Improved HNM-based Vocoder for Statistical Synthesizers

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

AHOLAB Signal Processing Laboratory
University of the Basque Country (UPV/EHU), Bilbao, Spain
{derro,inaki,eva,inma}@aholab.ehu.es

Abstract

Statistical parametric synthesizers have achieved very good performance scores during the last years. Nevertheless, as they require the use of vocoders to parameterize speech (during training) and to reconstruct waveforms (during synthesis), the speech generated from statistical models lacks some degree of naturalness. In previous works we explored the usefulness of the harmonics plus noise model in the design of a high-quality speech vocoder. Quite promising results were achieved when this vocoder was integrated into a synthesizer. In this paper, we describe some recent improvements related to the excitation parameters, particularly the so called maximum voiced frequency. Its estimation and explicit modelling leads to an even better synthesis performance as confirmed by subjective comparisons with other well-known methods.

Index Terms: vocoder, statistical parametric speech synthesis, harmonics plus noise model, speech parameterization

1. Introduction

During the last years, statistical parametric speech synthesis [1] has gained ground over other synthesis technologies such as unit selection [2], which had been the most popular for years. Statistical synthesizers model the acoustic characteristics and duration of the phonemes through context-dependent (CD) hidden Markov models (HMMs). During synthesis, given the phonetic and linguistic context descriptors of the sentence to be generated, they build a sentence-level HMM by concatenating appropriate phoneme-level CD-HMMs. Then, they reconstruct the speech signal from the acoustic vector sequence that shows maximum likelihood with respect to the sentence-HMM. Among the advantages of such a statistical framework, one can mention the following: 1) the low footprint of the resulting synthesizers, which is adequate for small devices; 2) their enormous flexibility: as speech is generated from models, the systems benefit from any technique that can modify these models (adaptation, interpolation, etc.) and therefore the acoustic characteristics of the voice, the speaking style or its emotional content; 3) it is relatively easy to adapt the system to new languages; 4) the speech generated by statistical synthesizers is quite smooth, in contrast to that generated through unit selection, which often shows annoying concatenation artifacts.

This paper is devoted to one of the main challenges concerning statistical parametric speech synthesis: the design of a high-quality vocoder. In this context, vocoders are used to translate the speech signals in the training corpus into the vectors from which the models are learnt, and also to reconstruct speech from the parameter vector sequences generated by the system. The performance of the whole synthesizer is strongly dependent on the quality of the underlying vocoder. In fact, this is one of the main reasons why the best examples of unit-selection synthesis are often preferred by listeners rather than the best examples of statistical synthesis [3][4].

In general, the speech signals are parameterized using two or three different vector streams: one for f0, one for the spectrum, and optionally one for any information related to the glottal source or excitation. Regarding the spectrum, Mel-frequency cepstral coefficients (MFCCs) and line spectral pairs (LSPs) are widely used. The way these coefficients are obtained varies depending on the system, although most of them use the Straight spectrum [5] and/or the Mel-generalized cepstral analysis [6]. Regarding the parameterization of the excitation component, many different proposals have been made during the last years: mixed excitation considering voicing strengths [7][8] or aperiodicity measures and phase manipulation [9], state-dependent filters for pulses and noise [10], deterministic plus stochastic model of the residuals [11], glottal source modelling [12], etc.

In [13], a vocoder based on the harmonics plus noise model (HNM) was presented. Taking benefit from the advantages of HNM [14], it succeeded at parameterizing speech using two streams, f_0 and spectrum, and achieved highly satisfactory performance scores in synthesis although it did not consider any parameter related to the excitation. In this paper we show that its performance can be improved by adding one more single parameter: the maximum voiced frequency (MVF), which is defined in the HNM framework as the frequency that splits the spectrum into a harmonic/voiced lower band and a noisy/unvoiced upper band. To some extent, this is equivalent to assuming a two-band excitation model [11][15][17], although in this case the parameterization and reconstruction procedures that deal with the MVF involve not only the excitation but the whole signal. Constant MVF was used in preliminary versions of the vocoder [15]. In a later version, we found that the perceived synthesis quality improved when the MVF was made somehow dependent on the energy of the signal [13]. This strategy alleviated the apearance of some metallic artifacts. As the MVF could be predicted from the energy, the system still needed only two parameter streams, and thus could be made compatible with other analyzers. However, the system was never compared with a 3-stream system including the MVF parameter. This paper addresses this subsequent step: it explores the usefulness of MVF estimation and explicit modelling in a speech synthesis context. The results presented in this paper confirm the HNM-based vocoder to be a good alternative to other state-of-the-art high-quality vocoders.

