Multi-Frame Amplitude Envelope Estimation for Modification of Singing Voice

Gilles Degottex, Luc Ardaillon and Axel Roebel

***Abstract*—Singing voice synthesis benefits from very high quality estimation of the resonances and anti-resonances of the Vocal Tract Filter (VTF), i.e. an amplitude spectral envelope. In the state of the art, a single frame of DFT transform is commonly used as a basis for building spectral envelopes. Even though Multiple Frame Analysis (MFA) has already been suggested for envelope estimation, it is not yet used in concrete applications. Indeed, even though existing attempts have shown very interesting results, we will demonstrate that they are either over complicated or fail to satisfy the high accuracy that is necessary for singing voice. In order to allow future applications of MFA, this article aims to improve the theoretical understanding and advantages of MFA-based methods. The use of singing voice signals is very beneficial for studying MFA methods due to the fact that the VTF configuration can be relatively stable and, at the same time, the vibrato creates a regular variation that is easy to model. By simplifying and extending previous works, we also suggest and describe two MFA-based methods. To better understand the behaviors of the envelope estimates, we designed numerical measurements to assess SFA and MFA methods using synthetic signals. With listening tests, we also designed two proofs of concept using pitch scaling and conversion of timbre. Both evaluations show clear and positive results for MFA-based methods, thus, encouraging this research direction for future applications.**

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

I. Introduction

Aesthetic properties being particularly important in synthesis of singing voice, the quality of the synthesized voice is a very important parameter. Accordingly, numerous research activities have been undertaken to improve different aspects of the quality in singing synthesis applications [1], [2], [3], [4]. Even though singing and speech syntheses have a strong common background, singing voice emphasizes perceived characteristics that require specific care. For example, for artistic needs, the reconstruction of the resonances of the Vocal Tract Filter (VTF) has to satisfy very high technical standards. This specificity in the singing voice allows us to study very specific problems, whose results can then benefit to voice processing in general. The analysis and representation of the VTF is addressed in this work, as it is a key element for intelligibility and also for controlling and creating many timbre-related effects in a singing voice. For synthesis, the basic issue is to estimate the VTF frequency response from recordings, in order to re-use this estimate, for example for pitch scaling or modification of the timbre. Independently of the used synthesis system (e.g. parametric or concatenative synthesis [3], [2]), this VTF estimate has to satisfy a few critical properties. Firstly, the accuracy of the estimate is obviously important for reproducing the phonetic content. Moreover, since the singer adapts the VTF's resonances when the pitch changes [5], it is also necessary to estimate the VTF on high pitch voices. This context is very challenging since the sampling of the VTF by the harmonics becomes extremely scarce [6]. Secondly, the differences between the estimates in different phonetic contexts is also important. Estimation techniques tend to underestimate the overall estimates' variance by *flattening* or averaging their shapes. Even though this phenomenon is well known in statistical modeling as *global variance* reduction [7], we will show that this phenomenon appears already during the estimation of the VTF. Thirdly, from a more practical point of view, since synthesis systems always involve large databases, it is preferable to ensure that the VTF estimate will not degenerate, i.e. create meaningless shapes and fake resonances. Otherwise, a manual and time consuming verification of the VTF estimates would be necessary.

Estimation methods of amplitude spectral envelopes are mainly designed to approximate the VTF frequency response [8], [9], [10], [11], [12]. Even though the glottal source contributes to the amplitude spectrum of the voice [13], [14], we model in this work both the glottal source contributions and the VTF as a whole, for reason of simplicity. All the estimation methods cited above estimate the spectral envelope using a single frame of frequency analysis (e.g. using the DFT of a short time window). With a single frame, the sampling of the harmonic structure provides only a very limited set of sampling points, one for each integer multiple of the fundamental frequency f_0. To address this issue, the Multi-Frame Analysis (MFA) has already been suggested [15]. Basically, this approach consists in gathering the information of different frames (non-successive in [15] and successive in [16]) for improving the estimation of the amplitude envelope. Analysis methods using 2D processing techniques have also been suggested [6] that follows the same principle. These approaches have shown interesting and promising results. However all of them present shortcomings that prevent any use in practical applications. The first method, the Discrete Cepstral Envelope using MFA (DCE-MFA) [15] is computationally heavy. However, as we will show in Section III, it can be largely simplified, thus, improving its efficiency and our understanding of it. The second one, the DLinear-MFA [16] shows a very good stability and time regularity, but do

Copyright (c) 2014 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.

G. Degottex, L. Ardaillon and A. Roebel are with Ircam, Paris, France. (e-mail: gilles.degottex@ircam.fr).

not compete with the DCE-MFA in terms of accuracy. The last one involves a quite sophisticated 2D technique, which seems to show no drastic improvement compared to a very simple Linear interpolation using multiple frames (Linear-MFA) [6]. Even though, this Linear-MFA exhibits good numerical properties, it has also irregularities that generate artifacts in voice transformations. Therefore, we conclude that a potential exists in the MFA approach for improving amplitude spectral envelopes. However, we also conclude that we are currently missing the necessary overall comprehension of MFA-based methods, in order to address all of these shortcomings and exploit MFA in concrete applications.

In this article, we first present theoretical results that aim at a better understanding of the MFA approach. To complete this theoretical study, we suggest simplified and extended versions of two existing MFA-based methods. We also present a comprehensive numerical evaluation of these methods using specific error measurements designed for the properties cited above. Based on our understanding of the state of the art, addressing most of current questions about MFA for both spoken and sung voice would be too wide for a single article. For this reason, we would like to take advantage of our research context to focus on stationary segments, namely sustained vowels with a pseudo-constant vibrato, where we assume the VTF response to be fixed over time. Focusing on vibrato makes also sense since it has been shown to be useful for the perceptual reconstruction of the VTF response [17]. Because the analysis of segments with vibrato is already quite difficult, managing dynamic VTF response and integrating the methods in a singing voice synthesizer is left for a forthcoming publication. Additionally, in this article, we focus only on cepstral representations [10], [11], [12], which make no assumption about the filter properties, conversely to ARMA models [8], [9]. The cepstral models have also the advantage of having an order that is related to the sampling scheme, which makes them more convenient for studying MFA sampling.

In a Single Frame Analysis (SFA) approach, the cepstral models are known to have stability issues [11]. If the Least Squares (LS) system is not properly conditioned, the spectral envelope can have ripples or it can completely degenerate. The current solution for SFA is to add a regularization term (an additional constraint) to the LS system. In MFA context, we will see in Sec. III-A2 that the used cepstral order is always far below its theoretical limit. Thus, the system is always better conditioned and there is no need of extra regularization term. In this context, one can see the MFA approach as a way to add extra constraints by adding extra spectral peaks of neighbor frames (extra information), instead of adding an artificial constraint through a regularization term. Thus, theoretically, MFA is globally very promising. Whereas a regularization term forces the smoothness of the envelope and might also flatten the envelope (reduce the global variance), the MFA approach adds useful information that solves the conditioning issue and improves the accuracy of the envelope estimate.

Next Section II discusses the MFA from a theoretical point of view in order to better understand this approach. Theoretical observations are made that should encourage further research on this topic. Then, Sec. III adds descriptions of two simple methods that are based on previous works. Finally, the Evaluation Section IV presents a comprehensive evaluation of properties the spectral envelopes should meet. In order to evaluate the perceived impact of the two MFA methods described, we also carried out two proof-of-concept listening tests, one for pitch scaling and another one for conversion of intensity. The results are presented at the end of Sec. IV.

II. Theoretical analysis of Multi-Frame Analysis

This section presents first theoretical elements about the MFA in the context of our study. The condition for a good reconstruction of the spectral envelope is basically dependent on two elements: The sampling scheme (SFA, MFA) and the cepstral properties of the sampled envelope. The former is addressed below and the next sub-section addresses the later.

For the sake of simplicity we assume that the glottal source samples the VTF with a strict harmonic grid $f_h = h \cdot f_0$. Within this assumption, the envelope estimation using SFA is equal to a usual signal reconstruction based on uniform sampling. According to the Nyquist theorem, for a sampling rate f_s, the envelope can be estimated up to the maximum cepstral order:

$$P^* = \left\lfloor \frac{0.5 \cdot f_s}{f_0} \right\rfloor \tag{1}$$

which will be called *usual order* in the following [18]. The cepstral order basically limits the cepstral representation of the envelope to estimate, assuming that all cepstral coefficients above the order are zeros. Top plot of Fig. 2 depicts the distribution of cepstral magnitude and shows where the limit of the cepstral order occurs with respect to f_0 and (1).

