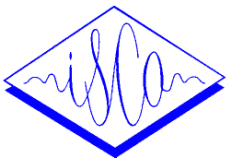

3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

AN ANALYSIS OF THE PERFORMANCES OF THE MBE MODEL WHEN USED IN THE CONTEXT OF A TEXT-TO-SPEECH SYSTEM.

Thierry Dutoit, Henri Leich

Faculté Polytechnique de Mons, 31, Boulevard DOLEZ, B-7000 Mons, Belgique.
Tel : /32/65/374133. Fax : /32/65/374300.

ABSTRACT.

The use of a hybrid Harmonic/Stochastic model, such as the MBE one, is examined in the context of a High Quality TTS system (Fs=16 kHz). Analysis errors are studied in case of a direct analysis-synthesis scheme, and the exact responsibility of the analysis algorithm, rather than the model itself, is investigated. Through its application on well-known signals, it is found that:

- Among the available analysis criteria, the Abrantes et al [2] approach slightly emerges, even though little audible improvements are obtained on real speech.
- The MBE analysis of a single cosine with slowly time-varying fundamental frequency introduces severe biases on its estimated amplitude and phase, especially for high central frequencies, while amplitude variations result in frequency-independent amplitude biases only. These effects are due to the fact that a constant frequency and amplitude is assumed during the whole analysis frame. They are responsible for the existence of HF noise in synthesized speech.

Keywords *: Multi-Band analysis, Analysis Criteria, Time-Varying Parameters*

1. INTRODUCTION.

Since their presentation in [1] and [2], hybrid models have been increasingly used in speech analysis, both in their Sinusoidal/Stochastic and Harmonic/Stochastic versions. They transfer Voiced/UnVoiced (V/UV) decisions to frequency bands, or even transform them into more flexible frequency-dependent V/UV ratios. The resulting additional degrees of freedom allow a better simulation of mixed sounds, for which fricative noise and periodic vibration of the vocal folds are not mutually exclusive.

They were originally developed for low bit rate voice coding purposes (4.8 kbits/sec with F_S = 8 kHz) [7]. As such their segmental quality has not really been optimized, at least not for their use in a high quality Text-To-Speech (TTS) system, which moreover runs at a sampling frequency of 16 kHz. The use of a Multi-Band Excited (MBE) model has recently been introduced in the context of a Text-To-Speech system presented by the authors in [3], and further refined in [4]. The resulting Multi-Band Re-synthesis Pitch Synchronous Overlap-Add (MBR-PSOLA) synthesis algorithm combines the simplicity of the original TD-PSOLA approach [5] with the flexibility of parametric synthesizers, as the MBE one [1]. Since the related TTS system is based on an MBE analysis / Re-Synthesis (with no coding block) of the segments database, its performances are mainly conditioned by the segmental quality provided by the Harmonic/Stochastic model underlying Multi-Band analysis. Further investigations have thus been made, so as to loose a minimum of naturalness during this operation.

2. THE HYBRID HARMONIC/STOCHASTIC (H/S) MODEL.

Hybrid models basically express speech signals as the summation of slowly varying harmonic and stochastic components :

$$\tilde{s}(t) = \tilde{s}_p(t) + \tilde{s}_r(t) = \sum_i a_i(t)\cos(\phi_i(t)) + s_r(t) \quad (1)$$

in which $a_i(t)$ and i $\omega_0(t)$ are the amplitudes and instantaneous frequencies of the harmonics and $\tilde{s}_r(t)$ is completely determined by its power density spectrum $\tilde{S}_{rr}(\omega)$. As for sinusoids amplitudes and frequencies, $\tilde{S}_{rr}(\omega)$ is supposed to vary slowly with time, according to the quasi-stationarity hypothesis. Parameters are therefore estimated and stored about once every 10 ms, and temporal linear interpolation on the resulting spectral samples is assumed to describe speech in a realistic way.

