Accurate estimation of sinusoidal parameters in an Harmonic+Noise Model for speech synthesis

Gérard Bailly
Institut de la Communication Parlée
46, av. Félix Viallet 38031 Grenoble CEDEX - FRANCE
bailly@icp.inpg.fr - http:://www.icp.inpg.fr/~bailly

Abstract

We present here an Harmonic+Noise Model (HNM) for speech synthesis. The noise part is represented by an autoregressive model whose output is pitch-synchronously modulated in energy. The harmonic part of the signal is represented by a sinusoidal model. This paper compares different methods for separating these two components. We then propose a method for the estimation of the sinusoidal parameters derived from the ABS model [8] and evaluate different models for the analysis/synthesis of the stochastic part proposed in the literature.

INTRODUCTION

Most text-to-speech systems (TTS) use nowadays concatenative synthesis where segments of natural speech are distorted by analysis-synthesis techniques in such a way that the resulting synthetic signal conforms to desired prosodic characteristics. Classically these characteristics include melody, sounds energy and duration. Modifications of these characteristics by Speech Synthesis Coders (SSC) are often accompanied by distortions in other temporal/spectral dimensions that do not necessarily reflect covariations observed in natural speech. Contrary to to synthesis-by-rule systems where such observed covariations may be described and implemented [9], SSC should intrinsically exhibit properties that guaranty an optimal extrapolation of temporal/spectral behaviour given only a reference sample. One of these such desired properties is shape invariance in the time domain [11, 13]. Shape invariance consists in maintaining the signal shape in the vicinity of vocal tract excitation (pitch markers). OLA techniques [4, 6] are such techniques that preserve the signal shape by centering short-term signals on pitch markers.

In 1986 McAulay and Quatieri [11] first proposed a new method for representing the speech signal $s(t)$ that preserves shape-invariance. $s(t)$ equals a sum of sinusoids $s(t) = \sum_1^{L(t)} A_l(t)exp(j\psi_l(t))$ where $A_l(t)$ and $\psi_l(t)$ are the amplitude and phase of the lth sine wave along the frequency track $\omega_l(t)$. Several major contributions have improved this initial proposal: Serra [16] introduced the necessary separation of deterministic/noise components, where the noise was considered a posteriori as the deterministic residual; Stylianou et al [17] introduced the first HNM by imposing the harmonicity on the deterministic component and limiting the energy of the modelling error by introducing a Maximum Voicing Frequency; more recently, d'Alessandro et al [5] overcome this limiting assumption by a prior deterministic/stochastic decomposition of the speech signal using continuous spectral interpolation [12].

1. ESTIMATING THE SINUSOIDS

Most HNMs use the FFT as a front-end for estimating the frequency, amplitude and phase of each individual sinusoid. The values of the parameters obtained by this method are however not directly related to $A_l(t)$ and $\psi_l(t)$. This is mainly due to the windowing and energy leaks due to the discrete nature of the computed spectrum (see for example the chapter 2 of Serra's thesis which is dedicated to the optimal choice of FFT length, hop size and window). George and Smith [8] proposes an analysis-by-synthesis method (ABS) for the sinusoidal model based on an iterative estimation and subtraction of elementary sinusoids. When imposing harmonic $\omega_l(t)$ [8, p.395], the iterative procedure can be performed in the ascending order of harmonics. We also improved slightly the algorithm by estimating the parameters on a time window centred on pitch marks and equal to exactly twice the local pitch period. PS-ABS enables a further synchronization between the harmonic and the stochastic component of the speech signal. Figure 1 gives an example of the performance of the ABS on a synthetic vowel with different pitch periods. The spectral enveloppe is estimated by a regularized discrete cepstrum estimation (RDCE) [7, 3]. The number of cepstrum coefficients is set to $Fech/F0_{moy}/2$ ($Fech$ sampling frequency of the speech signal and $F0_{moy}$ the average pitch frequency) and the regularization coefficient is set to $\lambda = 10^{-4}$. The Figure 2 shows its robustness to amplitude changes.

