CONCATENATIVE TEXT-TO-SPEECH SYNTHESIS BASED ON PROTOTYPE WAVEFORM INTERPOLATION (A TIME FREQUENCY APPROACH)

Edmilson S. Morais *Paul Taylor* *Fábio Violaro*

CSTR / University of Edinburgh
2 Buccleuch Place Edinburgh EH8 9LW, UK
emorais@cstr.ed.ac.uk

Department of Eletrical Engineering
Universidade Estadual de Campinas
Campinas, SP, Brazil

ABSTRACT

This paper presents some preliminary methods to apply the Time-Frequency Interpolation technique - TFI [3] to concatenative text-to-speech synthesis. The TFI technique described here is a pitch-synchronous time-frequency approach of the well known Prototype-Waveform Interpolation technique - PWI [2]. The basic concepts of representing the speech signal in the Time-Frequency Domain as well as techniques to perform Time-Scale and Pitch-Scale modifications are described. Using the flexibility of TFI technique to perform spectral smothing, a method was developed to minimize the spectral mismatch at the boundaries of the Synthesis-Units - SUs. The proposed system was evaluated using SUs (Diphones) and prosodic modifications generated by the Festival system [1]. An informal subjective test was performed, between the proposed TFI system and the standard TD-PSOLA system, highligthing the superior quality of the proposed system in comparasion with TD-PSOLA.

1. INTRODUCTION

Some of the basic operations in a Concatenative Text-To-Speech - TTS - synthesis system are the concatenation and prosodic modifications of the Synthesis-Units - SUs. However, in order to guarantee good concatenations and prosodic modifications the system has to be able to perform the following three operations : (1) Spectral smoothing at SUs boundaries. (2) Time Scale Modifications -TSM, independently from Pitch Scales Modification - PSM, and vice-versa. (3) Continuous amplitude modifications of SUs through time.

The TD-PSOLA is the technique most used in commercial Concatenative TTS synthesis systems. However TD-PSOLA presents some drawbacks, mainly under large prosodic variations : (1) PSM introduces simultaneous TSM which needs to be appropriately compensated. (2) TSM can only be implemented in a quantized manner, with a resolution of one pitch period (...,1/2,3/4,...,4/3,3/2,2,...). (3) Duration lengthening for unvoiced segments introduces a periodic component that is reponsable for a metalic-like sounding of the synthesized speech. (4) The Overlap and Add procedure at the boundaries of the SUs does not guarantee a good spectral smoothing.

In order to overcome some drawbacks of the TD-PSOLA, this paper presents a method based on Time Frequency Interpolation - TFI [3]. The TFI method introduced here is a Pitch-Synchronous Time-Frequency approach of the well known Prototype Waveform Interpolation technique - PWI [2]. The goal of this paper is to show that the TFI technique presents some important advantages to concatenative TTS synthesis. It allows PSM independently of TSM, in a quite straightforward manner and with high-quality. TSM and PSM can be done in a continuous way, without any limitation of pitch period resolution. Moreover, the TFI technique allows simple, flexible and efficient procedures to smooth diphone (or any other kind of unit) boundaries.

Section 2 presents the basic concepts of signal representation on Time-Frequency domain. Section 3 describes the Analysis and Re-Synthesis TFI procedures. The prosodic modifications based on TFI synthesis are discussed in Section 4. Section 5 describes one possible technique to perform Spectral Smothing at the boundaries of the SUs. Finally, Sections 6 and 7 present the results of an informal subjective test and some final considerations.

2. TFI FRAMEWORK

In this paper, the representation of a discrete signal $\mathbf{x} = \{x(0), x(1), ..., x(N-1)\}$ in the time-frequency domain is based on the concept of short-time $per-sample$ discrete spectrum sequence [3]. Each time n on a discrete-time axis is associated with an $M(n)-point$ discrete spectrum $\mathbf{X}_n = \{X_n(0), X_n(1), ..., X_n(M(n)-1)\}$. This representation defines a sequence of spectra $\mathcal{X} = \{\mathbf{X}_0, \mathbf{X}_1, ..., \mathbf{X}_{N-1}\}$. The $n-th$ spectrum is calculated projecting the sub-sequence $\mathbf{x}_n^{n+M(n)-1} = \{x(n), x(n+1), ..., x(n+M(n)-1)\}$ over the base functions of a specific operator $\mathbf{T}\{\cdot\}$