The signal described by equation (1) can alternatively be understood as the output of the filter model of Fig. 2 [2], in

which $e_r(t)$ is a white gaussian random signal with unit variance, and $e_p(t)$ is a harmonic excitation with unitary amplitudes and zero phases.

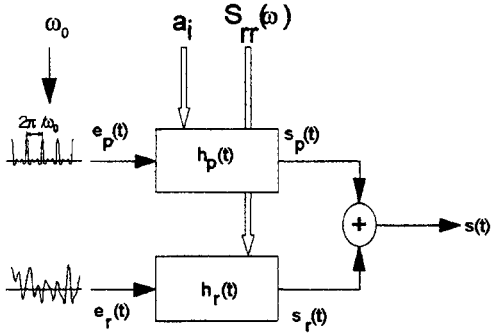

Fig. 2. The hybrid Harmonic/Stochastic model.

3. ANALYSIS CRITERIA - SELECTIVITY.

For a given speech frame, hybrid H/S analysis consists, as shown on Fig. 2, of simultaneously optimizing the fundamental frequency ω_0 and performing the approximation of both $h_p(t)$ and $h_r(t)$ filters, so as to minimize the error between the original frame and the one that would be synthesized from its model.

When no stochastic component is encountered, the problem of estimating ω_0 together with the harmonics amplitudes and phases a_i and ϕ_i, assuming they are constant in the analysis frame has typically been addressed by harmonic coding techniques [6]. Its solution is based on the convenient least-squares criterium of minimizing the cost function $e(t, h_p)$ for a given value of t :

$$e(t,h_p) = \int_{-\infty}^{\infty} w^2(t-\tau)\left|s(\tau)-\tilde{s}_p(\tau)\right|^2 d\tau = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left|S(t,\omega)-\tilde{S}_p(t,\omega)\right|^2 d\omega \quad (2)$$

in which $S(.,.)$ denotes a Short-Time Fourier Transform (STFT). Equation (2) obviously is non-linear in ω_0, but can easily be made linear in $\{a_i,\phi_i\}$ for a given value of ω_0. Since windowed exponential functions are not, in all generality, orthogonal to one another, the least squares approximation (3) is practically expressed by a Yule-Walker system of linear equations :

$$\boldsymbol{R}\underline{\boldsymbol{a}} = \boldsymbol{r} \quad (3)$$

which denotes, in the time or frequency domains, the orthogonality of the error function with the basis exponential functions. It is solved on a grid of frequency values, and the set of $\{\omega_0, a_i, \phi_i\}$ parameters with the lowest error is retained.

Similarly, if ones assumes that no harmonic component exists in $\tilde{s}_r(t)$, the optimization of $h_r(t)$ can be done in the Least Mean Squares sense, that is by minimizing the expected value of the squared difference between the analysis and synthesis samples :

$$e(t,h_r) = \frac{1}{2\pi}\int_{-\infty}^{\infty} E\left(\left|S_r(t,\omega)\right|^2\right) - E\left(\left|\tilde{S}_r(t,\omega)\right|^2\right) d\omega \quad (4)$$

In practice, $\tilde{s}_r(t)$ is roughly approximated as the superposition of N_b band limited noises with contiguous spectra, the central frequencies and bandwidths of which are either linearly distributed on the frequency axis ([1], [2]) or adjusted on the critical bands of the human ear, so as to maintain a constant psycho-acoustical low distortion on the frequency range [9].

The point is that both harmonic and stochastic components coexist in $s(t)$ and there is no *a priori* means of separating them, so that the optimization of $e(t, h_p)$ + $e(t, h_p)$ with equations (2) + (4) is impossible. Sub-optimal analysis criteria have thus been introduced.

