SPEECH SYNTHESIS SYSTEM BASED ON A VARIABLE DECIMATION/INTERPOLATION FACTOR

F. M. Giménez de los Galanes, M.H. Savoji† and J. M. Pardo
Dpto. Ingeniería Electrónica. U. Politécnica de Madrid.
E.T.S.I. Telecomunicación. Ciudad Universitaria. 28040. Madrid. SPAIN.
†.Dpto. Electrónica. Universidad de Cantabria. 39005. Santander. SPAIN.

ABSTRACT

In this paper we present a modification of the usual decimation-interpolation steps for resampling of speech signals which is especially adapted to arbitrary modification of fundamental frequency and duration of speech segments. The modification is intended to overcome the time and frequency domain limitation that such a resampling scheme imposes so it can be used in a speech synthesis system. The performance of this resampling method for prosody modification is better than the equivalent PSOLA (Pitch-Synchronous Overlap-Add) method when working at a sampling frequency of 8 to 10 kilohertz so the source spectrum of the voiced allophones can be said to be completely harmonical. An optimization of the proposed algorithm that allows a real time implementation is also discussed.

1. THE SPEECH SYNTHESIS SYSTEM

The speech is produced by unit concatenation. The units are diphone-type such as diphones, triphones, etc., and some subphone units. This corpus was extracted from a set of non-sense words (logatomes) and the PCM waveform is stored in about 2.5 Mbytes. In this paper the linguistic module is not discussed in detail, and it is assumed that the set of units with their prosodic parameters are available.

The units database has been preprocessed in order to mark the pitch epoch. In the unvoiced regions of the speech, these marks were positioned every 10 ms. while in the voiced regions where positioned at the closure of the glottis. This was done automatically following a procedure based of the maxima of the Hilbert envelope [1]. As we will see, these marks are used as reference points for the analysis module and for the prosody modification subsystem.

The database was synchronously analyzed and for every speech unit, the set of synchronous Linear Prediction coefficients and the residual signal after the LPC analysis filter were stored. Therefore, our unit corpus is built of three blocks: the residual samples, the linear prediction coefficients and the pitch marks.

The Prosody Modification module has to be able to modify the pitch and duration of the processed units independently and arbitrarily while preserving the formant structure of speech. These pitch modifications vary within the unit: the fundamental frequency can follow a rising or falling line -more complex functions are not used.

2. THE DEVELOPED METHOD

The developed scheme is similar in some ways to the system proposed in [2]. The main idea is to modify the pitch by interpolation of the speech signal, i.e. by expanding or compressing the speech waveform as required. Here the term interpolation refers to both up and down sampling needed for the expansion and compression of the waveform. In order to preserve the formant structure of the original speech, the interpolation is performed on the stored residual signal after LPC analysis. Once the residual is modified, the spectral envelope responsible for the formant structure will be superimposed by LPC synthesis filtering.

We tried three different configurations for our system: 1) an FD-PSOLA-like approach, where the windowed signals are modified, 2) an LP-PSOLA-like approach, modifying the windowed residual and 3) what we call the direct approach. No differences were found between these three systems so the simplest one (the direct approach) was therefore implemented. The three systems are outlined in Figure 1.

0-7803-2431-5/95 $4.00 © 1995 IEEE

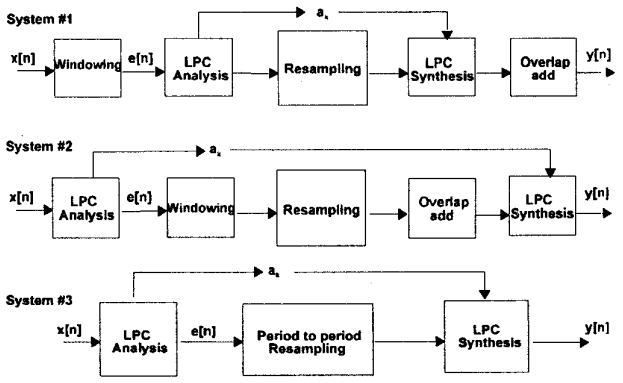

Figure 1. Block diagram of the three possible configurations for the proposed method: 1) FD-PSOLA-like, 2) LP-PSOLA-like and 3) direct method.