Section 2 contains an overview of existing MVF estimation techniques, and describes the one which was integrated into our HNM-based vocoder. Section 3 gives some details about the mentioned vocoder. The results of a perceptual evaluation of the system are shown and discussed in section 4. Finally, the conclusions are summarized in section 5.

2. MVF Estimation

HNM [13] assumes that speech signals are the sum of a harmonic component, which results from the vocal fold vibration and can be considered periodic in short-time frames, and a noise-like component, which contains the remaining parts of the signal. In voiced segments, the MVF splits the spectrum into a lower harmonic band and an upper noisy band. In practice, setting a constant MVF value around 4 kHz is quite reasonable [14][11]. The preliminary version of our vocoder followed this approach [15]. In [13], the time-varying MVF used during speech reconstruction was estimated from the 0th cepstral coefficient (the one carrying the energy) through a simple mapping. This strategy avoided the appearance of metallic artifacts near low-energy segments, particularly in sentence endings. Nevertheless, even though the two-band-excitation model is quite simplistic, handling the MVF in an independent stream would provide a more realistic parameterization of the signal. Apart from that, it would help to reduce the acoustic buzziness of the reconstructed signals around f_0 detection errors, which occur mainly at the beginnings and endings of voiced segments. As this type of errors often lead to low MVF estimated values, the involved frames would be treated as almost-unvoiced, which would make f_0 errors less audible.

Several MVF estimation methods have been proposed until present. In general, the common basic idea is measuring the degree of harmonicity of the peaks in the short-time spectrum. To estimate it, some methods calculate the spectral distortion between the spectral peaks and those that would have been produced by a sinusoid [18][19]. Some others take into account the relative amplitude of the peaks with respect to their adjacent valleys [13][20]. We can also mention the autocorrelation-based method in [15], although we found it to be too sensitive to the f_0 detection accuracy and also to the position of the formants.

Our proposal is based on a sinusoidal likeness measure (SLM) used to classify spectral peaks in music analysis [21]. First, the N-point complex spectrum $S[k]$ is computed on the current frame using a 3-period-length Hanning window (the pitch must be known in advance). N is the first power of two greater than $4L$, where 4 is the zero-padding factor and L is the frame length. The frequencies of the magnitude spectrum peaks, $\{f_i\}$, are determined through parabolic fitting around the maxima, and their SLM is calculated through local normalized cross-correlation [21]:

$$L_i = \frac{\left|\sum S[k]\cdot W_i^*[k]\right|}{\sqrt{\sum|S[k]|^2\cdot\sum|W_i[k]|^2}} \quad \forall k, \left|k\tfrac{f_s}{N} - f_i\right| < \tfrac{f_0}{2} \qquad (1)$$

where W_i is the Fourier transform of the analysis window multiplied by a cosine function at f_i frequency, operator * denotes complex conjugation, f_0 is the local pitch, and f_s is the sampling rate. W_i can be efficiently approximated using analytical expressions. The SLM ranges from 0 to 1. Values close to 1 indicate a pure sinusoid, and smaller values may indicate the presence of noise or sinusoids showing significant time-variation inside the analysis frame. Once the SLM has been calculated for all the peaks in the analysis band, the error of assuming the MVF to be placed at each of these peaks is computed as

$$\varepsilon_i^2 = \frac{1}{I}\left(\sum_{j<i}\left(1-L_j\right)^2 + \sum_{j\geq i}\left(\max\{L_j,\lambda\}-\lambda\right)^2\right) \qquad (2)$$

where I is the number of spectral peaks and λ can be understood as the voicing threshold. In our implementation, it was empirically set to 0.85. Such an error measure can be interpreted as the distance between the actual SLM contour and the one given by an ideal two-band signal whose MVF is equal to the frequency of the current peak. The frequencies showing relative minima of this error funcion (usually there is more than one) are taken as MVF candidates. The final decision is made after a Viterbi search of the MVF trajectory (over time) $\{f_{t,i(t)}\}$ that minimizes the following cost function:

$$C\left(\{f_{t,i(t)}\}_{t=1}^{T}\right) = \sum_{t=1}^{T}\varepsilon_{t,i(t)}^2 + \gamma\sum_{t=2}^{T}\left(f_{t,i(t)} - f_{t-1,i(t-1)}\right)^2 \qquad (3)$$

where t denotes the time instant, $i(t)$ is the index of the peak under consideration at time t, and γ stands for the relative weight of the second term of (3) (in our experiments, good results were obtained for $\gamma = 5\cdot10^{-4}\cdot r/f_s^2$, being r the frame rate). In unvoiced frames, a single candidate at 0 Hz is considered.

