Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis

Daniel Erro, Iñaki Sainz, Eva Navas, and Inma Hernaez

***Abstract*—This article explores the potential of the harmonics plus noise model of speech in the development of a high-quality vocoder applicable in statistical frameworks, particularly in modern speech synthesizers. It presents an extensive explanation of all the different alternatives considered during the design of the HNM-based vocoder, together with the corresponding objective and subjective experiments, and a careful description of its implementation details. Three aspects of the analysis have been investigated: refinement of the pitch estimation using quasi-harmonic analysis, study and comparison of several spectral envelope analysis procedures, and strategies to analyze and model the maximum voiced frequency. The performance of the resulting vocoder is shown to be similar to that of state-of-the-art vocoders in synthesis tasks.**

***Index Terms*—Harmonics plus noise model, statistical parametric speech synthesis, vocoder, voice transformation.**

I. Introduction

MODERN speech synthesizers and voice conversion systems are based on statistical modeling of sets of feature vectors extracted from speech signals. In speech synthesis, for instance, there is a training phase in which speech features are modeled at phoneme level using context-dependent hidden semi Markov models (HSMMs). Then, during synthesis, given a sequence of phonemes and contexts, their corresponding HSMMs are concatenated to form a sentence-level model and the system calculates the sequence of feature vectors that shows maximum likelihood with respect to that model. A more detailed description of this increasingly popular type of synthesizers can be found in [1]. In voice conversion systems, the correspondence between vectors belonging to the source and target speakers is often captured by means of Gaussian mixture models [2]–[4]. In both cases, vocoders play a crucial role: they are responsible for translating the involved speech signals into tractable sets of vectors with good properties for statistical modeling and also for reconstructing speech waveforms from vectors at the highest possible quality and naturalness. Vocoders performance is one of the main limitations of statistical parametric systems. This is particularly true for speech synthesis systems since they generate speech exclusively from models—voice conversion systems do not create new speech signals but just transform existing ones. This is the reason why this paper has been focused on synthesis.

Understood in this context, speech vocoding is different from speech coding. The main goal of speech coding is to achieve the highest possible resynthesis quality using the lowest possible number of bits to transmit the speech signal [5]. Real-time performance during analysis and reconstruction is also one of its typical requirements. In the statistical parametric frameworks mentioned above, vocoders must have not only these high resynthesis capabilities but also provide parameters that are adequate to statistically model the underlying structure of speech, while information compression is not a priority. Depending on the application, the efficiency requirements can also be relaxed, mainly in analysis mode. Nevertheless, both technologies have many points in common. Indeed, well known spectral representations used in speech coding (and also in other fields of speech technologies), i.e. Mel-frequency Cepstral Coefficients (MFCCs) and Line Spectral Frequencies (LSFs), are widely used in this context as well.

In recent years, the development of statistical parametric speech synthesizers and voice conversion systems has also pushed research towards vocoding techniques. In the first release of HTS (the publicly available HMM-based Speech Synthesis System [6]), speech was parameterized at certain frame rate into two streams: $\log f_0$ and spectral envelope [7]. Mel-generalized cepstral analysis [8] was applied to calculate the spectral (typically Mel-cepstral, MCEP) coefficients. During waveform reconstruction, a simple pulse/noise excitation (linked to f_0) was filtered through the so called MLSA filter [9] (linked to the spectral coefficients). Due to the simplicity of the excitation, the resulting speech showed an annoying buzziness. Subsequent works attempted to analyze the residual signal obtained via inverse filtering after spectral parameterization by measuring the voicing degree in different frequency bands and then using a mixed excitation to reconstruct speech [10], [11]. More recent HTS releases [12] included a vocoder based on STRAIGHT (see [13] for a description of this high-quality speech analysis, modification and reconstruction tool), which increased the solidness of the spectral estimation and provided aperiodicity measurements to characterize the mixed excitation in frequency. Maia *et al.* [14] used a sophisticated trainable mixed excitation based on state-dependent filters for pulses and noise. In [15], Drugman *et al.* used a two-band mixed excitation

Manuscript received March 06, 2013; revised July 27, 2013; accepted September 11, 2013. Date of publication September 25, 2013; date of current version March 11, 2014. This work was supported in part by the Spanish Ministry of Economy and Competitiveness (SpeechTech4All, TEC2012-38939-C03-03), the Basque Government (Ber2tek, IE12-333), and Eurorregion Aquitaine-Euskadi (Iparrahotsa, 2012-004). The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Keiichi Tokuda.

D. Erro is with the AhoLab Signal Processing Laboratory, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain, and also with the IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain (e-mail: derro@aholab.ehu.es).

I. Sainz, E. Navas, and I. Hernaez are with the AhoLab Signal Processing Laboratory, University of the Basque Country (UPV/EHU), 48013 Bilbao, Spain (e-mail: inaki@aholab.ehu.es; eva@aholab.ehu.es; inma@aholab.ehu.es).

Digital Object Identifier 10.1109/JSTSP.2013.2283471

in which the upper band was treated as noise and the lower band was modeled through a set of deterministic waveforms. In all these works, the inclusion of this third parameter stream related to the excitation resulted in noticeable improvements with respect to the two-stream baseline vocoder. Several other works assumed signal models inspired by speech production theories, thus replacing this spectral envelope + excitation scheme by a vocal tract + glottal source scheme. Some examples can be found in [16]–[18]. In these systems, the vocal tract was parameterized as was traditionally done with spectral envelopes, while the glottal source was characterized at waveform level. Despite the recent improvements, STRAIGHT is still the most widespread state-of-the-art method.

The vocoder described in this paper exploits the properties of the harmonics-plus-noise model (HNM) [2], [19], which assumes that locally stationary speech signal segments can be decomposed into a lower harmonic band and an upper noise-like band. Note that this signal model implicitly assumes a two-band mixed excitation, though it handles the whole speech signal rather than only the excitation. Unfortunately, HNM features (amplitudes, phases, etc.) cannot be applied directly to feed a statistical system for practical reasons [20], mainly the time-varying number of harmonics and their enormous sensitivity to variations in f_0. For this reason we choose to parameterize speech into three different streams, i.e. $\log f_0$, MCEP coefficients for spectrum, and maximum voiced frequency (MVF), the hallmark of our vocoder being that it uses HNM features and procedures to facilitate and improve parameter extraction and waveform reconstruction.

Although a similar strategy had already been followed in [21], the system implementation was not fully optimized at some levels: treatment of the phase, choice of the parameterization, etc. During the development of our vocoder, these issues have been studied in more depth and more recent advances regarding HNM, such as quasi-harmonic modeling (QHM) [22], have also been explored. Another previous attempt in the same direction can be found in [20], where modified HNM procedures capable of operating at constant frame rate were applied to synthesis. As the harmonic part and the noise part were parameterized into independent streams, some discontinuities appeared at the voicing boundaries. Therefore, in this work we suggest the use of a single stream devoted to spectral envelopes. This article extends our previous works in [23]–[25] by presenting a deeper experimental study of all the different aspects involved in the design of an HNM-based vocoder. Specifically: (a) it describes a QHM-based pitch refinement technique that enhances the performance of the subsequent analysis steps; (b) it explores different spectral envelope estimation techniques while considering the impact of parameterization; (c) it studies the convenience of MVF analysis and modeling in comparison with simpler predictive techniques. We will finally show that the performance of the proposed HNM-based vocoder is quite similar to that of the most popular state-of-the-art vocoder, namely the one based on STRAIGHT.

The rest of the paper is structured according to the different steps involved in speech analysis, parameterization and reconstruction procedures. Sections II, III and IV describe the way the three different parameter streams ($\log f_0$, MCEP coefficients and MVF, respectively) are obtained and discuss different issues arose during the development of the vocoder. Section V describes the HNM-based reconstruction procedure. Section VI is devoted to the optimization of several aspects of the vocoding process in resynthesis tasks. The performance of the optimized vocoder in synthesis is evaluated in Section VII. The final conclusions are summarized in Section VIII.

II. Pitch Detection and Refinement

The first analysis step to be performed is pitch detection. Many pitch detection algorithms (PDA) have been proposed in the literature, most of them exhibiting very good performance when applied to clean signals (note that the signals involved in speech synthesis usually show high signal-to-noise ratio). The vocoder presented in this paper includes an implementation of the autocorrelation-based algorithm described in [26].

