A Comparison of Four Candidate Algorithms in the context of High Quality Text-To-Speech Synthesis

Thierry Dutoit, Henri Leich

Faculté Polytechnique de Mons, TCTS-Multitel
31, Boulevard DOLEZ, B-7000 Mons, Belgique.
Tel : /32/65/374133. Fax : /32/65/374300.
Email : dutoit@fpms.tcts.ac.be, WWW : http://tcts.fpms..ac.be

Draft version of my ICASSP'94 paper, Adelaide, Australia.

ABSTRACT

In this paper, we investigate the use of four candidate speech models in the context of High Quality Text-To-Speech systems (HQ-TTS), address problems typically encountered by their prosody matching and segment concatenation modules, and compare their performances regarding : the segment database compression ratio they allow, the computational load of the related synthesis algorithms, as well as their intelligibility and subjective segmental quality; The models addressed are : the classical Auto-Regressive (LPC) one [1], the hybrid Harmonic/Stochastic (H/S) model proposed in [2] and [3], the 'null' model, as implemented by the Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) synthesis algorithm [5], and the Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add (MBR-PSOLA) model [6].

INTRODUCTION

The segmental quality provided by HQ-TTS systems, for which a sampling frequency of 16 kHz or higher is generally accepted as a must, is clearly subordinated to :

1. The type of segments chosen.
2. The corpus they were extracted from.
3. The corpus segmentation quality.
4. The speech signal model, to which the analysis and synthesis algorithms refer.
5. The amount of degradation introduced by the speech coding phase.
6. The prosody matching efficiency, which is strongly related to the model.
7. The capabilities of the segments concatenation algorithm.

In [8], we have investigated the use of four leading models on the basis of practical software implementations of the four related HQ-TTS systems, for which the same segments database and input data (phonemes and prosody) were used, so as to put the contributions of the models in evidence.

The models have addressed are :

1. The classical AutoRegressive (LPC) one [1], with a prediction order of 18, which was taken as ground quality.
2. The hybrid Harmonic/Stochastic (H/S) model proposed in [2] (denoted as the MBE model) and [3], which basically expresses speech signals as the summation of slowly varying harmonic and stochastic components, therefore transferring Voiced/UnVoiced (V/UV) decisions to frequency bands, or even transforming 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.
3. The 'null' model, as implemented by the Time-Domain Pitch-Synchronous OverLap-Add (TD-PSOLA) synthesis algorithm [5], which has recently drawn considerable attention, given its exceptional segmental and supra-segmental efficiency, associated with a virtually unequalled simplicity.
4. The Multi-Band Re-synthesis Pitch-Synchronous OverLap-Add (MBR-PSOLA) model, based on an original and efficient hybrid H/S re-synthesis of the segments database with constant synthesis pitch and constant initial phases [6].

Their peculiarities in the context of HQ-TTS synthesis are addressed in the next Section. It is followed by the presentation of the results of intelligibility and subjective quality tests. The paper is concluded by a comparative summary.

Our Four SYNTHESIZERS.

1.1. The LP model.

We have implemented a classical LP TTS system, with order 18. Prosody matching was straightforward, since pitch and duration are explicit parameters of the model. Both PARCOR's and LSP's have been tested as concatenation parameters. Even though we have noticed a small theoretical superiority of LSP's over PARCOR's (see Fig. 1), we found both parameters indistinguishable in currently synthesized speech. Synthesis was performed with a lattice filter, the coefficients of which were interpolated every 5 ms. The resulting computational load was equal to 70 operations per sample. Regarding database compression capabilities, storage rates of about 4000 bps are common.

1.2. The hybrid H/S model.

Analyses biases in the context of HQ-TTS have already been addressed in [4], where we have highlighted the influence of time-varying pitch and sinusoid amplitudes on the analysis accuracy (whatever analysis criteria used) and shown that the

resulting High Frequency error appears as typical additive HF noise in synthesized speech. In order to face this problem, the segments database we have used throughout our tests was recorded with the most constant pitch possible.
Synthesis was not performed as in [2] or [3], neither as in [11] (which is to our knowledge to only other TTS implementation of the hybrid model). An original and faster method was preferred, which computes samples by OverLap-Adding (OLA) the IFFT of spectral frames, obtained by summing stochastic components (in the form of FFT bands with constant amplitudes and random phases) and harmonic ones (in the form of the most significant samples of their Dirichelet kernels). The resulting computational load can be reduced to about 100 operations per sample (i.e. about 50% more than with LP synthesis)