The average modelling error on the harmonic component of the synthetic signals used by d'Alessandro et al [5] is 33 dB for PS-ABS. Figure 3 gives the amplitude spectrum for a natural sample.

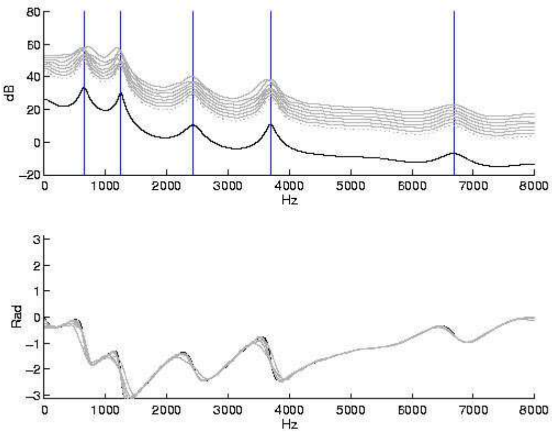


Figure 1: Determination of amplitudes and phases of harmonics of a synthetic vowel [a] with different pitch periods (from 51 to 244 Hz) using PS-ABS. Reference transfer function is shown with dark lines. Amplitudes are raising with F0 : $dB = 8.6 \log(Hz) - 16.7$.

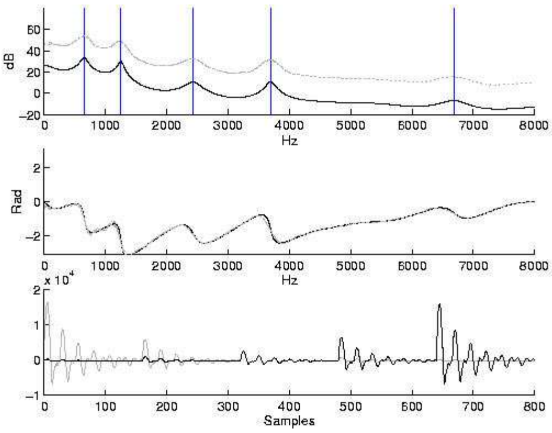


Figure 2: Same as Fig. 1 but with different global amplitude slopes (from -700 to +700 dB/s). The signals with maximal slopes are displayed at the bottom. Only the center period is analyzed.

2. H/S ESTIMATION

The residual of the above analysis/synthesis sinusoidal model has a large energy especially in "noise-like" sounds. Furthermore the sinusoidal model is not appropriate for their arbitrary expansion that result - as for TD-PSOLA techniques - in a periodic modulation of initial noise structure. Contrary to Almeida and Silva [2], Serra [16] considers the residual of sinusoidal modelling as a stochastic signal whose spectrum should be modelled globally. This stochastic signal however collects aspiration and friction noise, plosion ... but also modelling errors partly due to the procedure extracting sinusoidal parameters. In order to cope with this, Stylianou [17] introduces the notion of maximal voicing frequency (MVF). As quoted by d'Alessandro et al [5], this assumption is however unrealistic: the aspiration and friction noises cover the entire speech spectrum even in the case of voiced sounds.

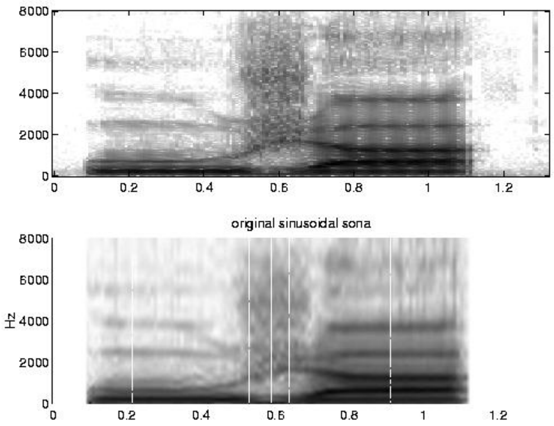


Figure 3: Comparison of the sonagram and PS-ABS amplitude spectrum for the logatom [uʒa].