$$\mathbf{X}_n = \mathbf{T}\{\mathbf{x}_n^{n+M(n)-1}\} \tag{1}$$

In this paper, the spectrum $\mathbf{X}_n$ is calculated using the Discrete Fourier Transform ($\mathbf{DFT}$) operator :

$$\mathbf{X}_n = \sum_{m=n}^{n+M(n)-1} \mathbf{x}(m) \cdot \exp^{-j\cdot\omega(n)\cdot k\cdot(m-n)} \tag{2}$$

where

$$\omega(n) = \frac{2\cdot\pi}{M(n)} \tag{3}$$

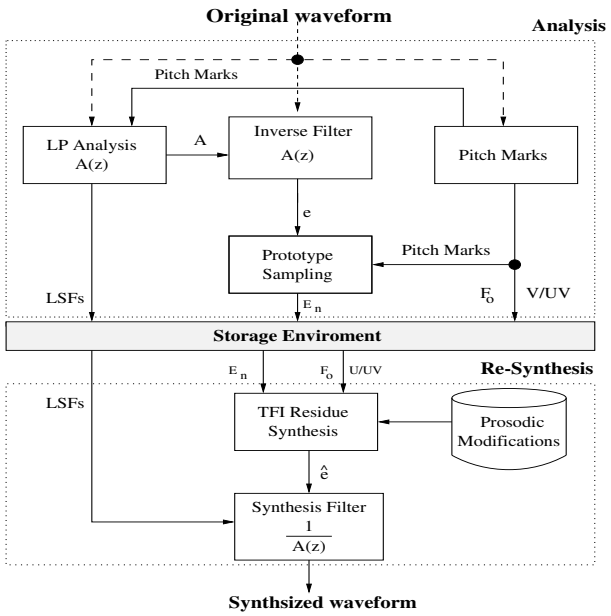

Figure 1: Diagram of the proposed system : Analysis and Re-Synthesis.

We will denote by $\mathbf{T}^{-1}\{\cdot\}$ the inverse operator that enables to transform the sequence of spectra $\mathcal{X}$ into the original signal $\mathbf{x}$.

3. ANALYSIS AND RE-SYNTHESIS

This section will described the procedures of Synchronous Analysis and Re-Synthesis using TFI. Figure 1 presents the block diagram of these two stages.

3.1. Analysis

Pitch-Marks The first step in the speech analysis is to provide a sequence of pitch-marks and a Voiced/Unvoiced classification for each segment between two consecutives pitch-marks. In this work this task was solved in a very precise way using the corresponding laryngograph and differencial laryngograph signals. Figure 2 shows the pitch-marks in a voiced segment. The distance between n_i and n_j pitch-marks will be called N_{ij}. In the original signal N_{ij} is equal to the fundamental period T_{o_i}. However, after prosodic modifications N_{ij} may be different from T_{o_i} as will be seen in Section 4. The fundamental frequency associated to the n_i pitch-mark is then $F_{o_i} = 1/T_{o_i}$.

Linear Prediction Model and Inverse Filter The proposed system performs a Linear Prediction - **LP** - analysis around each pitch-mark using an asymmetrical Hamming window starting at the previous pitch-mark and finishing at the next one, as shown in Figure 2. The LP coefficients are transformed into Line Spectral Frequency Coefficients - **LSF** and for each two consecutives pitch-marks, n_i and n_j, the corresponding pair of **LSFs** are upsampled to $8 \cdot F_{o_i}$ by linear interpolation (one **LSF** for each $N_{ij}/8$ samples; this set of 8 **LSF** will be denoted by $\mathcal{LSF}_{n_i}$). Using these interpolated **LSFs** ($\mathcal{LSF}_{n_i}$) new LP coefficients are derived