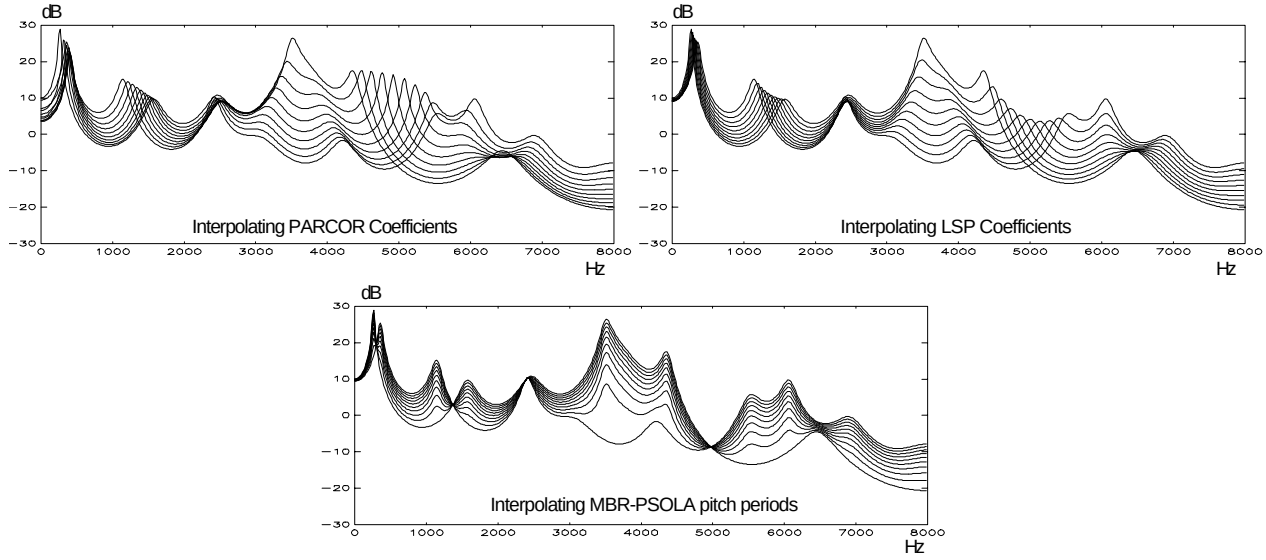

Figure 1. Linear interpolation of PARCOR's, LSP's , hybrid H/S spectral parameters (as well as MBR-PSOLA frames) between [l] phones, respectively encountered in diphones [al] and [lo].

Phase continuity was ensured during prosody matching, by propagating phase changes (due to pitch and duration variations) rightward, throughout segments.
As for segment concatenation, linear smoothing was applied on spectral amplitude parameters, which is equivalent to fading in/out spectral differences (see again Fig. 1). It is theoretically less adequate than the more realistic formant movements obtained with PARCOR's, but we found it of little perceptual importance, provided segments are sufficiently similar. As a matter of fact, few concatenation points, if any, could be heard. On the other hand, we have faced the problem of phase discontinuities at segment boundaries by applying offsets to all the phases of right segments. This results in maximally continuous speech, but inescapably produces some typical buzzyness in voiced fricatives and semi-vowels, which we understand as the loss of some long term phase coherency.
Finally, storage bit rates of about 10000 bps can reasonably be achieved with H/S models.

1.3. The TD-PSOLA algorithm.

Our implementation of the TD-PSOLA TTS synthesis system is similar to the one of [5]. Its drawbacks have been underlined in [7] : optimal pitch-marking is not fully automatic (it was done by hand in our case), and pitch, phase, and spectral amplitude mismatches prevent concatenations from being adequately smoothed. What is more, it offers few database compression possibilities (a 80,000 bps storage rate can be achieved with a zero-tap DPCM coder).
It does, however, lead to a good segmental quality, and its computational load is remarkably low : 7 operations per sample.

1.4. The MBR-PSOLA model.

It has been shown in [6] and [7], that the harmonic re-synthesis of the voiced part of the segment database provided by MBR-PSOLA happens to get rid of all the drawbacks of TD-PSOLA at the same time. This results from the fact that all the voiced periods in the database are imposed identical pitch and initial phases. Pitch mismatches thus no more exists,

pitch marking becomes implicit, and a simple temporal linear smoothing of frames to be concatenated is equivalent to the spectral smoothing performed with the hybrid H/S model (see again Fig.1). What is more, this is achieved with no increase in complexity during TTS synthesis itself, since an overall computational cost (including temporal linear smoothing) of 7 operations per sample is maintained.
As a result, maximal fluidity is ensured with MBR-PSOLA. In counterpart, speech suffers from some slight buzzyness, as with the hybrid H/S model.
Another important interest of the MBR-PSOLA approach resides in its potential database compression performances. As a matter of fact, the constant pitch re-synthesis operation ensures maximal pitch period similarity, so that the DPCM technique mentioned above can now be applied on a pitch period basis, rather than on a sample one. Since pitch period waveforms evolve rather slowly with time, and given the fact that voiced OLA frames fill up to 75% of the segment database, important storage reduction can be expected, while maintaining the simplicity of the coding technique (which is in a par with the explicitness of the synthesis algorithm). Tests are being performed, the results of which will be disclosed in the oral presentation.

