MBR-PSOLA : TEXT-TO-SPEECH SYNTHESIS BASED ON AN MBE RE-SYNTHESIS OF THE SEGMENTS DATABASE.

T. DUTOIT, H. LEICH

Faculté Polytechnique de Mons, 31, Boulevard DOLEZ, 7000 Mons, Belgique.

Tel : /32/65/374133. Fax : /32/65/374300.

ABSTRACT-: The use of the TD-PSOLA algorithm in a Text-To-Speech synthesizer is reviewed. Its drawbacks are comprehensively underlined and three conditions on the speech database are examined. In order to satisfy them, a previously described high quality re-synthesis process is developed and enhanced, which makes use of the well-known MBE model. An important by-product of this operation is that Pitch Marking turns out to be useless. The temporal interpolation block is finally refined. The resulting synthesis algorithm supports spectral interpolation between voiced parts of segments, with virtually no increase in complexity. It provides the basis of a high-quality Text-To-Speech synthesizer.

INTRODUCTION

Among the recently developed Text-To-Speech synthesis techniques, the PSOLA algorithms have drawn considerable attention, given their exceptional segmental and supra-segmental efficiency, associated with a virtually unequalled simplicity in the case of TD-PSOLA (MOULINES & CHARPENTIER, 1990). They are now widely adopted for concatenation synthesis.

Synthesized speech, however, is still not perfect. Simplicity is achieved at the expense of the ease of spectral matching between adjacent segments, a problem which is naturally solved with parametric synthesizers, as the highly efficient (and computationally expensive) MBE one (GRIFFIN, 1987). After a brief recall of the drawbacks of the TD-PSOLA algorithm (namely pitch, phase, and spectral amplitude mismatches), this paper will develop an original and efficient MBE re-synthesis operation, which happens to get rid of all of them at the same time.

THE DRAWBACKS OF TD-PSOLA.

The TD-PSOLA algorithm is a very efficient mean to alter the duration and/or intonation features of continuous speech. Its pitch modification capabilities are illustrated in fig 1, on a signal that was chosen (on purpose) perfectly periodic. It consists of summing windowed data, extracted pitch-synchronously from the original signal, and changing the time-shift between windows from the original pitch T_0 to the desired one T:

When $T=T_0$, it should be clear, from the Poisson summation formula, that the reconstructed signal is approximately proportional the original one. If $T<>T_0$, and when the original signal exhibits some strong periodicity as in our example, the operation results, according to the same formula, in a re-harmonization of the envelope of the spectrum of the original signal with the fundamental frequency $1/T$. Moreover, it is observed that, when the length of the window is changed from T_0 to 4 T_0, 2 T_0 is a surprisingly good compromise : if more, spectral lines appear in the spectrum of each frame, preventing it from being re-harmonized; if less, this spectrum gives a too coarse approximation of the envelope to re-harmonize.

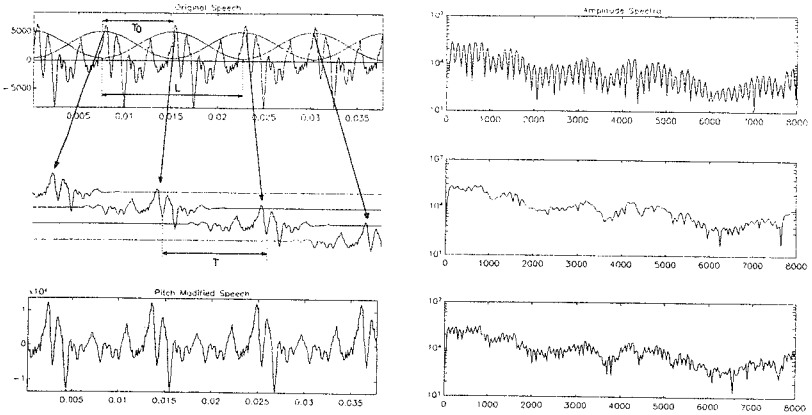
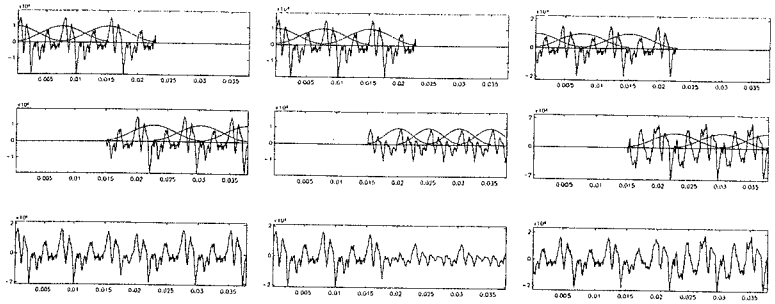