The interpolation of the residual has the same spectral considerations as a change in the sampling rate by a fractional factor, which can always be decomposed into two integer numbers: the number of original samples, and the number of interpolated samples. The interpolation problem is reduced to a resampling problem.

In [2] and [3], the process is pitch-asynchronous working on fixed-length frames, and it is therefore limited in quality. In our work, the LPC analysis and the resampling is done pitch-synchronously, taking advantage of the pitch marks obtained in the preprocessing. A set of LPC coefficients is associated to each pitch mark. The LPC frame is defined from the center point between two consecutive pitch marks (i, i+1) to the center point of the next consecutive marks (i+1, i+2), as shown in Figure 2, while the resampling frame is defined between pitch marks.

Control of the duration will be accomplished through duplication or elimination of pitch periods in the signal, in a similar way as in the PSOLA algorithm [4]. What is duplicated or eliminated is not only the residual signal associated to a signal period, but the associated set of Linear Prediction coefficients as well.

The pitch period to be eliminated or duplicated is determined by a linear mapping between the analysis and synthesis time axes. These axes contain the positions of the original and synthetic pitch marks. The portion of signal considered as a period is the portion between two consecutive marks, but the resampling module needs to know several samples from before the starting marks and several samples from beyond the second mark, in order to perform the filtering in a non-causal way to avoid the causal FIR filter delay. This derives from the fact that we are performing joint pitch-duration modification of the speech signal.

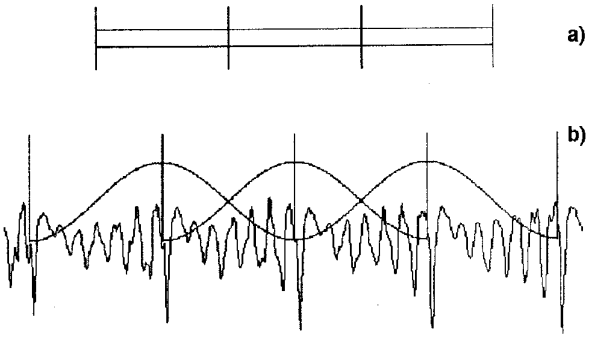

Figure 2. a) LPC frames considered in the analysis. b) A segment of speech with pith marks showing what samples are considered in the analysis (windowed).

There is no overlap between consecutive synthesized periods: they are concatenated directly as obtained after the resampling module. That means that a discontinuity may appear in the synthetic waveform when duplicating or eliminating a period, but the relative importance of these discontinuities is negligible since they are smoothed by the LPC synthesis filtering and, also, this discontinuity appears at the beginning of a pitch period, where there is a strong discontinuity due to the pulse-like nature of the excitation.

3. DECIMATION/INTERPOLATION MODULE

There is an established method for changing the sampling rate by a fractional factor $f=L/M$. This method has three steps: first we have to oversample the signal by a factor L. Then we need a low pass filtering stage with cut-off frequency equal to the minimum of π/L and π/M. The last step is a downsampling by the factor M.

Within this framework, the modification of the fundamental frequency would be done by resampling the residual by a factor >1 when lowering the pitch and by a factor <1 when raising it. In the case of raising the pitch, the filter (π/M) included in the resampling module just filters out the frequencies

which would cause aliasing, but in the case of lowering the pitch, the filter (π/L) reduces the information in the high frequencies area, resulting in a spectrum with almost no high frequency energy.