Since the next analysis step is HNM-based spectral envelope estimation, we found the accuracy of the PDA to have a non-negligible influence on spectral estimation. In order to optimize the pitch estimation towards an accurate harmonic analysis we explored pitch refinement algorithms based on QHM theory. QHM [22] assumes that the voiced speech segments can be locally approximated by a sum of quasi-harmonic sinusoids:

$$s(t) = \sum_{i=1}^{I} (a_i + t b_i) \exp(j 2\pi f_i t) \tag{1}$$

where $\{f_i\}$ are initial estimates of the frequencies of the I quasi-harmonic components, $\{a_i\}$ are their complex amplitudes, and $\{b_i\}$ are the complex slopes of these complex amplitudes over time. As detailed in [22], given $\{f_i\}$ and the samples of the current frame, both $\{a_i\}$ and $\{b_i\}$ can be calculated via least squares optimization within a 3-period frame. Two ways of correcting the frequencies $\{f_i\}$ using these coefficients can be found in the literature. The first one, proposed in the context of the so-called deterministic plus stochastic model [2], is based on combining the derivatives of the instantaneous phases of the quasi harmonic sinusoids. The second and most recent one, which was proposed in [22] and can be referred to as QHM-based frequency correction, is formulated as follows:

$$\Delta f_i = \frac{\mathrm{Re}\{a_i\}\mathrm{Im}\{b_i\} - \mathrm{Im}\{a_i\}\mathrm{Re}\{b_i\}}{2\pi |a_i|^2} \tag{2}$$

At the beginning, the frequencies can be initialized as $f_i = i f_0$, being f_0 the pitch value yielded by the PDA. Then, a single pitch correction term can be obtained as the weighted average contribution of the individual quasi-harmonics:

$$\Delta f_0 = \frac{\sum_{i=1}^{I} w_i \cdot \Delta f_i / i}{\sum_{i=1}^{I} w_i} \tag{3}$$

I depends on the bandwidth of the QHM analysis. The weights $\{w_i\}$ can be either constant over i (constant-weighting approach) or somehow proportional to the amplitudes (weighted-by-amplitude approach). The refinement can be iterated more than once. The effectiveness of this pitch refinement algorithm in the context of a parametric vocoder is studied in Section VI. As shown in Fig. 1, there are visible local differences between original and corrected f_0 contours.

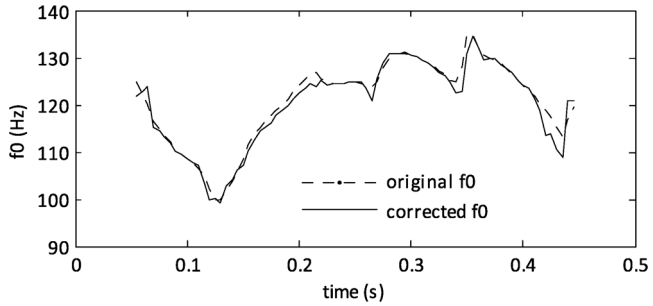

Fig. 1. Differences in the f_0 contour of a voiced segment uttered by a male speaker before and after QHM-based refinement (2 iterations).

III. Spectral Envelope Parameterization

Assuming a simplified speech production model in which a pulse-or-noise excitation passes through a shaping filter, the term *spectral envelope* denotes the amplitude response of this filter in frequency. Such an envelope contains not only the contribution of the vocal tract but also the contribution of the glottal source. In unvoiced frames, the spectrum of the noise-like excitation is flat, which means that the response of the filter coincides with the spectrum of the signal itself (except for a scaling factor). In voiced frames, the spectrum of the pulse-like excitation has the form of an impulse train with constant amplitude and linear-in-frequency phase placed at multiples of f_0. Therefore, the spectrum of the signal shows a series of peaks that result from multiplying the impulses of the excitation by uniformly spaced spectral samples of the filter response. Assuming local stationarity, full-band harmonic analysis returns these discrete samples of the spectral envelope. Then, a continuous envelope can be estimated via interpolation.

The assumption of a pulse-or-noise excitation implies ignoring the noise component of voiced segments during spectral estimation, thus using a harmonics-or-noise model—either harmonics or noise, but not both of them—instead of HNM. Given that separate analysis, parameterization and modeling of time/frequency-overlapping harmonic and noise components are known to produce some discontinuities at voicing boundaries during synthesis [20], this simplification is beneficial because it allows parameterizing all spectral information in a single continuous stream. Besides, full-band harmonic analysis is known to provide good estimates of the spectral content even at frequencies within noisy bands [27]. We therefore use a harmonics-or-noise analysis model and we tolerate a slight inconsistency between the signal models assumed during analysis and reconstruction in favor of a stable and smooth synthesis. Next we describe the way we extract and parameterize the spectral envelope.

A. Unvoiced Frames

First, the N-point log-amplitude envelope at frame k is calculated as follows:

$$S^{(k)}[m] = \log\left(\frac{1}{\sqrt{Lf_s}}\left|FFT_N\left\{s[n]\cdot w[n-n_k]\right\}\right|\right) \quad (4)$$

where $\mathrm{FFT}_N\{\cdot\}$ denotes the N-point fast Fourier transform $(N > L)$, $w[n-n_k]$ is an L-point Hamming window centered at the current analysis instant n_k, and f_s is the sampling frequency. For $f_s = 16$ kHz and for L corresponding to a 20 ms time span, N is typically set to 1024. The scaling term accompanying the FFT module in (4) normalizes its amplitude with respect to the implicitly assumed f_0, i.e. the effective frequency resolution of the FFT, equal to f_s/L. This amplitude normalization, carried out not only in the unvoiced case but also in the voiced case, enables energy-invariant signal reconstruction at different pitch and voicing conditions.

To translate $S^{(k)}[m]$ into a pth-order MCEP representation $\{c_k\}_{k=0\ldots p}$, the traditional cepstrum given by $\mathrm{FFT}_N^{-1}\{S^{(k)}[m]\}$ is warped in frequency by means of the recursion in [8]. This recursion is based on all-pass transforms and maps the frequency scale of the cepstral sequence according to

$$\beta_\alpha(\omega) = \tan^{-1}\frac{(1-\alpha^2)\sin\omega}{(1+\alpha^2)\cos\omega - 2\alpha} \quad (5)$$

where $\omega = 2\pi f/f_s$ (rad) and α can be adjusted for $\beta_\alpha(\omega)$ to match the Mel scale (typically, $\alpha = 0.42$ for $f_s = 16$ kHz) [8].

B. Voiced Frames

In short, the analysis procedure at voiced short-time segments consists of harmonic analysis followed by interpolation between harmonic amplitudes and transformation into a Mel cepstral representation. In this context, least squares based harmonic analysis [2] is advantageous with respect to other alternative methods such as peak picking from STFT spectrum [28] because it requires a shorter analysis window (2 periods), thus enabling quasi-stationary analysis even for rapidly varying signals, while showing very high accuracy.

Once the amplitudes of the harmonics at frame k, $\{A_i^{(k)}\}$, have been computed on the full analysis band (0, $f_s/2$), interpolation techniques can be applied to recover the underlying continuous log-amplitude spectral envelope $\mathrm{S}^{(k)}(f)$. In a preliminary work [24] we suggested to do this by placing weighted replicas of an f_0-dependent interpolative function $B_q(f)$ at the harmonic positions in the frequency domain. Omitting the frame index k for clarity, this can be formulated as follows:

$$S(f) = \hat{A}_1 B_q(f) + \sum_{i=1}^{I+q}\hat{A}_i\cdot[B_q(f-if_0) + B_q(f+if_0)] \quad (6)$$

where f_0 is the local fundamental frequency, I is the number of harmonics in the band $0-f_s/2$, and $\{\hat{A}_i\}$ are the f_0-normalized log-amplitudes given by

$$\hat{A}_i = \log\left(\frac{A_i}{2\sqrt{f_0}}\right), \quad 1 \le i \le I \quad (7)$$

Given that some harmonics at frequencies higher than $f_s/2$ are necessary for a correct interpolation of $S(f)$ in (6), we simply make $\hat{A}_i$ equal to $\min\{\hat{A}_1,\ldots,\hat{A}_I\}$ when $i > I$. As in the unvoiced case, normalization by f_0 removes the effect of periodicity from $S(f)$ (factor 2 is included to consider only the positive semispectrum, for consistency with the unvoiced case). Since the number of harmonics in the analysis band is inversely proportional to f_0 and their energy has to be preserved even after pitch modification, $f_0^{-1/2}$ is the appropriate normalization term. Regarding $B_q(f)$, a triangular function would implement

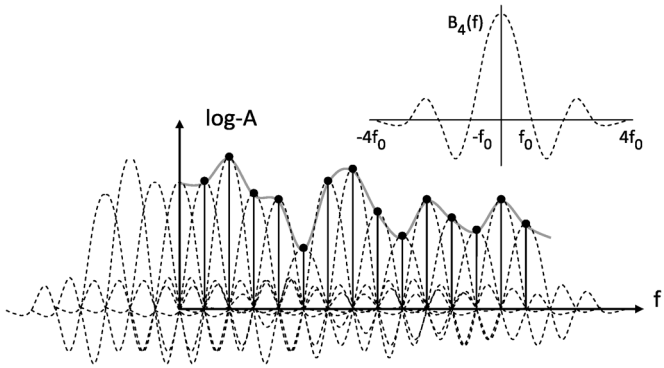

Fig. 2. Sinc-based interpolation between harmonic amplitudes.