The second assumption used in this work is the stationarity of the amplitude spectral envelope within a time window of two vibrato periods, according to the singer's mean vibrato frequency, e.g. 400ms for a 5Hz vibrato. Additionally, we assume constant vibrato within this window, which leads to the $f_0(t)$ model:

$$f_0(t) = f_c \cdot 2^{a_{\text{FM}} \sin(2\pi \cdot f_{\text{FM}} \cdot t)} \tag{2}$$

where f_c is the carrier frequency of the vibrato (the average f_0), $a_{\text{FM}} = \frac{\gamma}{1200}$ is the FM's amplitude of the vibrato[1], $f_{\text{FM}} \approx 5$Hz is the FM's frequency. We defined the FM component on a logarithmic scale, for the sake of simplicity for the following discussions. As shown in the following, the span of the f_0 curve in the time window is the most important characteristic. Thus, (2) is chosen mainly as a convenience for the following discussions. We do not need to assume that the actual f_0 follows this model strictly. The f_0 model could be slightly different (e.g. time-varying properties of the vibrato), the following results would still hold, but might be more difficult to express.

In the MFA approach, considering multiple frames is equivalent to sampling the same VTF frequency response multiple times using harmonic grids corresponding to different f_0 values along the time axis (see plots of Fig. 1).

[1]Because the pitch perception is logarithmic, it is convenient to express a frequency difference on a logarithmic scale. Thus, in music, a cent is defined as a 1200^{th} of an octave, i.e. a frequency Y, which is γ cents above another frequency X, is defined by: $Y = X \cdot 2^{\frac{\gamma}{1200}}$

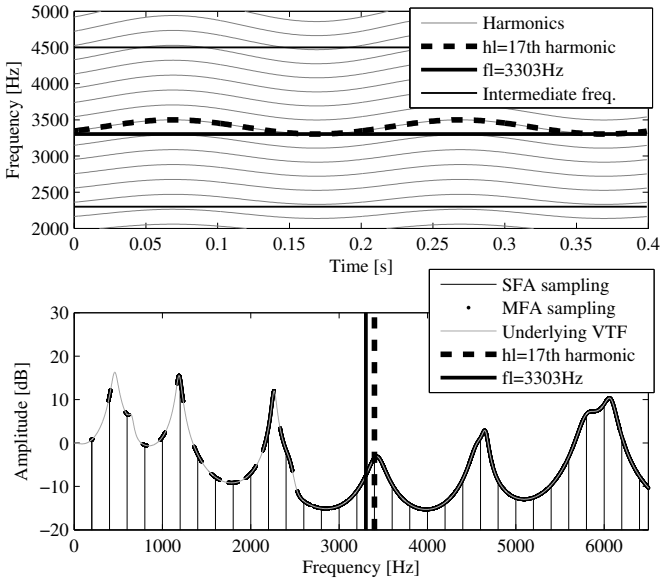

Fig. 1. Example of MFA sampling with average $f_0 = 200$Hz, vibrato's frequency=5Hz, vibrato's extent=$\pm$50cents. On bottom plot, the amplitude values sampled by the MFA are shown for a full vibrato period.

A. From sparse to complete sampling across frequency

For most methods, the estimation methods starts by extracting the sinusoidal components (i.e. pairs of frequency and log amplitude) of each frame, using peak picking [19] or harmonic models [20]. Using a synthetic VTF, bottom plot of Fig. 1 shows the resulting peaks for SFA and MFA sampling for a full vibrato period, using $f_c = 200$Hz, $f_{\text{FM}} = 5$Hz, $a_{\text{FM}} = 50$cents and a time step of 5ms between each frame. From Fig. 1, one can see that above a given frequency ($\approx$ 3300Hz) the frequency gaps between two frames are not visible. This is due to the fact that, within a period of vibrato, there exists a harmonic whose highest frequency reaches the lowest frequency of the next harmonic, as shown by the middle line of the upper plot of Fig. 1. Therefore, given an f_0 and a vibrato extent γ, there is a frequency limit above which the VTF can be entirely sampled with an arbitrary frequency resolution by reducing the distance between the frames. This condition can be expressed as:

$$h \cdot f_0 \cdot 2^{\frac{\gamma}{1200}} = (h+1) \cdot f_0 \cdot 2^{-\frac{\gamma}{1200}} \tag{3}$$

which can be solved for h:

$$h_l = \lceil (2^{\frac{2\gamma}{1200}} - 1)^{-1} \rceil \tag{4}$$

The frequency above which the VTF can be fully represented is the lowest frequency reached by the h_l-th harmonic. This frequency limit is:

$$f_l = h_l \cdot f_c \cdot 2^{-\frac{\gamma}{1200}} \tag{5}$$

With a common vibrato extent of 50cents, (4) leads to $h_l = 17$ and $f_c = 200$Hz leads to $f_l = 3303$Hz.

To summarize, using the MFA sampling scheme for analyzing vibrato segments, one can consider that the VTF can be fully recovered above a given frequency limit. Above this frequency limit, the theoretical limitations of the frequency resolution are determined by the time gap between two frames, which can be reduced arbitrarily by simply increasing the frame rate, and, by the accuracy of the sinusoidal components estimation of the spectral peaks. Therefore, using an MFA approach, the main reconstruction problem lies in the frequency band below f_l. This is obviously very encouraging compared to an SFA sampling scheme where the frequency sampling is always scarce up to Nyquist frequency.

B. VTF Cepstrum and Aliasing

The cepstral properties of the VTF are of major importance for its good estimation. Because the cepstrum of the actual VTF is not limited, whatever the used sampling scheme, aliasing effects arise in the spectral domain. Fig. 2 illustrates the cepstral distribution of 1000 VTFs generated by an acoustic tube simulator [21] (see Sec. IV-B for more details). This

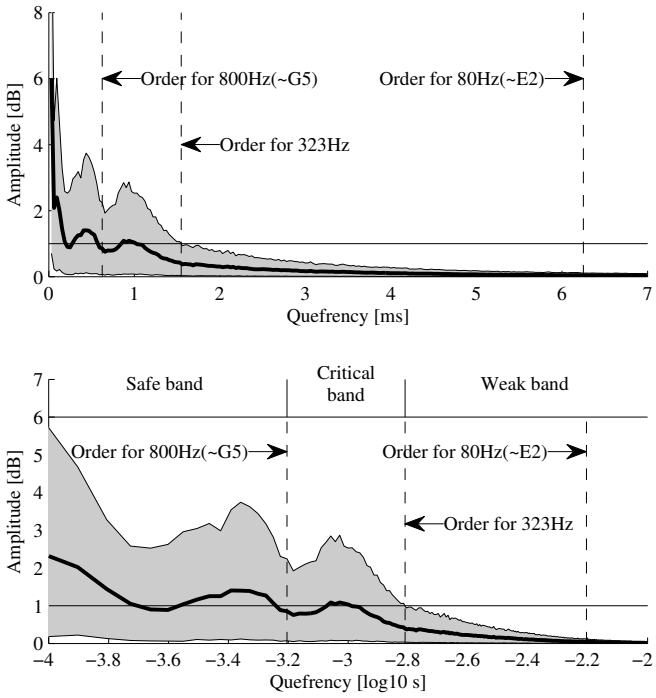

Fig. 2. The distribution of the cepstral magnitudes of 1000 synthetic VTFs with respect to the quefrency. The lowest, median and highest lines shows the 5%, 50% and 95% percentiles, respectively. The horizontal line corresponds to a hypothetical perception threshold of 1dB.

Figure shows in gray the possible values of the cepstral magnitudes between the 5% percentile (the lowest line) and the 95% percentile (the highest line). The 50% percentile is shown in thick black line. From this figure, we can see that from 1ms, the cepstrum diminishes almost in an exponential manner. This actually corresponds to the cepstral representation of poles and zeros [22], which decays in this same manner. Regarding aliasing effects, this shows that the VTF cepstrum is not limited and the aliasing effects might have an important impact on the estimations. Nevertheless, the decaying queue is encouraging since it implies that the estimation error due to aliasing is negligible for some cepstral order P. On the contrary, with an f_0 above 800Hz[2], the aliasing effects will be

[2]Even though 800Hz may look high compared to spoken voice, it is close to A5, a common upper bound for soprano singers.

substantial and recovering accurately the frequency response of the VTF seems highly compromised.

For the following discussions and to discuss the numerical evaluation, we split the cepstrum into 3 bands (see Fig. 2). Assuming a range of possible $f_0 \in [80, 800]Hz$, the cepstral coefficients below the lowest usual order are independent on the f_0. Additionally, they will always be better estimated than the coefficients above, because the sampling resolution given by f_0 will always be enough for their estimation. For this reason, we call *safe cepstral band* the quefrency interval from 0 up to the lowest usual order ($0.5 \cdot f_s/800$ in our case). The just noticeable amplitude difference is around 1dB [23], [24], [25]. In Fig. 2, one can also see that, above 1.5ms (≈ -2.81[log10 s]), the 95% of the cepstrum's magnitude is below this threshold. Below this quefrency limit, any single coefficient matters in the perception of the envelope's timbre, because it will easily generate spectral differences for more than 1dB over the whole spectrum. For this reason, in the following, we call *critical band*, the quefrency interval from the lowest possible usual order up to this noticeable limit. Finally, we call *weak band*, the quefrency band above the noticeable limit, because each single coefficient in this band should not impact the perception noticeably, even though a combination of them might be noticeable.

