time regularization technique for discrete spectral envelopes through frequency derivative," *Signal Processing Letters, IEEE*, vol. 22, no. 7, pp. 978–982, July 2015.

[14] R. Mignot and V. Valimaki, "True discrete cepstrum: An accurate and smooth spectral envelope estimation for music processing," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 7465–7469.

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

[16] 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.

[17] 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.

[18] 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.

[19] FLUX and Ircam, "Ircam Trax v3 [Online]," `http://www.fluxhome.com/products/plug_ins/ircam_trax-v3`, 2015.

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

[21] Walt Jesteadt, Craig C. Wier, and David M. Green, "Intensity discrimination as a function of frequency and sensation level," *The Journal of the Acoustical Society of America*, vol. 61, no. 1, pp. 169–177, 1977.

[22] M. R. Schroeder, B. S. Atal, and K. H. Kuttruff, "Perception of coloration in filtered gaussian noiseshorttime spectral analysis by the ear," *The Journal of the Acoustical Society of America*, vol. 34, no. 5, pp. 738–738, 1962.

[23] The ITU Radiocommunication Assembly, "ITU-R BS.1284-1: En-general methods for the subjective assessment of sound quality," Tech. Rep., ITU, 2003.

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

[25] 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, pp. 7924–7928.

[26] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 38, 2014.

[27] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, "COVAREP - a collaborative voice analysis repository for speech technologies," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, `http://covarep.github.io/covarep/`, 2014.

[28] M. Liuni and A. Roebel, "Phase vocoder and beyond," *Musica/Tecnologia*, vol. 7, pp. 73–89, 2013.

[29] A. Roebel, "Supervp software," `http://anasynth.ircam.fr/home/english/software/supervp`, 2015.

[30] Axel Roebel, "Shape-invariant speech transformation with the phase vocoder," in *Proc. Interspeech*, 2010, pp. 2146–2149.