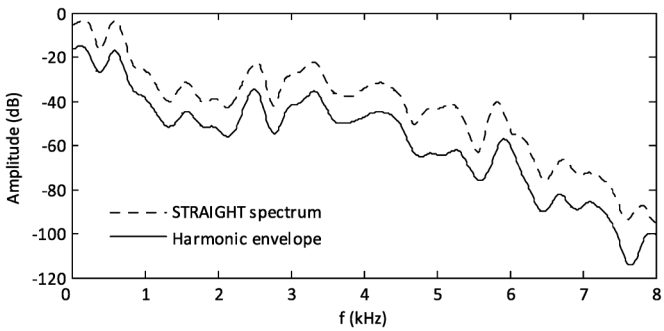

Fig. 3. Comparison between the spectral envelope interpolated from the harmonic log-amplitudes and the STRAIGHT spectrum in a real speech frame. The latter has been artificially shifted for clarity.

linear interpolation between log-harmonics. However, more accurate interpolation is obtained when $B_q(f)$ has the form of an f_0-dependent sinc function bandlimited by a Hanning window (see Fig. 2):

$$B_q(f) = \frac{1}{2}\left(1 + \cos\frac{\pi f}{q f_0}\right) \cdot \frac{\sin(\pi f/f_0)}{(\pi f/f_0)}, \quad |f| \le q f_0 \quad (8)$$

A good tradeoff between interpolation capability and complexity was found for $q = 4$. In (6), the term $\hat{A}_1 B_q(f)$ plays the role of keeping $S(f)$ almost constant below f_0, which gives good perceptual results according to prior research on HNM-based pitch modification [29]. In practice, a discrete version of the log-amplitude envelope $S(f)$ (6) is computed at frame k, $S^{(k)}[m]$, and the corresponding MCEP representation is obtained in the same way as in the unvoiced case (see Section III-A). $S^{(k)}[m]$ can be expected to be similar to the so called STRAIGHT spectrum [13], as confirmed by Fig. 3.

Alternatively, interpolation and cepstral parameterization can be carried out in a single step. The technique known as regularized discrete cepstrum (RDC), originally proposed in [30], allows direct fitting of cepstral coefficients to discrete points of the log-amplitude spectrum using any frequency scale. The Mel-RDC spectral representation is calculated as follows [2], [31]:

$$\begin{aligned} \mathbf{c} &= (\mathbf{M}^{\mathrm{T}}\mathbf{M} + \eta\mathbf{R})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{a} \\ \underset{\substack{i=1,\ldots,I \\ j=0,\ldots,p}}{\mathbf{M}_{i,j}} &= \begin{cases} 1, & j = 0 \\ 2\cos j\beta_\alpha(i\omega_0), & j > 0 \end{cases} \\ \underset{i,j=0,\ldots,p}{\mathbf{R}_{ij}} &= \delta[i-j]\cdot 8\pi^2 i^2, \qquad \underset{i=1,\ldots,I}{\mathbf{a}_i} = \hat{A}_i \end{aligned} \quad (9)$$

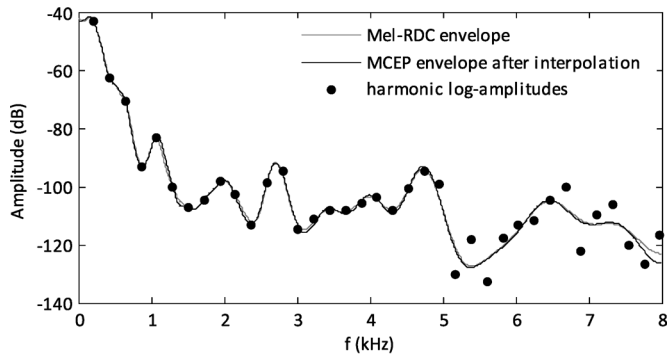

Fig. 4. Comparison between the MCEP envelopes obtained via sinc-based interpolation and Mel-RDC for the same set of harmonic amplitudes.

where η is the so called regularization coefficient, $\{\hat{A}_i\}$ are the f_0-normalized log-amplitudes given by (7), I is the number of harmonics, p is the desired cepstral order, $\beta_\alpha(\omega)$ is given by (5), and $\omega_0 = 2\pi f_0/f_s$. These equations yield the set of MCEP coefficients whose underlying envelope is closest to the discrete log-spectral observations $\{\hat{A}_i\}$ while considering a regularization term $\eta\mathbf{R}$. For $\eta = 2\cdot 10^{-4}$, this term (properly detailed in [30]) penalizes unnaturally abrupt MCEP envelopes while preserving a good match at the harmonic log-amplitudes.

Both methods, namely harmonic interpolation followed by MCEP analysis and Mel-RDC, have similar computational requirements and yield very similar MCEP envelopes as shown in Fig. 4. Their relative performance is studied and discussed in Section VI.

IV. Maximum Voiced Frequency Estimation

In earlier implementations of the vocoder [23], MVF was given a constant value. According to previous research works, this strategy yields sufficiently good quality in synthesis applications [2], [15]. However, we found that some buzziness could be perceived in some parts of the synthetic signals exhibiting a more breathy phonation, especially in low-energy segments and voiced sentence endings, where lower MVF was required. This phenomenon was alleviated when the local MVF was made energy-dependent by means of a heuristic linear relationship [24]:

$$v^{(k)} = v^{(\max)} \cdot \frac{c_0^{(k)} - c_0^{(\min)}}{c_0^{(\max)} - c_0^{(\min)}} \quad (10)$$

where $c_0^{(k)}$ is the 0th cepstral coefficient at frame k, $c_0^{(\max)}$ and $c_0^{(\min)}$ are the maximum and minimum values of c_0 over k, respectively, $v^{(k)}$ denotes MVF at frame k, and $v^{(\max)}$ (equal to 4.5 kHz in our implementation) is the maximum expectable MVF. This method will be referred to as MVF prediction in Sections VI and VII. Given the rough correlation we observed between the degree of harmonicity—linked to MVF—and the local intensity—linked to c_0—in voiced speech segments, this prediction method succeeded at alleviating the buzziness without any explicit extra parameter. Its main disadvantage is its incompatibility with real-time speech waveform generation because the c_0 range within the whole utterance has to be known in advance.

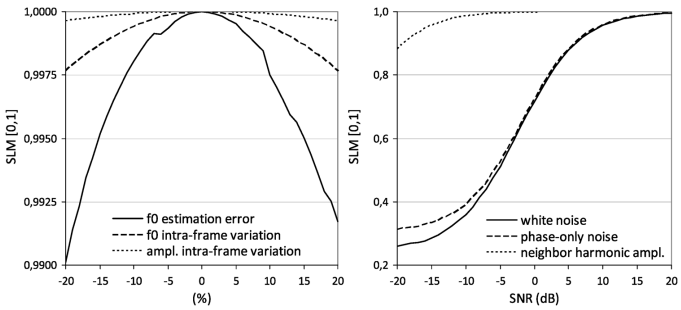

Fig. 5. Sensitivity of the SLM of a pure sinusoid to different sources of inharmonicity: pitch estimation errors, pitch and amplitude variation inside the analysis frame, overlapped white/phase noise, and interfering neighbor tones.