We can now investigate the effect of the sampling scheme on the VTF's cepstral response. The sampling function of the SFA scheme is:

$$S_{\text{SFA}}(f) = \sum_{h=-H}^{H} \delta(f - h \cdot f_0) \tag{6}$$

where H is the maximum number of harmonics up to Nyquist frequency. By stacking up the sampling of multiple frames, the sampling function of the MFA scheme is:

$$S_{\text{MFA}}(f) = \frac{1}{K} \sum_{k=0}^{K-1} \sum_{h=-H}^{H} \delta(f - h \cdot f_c \cdot 2^{a_{\text{FM}} \sin(2\pi f_{\text{FM}} k \cdot \Delta t)}) \tag{7}$$

with K the number of frames used and Δt the step size between two frames. In Fig. 3, using $f_c = 440$Hz, $f_{\text{FM}} = 5$Hz, $a_{\text{FM}} = 30$cents and $\Delta t = 5$ms, the 2nd plot from the top illustrates the power cepstral density of $S_{\text{SFA}}(f)$ and $S_{\text{MFA}}(f)$. It is the power of the continuous time inverse Fourier transform of the sampling functions (i.e. their continuous cepstra). It illustrates where the repetitions of the VTF's cepstra will occur through their convolutions with the sampling function. The 3rd plot shows the convolution of the VTF's cepstrum with that of the sampling function. Conversely to a conventional uniform sampling, the harmonic sampling is always truncated by the Nyquist frequency which rarely corresponds to a perfect integer multiple of f_0. Thus, the SFA sampling exhibits main lobes around the repetitions of $1/f_0$ instead of the standard Dirac delta functions. Globally, one can see that the main lobes of the MFA are actually lower than that of the SFA. The bottom plot shows a drop of ≈5dB on the first repetition. Moreover, with the MFA scheme, the repetitions continues to decrease with quefrency. This improvement can be explained by the fact that the cepstral peaks (2nd plot from the top) of each frame sums up perfectly at zero quefrency in the

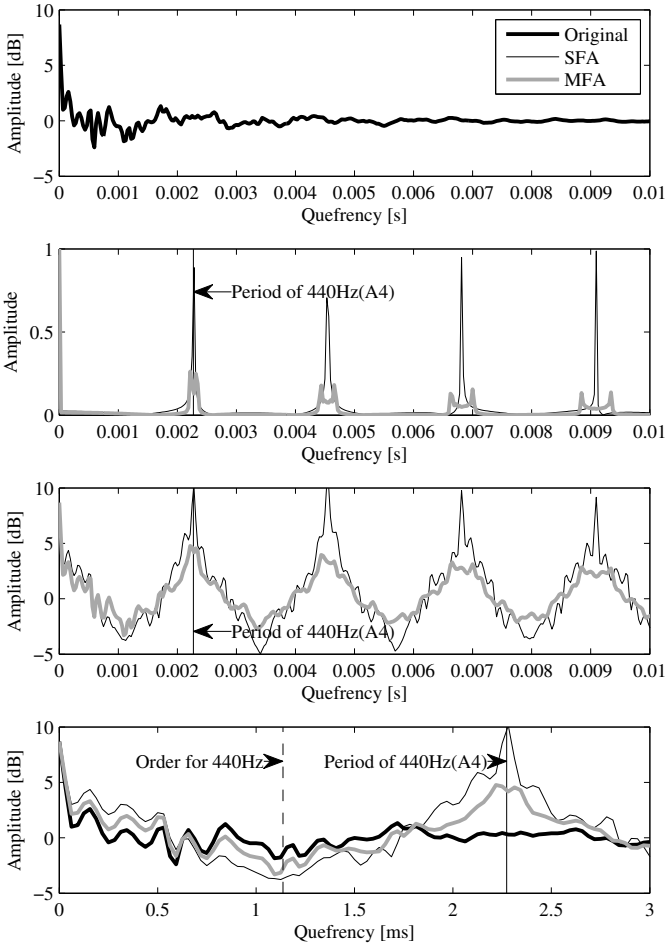

Fig. 3. An example of aliasing effects of the VTF's quefrency response through SFA and MFA sampling schemes. From top to bottom plots: The VTF cepstral response; The *power cepstral density* of the sampling schemes (using $f_c = 440$Hz $a_{\text{FM}} = 50$cents); The convolution of the VTF cepstral response with each sampling scheme; A zoom on the first period of the 3rd plot, superimposed with the original VTF's cepstral response.

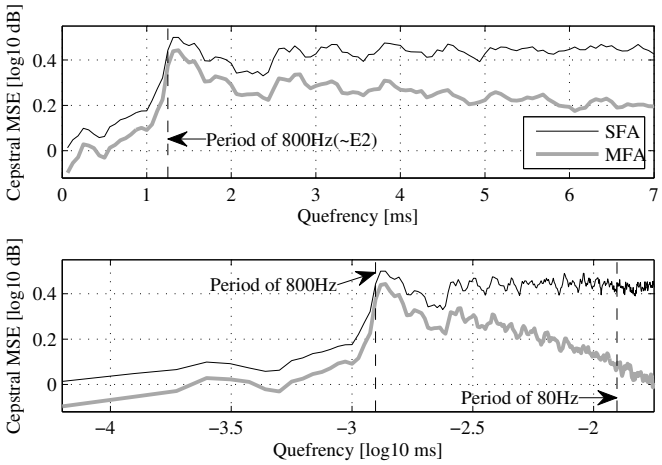

Fig. 4. Average of cepstral Mean Squared Error (MSE) of SFA and MFA sampling schemes, using 1000 synthetic VTFs.

SFA and MFA case. However, among the frames of MFA, the repetitions of this peak are not aligned, since they are delayed by different f_0. As a consequence, they do not sum up perfectly and their amplitudes decrease as the quefrency

increases. This is encouraging for the MFA approaches, as the stacked harmonic structures seems to help reducing the repetition of the VTF cepstrum, and consequently, the aliasing effects.

The bottom plot of Fig. 3 shows that the cepstra for both SFA and MFA are highly distorted compared to the original cepstrum. In order to assess this distortion more precisely, we computed the average cepstral error for the 1000 VTFs of Fig. 2 with random values of $f_c \in [80, 800]$Hz, $f_{\text{FM}} \in [4, 6]$Hz and $a_{\text{FM}} \in [10, 50]$cents, whose results are averaged and shown in Fig. 4. Fig. 4 shows that the averaged cepstral error of SFA is more important than that of the MFA, in all quefrency bands. This confirms that an MFA-based method should create an estimate where the aliasing effects have been reduced compared to an SFA-based method.

In conclusion, using MFA on vibrato segments, we have seen that the main issue is to interpolate the frequency gaps that are present only in the lowest frequencies, the highest being fully covered by the harmonics (See Fig. 1). Additionally, MFA-based methods should, in overall, provide more accurate estimates since the aliasing effects are reduced compared to the SFA sampling scheme.

III. Methods

In this section, we describe two MFA-based methods which are new extensions of existing methods. The first is based on the works of Shiga et al.[15] and the second on Wang et al.[6], using a cepstral LS solution and a simple linear interpolation technique, respectively.

A. Simple Discrete Cepstral Envelope for MFA (SDCE-MFA)

For both SFA and MFA, the cepstral model of the log amplitude spectral envelope is [10]:

$$E(f) = c_0 + 2\sum_{n=1}^{P} c_n \cos(n2\pi f/f_s) \tag{8}$$

where c_n are the cepstral coefficients, P the cepstral order. Given a set of sinusoidal parameters, the problem is to estimate c_n, such as $E(f)$ is smooth and it passes as close as possible to the harmonic amplitudes. For SFA, many solution have been suggested [10], [11], [12], [26]. The MFA solution suggested in [15] minimizes the error function:

$$\epsilon = \sum_{k=1}^{K} \|\boldsymbol{W}_k \cdot (\boldsymbol{a}_k - d_k\boldsymbol{u}_k - \boldsymbol{B}_k\boldsymbol{c})\| \tag{9}$$

where k and K are the frame index and number, $\boldsymbol{a}_k$ contains the partials log amplitudes at frame k, 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:

$$\boldsymbol{B} = \frac{1}{H}\begin{bmatrix} 1 & 2\cos(1\omega_1) & 2\cos(2\omega_1) & \cdots & 2\cos(P\omega_1) \\ 1 & 2\cos(1\omega_2) & 2\cos(2\omega_2) & \cdots & 2\cos(P\omega_2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos(1\omega_h) & 2\cos(2\omega_h) & \cdots & 2\cos(P\omega_h) \end{bmatrix} \tag{10}$$

with $\omega_h = 2\pi h f_0(t)/f_s$ and H the number of harmonics extracted in the frame. In this work, the weighting term $\boldsymbol{W}_k$ is set to a diagonal matrix whose terms corresponds to a Gaussian window centered at DC and 3kHz of standard-deviation, similarly to [15]. Compared to [11] and [26], (9) has a weighted term, but no regularization term. In this work, we continued to use (9), since the motivations for a regularization term drop using MFA, as argued in the introduction and shown in Sec. III-A2.