QUALITY ASSESSMENT.

The methodology we have followed for our CVC tests is modelled on [10]. Phonetically balanced lists of fifty CVC nonsense words were used. Semi-consonants were omitted, as well as non existing diphones. CVC words were synthesized with a fixed prosodic pattern, in which pitch was maintained constant (110 Hz) and durations were imposed as follows : [a, E, 9, i, O, y, u] = 70 ms, [e, 2, o, â, ô, ê] = 170 ms, fricatives = 100 ms, liquids = 80 ms, nasals = 80 ms, plosives = 100 ms. They were played directly by users, by depressing a key on a computer keyboard. Four CVC lists (one per model) were presented in a random order to each of the 17 listeners (17 CVC lists were therefore generated by each synthesizer), through headphones, in a sound treated booth. Each stimulus was presented once.
MOS tests were also used to assess the naturalness of synthesizers, as well as their segment concatenation efficiency. Six long sentences were synthesized by each system, the prosody of which was copied from a human reading and passed to TTS systems in the form of a sequence of phonemes with their durations and the position and values of pitch pattern points, which provide a piecewise linear description of intonation. Micro-prosody was not taken into account. The twenty-four resulting stimuli were presented randomly to each listener, who was asked to rate their overall naturalness, as well as, more specifically, the perceived *fluidity* of speech, strongly related to the amount of concatenation points that can still be perceived in synthetic utterances. Results are presented in Table 1.

	LPC	**hybrid H/S**	**TD-PSOLA**	**MBR-PSOLA**
CVC intelligibility	54.6 %	65.7 %	78 %	72.8 %
MOS fluidity	50.4 %	73.5 %	65 %	75.6 %
MOS naturalness	44.5 %	65 %	68.3 %	68.3 %

Table 1. CVC and MOS tests results.

As expected CVC tests reveal a clear superiority of hybrid H/S, TD-PSOLA, and MBR-PSOLA synthesizers on the LPC one. After a deeper examination of the CVC answers, we found that most of the errors originated from Voiced/Unvoiced confusions, mainly in the case of plosives. Confusions were also encountered for vocalic aperture ([eE],[ao], ...) and nasality ([âo],[aî], ...).
Among the most intelligible systems, TD-PSOLA still has a slight advantage in comparison with MBR-PSOLA. This is likely to be due to the fact that TD-PSOLA is much less sensitive to analysis V/UV errors than MBR-PSOLA. As a matter of fact, both algorithms were submitted to the same V/UV errors (in the sense that a common pitch analysis algorithm was used for both). However, erroneously considering an OLA voiced frame as unvoiced, or conversely, simply results, with TD-PSOLA, in applying a wrong time-shift between frames. Voicing itself is therefore not affected. In contrast, considering an unvoiced frame as voiced leads MBR-PSOLA to re-synthesize it as a sum of harmonically related sinusoids.
Finally, MBR-PSOLA is itself slightly better understood than its hybrid H/S counterpart. This naturally results from the fact that unvoiced frames are left untouched by the MBR-PSOLA re-synthesis algorithm (so that there is no difference between TD- and MBR-PSOLA as far as unvoiced frames are concerned), while they are re-synthesized in the hybrid H/S approach.
As for naturalness, one would have expected MOS results to follow the same trend as CVC ones : the more a synthesizer makes use of speech models, the less natural it appears. However, hybrid H/S and MBR-PSOLA synthesizers clearly prevail over TD-PSOLA and LPC ones with regard to their fluidity, given their improved concatenation capabilities.

Consequently, TD-PSOLA and MBR-PSOLA are perceived as equally natural, closely followed by hybrid H/S, and far before LPC.

CONCLUSIONS.

Our results are summarized in Table 2, which reads as follows :

1. LPC, hybrid H/S, and MBR-PSOLA are superior to TD-PSOLA regarding the availability of automatic analysis procedures, a key point for developing multi-lingual TTS systems.
2. Prosody matching gives comparable results with all four models.
3. As far as segments concatenation capabilities are concerned, which are essential features in TTS synthesis, LPC is slightly superior to hybrid H/S, which is itself approximately equivalent to MBR-PSOLA. TD-PSOLA virtually exhibits no segments concatenation capabilities.