Explicit MVF analysis was found to yield even more natural synthetic speech while preserving the real-time generation capability of the system. The analysis algorithm described next is a reformulated version of the one originally presented in [25]. It is based on the sinusoidal likeness measure (SLM) used to classify spectral peaks in music analysis [32]. The method consists of the following steps. First, using a 3-period-length Hanning window, we compute the N-point FFT spectrum of the current frame, $X^{(k)}[m]$. Since the SLM calculation requires a high zero-padding factor, N is set to the lowest power of two that verifies $N \geq 4L$, where L is the frame length. Then, the frequencies of the spectral peaks, $\{f_i^{(k)}\}$, are determined by parabolic fitting around the local maxima of $\log|X^{(k)}[m]|$, and each peak is given an SLM score $\lambda_i^{(k)}$. Omitting the frame index k for clarity, this score can be computed as

$$\lambda_i = \frac{|\sum X[m] \cdot W_i^*[m]|}{\sqrt{\sum |X[m]|^2 \cdot \sum |W_i[m]|^2}} \quad \forall m, \left| m\frac{f_s}{N} - f_i \right| < \frac{f_0}{2} \tag{11}$$

$W_i[m]$ is the N-point FFT of a cosine at frequency f_i multiplied by the same analysis window as in $X[m]$ ($W_i[m]$ can be efficiently approximated using analytical expressions). SLM can be seen as a localized cross-correlation between the ith spectral peak and the spectrum of a pure sinusoid at f_i. Therefore, it ranges from 0 to 1, where $\lambda_i = 1$ indicates a pure sinusoid and smaller values indicate the presence of noise, transients, or sinusoids showing significant time-variation inside the analysis frame. The influence of some of these phenomena on the SLM value is explored in Fig. 5. Interestingly, SLM remains very close to 1 even when the parameters of the sinusoid vary over time (left plot) while it decreases rapidly in the presence of noise (right plot), which makes SLM robust against voiced transients. A nonlinear scaling function $g(\lambda)$ like the one shown in Fig. 6 is used to map the measured SLM onto a sort of probability of voicing in such manner that $g(1) = 1$, $g(0) = 0$, and $g(\lambda) = 0.5$ for λ equal to an empirically estimated sinusoid/noise threshold. As a result, SLM values indicating pure sinusoids remain close to 1 while SLM values indicating the dominance of noise (SNR below 0 dB in Fig. 5) are pushed strongly towards 0. In an ideal case, according to the assumptions of HNM, i.e. harmonics below the local MVF and noise above it, MVF would be equal to f_i if $g(\lambda_j) \approx 1$ for $j < i$ and $g(\lambda_j) \approx 0$ for $j \geq i$. In a realistic case the separation between harmonics and noise is not abrupt and the error of assuming MVF to be equal to f_i at frame k can be estimated as

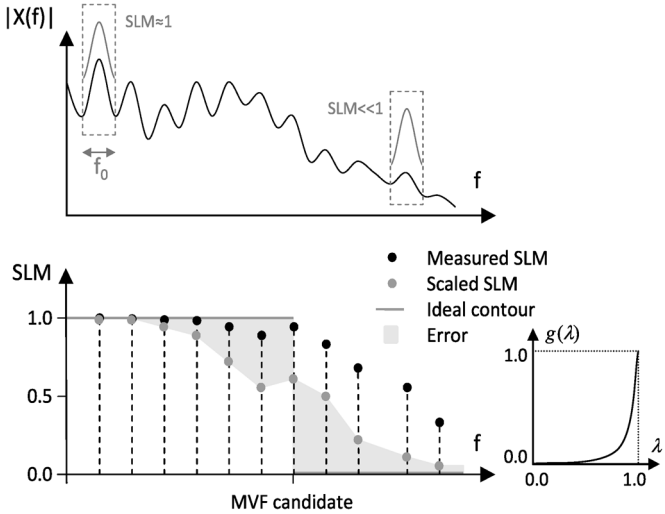

Fig. 6. Graphical explanation of the SLM-based MVF estimation method. Top: measuring the SLM of different spectral peaks. Bottom: nonlinear SLM scaling and calculation of the error for a specific MVF candidate.

$$\varepsilon_i^{(k)} = \frac{1}{I}\left[\sum_{j=1}^{i-1}\left(1 - g\left(\lambda_j^{(k)}\right)\right)^2 + \sum_{j=i}^{I} g\left(\lambda_j^{(k)}\right)^2\right] \tag{12}$$

where I is the total number of spectral peaks at the current analysis instant. The use of this SLM-based error criterion is illustrated in Fig. 6 for an exponential scaling function $g(\lambda)$; in our experiments we used a simpler empirically-adjusted piecewise linear $g(\lambda)$ given by the points (0, 0), (0.85, 0) and (1, 1). The frequencies showing relative minima of this error function are taken as MVF candidates for each k. Once the MVF candidates have been selected for all frames, $k = 1 \ldots K$, the final decision is made by determining the sequence of peak indices $\{i_1, i_2, \ldots, i_k, \ldots, i_K\}$ that minimizes the following cost function through a dynamic programming search:

$$C\left(\{i_1 \ldots i_K\}\right) = \sum_{k=1}^{K} \varepsilon_{i_k}^{(k)} + \gamma \sum_{k=2}^{K}\left(\frac{f_{i_k}^{(k)} - f_{i_{k-1}}^{(k-1)}}{\frac{f_s}{2}}\right)^2 \tag{13}$$

where $f_{i_k}^{(k)}$ is the i_kth candidate at frame k and γ has to be adjusted according to the analysis rate (in our experiments, good results were obtained for $\gamma = 1$ at 5 ms analysis rate). This cost function penalizes high values of ε and abrupt MVF variations over time (see Fig. 7). The final MVFs are given by $v^{(k)} = f_{i_k}^{(k)}$.

V. HNM-Based Waveform Reconstruction

Unlike other systems based on HNM, which reconstruct signals at pitch-synchronous frame rate and frame length, in this case the reconstruction procedure has been designed to operate at constant frame rate, thus being fully compatible with the output of statistical synthesizers. The process can be described through the following general formula:

$$s[n] = \sum_{k=1}^{K} t[n - n_k] \cdot s^{(k)}[n - n_k] \tag{14}$$

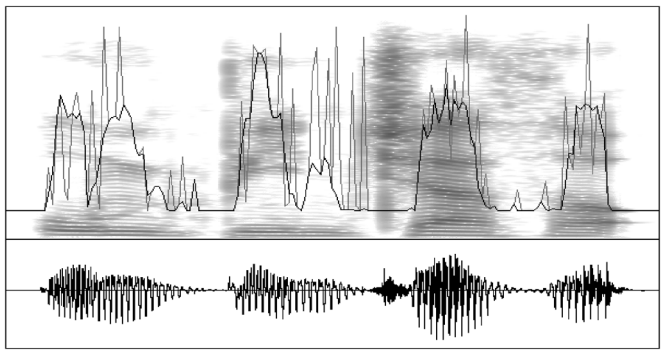

Fig. 7. Importance of the dynamic programming search in MVF estimation. Bottom: speech waveform. Top: spectrogram and MVF contour with (black line) and without (gray line) dynamic programming.

where n_k is the central sample of the kth synthesis frame, $t[n]$ is a triangular overlap-add (OLA) window, and $\mathrm{s}^{(k)}[n]$ is given by a sum of harmonic sinusoids and a time-modulated noise:

$$s^{(k)}[n] = \sum_{i=1}^{I^{(k)}} A_i^{(k)} \cos\left(i\omega_0^{(k)} n + \varphi_i^{(k)}\right) + \rho \cdot r^{(k)}[n] \cdot e^{(k)}[n] \tag{15}$$