1) Simplification of the DCE-MFA: The voice source is likely to have an amplitude modulation component. In singing voice, during vibrato, a tremolo component is often present. 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 the original DCE-MFA [15], the correction factors d_k are estimated for this purpose. In [15], d_k and $\boldsymbol{c}$ are estimated jointly in an iterative way. Here below, we show that these gain correction can actually be dropped, thus leading to a simpler solution. In [15](9), the estimate of $\boldsymbol{c}$ is given by:

$$\Big(\sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{B}_k\Big)\boldsymbol{c} = \sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k(\boldsymbol{a}_k - d_k\boldsymbol{u}_k) \tag{11}$$

whose right-hand side can be distributed:

$$\begin{aligned}\Big(\sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{B}_k\Big)\boldsymbol{c} &= \sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{a}_k \\ &\quad - \sum_{k=1}^{K} \boldsymbol{B}_k^T\boldsymbol{W}_k(d_k\boldsymbol{u}_k)\end{aligned} \tag{12}$$

Distributing the inverse of the left-hand side of this equation on the right leads to:

$$\begin{aligned}\boldsymbol{c} &= \sum_{k=1}^{K}\Big(\sum_{l=1}^{K} \boldsymbol{B}_l^T\boldsymbol{W}_l\boldsymbol{B}_l\Big)^{-1} \cdot \Big(\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{a}_k\Big) \\ &\quad - \sum_{k=1}^{K} d_k \cdot \Big(\sum_{l=1}^{K} \boldsymbol{B}_l^T\boldsymbol{W}_l\boldsymbol{B}_l\Big)^{-1} \cdot \Big(\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{u}_k\Big)\end{aligned} \tag{13}$$

The second term of (13) containing the cepstral coefficients of the constant spectrum $\boldsymbol{u}_k$, we can write:

$$\begin{aligned}\boldsymbol{c} &= \sum_{k=1}^{K}\Big(\sum_{l=1}^{K} \boldsymbol{B}_l^T\boldsymbol{W}_l\boldsymbol{B}_l\Big)^{-1} \cdot \Big(\boldsymbol{B}_k^T\boldsymbol{W}_k\boldsymbol{a}_k\Big) \\ &\quad - \sum_{k=1}^{K} d_k \cdot [1, 0, \ldots, 0]^T\end{aligned} \tag{14}$$

Therefore, as shown by (14), the solution of $\boldsymbol{c}$ in (11) is basically an average of the solutions of the different frames, plus, a bias on the first cepstral coefficient due to the sum of the correction terms d_k. What is important to note is that the cepstral coefficients with $n > 0$ are not influenced by the choice of d_k. Therefore, the estimation of the shape of the envelope is independent of the frames alignment. Additionally, after estimation, the envelope's shape has to be re-aligned on the central frame (of index $(K-1)/2$). Thus, the last term of (14) will be replaced, which makes the estimation of d_k

and any pre-alignment of the frames useless. This theoretical result shows that, even though the global optimal solution of (9) might still need a joint optimization of shape and alignment, eq. (14) suggests a very convenient sub-optimal solution. For the following, we call the Simplified DCE-MFA (SDCE-MFA), the solution given by (14) while disregarding the alignment corrections.

Finally, it is worth noting that if we add a regularization term in the left-hand term, as suggested for SFA in [11], the implication (13)→(14) is no more valid. The alignment will impact the envelope's shape, and the iterative procedure described in [15] would become necessary.

2) Order selection: Since the frame alignment can be discarded, the remaining critical parameter is the cepstral order P. The existence of the LS solution in (14) is conditioned by the invertibility of the sum of the left-hand term of (11). With the SFA scheme ($K = 1$), this matrix is invertible if there are at least P harmonics. With the MFA scheme, since there are K times more partials than with the SFA scheme, there is a theoretical potential to multiply the usual order by K. This first seems very promising as the envelope could be estimated up to an arbitrary cepstral order by simply increasing K. However, within a time window, many partials

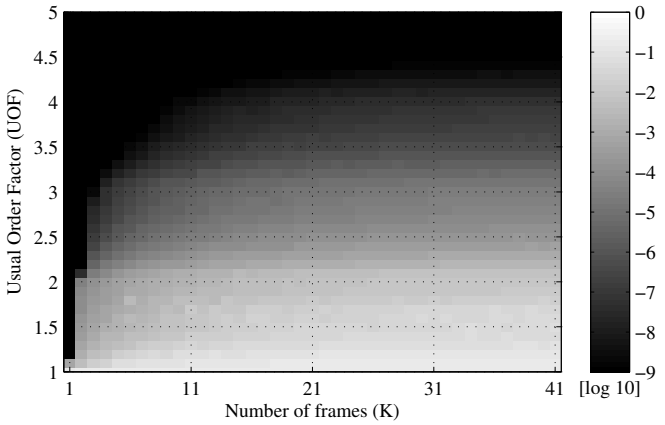

Fig. 5. Average condition number of the left-hand term of (11) over 1000 random vibrato settings, with respect to a factor of the usual order.

of different frames are actually very close to each other, thus, providing similar sampling points, as shown in Fig. (1) in the low frequencies. To illustrate the actual limitation and true potential for extending the order, Fig. 5 shows the condition number of the matrix to inverse in (11), averaged using 1000 random $f_c \in [80, 800]$Hz, random $a_{\text{FM}} \in [0, 100]$cents and random $f_{\text{FM}} \in [0, 5]$Hz (the time window always covers two periods of vibrato). This condition number is expressed in terms of a factor of the usual order, i.e. the *Usual Order Factor* (UOF). This figure shows that for UOF>4.5, the condition number is already smaller than 10^{-9}. Thus, within a vibrato segment, there is little hope for increasing the usual order further than a factor of 4.

The gaps between the partials play also an important role in the stability of the envelope estimate. The NyquistShannon theorem tells us that the stability of the reconstruction is ensured for uniform sampling (i.e. SFA in our case). With MFA, the stability is expected with UOF=1, because the system is simply more constrained than in the SFA scheme. The question about the stability remains for UOF>1. Answering this question theoretically goes beyond the scope of this presentation, even though it would be interesting to know the theoretical behavior of this *fan-harmonic*-sampling that appears with MFA sampling scheme. In this work, we do not actually need to answer this question since, as shown in the evaluation, the UOF cannot be increased further than 1.4 without creating meaningless variations on the envelope estimate.

In terms of computational efficiency, it is important to note that each computation of the SDCE-MFA needs 3K matrix multiplications and one matrix inversion. Even though the dimensions of these matrices are relatively small (~ 300 for the order and 81 for the number of frames in the following experiments), this currently prevents real-time analysis. However, this is not a major issue for concatenative synthesis, since the spectral envelope estimation needs to be run only once on the database, which can be done offline, independently from the synthesis stage.

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

The second MFA-based method described in this work is a simple extension of the traditional linear interpolation [27] for MFA analysis, which has already been used for comparison in [6] (called Linear-MFA in the following). 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. Note that this interpolation adds an overall constraint on the estimate, conversely to the DCE-MFA approach, where we assume that the LS solution does not need any other constraint than the sole peaks for reconstructing the intermediate points.

Erratic shapes appear on the estimate because of the noise in the sinusoidal parameters and the very close proximity of harmonics of different frames (See the example in Fig. 6). This issue can become even more problematic, with high f_0, when the gaps between the harmonics are more substantial. In this situation, these erratic shapes can be as wide as a formant, thus, resulting in *musical* sounds in synthesis. This problem makes the Linear-MFA practically unusable without further processing. For example, in [6], they applied an AR model on top of the Linear-MFA, which is then used for estimation of formants' position.

In this article, we suggest the Linear-MFA-LIFT method, which consists in a simple low-pass lifter of the Linear-MFA envelope. The choice of the cepstral order for this liftering is far from straightforward and will be studied in the Evaluation Section. Even though this approach first seems simplistic, it actually provides interesting results, as shown in [6] for estimation of AR models and in [16] for time regularization and in the evaluations of this article. The energy alignment might also look simplistic. However, as shown in Fig. 1, we have seen that the harmonics of different frames are very close to each other in the lowest frequencies. Thus, between successive frames, it is likely to find partials that are expected to have the same amplitude and would help in the alignment of

the frames. Because partials overlap above a frequency limit, as shown in Sec. II, one might argue that higher frequencies should be used for amplitude alignment. However, in singing voice, as in speech, noise might also appear above 4-5kHz, which would definitely impact and degrade the alignment. For this reason, we chose to use the first 4kHz, similarly to the evaluations in [6] and [16], while keeping in mind that it could be better studied and improved in future works.

In terms of computational efficiency, the Linear-MFA-LIFT is obviously faster than the SDCE-MFA. It requires only one linear interpolation and two FFT for liftering the cepstrum. This also means that this method can be used in real-time.

IV. Evaluation

In this section, we assess the advantages of the two extended methods described above, while comparing with other state-of-the-art methods, using first numerical measurements, then using subjective listening tests.