The described method is characterized by two adjustable parameters: λ and γ. Since the MVF is not a physical speech characteristic but results from the assumption of a simplified speech model, there is no labelled database available. Therefore, the parameters were manually optimized by means of informal listening experiments. Regarding the behaviour of the method when the local harmonicity condition is not met, pitch variations inside the analysis window decrease the SLM, especially in low-pitched signals (due to the pitch-dependence of the window length). The choice of an adequate λ partially compensates this effect, though a more sophisticated solution should consider using a pitch-dependent λ value.

3. Vocoder Description

The improved version of the vocoder parameterizes the speech frames in three different streams: f_0, MVF and spectrum. The next paragraphs describe how these parameters are extracted from the signal frames and how speech signals are reconstructed from them. A more detailed description is available in [13].

Both f_0 and MVF are scalars: f_0 is given by any accurate pitch detection algorithm [22], while the method in section 2 yields the MVF values at the centre of the analysis frames. The spectrum is represented by p+1 cepstral coefficients (including the one related to the energy). Voiced and unvoiced frames are treated in a different way to extract their cepstral representation. If the input frame has been classified as voiced by the pitch detector, a harmonic analysis based on least squares optimization [14] is performed on the full analysis band to get the amplitudes of the harmonics at frequencies multiple of f_0. These amplitudes are assumed to be discrete samples of the actual spectral envelope even at high frequencies, where the harmonics-to-noise ratio is low. Unvoiced frames are analyzed through a simple fast Fourier transform (FFT), which can be viewed also as a harmonic analysis for f_0 equal to the FFT resolution. In order to homogenize the discrete representation of the spectrum, the envelope given by the harmonic amplitudes at voiced frames is normalized in amplitude and then resampled at the FFT resolution via interpolation [13]. During the last step of the analysis, cepstral coefficients are extracted from the amplitude spectra as follows. First, a traditional cepstrum is obtained as the inverse FFT of the log-amplitude spectrum. Then, the cepstrum is warped in frequency to match the Mel scale using the recursion described in [6].

The reconstruction of the signal is carried out via overlap-add (OLA) after generating the samples of the individual

frames from their corresponding parameters. Each frame is built using HNM synthesis procedures. First, the noise part of the frame (which is assumed to be present in both voiced and unvoiced segments) is generated in the frequency domain; the module of the noise spectrum results from sampling the cepstral envelope, and the phase is given random values. Unvoiced frames are given by the inverse FFT of the noise spectrum. In voiced frames, the noise spectrum is multiplied by the frequency response of a high-pass filter given by the MVF before computing the inverse FFT. Next, the harmonic component is generated in the time-domain. The harmonic amplitudes are obtained by sampling the cepstral envelope at multiples of f_0; the phases are obtained through a minimum phase approach, and the phase coherence between adjacent frames is ensured by adding a controlled linear-in-frequency phase term.

4. Evaluation

Due to the lack of a reference labelled database to evaluate the MVF estimation method, the whole 3-stream vocoder was evaluated through subjective listening experiments. After informally verifying its resynthesis capabilities, the described vocoder was evaluated in a speech synthesis context using HTS, the open-source software toolkit publicly released since 2002 by the so called HTS working group [23]. HTS 2.1.1 includes demo scripts for training speaker-dependent and speaker-adaptive systems. It provides two different vocoding methods: the basic one, based on mel-cepstral analysis and a simple pulse/noise excitation model, and the Straight-based method, which is known to have very good performance [9].

An HMM-based synthesizer was built by combining HTS with the linguistic analyzer of AhoTTS (the Aholab synthesizer) [24]. The synthetic speech was generated using three different vocoders: the improved HNM-based vocoder described in this paper (denoted as H++), the previous version of the same vocoder (denoted as H, which used no explicit model for MVF but considered some energy-dependence), and the one based on Straight. All of them were configured to use the same number of cepstral coefficients (39+energy), whereas the number of excitation parameters was different for each: no excitation in H, just one parameter in H++ (the MVF), and 5 band-aperiodicities in Straight.