For each k, I is the number of harmonics, $\omega_0 = 2\pi f_0/f_s$, $\{A_i\}$ and $\{\varphi_i\}$ are the synthetic harmonic amplitudes and phases, ρ is a constant scaling factor and $r[n]$ is the temporal envelope of the synthetic noise component $e[n]$. The different terms that take part in (15) are detailed in the next paragraphs. The number of harmonics, $I^{(k)}$, depends on the local MVF. In unvoiced frames, $I^{(k)} = 0$ and the speech waveform is formed only by noise, while in voiced frames

$$I^{(k)} = \left\lceil \frac{v^{(k)}}{f_0^{(k)}} \right\rceil - 1 \tag{16}$$

MVF also determines the cut-off frequency of the high-pass filter $H_n^{(k)}(f)$ that delimits the noisy band. The amplitude of this real-valued filter equals 1 at frequencies above $v^{(k)}$ and its attenuation below $v^{(k)}$ is defined in a deterministic way. In our implementation, a piecewise linear function is used with 32 dB attenuation at $f = 0$, 20 dB at $f = 0.8v^{(k)}$ and 0 dB at $f = v^{(k)}$. A complementary low-pass filter for the harmonic part is built as

$$H_h^{(k)}(f) = \sqrt{1 - {H_n^{(k)}}^2(f)} \tag{17}$$

A set of complex harmonic log-amplitudes $\{C_i^{(k)}\}$ is obtained by resampling the MCEP envelope at the Mel-harmonic positions using the same frequency scale as in expression (5):

$$C_i^{(k)} = c_0^{(k)} + 2\sum_{q=1}^{p} c_q^{(k)} \exp\left(-jq\beta_\alpha\left(i\omega_0^{(k)}\right)\right) \tag{18}$$

The corresponding real amplitudes $\{A_i^{(k)}\}$ are then given by

$$A_i^{(k)} = 2\sqrt{f_0^{(k)}} \cdot H_h^{(k)}\left(if_0^{(k)}\right) \cdot \exp\left(\mathrm{Re}\left\{C_i^{(k)}\right\}\right) \tag{19}$$

This expression includes an f_0-denormalization term (the opposite operation was performed during analysis to enable reconstruction at different pitch values) and also the contribution of the low-pass harmonic filter. Phases $\{\varphi_i^{(k)}\}$ are the result of summing the minimum-phase response given by the MCEP envelope and a linear-in-frequency term which can be attributed to the underlying excitation.

$$\varphi_i^{(k)} = \mathrm{Im}\left\{C_i^{(k)}\right\} + i\phi^{(k)} \tag{20}$$

Although the inclusion of non-minimum phase terms [33], [34] was also considered during our investigation, its contribution to the final perceptual quality was unclear and this aspect was postponed for further investigation. The linear phase term is essential to operate at constant frame rate because it ensures the phase coherence between adjacent frames. It is closely linked to f_0 and can be obtained through the following recursion:

$$\phi^{(k)} = \phi^{(k-1)} + \frac{1}{2}\left(\omega_0^{(k)} + \omega_0^{(k-1)}\right)(n_k - n_{k-1}) \tag{21}$$

The noise component $\mathrm{e}^{(k)}[n]$ is generated by inverse FFT. The positive amplitude semispectrum of the noise is obtained in the same way as the amplitudes (19) for f_0 equal to the frequency resolution of the inverse FFT and replacing $H_h^{(k)}(f)$ by $H_n^{(k)}(f)$. The phases are given random values and the negative semispectrum is built by symmetry. In voiced frames, after generation, the noise $e^{(k)}[n]$ is modulated by a time-domain window $r^{(k)}[n]$ that synchronizes its energy with respect to the harmonic component, as suggested in [2]. In our implementation, the following window is used:

$$r^{(k)}[n] = \sqrt{\frac{2}{2b^2+1}} \cdot \left(b - \cos\left(\omega_0^{(k)} n + \varphi^{(k)}\right)\right), \quad b > 1 \tag{22}$$

This window has the form of a raised cosine multiplied by an energy normalization term. For a correct synchronization, the linear phase term has been included in such manner that the minimum values of $r^{(k)}[n]$ coincide with the peaks of the excitation [2]. Satisfactory results are obtained for $2 \leq b \leq 3$. Finally, factor ρ compensates for the energy reduction due to the interference between adjacent OLA noise frames with random phases. Assuming uniformly distributed phase mismatches and considering the triangular shape of $t[n]$, ρ was analytically set to 1.21 in our implementation.

It is worth mentioning that the pitch, duration and vocal tract length can be easily controlled through $\{f_0^{(k)}\}$, $\{n_k\}$ and $\beta_\alpha(\omega)$, respectively. This means that the proposed vocoder can act as a very flexible standalone speech manipulation system.

VI. Optimization in Resynthesis

As mentioned at the beginning, vocoders must be capable of (i) reconstructing signals at the highest possible quality and naturalness and (ii) translating signals into tractable sets of vectors with good properties for statistical modeling. Previous works have already shown that the parameters yielded by the vocoder presented here, especially the MCEP coefficients, are suitable for statistical modeling [2]. In our experience, however, small variations in the way these parameters are extracted may have a noticeable impact on the naturalness of the resulting synthetic voice. This section reports several experiments that were carried out to determine the best configuration of the vocoder from an analysis/reconstruction point of view. Later, in Section VII,

we will show that analysis/reconstruction improvements lead to improved modeling/synthesis. Three specific aspects are studied here in correspondence with the contents of Section II, III and IV, respectively: pitch refinement, spectral envelope analysis and MVF.

In order to cover a wide variety of voices, a database consisting of 53 different voices (25 female + 28 male) and 2 utterances per voice was collected for this first set of resynthesis experiments. The voices were taken from different speech synthesis and recognition databases in English, Spanish and Basque. The specific utterances representing each voice were chosen randomly among candidates with a suitable duration (around 5 s). Although the recording conditions were database-dependent, in all cases the sampling frequency was 16 kHz and the signal-to-noise ratio was checked to be high enough for analysis-synthesis purposes.

A. Pitch Refinement

Pitch detection errors result in less precise harmonic analysis. In this experiment we have used the energy of the harmonic modeling error to assess the accuracy of the pitch refinement method described in Section II. We studied two particular aspects: the type of weighting applied in (3) and the bandwidth of the underlying QHM analysis. For every configuration, the refinement was iterated twice. The experimental procedure was the following. All the utterances in the database were pitch-analyzed under several configurations of the PDA. Next, full-band least squares harmonic analysis was performed at every voiced frame for all the signals in the database. The energy of the error between the original voiced frames and their synthetic harmonic counterparts was computed and divided by the frame length (equal to 2 periods at the local f_0). Finally, the errors at all voiced frames and signals were summed up together for each method under comparison.

Table I-A shows the increment of the global harmonic modeling error (negative values mean error reduction, thus better modeling) achieved by means of each refinement method with respect to the default configuration of the PDA (without any refinement). Although the constant-weighting approach reduces the error at many bands, it is worse than the weighted-by-amplitude approach within 0–2 kHz. It actually increases the modeling error in the first band. The lowest total error is achieved by the weighted-by-amplitude approach when the bandwidth of the QHM analysis coincides with the MVF yielded by the method in Section IV. Constant 4 kHz bandwidth gave also satisfactory results. Lower bandwidths of 1 or 2 kHz imply notably lower computational effort but are more prone to inaccuracies. Taking all this into consideration, we applied a weighted-by-amplitude approach with MVF-dependent bandwidth in the remaining experiments. It is worth mentioning that the results in Table I-A were found to be consistent for female and male voices.

Table I-B shows that the final configuration of the pitch refinement method can also reduce the error of full-band stationary QHM analysis, i.e. least squares based sinusoidal analysis at quasi-harmonic frequencies $f_i = if_0 + \Delta fi$ with Δfi given by (2). The overall superiority of QHM over harmonic analysis was already reported in [22].