In this evaluation, most of the methods make use of a set of sinusoidal parameters (frequency and amplitude). These parameters are extracted from spectral peaks picked from a 4096 bins DFT [19], using a Blackman window of 3 periods duration, estimated each 5ms. The amplitude of each peak is fit by quadratic regression of the closest bins to $h \cdot f_0$ in the DFT, as described in [27]. There is no such peak at DC frequency since the voice signal is an acoustic signal. Partials are also missing at right before Nyquist since the f_0 is never a perfect divisor of the sampling frequency. At DC, because the gap is twice as big as the expected sampling distance corresponding to the usual order, the amplitude envelope is likely to be unstable, which might lead to a main ripple at DC and a Gibbs phenomenon. There is no straightforward solution to deal with this issue since the signal itself is not supposed to have a DC. In this work, to alleviate this problem, artificial partials are added at DC using the first harmonic's amplitude. Before Nyquist, artificial partials are also added up to Nyquist, using the last available amplitude.

For all the MFA-based methods compared below, a window duration of 400ms was used ($\approx$ 2 vibrato periods for a 5Hz vibrato, i.e. 400/5+1=81 successive frames). Theoretically, a window duration of a single vibrato period would be sufficient. However, it is safer to increase this window in order to minimize the modulation effects due to the position of the window in the vibrato period. A longer window is not advised as it would extend the computation time for no practical benefits.

A. Compared methods

For the sake of the comparison, we also compared with the following methods. First, two SFA-based methods are used. The "True-Envelope" (TE) [28], [12], [29], which is based on a cepstral model and does not use sinusoidal parameters. It makes use of an iterative algorithm minimizing the distance between the spectral peaks and the envelope. The traditional Discrete Cepstral Envelope (DCE) is also used, in order to evaluate the gain offered by its MFA counterpart, namely the SDCE-MFA. However, conversely to the SDCE-MFA, we used a regularization term in the DCE, as described in [11], in order to prevent envelope instability, since this problem is far more likely to happen in SFA. The regularization coefficient is set to 0.035, as in [11], [16]. Finally, in order to compare with another MFA-based method, the DLinear-MFA [16] is also used in the evaluation. Basically, this method consists in averaging the frequency derivative of a basic linear interpolation envelope. The final envelope is retrieved through cumulation of the averaged derivative. It has shown very good time regularity, less over-smoothing effect than using a simple low-pass 2D filtering of the envelope, but less accuracy than other MFA-based methods. Fig. 6 shows examples of the 6 methods compared in this evaluation.

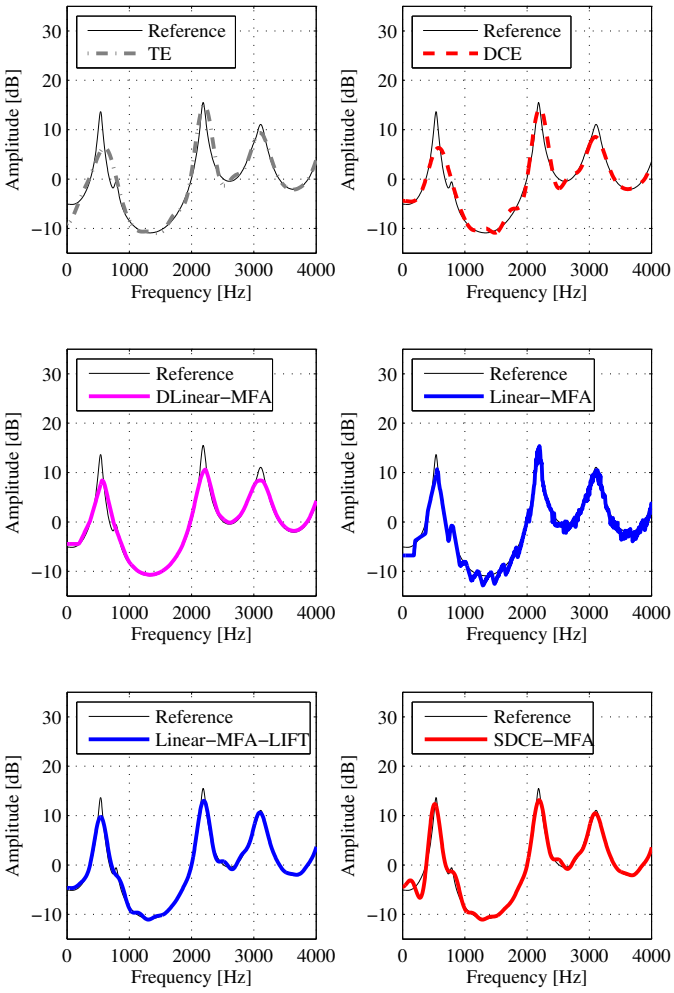

Fig. 6. Examples of estimate of an amplitude spectral envelope using the methods compared and a synthetic signal using a known reference frequency response.

B. Numerical evaluation using synthetic signals

Since the ground truth of the VTF of the voice signals is unknown, we first evaluate the methods numerically using synthetic signals and, thus, with known reference VTFs. For this evaluation 1000 samples are generated using the following procedure. First, the synthetic $f_0(t)$ follows (2) with f_c a random value in $[80, 800]$Hz, a_{FM} a random value in $[0, 150]$cents and f_{FM} a random value in $[4, 6]$Hz. A Dirac

impulse train is then generated according to the $f_0(t)$, which is then modulated in amplitude, in order to reproduce a tremor. This amplitude modulation is generated by low-pass filtering a noise of Gaussian distribution so that its final standard-deviation is equal to 0.5dB. The low-pass filtering is made by using a Butterworth IIR filter of order 4 with a cutoff frequency of 5Hz. Finally, each synthetic source is convolved by a stationary VTF generated using the same digital acoustic synthesizer [21] as in the theoretical sections of this article. A different set of uniform random articulatory parameters was used for each of the 1000 synthetic samples. These articulatory parameters include: the jaw position, 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, as in French) [21]. The length of the vocal tract was also set randomly in $[13, 18]$cm for each sample.

Below we describe four measurements which allowed us to assess various properties that a spectral envelope estimate needs to satisfy in the context of our study. Each measurement is computed for each sample. Then, the measurements of the samples are averaged in order to produce the figures 7 and 8.

1) Absolute cepstral error (AC Error): Knowing the reference cepstrum c_n^*, we assess the overall error of the envelope estimate for each voice sample using the absolute cepstral error:

$$\epsilon_n = \frac{1}{M}\sum_{m=1}^{M} |c_n^* - c_{m,n}| \tag{15}$$

where M is the number of frames in the voice sample and $c_{m,n}$ is the nth-cepstral coefficient of the frame m in each sample.

2) Cepstral Variance: As we have seen from Fig. 2, the amplitude distribution of the cepstrum's magnitude contains interesting information. For example, we will see in the following results that the reconstruction of the global variance of the 1000 synthetic VTFs is important. In order to assess the capacity of a method to reproduce this *global* variance, we suggest to compute the following ratio:

$$\bar{\sigma}_n = \frac{\text{std}_i(\bar{c}_{n,i})}{\text{std}_i(\bar{c}^*_{n,i})} \quad \text{where} \quad \bar{c}_{n,i} = \frac{1}{M}\sum_{m=1}^{M} c_{m,n,i} \tag{16}$$

where $\bar{c}_{n,i}$ is the average cepstrum over all the frames of each sound sample i and $\text{std}_i(.)$ computes the standard-deviation over the 1000 samples i. If $\bar{\sigma}_n < 1$, it means that the variance of the envelope estimates is smaller than it should be, compared to the variance of the synthetic VTFs. This corresponds to an averaging effect and, in the frequency domain, this can be interpreted as a flattening of the envelope estimates. If $\bar{\sigma}_n > 1$, it means that the envelopes are more different than it should be. Unnatural resonances might be expected during synthesis.

3) Relative cepstral error (RC Error): In order to better distinguish differences between methods, we also use the relative cepstral error:

$$\rho_n = \frac{1}{M}\sum_{m=1}^{M} \left| \frac{c_n^* - c_{m,n}}{c_n^*} \right| \tag{17}$$

The relative error allows to know how far we are from estimating a cepstral coefficient. To avoid divisions by zero, we forced values smaller than an epsilon to the median values of its neighbors.

4) Relative Cepstral Excess (RC Excess): If an envelope estimate is too loose, not constrained enough, it risks to degenerate and create incoherent resonances (sometimes perceived as *musical sounds* during synthesis). With the following measurement, we want to represent this property of *inventing* incoherent, erratic, unstable or meaningless shapes. Using the relative error, this property can be assessed by measuring the error excess above 1:

$$\chi_n = \frac{1}{M}\sum_{m=1}^{M} \max(\{\rho_{m,n}, 1\}) - 1 \quad \text{where} \quad \rho_{m,n} = \left| \frac{c_n^* - c_{m,n}}{c_n^*} \right| \tag{18}$$

Fig. 7, shows the results of the measurements for the six methods compared (legend in the left bottom plot). With 1000 sound samples, we assume that the measurements are accurate enough, so that there is no need to show confidence intervals or run significance tests. In these plots, the UOF has been set to 1.4 for the SDCE-MFA and the Linear-MFA-LIFT. This value will be justified in Sec. IV-B5.