In [1], a pragmatic two steps approach is adopted. The parameters of the periodic part are first deduced from equation (2), and $s_r(t)$ is sub-optimally introduced in equation (4) as

$$s_r(t) \approx s(t) - \tilde{s}_p(t) \quad (5)$$

In [2], it was proposed to perform a frequency domain Least Squares decomposition of $S(t,\omega)$ on two sets of basis functions simultaneously : the STFT of finite duration weighted cosines as before, and the square rooted power density spectrum of the aforementioned band limited noises, weighted in the same way (see equation (6), in which $\tilde{S}_{rr}(\omega)$ denotes a weighted sum of power density spectra, and n=1). As we shall see, the key point of this method is that, since both sets of basis functions are far from being orthogonal, some components previously wrongly described by sinusoids are given the opportunity to reduce their contribution to harmonics.

$$e(t,h_p,h_r) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(\left|S(t,\omega)\right|^n - \left|\tilde{S}_p(t,\omega)\right|^n - \left|\sqrt{\tilde{S}_{rr}(\omega)}\right|^n\right) d\omega \quad (6)$$

We have examined the use of an expanded version of these criteria, on artificial signals first (a sum of sinusoids with equal amplitudes, random initial phases and constant frequency, and white uniform noise), and on real speech. We shall refer to the approach of [1] as Criterion 1 : (2) , (5) and

(4). Criteria 2 and 3 also make use of (5) and (4), but (2) is replaced by (7), in which n is respectively set to 1 and 2.

$$e(t,h_p) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(|S(t,\omega)|^n - |\tilde{S}_p(t,\omega)|^n\right)d\omega \qquad (7)$$

Criterion 4 is a complex version of the approach of [2] : (8). Finally, criteria 5 and 6 are given by equation (6), for n respectively equal to 1 and 2.

$$e(t,h_p,h_r) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left(S(t,\omega) - \tilde{S}_p(t,\omega) - \left|\sqrt{\tilde{S}_{rr}(\omega)}\right|\right)d\omega \qquad (8)$$

Analysis was performed at Fs=16000 kHz, with a Hamming weighting window (480 samples). As expected, all criteria successfully computed sines amplitudes, for which no additional noise was found. It clearly appeared that criteria 5 and 6 were more efficient, since all other analysis methods resulted in finding high amplitude sinusoids in white noise. More precisely, the average sinusoid amplitude extracted with criterion 5 was 14 dB lower than with criterion 1, while the difference decreased to 5 dB with criterion 6. We conclude that harmonics rejection is enhanced with the approach of [2]. However, noise levels were mostly identical in all cases, even with criterion 1. Synthesis surprisingly revealed that the strong harmonic components obtained with the Griffin method only moderately contributed to buzzyness, given their strongly time-varying frequencies and amplitudes. What is more, they were partly masked by the average noise. Experiments on real speech confirmed this effect, so that we have kept using criterion 1, given the important simplifications it allows in the resolution of (3).

4. TIME DEPENDENCY.

When performing MBE analyses of real speech digitized at 16 kHz, we noticed that most of the high frequency part of the spectrum (i.e. above 4 kHz) was interpreted as noise components, even when analysing steady state parts of vowels. This resulted in unpleasant breathy synthesis. We then investigated the possible causes of this effect, in order to find if the errors were due to the hybrid H/S model itself or to some inadequacy of the analysis algorithm. As a matter of fact, the H/S model describes speech with slowly time-varying parameters, while MBE analysis assumes they are constant on the analysis frame. Time dependency is simply re-introduced during synthesis by interpolating frame parameters from sample to sample.

It is possible to analytically estimate the effect of a first-order time dependency of the parameters on the harmonic analysis accuracy, as well as on the residual error. As a matter of fact, let us try to estimate the best least squares decomposition of a windowed cosine with initial phase ϕ, time-varying amplitude $A(t)$, *and* time-linear frequency $\omega(t)=\omega_0+\alpha\omega_0 t$, on a windowed imaginary exponential function with zero initial phase, constant frequency ω and unitary amplitude. We compute :

$$\tilde{\underline{A}}(0) = \frac{\int_{-\infty}^{\infty} A(t)w^2(t)\cos(\frac{\alpha\omega_0}{2}t^2 + \omega_0 t + \phi)\exp(-j\omega t)dt}{\int_{-\infty}^{\infty} w^2(t)dt} \qquad (9)$$