Two databases were used to build the voices tested in this evaluation: the first one contained 2K short sentences (>2 hours of speech) spoken by a female speaker in standard Basque and the second one contained 1.2K sentences (2 hours) uttered by a native male speaker in Spanish. Both of them were emotionally neutral.

Two comparisons were made through comparative mean opinion score (CMOS) tests. The first one involved H and H++, and the second one involved Straight and H++. Given 12 randomly selected sentence pairs (the sentences in each pair where also displayed in random order), the listeners were asked to 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”. Some recordings of the original 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. In both tests, the numeric values were assigned in such manner that 2 points would indicate strong preference for H++. The number of participants in each test was 45 and 30, respectively (including 6 experts).

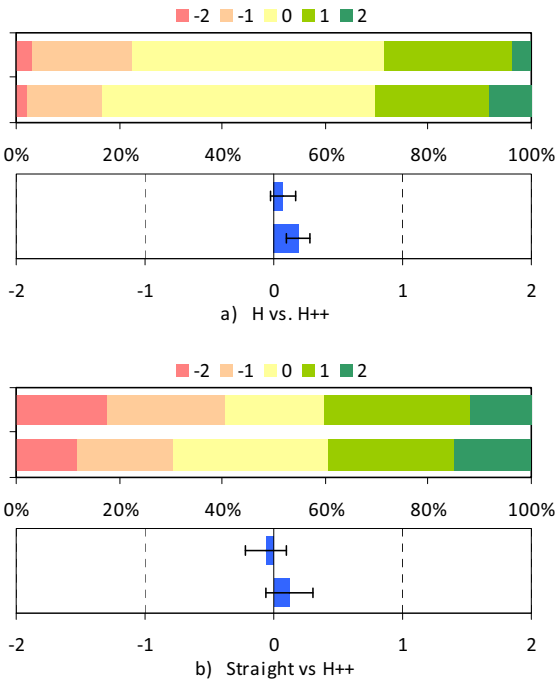

Figure 1: *Score distribution and CMOS for two different method pairs and voices. (a) H vs. H++; (b) Straight vs. H++. Bottom: Basque female voice; top: Spanish male voice.*

As it can be seen from Figure 1a, H++ is judged to be just slightly better than H (around 30% preference for H++ and 20% for H). However, the differences are not easily perceived by listeners. Two conclusions can be extracted from here. On the one hand, the explicit modelling of the MVF results in improved speech quality during synthesis, which justifies the inclusion of the MVF estimation method described in section 2 into the HNM-vocoder. On the other hand, this means that the approach followed in H (predicting the MVF from the energy during synthesis) would be enough for practical applications. The differences are less audible in the male voice, which can be partially due to the way the parameters of the MVF estimation method were adjusted (read the last paragraph of section 2 for details).

With regard to the second preference test in Figure 1b, the score distributions show that the differences between methods were quite clearly perceived by listeners, but the average preference remains not clear. Despite the number of listeners, the results are not significant enough to draw conclusions. The 95% confidence intervals include 0 (no preference) for both voices. Nevertheless, it is worth mentioning that the average scores achieved by the proposed method are higher than those of Straight for the female voice (for which the impact of modelling the MVF explicitly was more noticeable according to Figure 1a). This is coherent with the results reported in [13].

It can be concluded that the improved HNM vocoder presented in this paper is an interesting alternative to the well known Straight vocoder, al least for some voices. In the interest of the scientific community, we plan to make it publicly available in the near future.

5. Conclusions

Continuing our research into HNM-based vocoding, this research work explored the impact of estimating and modelling explicitly the maximum voiced frequency, which can be defined as the frequency that splits the spectrum into a harmonic band and a noisy band. The resulting system achieved slightly better scores than its predecessor. The improvements were more audible for one of the two voices under study, for which the proposed vocoder even outperformed the state-of-the-art Straight-based vocoder. Therefore, it can be taken into consideration during the development of high-quality synthesizers.

Future works on the MVF estimation method should consider some pitch-dependency of its threshold values. With regard to the vocoder itself, we are currently working to improve the efficiency of the speech analysis step, which will make it more adequate for online tasks such as voice cloning. The vocoder will be publicly available in the near future.

6. Acknowledgements

This work has been partially supported by UPV/EHU (Ayuda de Especialización de Doctores), the Spanish Ministry of Science and Innovation (Buceador Project, TEC2009-14094-C04-02) and the Basque Government (Berbatek, IE09-262). The authors would like to acknowledge the HTS Working Group for making HTS publicly available, and also the participants of the listening tests for their time.