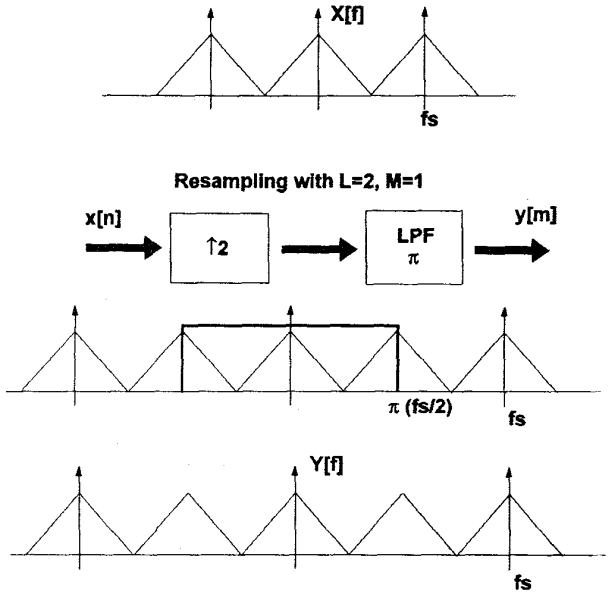

Figure 3. Resampling with a π/M cut-off frequency filter in the case of L=2 and M=1.

We propose a modification of the resampling chain. The low pass filter is not different in the case of factors bigger or smaller than one. By forcing this filter to have a cut-off frequency of π/M, some information of the alias spectrum will fill the empty area in the high frequencies when lowering the pitch and filters out the undesired harmonics when raising it. This is illustrated in Figure 3. The scheme is similar to duplicating the upper frequency spectrum in the high frequency area or discarding the high frequency information as proposed for the frequency domain PSOLA method [4].

4. FAST FORMULATION

It has been shown that the interpolation can be carried out by a FIR filter designed by windowing, as in [5]. As a matter of fact, depending on the value of the fractional factor f, the up/down sampling processes may be cumbersome as L and M could be large integers. In this case it is easier to resort to a *sinc* interpolation scheme combined with a simple windowing to limit the length of the FIR filter that carries out the *sinc* interpolation. This scheme is also more versatile because we can easily modify the filter for every different resampling factor. The window used is a 41 point Hamming window, i.e., it is a zero-centered, N=20, Hamming window. With this filter, the output samples of the resampling chain will be:

$$y(m)=\sum_{n=-N}^{N} x\left(n+\left\lfloor \tfrac{mM}{L} \right\rfloor\right)\operatorname{sinc}\left(\frac{mM-\left(n+\left\lfloor \tfrac{mM}{L} \right\rfloor\right)L}{M}\right)\cdot \operatorname{ham}\left(\frac{mM-\left(n+\left\lfloor \tfrac{mM}{L} \right\rfloor\right)L}{M}\right)$$

The window function *ham(·)* can be approximated by forcing it to be centered at the closest precedent original sample instead of being centered at the synthesis time point, which is equivalent to making $mM-nL=0$, without an important decrease of the performance. The window is then tabulated, so the main computation left is that of the *sinc* functions:

$$y(m)=\sum_{n=-N}^{N} x(n+n_o)\frac{\sin\pi(m-(n+n_o)L/M)}{\pi(m-(n+n_o)L/M)}\operatorname{ham}(n)$$

where $n_0=\left\lfloor \tfrac{mM}{L} \right\rfloor$ and we can see that (without the term n_o, for simplicity) :

$$\sin\pi(m-n\frac{L}{M})=$$

$$\sin\pi m\cos\frac{\pi L}{M}n-\cos\pi m\sin\frac{\pi L}{M}n$$

where

$$\cos\pi m=\begin{cases}+1\\-1\end{cases};\ and\quad \sin\pi m=0$$

so, the output sample can be written now as:

$$y(m)=\sum_{n=-N}^{N}\frac{(-1)^m x(n+n_0)M\sin\frac{\pi L}{M}(n+n_0)}{\pi(mM-L(n+n_0))}\cdot \operatorname{ham}(n)$$

where the denominator can also be tabulated in the form of the $1/\pi x$ function to avoid the division. Notice that every original sample is always multiplied by the same *sin* function. This property allows fast computation because there are a finite number of *sin* functions to be precomputed (every time we change the resampling factor).