Switching to more economical criteria, one notices that :

4. The availability of an efficient segment database compression algorithm is ensured for LPC and hybrid H/S synthesizers. It is currently being developed for MBR-PSOLA, but it is clear that the resulting compression ratio will be superior to the one obtained for TD-PSOLA, while remaining computationally simple.
5. As a result of the computational complexity of their respective synthesizers, the LPC and hybrid H/S approaches clearly cannot spare a DSP. In contrast, both TD and MBR-PSOLA run in real time on a PC-386 machine.

Finally, when comparing the quality and intelligibility test results of the four models, it appears that :

6. TD-PSOLA and MBR-PSOLA have the highest CVC-scores, with a slight advantage for TD-PSOLA. As expected, the hybrid H/S synthesizer is itself much more intelligible than the LPC one.
7. Fluidity is better ensured by MBR-PSOLA and hybrid H/S synthesizers, given their superior concatenation capabilities.
8. Regarding naturalness, TD-PSOLA prevails *de facto*, since it does not make use of any speech model. MBR-PSOLA, however, is found to be as natural as TD-PSOLA, given its increased fluidity. H/S follows, far before LPC.

We conclude that the MBR-PSOLA is an interesting alternative to TD-PSOLA, especially in the context of multi-lingual TTS systems, for which the ability to derive segment databases automatically, to store them in a compact way, and to synthesize high quality speech with a minimum number of operations per sample is of considerable interest.

REFERENCES

[1] J.D. MARKEL, A.H. GRAY Jr, Linear Prediction of Speech, Springer Verlag, New York, pp. 10-42, 1976.

[2] D.W. GRIFFIN, J.S. LIM, 'Multi-Band Excitation Vocoder', IEEE Trans. on ASSP, vol. ASSP-36, pp. 1223-1235, august 1988.

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

[4] T. DUTOIT, H. LEICH, "An analysis of the performances of the MBE model when used in the context of a Text-To-Speech system", *Proc. EUROSPEECH 93,* Berlin, September 93, pp. 531-534.

[5] E. MOULINES, F. CHARPENTIER, "Pitch Synchronous waveform Processing techniques for Text-To-Speech Synthesis using diphones", *Speech Communication*, Vol. 9, n°5-6. 1989.

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

[7] T. DUTOIT, H. LEICH, "MBR-PSOLA : Text-To-Speech Synthesis based on an MBE Re-Synthesis of the Segments Database", *Speech Communication*, Elsevier Publisher, december 1993.

[8] T. DUTOIT, *High Quality Text-To-Speech Synthesis of the French Language*, Ph.D. dissertation, october 1993.

[9] H.J.M. STEENEKEN, E. AGTERHUIS, *Speech intelligibility ans speech quality of seven speech coders for the HERMES project*, report IZF 1992 C-41, TNO Institute for Perception.

[10] H.J.M. STEENEKEN, *On measuring and predicting speech intelligibility*, Ph.D. dissertation, 1992, TNO Institute for Perception.

[11] E.R. BANGA, E. LOPEZ-GONZALO, C. GARCIA-MATEO, "A text-to-speech system for Spanish with a frequency domain based prosodic modification algorithm", *Proc. ICASSP 93*, Vol. 2, pp. 183-186.

	LPC	hybrid H/S	TD-PSOLA	MBR-PSOLA
Analysis	automatic, easy	automatic, requires a careful design	semi-automatic (pitch marking)	automatic, requires a careful design
Coding (bit rate, database size for 3 min of speech at 16 kHz)	≈4 kBits/s ≈100 kbytes	≈10 kBits/s ≈200 kbytes	≈80 kBits/s ≈1.7 Mbytes	customized coding strategies are being tested - better than TD-PSOLA anyway
Prosody matching	trivial	simple, though not trivial	PSOLA itself	PSOLA itself
Segments concatenation	linear smoothing of PARCOR's or LSP's ≈ natural transitions of formant frequencies and bandwidths	linear smoothing ≡ fade in / fade out	poor, due to pitch, phase, and spectral amplitude mismatches	cf. hybrid H/S for voiced sounds and TD-PSOLA for unvoiced ones
Synthesis	70 operations per sample	≈100 operations per sample, with the OLA/IFFT method	7 operations per sample	5 operations per sample (7 including linear smoothing)
Modelization quality (as revealed by copy synthesis)	poor	very good	perfect (no model)	very good, cf. hybrid H/S
CVC intelligibility	low	high	almost perfect	very high
MOS fluidity	low	very high	fair	very high
MOS naturalness	low	high	very high	very high

Table 2. A comparison of the LPC, hybrid H/S, TD-PSOLA, and MBR-PSOLA segment concatenation synthesizers.