TABLE I
A RELATIVE HARMONIC MODELING ERROR FOR DIFFERENT PITCH REFINEMENT CONFIGURATIONS. B RELATIVE ERROR OF QUASI-HARMONIC MODELING WITH RESPECT TO HARMONIC MODELING, WITH AND WITHOUT PITCH REFINEMENT

(a)

Weighting, Band	0-1kHz	1-2kHz	2-4kHz	4-8kHz	Total
Constant, MVF	2.1%	-14.2%	-28.0%	-14.4%	-4.5%
By ampl., MVF	-8.1%	-15.5%	-21.5%	-8.5%	-10.9%
By ampl., 4kHz	-6.7%	-16.1%	-23.5%	-12.6%	-10.4%
By ampl., 2kHz	-10.1%	-15.1%	-3.2%	4.2%	-9.8%
By ampl., 1kHz	-14.8%	2.1%	17.5%	12.7%	-7.7%

(b)

Weighting, Band	0-1kHz	1-2kHz	2-4kHz	4-8kHz	Total
No refinement	-45.8%	-59.1%	-63.2%	-55.5%	-50.3%
By ampl., MVF	-47.1%	-61.9%	-66.3%	-57.9%	-52.0%

B. Spectral Envelope Analysis

As spectral estimation is quite straightforward in unvoiced frames, this subsection is focused exclusively on voiced frames. As detailed in Section III-B, the proposed spectral envelope analyzer consists of harmonic analysis followed by interpolation and cepstral fitting. The accuracy of the spectral envelope therefore depends basically on three factors: the way of interpolating between harmonic log-amplitudes, the inherent loss of the cepstral parameterization, and the accuracy of the underlying harmonic analysis itself. In order to compare different configurations of the spectral envelope analyzer, an objective perceptual measure has been used: PESQ (ITU-T/P.862) [35]. Given a natural speech signal and a transcoded version of the same signal, PESQ predicts the mean opinion score (MOS) that the latter would achieve in a subjective listening test. Since PESQ is waveform-sensitive, we generated the transcoded signals as follows. First, harmonic analysis was performed on the voiced signal segments (the unvoiced segments were simply copied from the original signals). Second, MCEP coefficients were estimated from the amplitudes using any of the methods under study. Third, the amplitudes were replaced by those that resulted from resampling the MCEP envelope using expression (19). Finally, harmonic reconstruction was applied to regenerate the voiced segments using the original phases, thus preserving the waveform for PESQ measurements.

Fig. 8 shows the PESQ-MOS scores for three different ways of estimating the spectral envelope from harmonic log-amplitudes: linear interpolation (reference method), sinc-interpolation (6)–(8), and Mel-RDC (9). Two different cepstral orders are considered: 24, the one typically used in voice conversion applications, and 39, the one typically used in synthesis (both for $f_s = 16$ kHz). The results show the superiority of Mel-RDC regardless of the cepstral order. It can be remarked, however, that sinc-based interpolation yields practically identical results for order 39 while not involving any matrix inversion apart from harmonic analysis itself. In practice, the quality loss due to minimum phases [36] makes the effect of this gap absolutely irrelevant. We checked these results to be consistent for voices from

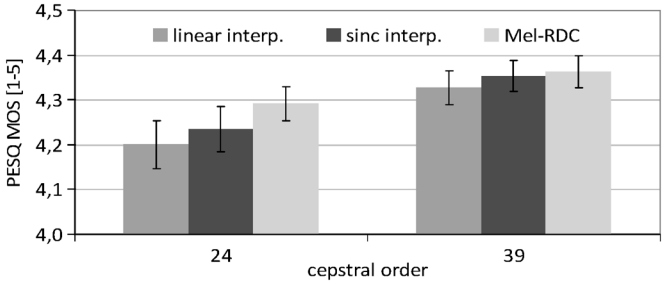

Fig. 8. Comparison of different spectral estimation methods using average PESQ-MOSs at 95% confidence intervals.

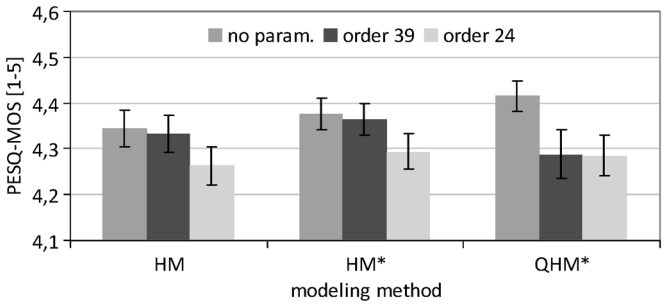

Fig. 9. Comparison of different harmonic modeling methods followed by Mel-RDC using average PESQ-MOSs at 95% confidence intervals. Symbol $*$ means that f_0 refinement is enabled.

both genders. Only two irrelevant differences were observed: the global PESQ-MOSs were slightly higher for male voices than for female voices, and the gap between order 24 and order 39 was smaller for female voices than for male voices.

In relation to accuracy of the underlying harmonic analysis, Table I-B suggests that the spectral envelope analysis could be improved if QHM analysis was applied before Mel-RDC instead of harmonic analysis. We performed PESQ measurements to validate this hypothesis. Interestingly, Fig. 9 shows that, although full-band QHM outperforms harmonic modeling when no cepstral fitting is applied, this is no longer true after MCEP parameterization. Its nonuniform frequency spacing and the corresponding amplitudes seem to mislead the Mel-RDC envelope analysis procedure, which is a valuable observation. Once again, we checked that this happens for both female and male voices, being more visible for female voices. On the other hand, the PESQ measurements in Fig. 9 confirm the positive effect of applying QHM-based f_0 refinement before harmonic analysis.

In summary, our final choice is QHM-based pitch refinement with MVF-dependent bandwidth followed by full-band harmonic analysis and Mel-RDC, whereas sinc-interpolation based 39th-order MCEP analysis can also be recommended because of its computational advantages. Although the differences between methods reported in Table I, Fig. 8 and Fig. 9 are hardly audible according to the numerous informal perceptual tests conducted throughout this work, inaccurate pitch and/or spectral estimation may result in slight inconsistencies between different instances of phonetically equivalent sounds, which means worse statistical modeling as we will show later.

C. Maximum Voiced Frequency

Since there is no database labeled at MVF level and we found PESQ scores to be uncorrelated with subjective perception for manual MVF variations, subjective tests are the most appropriate way to determine whether the explicit MVF analysis procedure in Section IV is advantageous or not. Preliminary informal listening tests involving resynthesized signals reveal that the differences between c_0-based MVF prediction (10) and explicit MVF analysis are very subtle, but both of them were found to reduce the buzziness produced by a constant MVF. This is partially related to phase: in standard HNM [2], the measured harmonic phases and their interframe variation convey some degree of information about the harmonicity at different bands even when constant MVF is used; in the proposed framework, however, phase information is discarded after the analysis (minimum phase is used during reconstruction) and an appropriate handling of MVF becomes more crucial.

We conducted a preference test to compare MVF prediction (10) with explicit MVF analysis. Eight expert evaluators listened to original utterances and their two vocoded counterparts and were asked about their preference, if any. Each evaluator rated 15 different voices randomly chosen among the 53 available ones. The results showed 38% preference for an explicit analysis of MVF, while prediction was preferred 17% of the times. These results were found to be consistent for different evaluators and voices, and no particular gender dependencies were observed.

VII. Evaluation in Synthesis

This section evaluates the proposed vocoder in the context of statistical parametric speech synthesis. For this purpose, an HMM-based synthesizer was built by combining HTS 2.1 (the well known open-source HMM-based speech synthesis system [6]) with the linguistic analyzer of AhoTTS [37]. HTS models the acoustic features provided by the vocoder by means of context-dependent 5-state left-to-right HSMMs [38]. Details about the context labels supplied by the linguistic analyzer can be found in [39].

As usual, log f_0 was modeled by means of multi-space distributions (MSD) [40] to deal with its discontinuous nature, while continuous HSMMs were applied to model the MCEP vectors. The inclusion of excitation-related features such as MVF implies adding one more stream to the learning process. This is a delicate point. Although the use of a MVF makes sense only in voiced segments, MSD-HSMM modeling can lead to undesired interactions with log f_0 during synthesis. For instance, very low MVF combined with a higher f_0 results into an unvoiced frame. Joint log f_0 and MVF modeling is not an option because their behavior over time is uncorrelated. If continuous HSMMs are used and null MVF is assumed during unvoiced segments, the smoothness of the parameter generation algorithm [41] can produce too low MVF values near the voiced-unvoiced transitions during synthesis. Taking all this into account, we finally decided to establish a minimum MVF value of 1 kHz during both analysis and synthesis, even in unvoiced segments, and then we used continuous HSMMs for modeling. We believe that a spectrum-to-excitation mapping applied only in voiced segments [42] might be adequate as well.