From this Figure, we can derive the following observations. Regarding the AC Error, we can first note that all methods reach almost the same error on the highest cepstral order. This makes sense since the highest possible frequency resolution given by the harmonic sampling does not allow to retrieve higher cepstral information, whatever the estimation method. The RC Error shows that, in the safe band, all the MFA methods divide the error by a factor slightly less than two, compared to the SFA methods. This is very encouraging for MFA approaches and corresponds to the expectation one can have from the theoretical observations in Sec. II. The biggest lack of estimation lies in the weak band. However, the absolute error shows that this quefrency band has a smaller error than that of the critical band, which is due to a very weak cepstral magnitude. Thus, future efforts should not focus on this band and still put the priority on the safe and critical bands. Globally for all methods, one can see that the Cepstral Variance is bigger than 1 in the safe band. The errors added by the estimation methods can surely explain this increase of Cepstral Variance. In the critical and weak bands, the Cepstral Variance is basically lower than 1, for most methods. This lack of cepstral variance demonstrates an averaging and *flattening* effect of the envelope estimates. The DLinear-MFA seems to suffer the most of this problem, whereas the SDCE-MFA seems to be the most capable of reconstructing the original variance. The erratic shapes present in the Linear-MFA (see Fig.6), which are liftered in the Linear-MFA-LIFT, appear clearly in the RC Excess and mainly impact the critical and weak bands. The RC Excess also shows that both SFA methods can produce substantial amount of erratic shapes, even in the safe band. However, because their cepstral order is lower than that of Linear-MFA-LIFT and SDCE-MFA, this behavior reduces drastically in the critical band. The Linear-MFA-LIFT shows a quite positive result as it clearly reduces the amount of shapes that might deteriorate a synthesis quality.

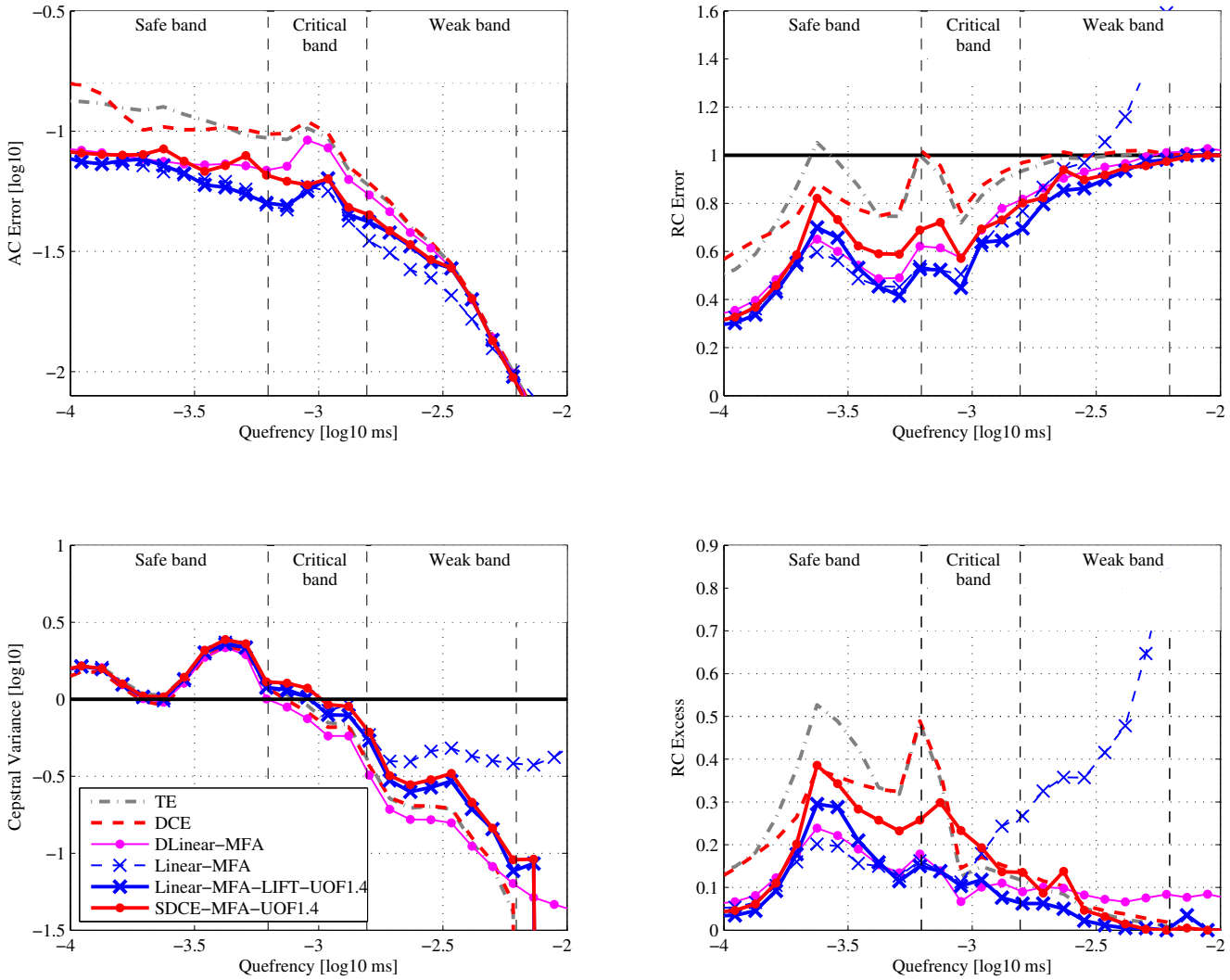

Fig. 7. Numerical evaluation of envelope estimation methods, using four measurements: The Absolute Cepstral Error (AC Error), the Relative Cepstral Error (RC Error), the Cepstral Variance (a measure of global variance of the estimates) and the Relative Cepstral Excess (RC Excess, a measure of presence of meaningless and erratic shapes).

5) Measurements with respect to UOF: Since the order selection is far from straightforward in MFA approaches, we show in this section the evolution of the four measurements with respect to the Usual Order Factor (UOF). In this experiment, the *safe band* should not be considered since it is independent of f_0 (and thus independent of UOF). We also consider that the *weak band* might bias the results with information which has very limited impact on the perception. Therefore, the results of this experiment are computed by averaging only the measurements in the critical band (See Fig. 8) (DLinear-MFA and Linear-MFA have no cepstral order and, thus, appear as constants). Based on this Figure, we can make the following observations. For Linear-MFA-LIFT and SDCE-MFA, it shows that the AC and RC Errors first decrease when UOF increases. The Cepstral Variance helps to better understand the nature of the RC Error. When the order is low, the resulting error is due to the flattening of the envelope estimates, which is over constrained. When the order increases, the Cepstral Variance increases and, thus, the error is more related to a lack of an overall precision of the formants' shape rather than a systematic flattening. With the SDCE-MFA, over $\approx$1.4 the Cepstral Variance continues to increase, and as a consequence, the AC Error increases again, which defines a local minimum of AC Error.

By looking at the Cepstral Variance, one can also see that the Linear-MFA-LIFT is bounded by the Linear-MFA. This makes sense since the Linear-MFA-LIFT is a low-pass liftered version of the Linear-MFA. The variance of the former is expected to be always lower than that of the later. On the contrary, the Cepstral Variance of SDCE-MFA continues to increase when UOF increases. The SDCE-MFA is actually the only method which is able to reconstruct the global variance that is observed in the critical band of the synthetic VTFs. According to this experiment, we observe that an UOF of 1.4 allows the SDCE-MFA method to reproduce the averaged variance in the critical band. This justifies the UOF value used for Fig. 7 and this same value will be used in the following listening tests. In this work, in order to be able to compare