It is found from [8], after some transformations and simplifications, that for $a>0$:

$$\int_{-\infty}^{\infty}\exp(-at^2 - 2bt)\exp(j(pt^2 + 2qt + r))dt = f(a,b,p,q,r) = \frac{\sqrt{\pi}}{\sqrt[4]{(a^2+p^2)}}\exp(\frac{ab^2 - aq^2 + 2bpq}{a^2+p^2}) \quad X$$

$$\exp\left[j\left(\frac{1}{2}\arctan(\frac{p}{a}) - \frac{pq^2 - p^2r - b^2p + 2abq - a^2r}{a^2+p^2}\right)\right] \qquad (10)$$

It follows that, if a and b are respectively chosen so that $\exp(-a\,t^2/2)$ approximates a given weigthing window, and $\exp(-2bt)$ approximates $A(t)$:

$$\tilde{\underline{A}}(0) \approx \frac{f(a,b,\frac{\alpha\omega_0}{2},\frac{\omega_0-\omega}{2},\phi) + f(a,b,\frac{-\alpha\omega_0}{2},\frac{-\omega_0-\omega}{2},-\phi)}{2f(a,0,0,0,0)} \qquad (11)$$

Results for reasonable values of α and b are given in Fig. 4 and 5, in which the effect of $A(t)$ and $\omega(t)$ are analysed separately. Preliminary computations showed that the best least squares decomposition is always encountered for $\omega=\omega_0$, i.e. for the frequency at the center of the frame. Consequently, all plots are given for this value only.

As far as pitch variation is concerned, it appears that biases arise in the estimation of the sinusoid amplitude and phase, even for small values of α, and mostly in high frequency. Errors of more than 6 dB (from the exact value of -6 dB) can be encountered when α approaches ±5, which corresponds to an increase of ±5 Hz in a 10 ms time at f_0=100 Hz. Phase errors are still more impressive : on can say that high frequency harmonics phases are almost always erroneous. This clarifies the existence of high frequency noise in re-synthesized speech, as shown on the bottom plot of Fig. 4.

Amplitude variations result in a frequency-independent bias in the estimated sinusoid amplitude, but have no effect on phase (a quick glance at formula (10) suffices to confirm it). As shown on Fig. 5, important amplitude variations (b=±150

corresponds to ±25 dB in a 10 ms time) can be responsible for considerable noise in the MBE model.

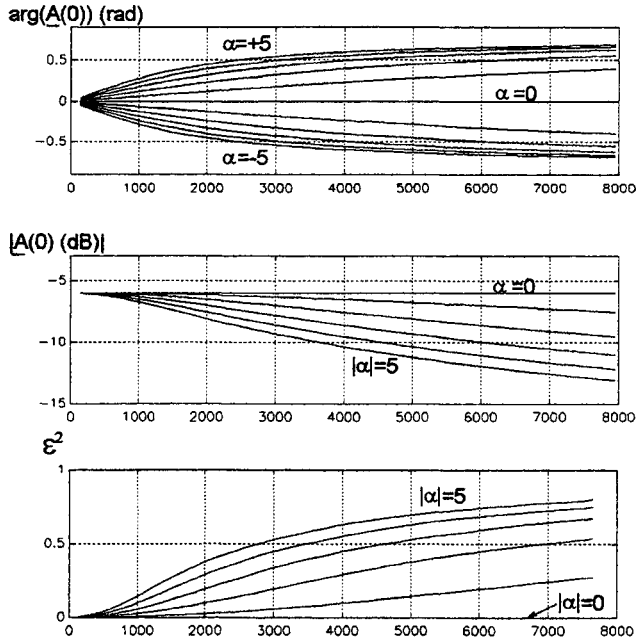

Fig. 4. Effect of time-linear frequency : best Least squares decomposition *($\omega=\omega 0$)* results *($\tilde{A}(0)$ and ε^2)* as a function of the center frequency, for *$\phi=0$, $a=25000$* (which approximately corresponds to a 30 ms Hamming window), *$b=0$*, and *α in $[-5,+5]$* .