7. References

[1] H. Zen, K. Tokuda, A. W. Black, “Statistical parametric speech synthesis”, Speech Communication, vol. 51(11), pp. 1039-1064, 2009.

[2] A. Hunt, A. Black, “Unit selection in a concatenative speech synthesis system using a large speech data-base”, Proc. ICASSP, pp. 373–376, 1996.

[3] [Online], http://festvox.org/blizzard

[4] F. Mendez, L. Docio, M. Arza, F. Campillo, “The Albayzin 2010 text–to–speech evaluation”, Proc. FALA, pp. 317-340, 2010.

[5] H. Kawahara, “Straight, exploration of the other aspect of Vocoder: perceptually isomorphic decomposition of speech sounds”, Acoustic Science and Technology, vol. 27(6), pp. 349-353, 2006.

[6] K. Tokuda, T. Kobayashi, T. Masuko, S. Imai, “Mel-generalized cepstral analysis - a unified approach to speech spectral estimation”, Proc. Int. Conf. Spoken Lang. Proc., vol. 3, pp. 1043-1046, 1994.

[7] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, “Mixed excitation for HMM-based speech synthesis”, Proc. Eurospeech, pp. 2263–2266, 2001.

[8] X. Gonzalvo, J. C. Socoro, I. Iriondo, C. Monzo, E. Martinez, “Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish”, Proc. 6th ISCA Speech Synthesis Workshop, pp. 362–367, 2007.

[9] H. Zen, T. Toda, M. Nakamura, K. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005”, IEICE Trans. Inf. Syst., E90-D(1), pp. 325–333, 2007.

[10] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, “An excitation model for HMM-based speech synthesis based on residual modeling”, Proc. 6th ISCA Speech Synthesis Workshop, pp. 131-136, 2007.

[11] T. Drugman, G. Wilfart, T. Dutoit, “A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis”, Proc. Inter-speech, pp. 1779-1782, 2009.

[12] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, P. Alku, “HMM-Based Speech Synthesis Utilizing Glottal Inverse Filtering”, IEEE Trans. Audio, Speech, & Lang. Proc., vol. 19(1), pp. 153-165, 2011.

[13] D. Erro, I. Sainz, E. Navas, I. Hernaez, “HNM-based MFCC+F0 extractor applied to statistical speech syn-thesis”, Proc. ICASSP, 2011.

[14] Y. Stylianou, “Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification”, PhD thesis, École Nationale Supèrieure de Télécommunications, Paris, 1996.

[15] D. Erro, I. Sainz, I. Saratxaga, E. Navas, I. Hernaez, “MFCC+F0 extraction and waveform reconstruction using HNM: preliminary results in an HMM-based synthesizer”, Proc. FALA, pp. 29-32, 2010.

[16] S.J. Kim, J.J. Kim, M. Hahn, “HMM-based Korean speech synthesis system for hand-held devices”, IEEE Trans. Consumer Electronics, vol. 52(4), pp. 1384-1390, 2006.

[17] H. Silen, E. Helander, J. Nurminen, M. Gabbouj, “Parameterization of vocal fry in HMM-based speech synthesis”, Proc. Interspeech, pp. 1775-1778, 2009.

[18] D.W. Griffin, J.S. Lim, “Multiband Excitation Vocoder”, IEEE Trans. Acoust., Speech & Sig. Proc., vol. 36(8), pp. 1223-1235, 1988.

[19] R. McAulay and T. Quatieri, “Sinusoidal Coding”, chapter in Speech Coding and Synthesis, Elsevier, pp. 121-173, 1995.

[20] K. Hermus, H. van Hamme, S. Irhimeh, “Estimation of the voicing cut-off frequency contour based on a cumulative harmonicity score”, IEEE. Sig. Proc. Letters, vol. 14(11), pp. 820-823, 2007.

[21] X. Rodet, “Musical Sound Signals Analysis/Synthesis: Sinusoidal+Residual and Elementary Waveform Models”, Applied Sig. Proc., vol. 4, pp. 131-141, 1997.

[22] I. Luengo, I. Saratxaga, E. Navas, I. Hernaez, J. Sanchez, I. Sainz, “Evaluation of pitch detection algorithms under real conditions”, Proc. ICASSP, pp. 1057-1060, 2007.

[23] [Online], “HMM-based Speech Synthesis System (HTS)”, http://hts.sp.nitech.ac.jp/

[24] D. Erro et al., “HMM-based speech synthesis in Basque language using HTS”, Proc. FALA, 2010.