Fig 1. The TD-PSOLA Re-Harmonization process.



a. **b.** **c.**

Fig 2. *Three types of mismatches that happen to affect the quality of the waveforms produced by a TDPSOLA synthesizer when used in a Text-To-Speech environment. (vowels 'a' sampled at 16 kHz; Top = Last frames of the first segment to be concatenated; Center = First frames of the second one; Bottom = after OLA; OLA windows are positioned as they would in the TTS segments database).*

a. *Phase Mismatch: Waveforms are identical but the OLA windows are not centered on the same relative positions within the period.*

b. *Pitch Mismatch: Both segments have exactly the same spectral envelope, but were pronounced with different pitches. OLA windows are positioned coherently. The synthetic signal is obtained after changing the interval between the frames of the right segment, in order to impose a constant pitch*

c. Spectral Envelope Mismatch: The left segment is the diphone 'ma', the right one is 'am'. The pitch is constant, and the windows are again positioned coherently. The spectral discontinuity is concentrated in one period.

Things may not be the same when the algorithm is used to concatenate speech segments. As a matter of fact, the OLA operation, when performed inside a quasi-stationary segment, approximately respects the theoretical case described above (where the added windows are all identical). Problems arise tat the border of two segments extracted from different words, due to three incoherence events, respectively related to phase, pitch, and overall spectral envelope mismatches (DUTOIT & LEICH, 1992), and illustrated in fig 2 .

Phase and pitch mismatches can hardly be avoided in a database of more than one thousand of segments. Automatic procedures for coherent positioning of the pitch markers (see DI FRANCESCO & MOULINES 1989, DUTOIT & LEICH 1991, ISAKSSON & MILLNERT 1989 for a review) are computationally intensive, and suffer from a lack of precision (manual corrections are needed). Spectral envelope mismatch is well known as the most important drawback of concatenative synthesis. It is mostly due to speaker's phonemic variability and to co-articulation effects proper to any language. Parametric synthesizers can naturally cope with it. So did the frequency domain equivalent of TD-PSOLA (CHARPENTIER & MOULINES, 1988). The current implementations of TD-PSOLA can't.

MULTI-BAND RESYNTHESIS.

In order to approach the optimum case of identical OLA windows, the speech database should exhibit the following features :

1. All the words should be pronounced with a constant pitch;

2. A coherent positioning of OLA windows should be possible;

3. Spectral interpolation between concatenated windows should be easy to perform, without an excessive increase of the computational load.

This incited us to re-synthesize the complete segments database, in order to adapt it to the requirements of TD-PSOLA.

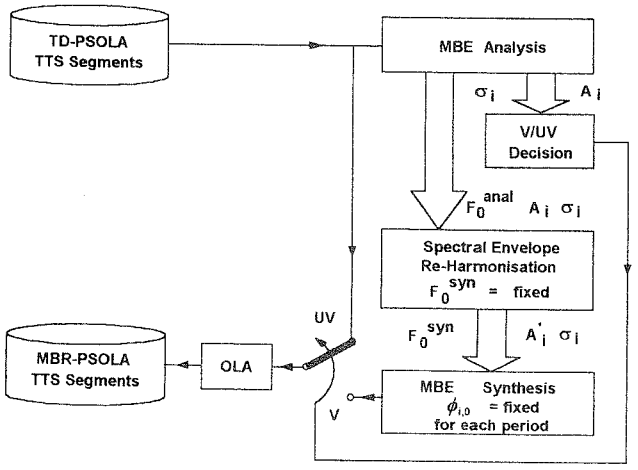


Fig 3. The MBE Re-Synthesis Operation.

The method was introduced in DUTOIT & LEICH (1992). It is based on Harmonic re-synthesis (McAULAY & QUATIERI 1986; ABRANTES & MARQUES 1992) and further enhanced by the use of voiced/unvoiced ratios during the analysis stage (Fig 3). The result is an improved TD-PSOLA synthesis scheme, called MBR-TDPSOLA (Multi-Band Re-synthesis Time-Domain Pitch Synchronous Overlap Add), which combines the computational efficiency of the original algorithm with the flexibility of the MBE model . It is performed once for all on the voiced frames of the segments database, with almost no increase in complexity during the TTS synthesis itself.