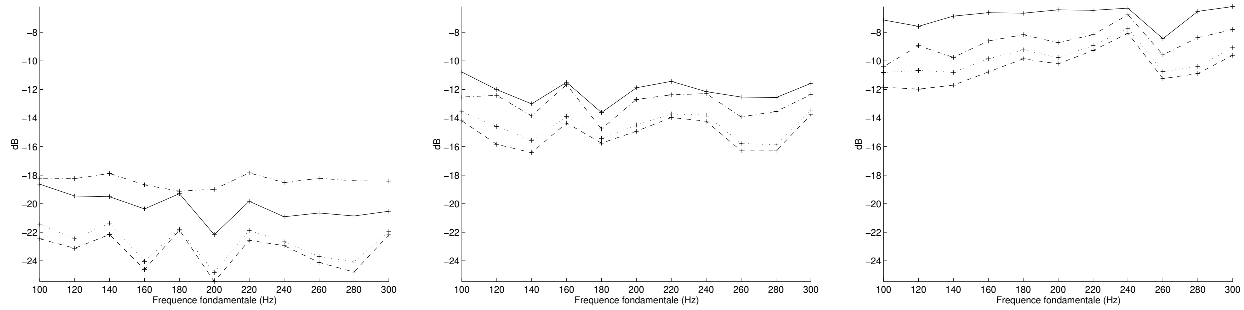
Using synthetic and natural stimuli, we compared three methods for the estimation of the stochastic component: (a) the a posteriori estimation by considering the PS-ABS residual (b) YAD : the a priori decomposition proposed by d'Alessandro et al [5] (c) AH: the a priori decomposition proposed by Ahn and Holmes [1] where both components are estimated jointly. We assessed our current implementation of YAD by using the synthetic stimuli from d'Alessandro et al. The results are summarised in Fig. 4. They show that the YAD and AH perform equally well and slightly better than the original YAD implementation. This is probably due to the stop conditions: we stop the convergence when successive interpolated aperiodic components differ by less than 0.1 dB. The average number of iterations for YAD is however 18.1 compared to 2.96 for AH. The estimation errors for PS-ABS are always 4dB higher. We further compared the performance between the three decomposition techniques using natural VCV logatoms where C is a voiced fricative. For YAD, AH and PS-ABS, the average differences between V's and C's HNR are 18, 18.8 and 17.5 respectively.

In summary the AH method seems to be the quickest and the most reliable method for the decomposition of harmonic/aperiodic components of speech.

3. HARMONIC SYNTHESIS

3.1. Time-scale modification

Most sinusoidal synthesis methods make use of the polynomial sinusoidal synthesis described by McAulay and Quatieri [11, p.750]. Systems that avoids a pitch-synchronous analysis-synthesis should separate system and excitation contributions to the phase spectrum [10]. However in the context of concatenative synthesis, it seems reasonable to have access to individual pitch cycles. In this case, the polynomial sinusoidal synthesis mentioned above has the intrinsic ability to interpolate between periods (see



(a) (b) (c)

Figure 4: Recovering a known deterministic component using four different algorithms: PS-ABS (plain), YAD (dashed), AH (dotted). The original YAD results have been added (dash dot). The figures show the relative error of the deterministic component at different F0 values for three increasing aperiodic/deterministic ratio: (a) 20dB, (b) 10dB and (c) 5dB.

for example fig. 5).

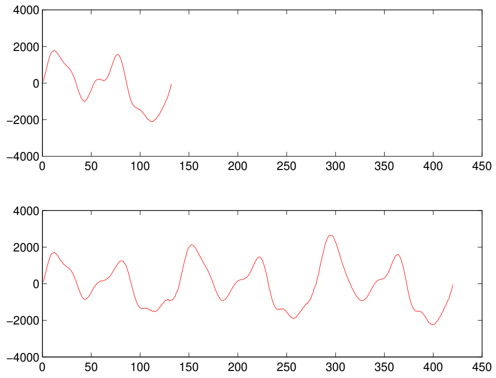


Figure 5: Intrinsic ability of the polynomial sinusoidal synthesis to interpolate periods. Top: synthesised period of length T=140 samples. Bottom: same adjacent sinusoidal parameters but with T=420 samples.