Two emotionally neutral databases were used to build the voices tested in this evaluation. The first one contained 2 k short sentences (>2 hours of speech) spoken by a female speaker in

TABLE II
AVERAGE LOG-LIKELIHOOD PER FRAME FOR DIFFERENT VOCODER CONFIGURATIONS

Method \ Voice	Female	Male
Sinc interp. + MCEP	$1.0095{\cdot}10^2$	$1.1034{\cdot}10^2$
Mel-RDC	$1.0339{\cdot}10^2$	$1.1176{\cdot}10^2$
f_0 ref. + Mel-RDC	$1.0446{\cdot}10^2$	$1.1519{\cdot}10^2$

TABLE III
SUBJECTIVE RELATIVE PREFERENCE FOR TWO MVF-RELATED STRATEGIES

Method \ Voice	Female	Male
MVF anal. & model.	30%	29%
No preference	53%	49%
MVF prediction	17%	22%

standard Basque; the second one contained 1.2 k sentences (2 hours) uttered by a native male speaker in Spanish. The sampling frequency was always 16 kHz and the order of the MCEP representation was set to 39.

A. Implications of Statistical Modeling

In the previous section, different configurations of the vocoder have been compared from an analysis/reconstruction point of view. Nevertheless, conclusions about the relative resynthesis performance of different configurations of the vocoder cannot be directly extrapolated to synthesis when statistical modeling is involved. This subsection aims at validating these conclusions from the synthesis side before proceeding to the evaluation of the vocoder.

First, the intermediate objective scores yielded by HTS during the training process can be used to verify the effectiveness of pitch refinement and spectral envelope analysis methods when statistical modeling is involved. At the beginning of training, HTS obtains one context-independent (CI) HMM for each phone in the database. Given this set of phone-specific CI-HMMs, the average log-likelihood per frame can be used as an indicator of how consistent the spectral estimation is when multiple acoustic realizations of phonetically equivalent sounds are available. Considering only the log f_0 and MCEP streams, we compared the average log-likelihood per frame that resulted from three different configurations of the vocoder: sinc-interpolation between harmonic log-amplitudes followed by MCEP analysis, Mel-RDC, and MVF-dependent pitch refinement followed by Mel-RDC. The results shown in Table II confirm the trends seen in resynthesis: Mel-RDC is slightly superior to (i.e. it shows higher average log-likelihood per frame than) interpolation-based techniques and pitch refinement enhances its performance even more. Informal listening tests carried out after synthesis with many training databases revealed that the differences between pitch and spectral envelope estimation configurations are audible only for some voices.

Given the aforementioned difficulties of MVF modeling and the lack of an appropriate objective measure, a subjective preference test was carried out to validate the usefulness of the MVF stream in synthesis. In this experiment, 45 listeners (including 6 experts) compared 12 synthetic utterance pairs each and chose the one they preferred, if any. The results shown in Table III indicate that explicit MVF analysis and modeling ($\sim$30% preference) is slightly better than MVF prediction ($\sim$20% preference) also in synthesis, though the differences are not easily perceived by listeners. Indeed, they are less audible in the male voice because it has lower quality as a result of the lower number of training sentences. Two conclusions can be derived from here. On the one hand, explicit MVF modeling results in improved speech quality during synthesis, which justifies the inclusion of the method described in Section IV. On the other hand, the prediction approach may be advantageous in practical applications with no headphones involved, because the resulting two-stream system has lower footprint and complexity than the one considering MVF as a third stream.

Before proceeding to the evaluation, one more experiment was conducted to analyze the perceptual impact of statistical modeling on segmental and suprasegmental speech parameters. Once having trained statistical models for the two voices, we randomly selected 10 sentences per voice from the training database and we regenerated them by three different methods: resynthesis, synthesis of Mel-cepstrum and MVF combined with original durations and f_0 contour, and full synthesis. In order to figure out to which extent the signal "survives" after parameterization and statistical modeling, 15 listeners (including 6 experts) were asked to rate the naturalness of 24 synthetic "utterances each in comparison with the original recorded signal using a 5-point scale: $1 =$ "unnatural", $\ldots, 5 =$ "natural." Fig. 10 shows the results of the test in terms of MOS and 95% confidence interval. We can observe that the MOS achieved via resynthesis is slightly lower than 4. In other words, the effect of full parameterization is clearly perceived by listeners. When Mel-cepstral vectors and MVF are replaced by their synthetic counterparts (combined), some extra loss of naturalness is observed. This behavior is logical given the limitations of the statistical learning process. A similar extra loss was also expected when replacing the natural durations and f_0 by synthetic ones. In our case, however, this happened only for the male voice. We believe that the consistent neutral prosodic patterns used by the female voice while recording the database, the higher number of training utterances and the possible inaccuracies during the combination of resynthesized and synthetic streams justify the observed scores. In brief, both the parameterization and the statistical learning process imply a certain loss of naturalness, although the latter can be minimized if the database is well recorded and sufficiently large.

B. Comparison With STRAIGHT

Earlier versions of the proposed vocoder [23], [24] clearly outperformed the basic one included in HTS [7], namely MLSA filter and pulse/noise excitation. In this experiment, the best configuration of our vocoder is compared with the STRAIGHT-based vocoder included in HTS 2.1 [12]. The speech representation handled by the original STRAIGHT toolkit [13] consists of three streams: fundamental frequency, spectral envelope sampled at high-resolution, and degree of aperiodicity (it can be seen as the harmonics to noise ratio) at each frequency bin. Its HTS-compatible version [12] transforms the spectral envelopes into MCEP coefficients and takes the average aperiodicity of 5

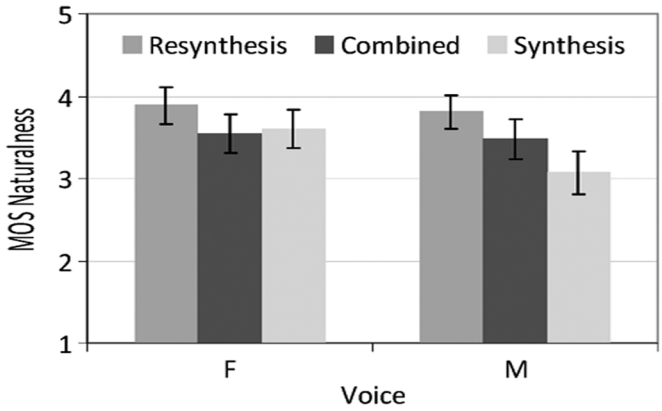

Fig. 10. Subjective naturalness of resynthesized speech, resynthesized prosody combined with synthetic MCEP and MVF, and fully synthetic speech. MOS at 95% confidence intervals. M: male voice. F: female voice.

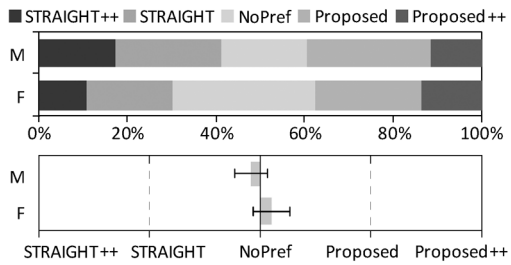

Fig. 11. STRAIGHT-based vs. proposed vocoder: score distributions and CMOSs at 95% confidence intervals. M: male voice. F: female voice. We denote strong preference by symbol ++.

bands: 0–1, 1–2, 2–4, 4–6 and 6–8 kHz. Despite the existence of some interesting alternative vocoders that could be used as baseline [18], this one is still considered the state-of-the-art by many researchers in the field. In this test, the number of participants was 30.

In our preliminary works [23], [24] we evaluated the system via MOS tests. At this point, a more discriminative test was conducted to make the differences between methods more visible. The evaluation had the form of a comparative MOS (CMOS) test. Given a number of randomly selected synthetic sentence pairs (the sentences in each pair, A and B, were also played in random order), several volunteer native evaluators were asked to listen to them using headphones and then rate their preference in a 5-point scale: "strong preference for A," "slight preference for A," "no preference," "slight preference for B," "strong preference for B." Recordings of the original natural voices were included as a reference. Each point in the scale was given an integer numeric value (-2 to 2), and the final CMOS was calculated by averaging the numeric values that correspond to the listeners' choices.

The score distributions in Fig. 11 show that the differences between methods are quite clearly perceived by listeners, but the average preference remains not clear. Despite the high number of listeners, the results are not significant enough to draw sententious conclusions. The no-preference point lies inside the 95% confidence intervals for both voices. Nevertheless, it is worth mentioning that the average scores achieved by the proposed vocoder tend to be higher than those of STRAIGHT for the female voice (for which the positive impact of modeling MVF explicitly was more noticeable according to Table III). Therefore, as a final conclusion, we can assert that the HNM-based vocoder presented in this article is an interesting alternative to the well known STRAIGHT-based vocoder, at least for some voices.