Pitch and phase mismatches are eliminated thanks to a pitch-constant re-synthesis, with harmonic phases reset to well chosen values at the beginning of each pitch period, so that concatenated waveforms are maximally similar to one another. An important by-product of this operation is that Pitch Marking turns out to be useless.

The quality of the resulting synthetic speech is highly dependent on the phase reset strategy used during re-synthesis. It is well known that constant or linearly distributed harmonic phases leads to somewhat metallic sounds. In DUTOIT & LEICH 1992, phases were given random (but fixed) initial values, resulting in a higher quality synthetic voice. An informal study of the influence of the initial pitch values on the final TTS quality led us recently to keep the original phases for high frequency harmonics. The threshold frequency was chosen as a compromise : if too high, no enhancement was perceived; it too low, the generated waveforms were no more under control, and maximal similarity of concatenated waveforms could not be preserved. It turned out that it was possible to satisfy both of these constraints with a threshold frequency of approximately 2000 Hz.

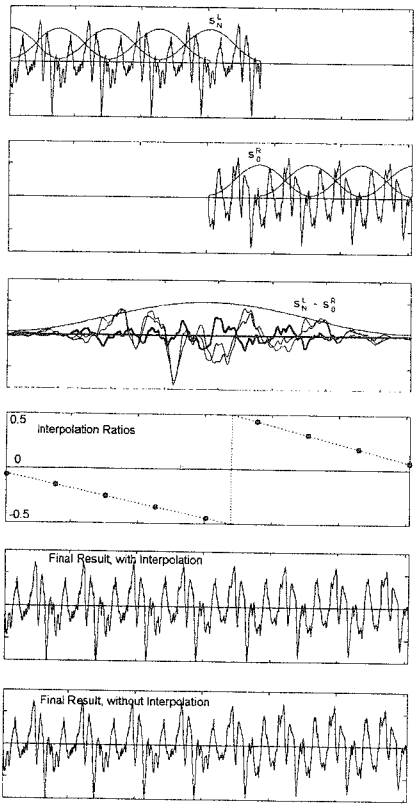


Fig 4. The Variable Length Linear Interpolation Process. $M_L = 5$; $M_R = 4$

What 's more, a crucial advantage of the phase reset, when combined with a constant re-synthesis pitch, is that spectral envelope interpolation becomes equivalent to direct temporal interpolation (temporal interpolation of signals is always equivalent to their complex spectral interpolation). This enables the TTS synthesizer to fulfil condition 3 by performing the desired spectral matching directly on neighbouring segments' waveforms, prior to the OLA itself.

The linear interpolation algorithm itself was still recently enhanced through the use of a segment-dependant number of interpolated periods (as opposed to the previous fixed length associated with yes/no decisions). This avoids the interpolation to be performed on transient parts, where it is neither desired nor suited. When no such control was performed, some vowels tended to be heard as two simultaneous sounds. Voiced stable states in each segment are automatically detected, from the V/NV energy ratio computed during the MBE complex least squares analysis stage. Possibly interpolated frames at the beginning and ending of each segment are restricted to this class.

If s^l_N and s^r_0 respectively denote the last and first OLA windows of two concatenated segments (the left and the right one), we thus modify a certain amount M_L and M_R of OLA

windows before s^l_N and after s^r_0 according to (Fig 4):

$$s^L_{N-i} = s^L_{N-i} + (s^R_0 - s^L_N)\frac{1}{2}\left(\frac{M_L - i - 0.5}{M_L}\right)$$

$$s^R_j = s^R_j + (s^L_N - s^R_0)\frac{1}{2}\left(\frac{M_R - j - 0.5}{M_R}\right)$$

for i=0...M_L-1 and j=0...M_R-1

Spectral envelope interpolation may thus be performed through a very simple time-domain linear interpolation of the voiced pitch periods. This may be seen as a (much simpler) temporal equivalent to the spectral linear interpolation proposed in CHARPENTIER & MOULINES (1988).

The resulting MBR-PSOLA algorithm provides the basis of the High Quality Text-To-Speech Synthesizer under development at the Faculté Polytechnique de Mons.

The presentation will include listenings of the final high quality synthetic speech.

REFERENCES

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

CHARPENTIER, F. & MOULINES, E. (1988), "TTS algorithms based on FFT synthesis", ICASSP 88, pp. 667-670.