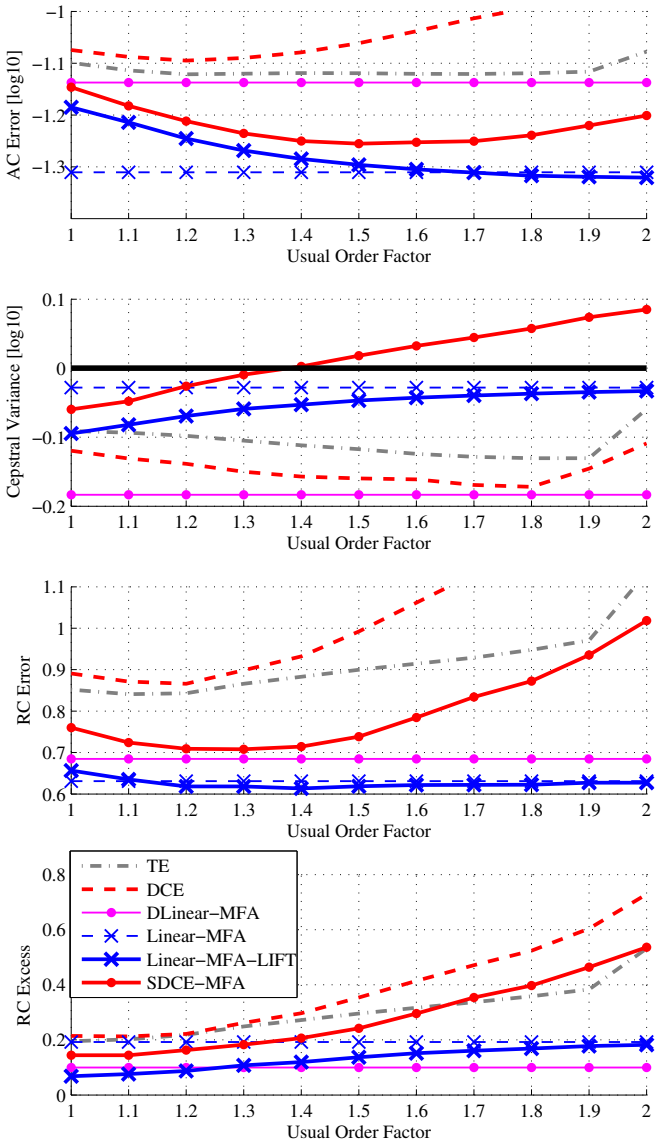

Fig. 8. From top to bottom, the Absolute Cepstral Error (AC Error), Relative Cepstral Error (RC Error), Cepstral Variance and Relative Cepstral Excess (RC Excess) with respect to Usual Order Factor (UOF), averaged in the critical cepstral band only.

the SDCE-MFA and the Linear-MFA-LIFT, we used the same UOF value between the two methods. Note that an UOF of 1.4 also corresponds to a local minimum of the Linear-MFA-LIFT's RC Error.

C. Listening tests

In this section, using comparative listening tests, we present the results of two proof-of-concept experiments specific to singing voice synthesis: pitch scaling and conversion of the intensity's timbre. Because this work focuses on the spectral amplitude envelope, these listening tests should be completely independent of the voice source. This is obviously not possible since a source has to be chosen for *listening to* the envelope. Depending on the synthesis technique, noise is always synthesized differently in the voice source resulting in perceptual differences for a same amplitude envelope. For example, a spectral amplitude modification using a vocoder might increase a noisy frequency band and degrade the overall perception of the signal, whereas this amplitude modification might be perfectly meaningful depending on the desired effect (e.g. formant increase, reduction of the spectral tilt). For this reason, because the perceived results are dependent on the used synthesis technique, we present results for two techniques. The first technique is using harmonic synthesis [20], [30], [31]. The used implementation is basically the one presented in [31] and available in COVAREP [32](v1.3.2). In order to limit the influence of the noise effects mentioned above, we chose to fix the characteristics of the voice source with a fixed voiced/unvoiced Frequency at 4kHz. The second technique is a phase vocoder [33], [34], [29] using shape preservation [35]. Because a listening session cannot last too long (for reason of focus and patience of the listeners) and comparison tests tend to grow rapidly as the number of compared methods grows, we compared only three methods: the SDCE-MFA, the Linear-MFA-LIFT and the TE. We chose the TE and not the DCE, because their numerical results are similar and the TE has been widely used through the vocoder used in these experiments.

Both tests make use of sustained sung vowels of duration between 1 and 3 seconds, containing natural vibrato, thus, corresponding to the f0 model used in this study. This work being part of the ChaNTeR national project, which aims at synthesizing singing voice in French, we used recordings of the 15 French vowels. In order to satisfy the assumptions and models used in this study, we took the opportunity of the project to record specific samples for these listening tests, from two singers, one male and one female. The setup is very specific and favorable to the methods, therefore, it cannot be used to generalise to arbitrary speech or singing signals. We note that the application for real world singing or speech signals will require further studies that investigate the performance in situations not fitting the underlying assumptions as well. We also hope that these two proofs of concept will encourage future research in MFA analysis. For both experiments, the overall procedure is the following: Using recordings of the 15 vowels mentioned above, samples are synthesized using the amplitude spectral envelope analyzed by means of the three methods and the specific procedures detailed in the following sub-sections. Then, using a web-based interface, listeners give their pairwise preferences for the three possible combinations of methods using a 7-points scale [36], based on *the clarity of the pronunciation of the phoneme*. Finally, the scores are aggregated into a Comparison Mean Opinion Score (CMOS) for each method [36]. The evaluation using harmonic synthesis and vocoder have no reason to be dependent on each other. For this reason, each listening test has been run in an independent web-page (i.e. one for the pitch scaling using harmonic synthesis, one for the pitch scaling using vocoder, and the same for intensity conversion). As a result, we will not compare the absolute results between the tests in the conclusions. We will rely only on the final conclusions of each test, which is sufficient for this study. Note that all the audio files used in the following tests can be found at:
http://gillesdegottex.eu/Demos/DegottexG2016mfaenvsing

1) Pitch scaling: In this test, for each of the two singers, 15 recordings of the 15 vowels in fortissimo intensity are pitch

scaled downwards and upwards, using 0.75 and 1.25 scaling factors, respectively. We selected these two scaling factors for the two following reasons: i) It avoids substantial factors which emphasize artifacts related to the synthesis technique. ii) It samples the estimated envelope at frequencies which were not present in the original signal (using 0.75 and 1.25, 3 harmonics out of 4 have frequencies that are not present in the original signal). For each of the 15 original recordings, 6 samples are generated (3 methods times 2 scaling directions). There is also two voices and two sounds per comparison pair. Thus, there is a minimum of 24 sounds to examine, which can take easily 15 minutes to assess properly. For this reason, each listener evaluates only 4 different vowels, one per singer for both scaling direction, i.e. 12 comparisons pairs. The 4 vowels are taken randomly among the 15, for each listener. 31 and 33 listeners took the tests for the harmonic synthesis and the vocoder-based modification, respectively (See Fig. 9). Based on these results we first conclude that, for both synthesis techniques, the MFA methods are clearly preferred compared to the TE envelope. One can also note that the very simple Linear-MFA-LIFT provides already a very interesting improvement compared to the TE. The SDCE-MFA is also preferred to the Linear-MFA-LIFT, but this difference is only shown on the harmonic synthesis for the female voice. Since the numerical results using synthetic signals between the SDCE-MFA and the Linear-MFA-LIFT methods are very similar, except for the Relative Cepstral Deviation (see Fig. 8), one can also assume that the preference shown for the SDCE-MFA is mainly due to the lack of cepstral variance in the Linear-MFA-LIFT. The trends suggest also that bigger improvements could be expected on female voices than on male voices. This also makes sense since the f_0 is higher with the female voice, the sampling is more scarce and, thus, the information from the neighbor frames bring even more improvement when using the MFA-based methods.

2) Intensity timbre conversion: We present in this section a simple experiment to evaluate the impact of the envelope estimates on the perceived quality, when replacing the original amplitude envelope by another *target* envelope. This target envelopes is estimated from a single sample of singing voice recorded with the target effect, i.e. fortissimo in this experiment. In this test, the source samples are similar to those of the previous test, but with a mezzo-forte intensity. The amplitude of the source is then replaced by the fortissimo envelopes from the same three envelope methods, using the same two synthesis techniques. In this test, each listener assesses also 4 random vowels among the 15, thus, leading also to a total of 12 comparisons pairs.

27 and 25 listeners took the tests for the harmonic synthesis and the vocoder-based modification, respectively (See Fig. 10). Based on these results one can conclude very similar results to the previous experiment. For both synthesis techniques, the MFA-based envelope estimates exhibit a clear preference compared to the TE envelope. With this test, we cannot, however, observe differences between the SDCE-MFA and the Linear-MFA-Lift methods. This suggests that the slight advantage of the SDCE-MFA for female voice seen in the previous test might be quite rare in concrete applications.

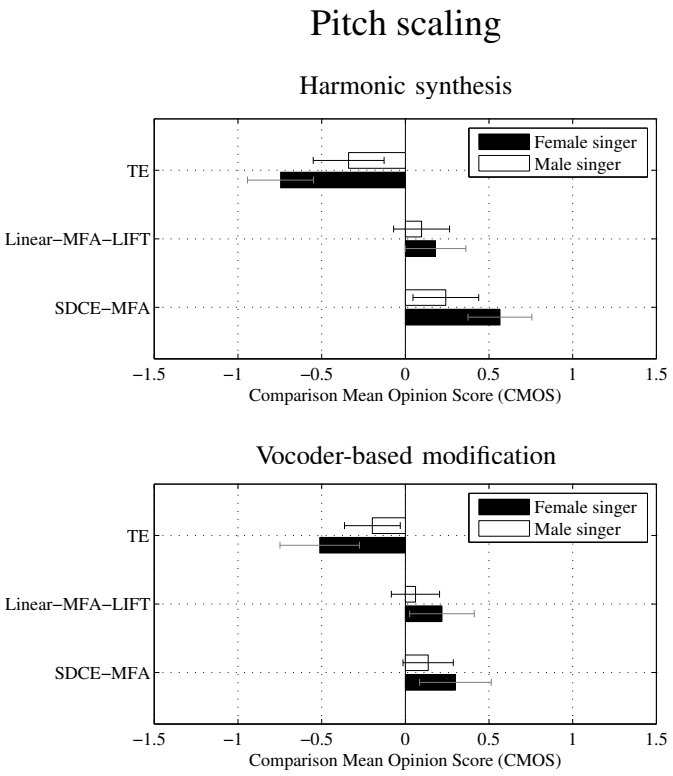

Fig. 9. Results of a listening test about pitch scaling, using harmonic synthesis on the top and vocoder-based modification on the bottom.