3.2. Pitch-scale modification

Changing the fundamental frequency of the speech signal while preserving the shape invariance and the spectral envelope consists of re-sampling the envelope at new harmonics. From the sinusoidal harmonics delivered by PS-ABS, the regularized discrete cepstrum (see section 1) delivers a robust and reliable estimate of spectral enveloppes (amplitude and phase). Analysis amplitude and phase spectrum can be then re-sampled at harmonics of the new fundamental frequency.

4. STOCHASTIC SYNTHESIS

Using a preference test, we compared essentially two methods actually proposed for stochastic analysis/synthesis: (a) modulated LPC versus (b) formantic waveforms (FOF) proposed originally by Rodet [15] and applied to stochastic signals by Richard and d'Alessandro [14].

The modulated LPC consists here in approximating by a polynomial the pitch normalized energy of

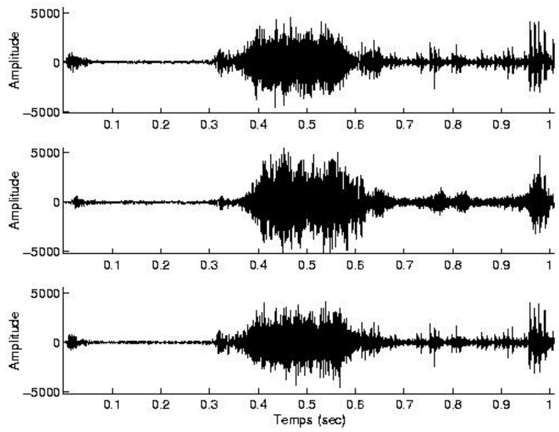


Figure 6: Stochastic modelling of a logatom [uʃa]. From top to bottom: original, modulated LPC and FOF.

the LPC residual of the stochastic component. This analysis is performed pitch-synchronously: the energy contour of each period - computed as the absolute value of its Hilbert transform - is fitted against a normalized time interval [0-1]. The multi-band FOF analysis is performed by associating a FOF at each maxima of the energy contour of each bandpass signal. After HNM decomposition by AH, the stochastic part of natural VCVs was analysed, resynthesized and added back to the original harmonic part. We performed a preference test. Although we improved notably the estimation of FOF parameters, the modulated LPC was still largely preferred to FOF for unvoiced sounds. Furthermore the modulated LPC has the advantage to deliver less parameters and the representation is easier to synchronise with the harmonic part.

Figure 8 shows an example of the complete harmonic and stochastic analysis-synthesis process. Samples can be retrieved from the Cost 258 evaluation server: `www.icp.inpg.fr/cost258/evaluation/server/cost258_coders.html` or from the CDROM (td = TD-PSOLA, sy = present

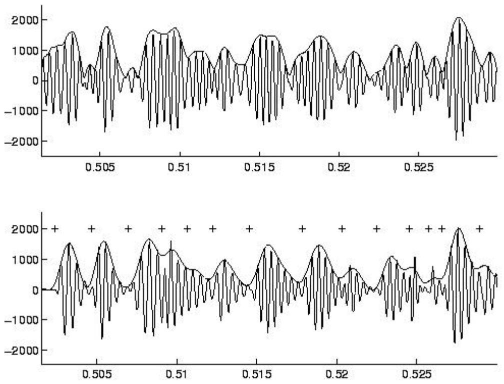


Figure 7: Stochastic modelling using FOF. Subband [1770-2800Hz]. Top: original stochastic signal with the energy contour superposed. Bottom: same for the synthesized FOF signal. Crosses indicate FOF excitation points

coder, na = natural stimuli).