5. SYSTEM INTEGRATION

This method of pitch modification can be integrated in a synthesis system where the signal is coded using a LPC-based method (CELP, MLPC). The reconstructed residual is modified by resampling prior to the envelope filter. However we report results obtained with residual used directly that will not mask the possible distortions.

In our system we have already integrated a spectral modification module for inter-unit smoothing [6]. This module could also be used for arbitrary spectral modification such as joint pitch-formant modification, etc., once the modifications rules are obtained, or for voice conversion.

6. RESULTS

We have carried out an evaluation test to compare the developed method with an equivalent PSOLA (reference) system.

The units where recorded originally at 16 Khz, so it was necessary to filter them to a bandwidth of 4 Khz and to decimate them afterwards. The filter used to low pass filter the unit corpus was designed to have a very narrow transition band, allowing the harmonic content of the signal to be preserved in almost the whole range of frequencies. In practice, this approach is necessary in order to avoid the "empty" high frequency spectrum obtained when recording directly at 8 Khz and using an analogue bandwidth-limiting filter. We think that this effect was partly responsible for the poor results reported by Moulines [3], but it can be easily removed and it should not be the limiting factor.

As we introduced in preceding sections, these units where LPC synchronously analyzed and filtered (where the frames were selected as in Figure 2) and the residual signal was stored. The PSOLA module, as well as the resampling module, would operate over this residual signal, in an LP-PSOLA scheme.

The test corpus were four sentences containing at least an example of every class of phonemes. For every sentence, two utterances (one by each system under test) were recorded and presented to the listeners in a random order. The listener had three options: no preference, first system or second system. The results of this preliminary test carried out on 16 native listeners indicated a slight preference of the resampling system over the LP-PSOLA system.

RESAMPLING	LP-PSOLA	NO PREFERENCE
41%	29%	30%

Table 1. Results of the subjective test comparing two prosody modification modules: LP-PSOLA and the proposed RESAMPLING method.

The quality of the synthesized speech obtained by this system is comparable to the quality obtained by FD-PSOLA modification methods, but it is much faster since spectral calculations are avoided. The distortion due to the resampling module is exactly the same as the distortion that appears in the frequency domain Compression-Expansion PSOLA techniques, and can be neglected in a text-to-speech application with telephone bandwidth.

REFERENCES

[1] F. M. Giménez de los Galanes, M. H. Savoji, J. M. Pardo. 1993. "Marcador automático de excitación glotal." *Proc. of URSI'93: 189-193. Valencia.*

[2] Stephanie Seneff. 1982. "System to Independently Modify Excitation and/or Spectrum of Speech Waveform without Explicit Pitch Extraction." *IEEE Trans. on ASSP, Vol 24, No. 4: 358-365.*

[3] Eric Moulines. 1990. *Algorithmes de codage et de modification des parametres prosodiques pour la synthèse de parole a partir du texte.* Thèse de Doctorat, ENST, Paris.

[4] Francis Charpentier. 1988. *Traitment de la parole par analyse-synthèse de fourier. Aplication à la synthèse par diphones.* Thèse de Doctorat, ENST, Paris.

[5] M. H. Savoji. 1987. "Resampling: A Software Implementation." *R18/G163 British Telecom Research Lab. Technical Report.*

[6] F. M. Giménez de los Galanes, M. H. Savoji, J. M. Pardo. 1994. "New Algorithm for Spectral Smoothing and Envelope Modification for LP-PSOLA Synthesis." *IEEE Proc. of Intern. Conf. on ASSP. Adelaide.*