VIII. Conclusion

In this research work we have studied the potential of the harmonics-plus-noise model in parametric speech vocoding tasks. Three aspects of the analysis have been investigated: pitch refinement, spectral envelope analysis and maximum voiced frequency estimation.

We have shown that the quasi-harmonic analysis model can be used to implement a pitch refinement algorithm which improves the accuracy of the subsequent spectral estimation.

While using a harmonic plus noise model to reconstruct the speech signals from parameters, we have discussed the convenience of using a full-band purely harmonic analysis in combination with cepstral fitting techniques to estimate the spectral envelope, thus avoiding the discontinuities arising when the spectral information is split into several streams. Remarkably, harmonic analysis yields more accurate cepstral envelopes than stationary sinusoidal analysis at quasi-harmonic frequencies.

We have presented a new method to estimate the maximum voiced frequency from speech signals and we have shown the advantages of an adequate explicit modeling of this parameter.

The result of this work, which we refer to as Ahocoder, is comparable with top state-of-the-art vocoders when integrated into an HMM-based speech synthesizer. It can be used in speech manipulation and voice conversion applications as well. Future works will aim at incorporating phase information into the analysis and modeling process.

Acknowledgment

The authors thank the HTS working group for making the code of HTS publicly available and Prof. Yannis Stylianou for his useful suggestions during his stay at UPV/EHU.

References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.
[2] Y. Stylianou, "Harmonic plus noise models for speech, Combined with statistical methods, for speech and speaker modification," Ph.D. dissertations, École Nationale Supèrieure de Télécommunications, Paris, France, 1996.
[3] A. Kain, "High resolution voice transformation," Ph.D. dissertation, OGI School of Sci. and Eng. at OHSU, Portland , OR, 2001.
[4] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.
[5] J. L. Flanagan, "Parametric representation of speech signals," *IEEE Signal Process. Mag.*, vol. 27, no. 3, pp. 141–145, 2010.
[6] HMM-Based Speech Synthesis System (HTS), [Online]. Available: http://hts.sp.nitech.ac.jp/
[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis," in *Proc. Eurospeech*, 1999, pp. 2347–2350.
[8] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis—A unified approach to speech spectral estimation," *Proc. Int. Conf. Spoken Lang. Process.*, vol. 3, pp. 1043–1046, 1994.
[9] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. ICASSP*, 1983, pp. 93–96.

[10] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Mixed excitation for HMM-based speech synthesis," in *Proc. Eurospeech*, 2001, pp. 2263–2266.
[11] X. Gonzalvo, J. C. Socoro, I. Iriondo, C. Monzo, and E. Martinez, "Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish," in *Proc. 6th ISCA Speech Synth. Workshop*, 2007, pp. 362–367.
[12] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 1, pp. 325–333, 2007.
[13] H. Kawahara, I. Masuda-Kasuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, pp. 187–207, 1999.
[14] R. Maia, T. Toda, H. Zen, Y. Nankaku, and K. Tokuda, "An excitation model for HMM-based speech synthesis based on residual modeling," in *Proc. 6th ISCA Speech Synth. Workshop*, 2007, pp. 131–136.
[15] T. Drugman, G. Wilfart, and T. Dutoit, "A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis," in *Proc. Interspeech*, 2009, pp. 1779–1782.
[16] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Glottal spectral separation for parametric speech synthesis," in *Proc. Interspeech*, 2008, pp. 1829–1832.
[17] P. Lanchantin, G. Degottex, and X. Rodet, "A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method," in *Proc. ICASSP*, 2010, pp. 4630–4633.
[18] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMm-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 153–165, 2011.
[19] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *Proc. ICASSP*, 1993, pp. 550–553.
[20] E. Banos, D. Erro, A. Bonafonte, and A. Moreno, "Flexible harmonic/stochastic modeling for HMM-based speech synthesis," in *Proc. V Jornadas en Tecnologías del Habla*, 2008, pp. 145–148.
[21] C. Hemptinne, "Integration of the harmonic plus noise model into the hidden Markov model-based speech synthesis system," M.S. thesis, IDIAP Research Inst., Martigny , Switzerland, 2006.
[22] Y. Pantazis, O. Rosec, and Y. Stylianou, "Iterative estimation of sinusoidal signal parameters," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 461–464, 2010.
[23] D. Erro, I. Sainz, I. Saratxaga, E. Navas, and I. Hernaez, "MFCC+F0 extraction and waveform reconstruction using HNM: Preliminary results in an HMM-based synthesizer," in *Proc. FALA*, 2010, pp. 29–32.
[24] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "HNM-based MFCC+F0 extractor applied to statistical speech synthesis," in *Proc. ICASSP*, 2011, pp. 4728–4731.
[25] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Improved HNM-based vocoder for statistical synthesizers," in *Proc. InterSpeech*, 2011, pp. 1809–1812.
[26] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in *Proc. Inst. Phon. Sci., Univ. of Amsterdam*, 1993, vol. 17, pp. 97–110.
[27] G. Degottex and Y. Stylianou, "A full-band adaptive harmonic representation of speech," in *Proc. Interspeech*, 2012.
[28] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, 1986.
[29] D. Erro, A. Moreno, and A. Bonafonte, "Flexible harmonic/stochastic speech synthesis," in *Proc. 6th ISCA Speech Synth. Workshop*, 2007, pp. 194–199.
[30] O. Cappé, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *Proc. IEEE Workshop Apps. Signal Process Audio Acoust.*, 1995, pp. 213–216.
[31] S. Shechtman and A. Sorin, "Sinusoidal model parameterization for HMM-based TTS system," in *Proc. Interspeech*, 2010, pp. 805–808.
[32] X. Rodet, "Musical sound signals analysis/synthesis: Sinusoidal + residual and elementary waveform models," *Appl. Signal Process.*, vol. 4, pp. 131–141, 1997.
[33] S. Ahmadi and A. S. Spanias, "Low bit-rate speech coding based on an improved sinusoidal model," *Speech Commun.*, vol. 34, pp. 369–390, 2001.
[34] X. Sun, F. Plante, B. M. G. Cheetham, and K. W. T. Wong, "Phase modelling of speech excitation for low bit-rate sinusoidal transform coding," in *Proc. ICASSP*, 1997, vol. 3, pp. 1691–1694.
[35] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001, pp. 749–752.
[36] I. Saratxaga, I. Hernaez, M. Pucher, E. Navas, and I. Sainz, "Perceptual importance of the phase related information in speech," in *Proc. Interspeech*, 2012.
[37] I. Sainz, D. Erro, E. Navas, I. Hernaez, J. Sanchez, I. Saratxaga, I. Odriozola, and I. Luengo, "Aholab Speech Synthesizers for Albayzin 2010," in *Proc. FALA*, 2010, pp. 343–347.
[38] H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis," in *Proc. ICSLP*, 2004, vol. II, pp. 1397–1400.
[39] D. Erro, I. Sainz, I. Luengo, I. Odriozola, J. Sanchez, I. Saratxaga, E. Navas, and I. Hernaez, "HMM-based speech synthesis in Basque language using HTS," in *Proc. FALA*, 2010, pp. 67–70.
[40] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. ICASSP*, 1999, vol. 1, pp. 229–232.
[41] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Trans. Inf. Syst.*, vol. E90-D, no. 5, pp. 816–824, 2007.
[42] H. Silen, E. Helander, and M. Gabbouj, "Prediction of voice aperiodicity based on spectral representations in HMM speech synthesis," in *Proc. Interspeech*, 2011, pp. 105–108.

Daniel Erro received his telecommunication engineering degree from UPNA in 2003 and his Ph.D. degree from UPC in 2008. Currently he is an Ikerbasque Research Fellow at Aholab, UPV/EHU.

Iñaki Sainz received his telecommunication engineering degree from UPV/EHU in 2005. Currently he is a research engineer at Aholab, UPV/EHU.

Eva Navas received her telecommunication engineering degree and her Ph.D. degree from UPV/EHU in 1996 and 2003, respectively. She is a tenured lecturer at UPV/EHU and a researcher at Aholab.

Inma Hernaez received her telecommunication engineering degree from UPC in 1987 and her Ph.D. degree from the UPV/EHU in 1995. She is a Full Professor at UPV/EHU. She is the founding member of Aholab.