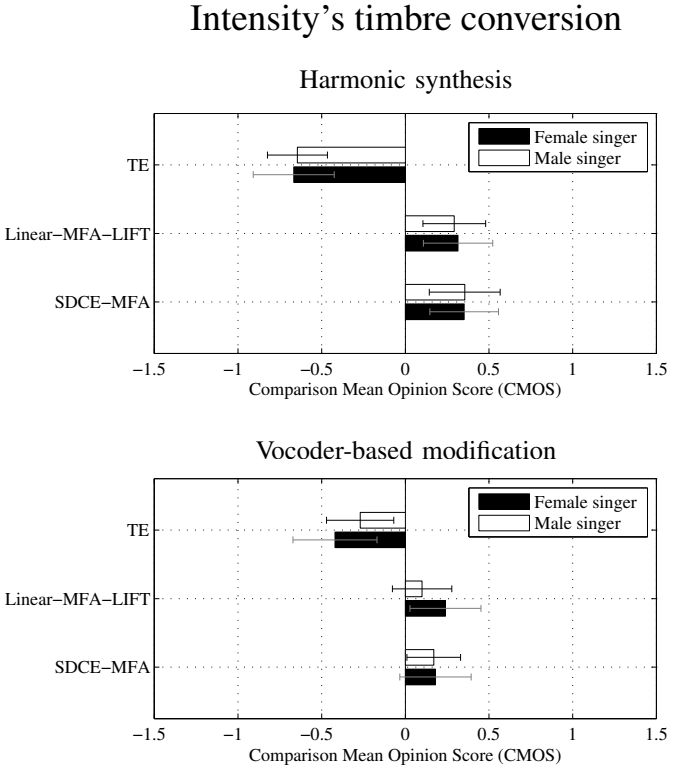

Fig. 10. Results of listening tests for intensity's timbre conversion. Using harmonic synthesis on the top and vocoder-based modification on the bottom.

This is also very encouraging for MFA-based approaches as the Linear-MFA-LIFT method is also extremely simple to implement and does not involve the computational complexity of the LS solution in the SDCE-MFA.

V. Conclusions

Whereas most estimation methods of amplitude spectral envelopes use only a Single Frame of frequency Analysis (SFA), we have brought, in this article, a few very encouraging elements about estimation methods that are using Multiple Frame Analysis (MFA). We did not design any solution specific to singing voice. However, because our target application is singing voice synthesis, we studied the MFA in the context of a sustained vowel with the presence of a constant vibrato.

We have shown that, above a certain frequency limit, the sampling of the Vocal Tract Filter (VTF) is fully sampled, conversely to SFA where the sampling is sparse up to Nyquist. We have also shown that the aliasing effects present in the envelope estimate are clearly reduced when using MFA sampling scheme compared to SFA.

Additionally, we have simplified and extended two MFA-based methods described in previous works. We have shown that the frame alignment, which seems necessary in MFA approaches, can be discarded when using the Discrete Cepstral Envelope (DCE-MFA), leading to a Simplified SDCE-MFA. We also suggested to simply lifter the cepstrum of the MFA-based linear interpolation (Linear-MFA), leading to a very simple and efficient Linear-MFA-LIFT. To evaluate the presented methods, we compared them to state-of-the-art methods using numerical evaluation. We noticed that the error is almost halved when using MFA-based methods compared to SFA methods. It was also shown that the SDCE-MFA is the only method that is able to recover the original global variance of the reference VTFs, by extending the usual cepstral order with a factor of 1.4. The MFA-based methods also reduce the erratic and inconsistent shapes, thus, providing globally a more accurate and reliable envelope estimate than the compared SFA methods. Finally, according to listening tests, which are dedicated to the material we are using for our research project, the MFA-based methods clearly improve the quality of pitch scaling and conversion of intensity's timbre.

VI. Acknowledgements

This work was supported by the ChaNTeR ANR project (ANR-13-CORD-0011). *Web link:* http://chanter.limsi.fr

References

[1] H. Kenmochi and H. Ohshita, "Vocaloid - commercial singing synthesizer based on sample concatenation," in *Proc. INTERSPEECH*, 2007, pp. 4009–4010.

[2] Jordi Bonada, *Voice Processing and Synthesis by Performance Sampling and Spectral Models*, Ph.D. thesis, Universitat Pompeu Fabra, Spain, 2008.

[3] K. Nakamura, K. Oura, Y. Nankaku, and K. Tokuda, "HMM-based singing voice synthesis and its application to japanese and english," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 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*, 2015.

[5] E. Joliveau, J. Smith, and J. Wolfe, "Vocal tract resonances in singing: The soprano voice," *The Journal of the Acoustical Society of America*, vol. 116, no. 4, pp. 2434–2439, 2004.

[6] 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.

[7] 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.

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

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

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

[11] 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.

[12] 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.

[13] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[14] 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.

[15] Y. Shiga and S. King, "Estimating the spectral envelope of voiced speech using multi-frame analysis," in *European Conference on Speech Communication and Technology, EUROSPEECH*, 2003, pp. 1737–1740.

[16] 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.

[17] S. McAdams and X. Rodet, *Basic Issues in Hearing*, chapter The role of FM-induced AM in dynamic spectral profile analysis, Academic Press, 1988.

[18] 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.

[19] 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.

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

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

[22] Alan V. Oppenheim and Ronald W. Schafer, *Digital Signal Processing*, Prentice-Hall, 2nd edition, 1978.

[23] W. Jesteadt, C.C. Wier, and D.M. Green, "Intensity discrimination as a function of frequency and sensation level," *The Journal of the Acoustical Society of America (JASA)*, vol. 61, no. 1, pp. 169–177, 1977.

[24] 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.

[25] R. R. Riesz, "Differential intensity sensitivity of the ear for pure tones," *Physical Review*, vol. 31, pp. 867–875, 1928.

[26] 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.

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

[28] 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.

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

[30] 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.
[31] 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.
[32] 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.
[33] M. Liuni and A. Roebel, "Phase vocoder and beyond," *Musica/Tecnologia*, vol. 7, pp. 73–89, 2013.
[34] A. Roebel, "Supervp software," http://anasynth.ircam.fr/home/english/software/supervp, 2015.
[35] Axel Roebel, "Shape-invariant speech transformation with the phase vocoder," in *Proc. Interspeech*, 2010, pp. 2146–2149.
[36] The ITU Radiocommunication Assembly, "ITU-R BS.1284-1: Engeneral methods for the subjective assessment of sound quality," Tech. Rep., ITU, 2003.

Gilles Degottex received the Diploma degree in computer science from University of Neuchâtel (UniNE), Switzerland. After a one-year specialization at École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, he obtained his Ph.D. degree in 2010 at the Institut de Recherche et Coordination Acoustique/Musique, IRCAM/UPMC, Paris, France. He did a postdoctoral position at University of Crete, Heraklion, Greece, on voice modeling, transformation and synthesis. He is currently holding a postdoctoral position at Ircam, Paris, France on singing voice synthesis in the national ChaNTeR project. His research interests include voice source features, sinusoidal and spectral modeling for speech and singing voice synthesis.

Luc Ardaillon graduated from the Institut Supérieur d'électronique de Paris (ISEP, Paris) in 2013 with a specialization in signal and image processing, after one year studying sound and music processing at the Queen Mary University of London (QMUL, London) as an Erasmus student. He also received a master degree from UMPC/IRCAM, Paris, in acoustic, signal processing and informatics applied to music (ATIAM master) in 2013. He is currently pursuing a PhD degree in the Analysis/Synthesis team at IRCAM, Paris, on the subject of expressive singing voice synthesis and transformation, with a focus on parameters generation and singing styles modeling.

Axel Röbel received the Diploma in electrical engineering from Hannover University in 1990 and the Ph.D. degree (summa cum laude) in computer science from the Technical University of Berlin in 1993. In 1994 he joined the German National Research Center for Information Technology (GMD-First) where he worked on adaptive modeling of time series of nonlinear dynamical systems. In 1996 he became assistant professor for digital signal processing in the communication science department of the Technical University of Berlin. In 2000 he obtained a research scholarship to work on adaptive sinusoidal modeling at CCRMA Standford University, and, he joined IRCAM for working in the Analysis/Synthesis team doing research on frequency domain signal processing. In summer 2006 he was Edgar-Varse guest professor for computer music at the Electronic studio of the Technical University of Berlin. Since 2011 he is head of the Analysis/Synthesis team of IRCAM. His current research interests are related to music and speech signal modeling, transformation and synthesis.