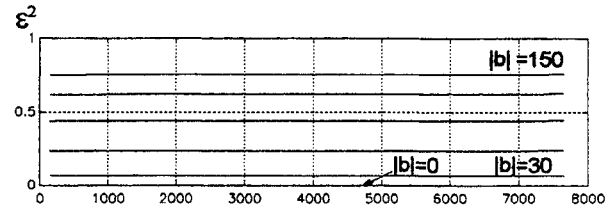

Fig. 5. Effect of time-exponential amplitude : best Least Squares decomposition ($\omega=\omega 0$) results (ε^2) as a function of the frequency, for *$\phi=0$, $a=25000$, $b=[-150,+150]$*, and *$\alpha=0$.*

5. CONCLUSION.

In the context of High Quality Text-To-Speech synthesis, we have examined two ways of improving the quality of speech submitted to a hybrid Harmonic/Stochastic analysis-synthesis. In a first step, we have browsed an enlarged set of analysis criteria, starting from the ones proposed in [1] and [2], and studied their pros and cons. We found that the approach of [2] was the most selective one, in the sense that smaller sinusoids were extracted from white noise. This advantage, however, turned out to be of minor importance when real speech was processed. We thus kept using the Griffin criterion. In a second step, we investigated the influence of time-varying frequency or amplitude on the accuracy of the MBE analysis of a single sinusoid and on the related analysis error (which is interpreted as noise in the MBE model). We developed analytical expressions for the results and showed that the analysis accuracy was as sensitive to moderate pitch changes as it is to important amplitude variations, mostly in high frequency. The resulting HF error appears as typical additive HF noise in synthesized speech.

First trials to adapt the analysis algorithm to such time-dependency gave poor results. Future research should be devoted to correct these effects in a post-processing stage.

REFERENCES

[1] *GRIFFIN, D.W.* (1987), "Multi-Band Excitation Vocoder", Ph.D. Dissertation, MIT.

[2] *A.J. ABRANTES, J.S. MARQUES, I.M. TRANSCOSO,* "Hybrid Sinusoïdal Modeling of Speech without Voicing Decision", EUROSPEECH 91, pp. 231-234.

[3] *DUTOIT, T. & LEICH, H.* (1992), "Improving the TD-PSOLA Text-To-Speech Synthesizer with a Specially Designed MBE Re-Synthesis of the Segments Database", EUSIPCO 92, 25-28 august 92, Brussels, pp. 343-347.

[4] *DUTOIT, T. & LEICH, H.* (1992), "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", Proceedings of the Fourth Australian International Conference on Speech Science and Technology, december 92, Brisbane pp. 202-206.

[5] *MOULINES, E. & CHARPENTIER, F.* (1990), "Pitch Synchronous waveform Processing techniques for Text-To-Speech Synthesis using diphones", Speech Communication, Vol. 9, n°5-6.

[6] *MARQUES, J., ALMEIDA, L., TRIBOLET, J.,* 'Harmonic coding at 4.8 kbps', Proc. ICASSP 90, vol. 1, pp. 17-20.

[7] *HARDWICK, J.C., LIM, J.S.*, 'A 4.8 Kbps Multi-Band Excitation Speech Coder', Proc. ICASSP 88, pp. 374-377.

[8] *GRADHSTEYN, I.S. & RYZHIK, I.M.* (1965), Table of Integral Series and Products, Academic Press, New York, p. 485.

[9] *HOLGER C., KOLPATZIK, B.*, 'Speech coding unsing nonstationary sinusoidal modelling and narrow-band basis functions', Proc. ICASSP 1991, pp. 581-584.