ABRANTES, A.J.& MARQUES, J.S (1992), "Hybrid Harmonic Coding of Speech", EUSIPCO 92, 25-28 august 92, Brussels, pp. 487-491.

McAULAY, R.J. & QUATIERI, T.F. (1986), "Speech Analysis /Synthesis based on a sinusoidal representation", IEEE Trans. on ASSP, vol 34, August 86, pp. 744-754.

GRIFFIN, D.W. (1987), Multi-Band Excitation Vocoder, Ph.D. Thesis, MIT.

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.

DI FRANCESCO, R. & MOULINES, E. (1989), "Detection of the glottal closure by jumps in the statistical properties of the signal", Eurospeech 89, Vol. 2, pp. 39-41.

ISAKSSON & MILLNERT (1989), "Inverse Glottal Filtering Using a Parameterized Input Model", Signal Processing, n°18, pp. 435-445.

DUTOIT, T. & LEICH, H. (1991), "Glottal Inverse Filtering : geometrical Interpretation leading to a faster algorithm", Proc. IEEE/ProRisc Symposium on Circuits, Systems, and Signal Processing, Veldhoven, april 1991, pp. 151-157.

SYNTHESISING FACIAL MOVEMENT: DATA CAPTURE

R.E.E.Robinson
Speech Hearing and Language Research Centre
School of English and Linguistics
Macquarie University

ABSTRACT - A method of data capture and analysis of lip movement is described. Cine film was digitised by computer and a frame by frame comparison was done to generate a codebook of lip shapes for the synthesis of facial movement. By various methods of data reduction, real-time image playback was achieved.

INTRODUCTION

The Speech Hearing and Language Research Centre (SHLRC) at Macquarie University has a project under way which will synthesise facial movements in concert with a speech synthesiser. This is a project to produce a realistic talking image coupled with a voice using the unlimited vocabulary Text To Speech (TTS) system also under development at SHLRC. The TTS system takes as input any typed text and has a set of inbuilt rules that provide a sequence of parameters, to control a speech synthesiser. The same sequence will be used to control a facial synthesiser when this project is complete.

The current approach uses a natural image that has been captured on film at 100 frames per second, digitised with a frame grabber on a Sun 4/330 computer under SunOs V4.1.1 and Version 2 of Openwindows. The frame grabber is a Data Translation DT1451 which provides 512x512 resolution with a 256 level grey scale, and uses the DT IRIS subroutine library with control programs written in the C programming language. The monitor is a Mitsubishi C-3910 RGB monitor and the video camera is a colour Hitachi FP7.

To overcome some of the computing limitations imposed by current technology, the disc storage was reduced, and synthesis speed increased by Area Of Interest (AOI) processing and frame by frame comparison, subtraction and storage. The captured images were analysed to look for the rapid movement associated with some fricatives. This method produced a large amount of redundant data for other slower articulations, especially steady state vowels.

The captured data has been partially analysed, and will form the beginning of a data base that will contain lip positions that correspond to speech. A code book will be devised to correlate the lip positions to the controlling parameters coming from the TTS controlling program. The code book will be manageable as estimates of different lip shapes vary from 15 (Storey and Roberts, 1988) to 21 (Benoit et al 1992).

FILM

The original cine film was shot on black and white film (AGFA GEVAERT GEVAPAN 36) with a Bolex 16mm cine camera that was modified to expose a new frame every 10ms (100 frames per second). A purpose built digital counter using high output Light Emitting Diodes (LEDs) was constructed and synchronised to the camera such that each frame had a unique number visible in the corner. The voice was recorded on audio cassette. The subject was prompted by cards which were placed on a floor stand and then revealed as each new word was to be spoken. A mirror was placed in the field of view so that the cue cards could be seen (in reverse) to show the viewer which word was being spoken. A head restraint allowed the subject to lean their forehead against it in a comfortable and stable position. A side mirror was angled at 45 degrees to show a profile view of the subject. A graduated rule was fixed next to the subjects mouth to allow scaling if required, and arranged to be visible in both views. The camera was arranged to show the front and side of the subjects face and moved as close as possible to fill the frame. The film was shot as a component part of a physiology study yet to be published in its entirety. At the same time the film was being shot, the subjects had EMG electrodes attached to several facial muscles which allowed the speech, muscle, and frame count to be recorded on computer magnetic tape. The speech consisted of 7 consonants (P B M S Z ZH SH) and 3 vowels