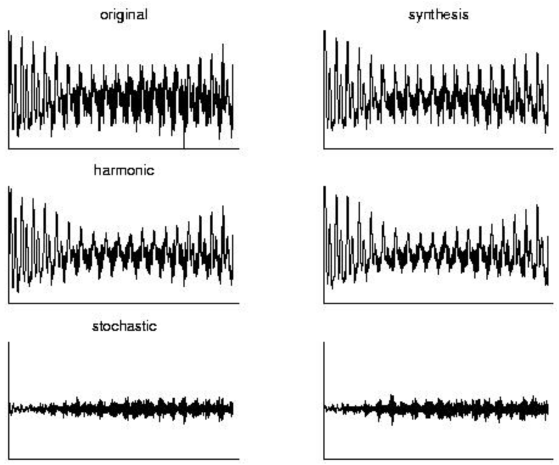


Figure 8: Analysis-resynthesis of the central part of [uʒa]. Left: original; Right: synthesis. From top to bottom: signal, harmonic and stochastic contributions.

CONCLUSIONS

Most experiments demonstrate that TD-PSOLA [4] is still the most simple and efficient analysis/synthesis scheme for speech synthesis. There is no evidence that the HNM proposed here could always compete with the robust shape-invariance of TD-PSOLA. With a comparable performance in basic prosodic modifications it however offers a more versatile speech synthesis system for waveform interpolation and voice transformation.

ACKNOWLEGMENTS

This work is supported by Cost 258 and ARC-B3 initiated by AUPELF-UREF. We acknowledge Yannis Stylianou for his help and Christophe d'Alessandro for providing us with the vocalic synthetic stimuli used in his papers.

REFERENCES

[1] Ahn, R. and Holmes, W.H. An accurate pitch detection method for speech using harmonic-plus-noise decomposition. In *Proceedings of the International Congress of SPeech Processing*, pages 55–59, Seoul - Korea, 1997.

[2] Almeida, L.B. and Silva, F. Variable-frequency synthesis: an improved harmonic coding scheme. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 27–5, 1984.

[3] Cappé, O., Laroche, J., and Moulines, E. Regularized estimation of cepstrum enveloppe from discrete frequency points. In *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Mohonk - USA, 1995.

[4] Charpentier, F. and Moulines, E. Pitch-synchronous waveform processing techniques for text-to-speech using diphones. *Speech Communication*, 9(5-6):453–467, 1990.

[5] d'Alessandro, C., Darsinos, V., and Yegnanarayana, B. Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Transactions on Speech and Audio Processing*, 6(1):12–23, January 1998.

[6] Dutoit, T. and Leich, H. Mbr-psola: text-to-speech synthesis based on an mbe re-synthesis of the segments database. *Speech Communication*, 13(3-4):435–440, 1993.

[7] Galas, T. and Rodet, X. Generalized functional approximation for source-filter system modeling. In *Proceedings of the European Conference on Speech Communication and Technology*, volume 3, pages 1085–1088, Genova – Italy, 1991.

[8] George, E.B. and Smith, M.J.T. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Transactions on Speech and Audio Processing*, 5(5):389–406, September 1997.

[9] Gobl, C. and Chasaide, N. Acoustic characteristics of voice quality. *Speech Communication*, 11(4-5):481–490, 1992.

[10] Macon, M.W. *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, 1996. Sous la direction de Mark Clements.

[11] McAulay, R.J. and Quatieri, T.F. Speech analysis-synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-34(4):744–754, august 1986.

[12] Papoulis, A. *Probability, random variables, and stochastic processes*. McGraw-Hill Book Company, New-York, 1986.

[13] Quatieri, T.F. and McAulay, R.J. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, march 1992.

[14] Richard, G. and d'Alessandro, C. Modification of the aperiodic component of speech signals for synthesis. In Van Santen, J.P.H., Sproat, R.W., Olive, J.P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 41–56. Springer Verlag, New York, 1997.

[15] Rodet, X. Time-domain formant wave function synthesis. *Computer Music Journal*, 8(3):9–14, 1980.

[16] Serra, X. *A system for sound analysis/ transformation/ synthesis based on a deterministic plus stochastic decomposition*. PhD thesis, Stanford University - Standford, CA, 1989.

[17] Stylianou, Y. *Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification*. PhD thesis, École Nationale des Télécommunications, Paris – France, 1996. sous la direction d'Eric Moulines.