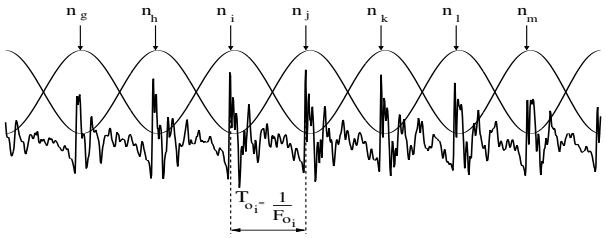

Figure 2: Indication of the pitch-marks and the postions of the LP window analysis

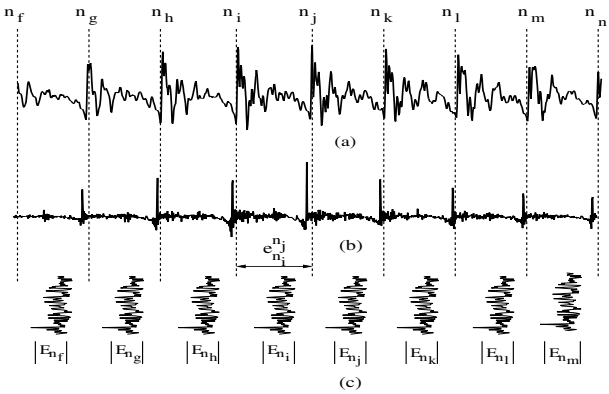

Figure 3: (a) - Speech signal. (b) Extration of the prediction error subsequences $\mathbf{e}_{n_i}^{n_i+T_{o_i}-1}$. (c) - Prototypes $\mathbf{E}_n$

and used to filter the speech signal $\mathbf{x}$, generating the Linear Predicion Error (residue) $\mathbf{e}$.

Sampling of Prototypes As shown in Figure 3, at each pitch-mark n_i a sub-sequence of the linear prediction error must be taken $\mathbf{e}_{n_i}^{n_i+T_{o_i}-1} = \{e(n_i), e(n_i+1), ..., e(n_i+T_{o_i}-1)\}$ and its spectrum $\mathbf{E}_{n_i} = \mathbf{T}\{\mathbf{e}_{n_i}^{n_j}\}$ must be calculated. This procedure defines a sequence of spectra $\bar{\mathcal{E}} = \{...\mathbf{E}_{n_h}, \mathbf{E}_{n_i}, \mathbf{E}_{n_j}, ...\}$, where

$$\mathbf{E}_{n_i} = \sum_{m=n_i}^{n_i+T_{o_i}-1} e(m) \cdot \exp^{-j \cdot \omega(n_i) \cdot k \cdot (m-n_i)} \quad (4)$$

with

$$\omega(n_i) = \frac{2 \cdot \pi}{T_{o_i}} \quad (5)$$

Defining $\mathcal{E} = \{\mathbf{E}_0, \mathbf{E}_1, ..., \mathbf{E}_N - 1\}$ as the Time-Frequency representation of the whole linear prediction error $\mathbf{e} = \{e(0), e(1), ..., e(N-1)\}$, thus the sequence of spectra $\bar{\mathcal{E}} = \{...\mathbf{E}_{n_h}, \mathbf{E}_{n_i}, \mathbf{E}_{n_j}, ...\}$ can be understood as a Pitch-Synchronous time-domain decimation of the sequence of spectra $\mathcal{E}$ (where only the spectra at the pitch-mark positions were mantained). Henceforth, each spectrum $\mathbf{E}_n \in \{...\mathbf{E}_{n_h}, \mathbf{E}_{n_i}, \mathbf{E}_{n_j}, ...\}$ will be called **prototype**. Briefly, for the $n-th$ pitch-mark, the following parameters must be stored for a $posteriori$ decoding : (1) Prototype - $\mathbf{E}_n$, (2) Voiced and Unvoiced decision - $\mathbf{V/UV}$, (3) Fundamental frequency - $F_o = \frac{1}{T_o}$ and (4) **LSF** coefficients - $\mathbf{LSF}$.

3.2. Re-Synthesis

The **TFI** synthesis assumes a slow evolution of the spectra and reconstructs the full time-frequency representation $\mathcal{E}$ by calculating a linear interpolation of the sequence of prototypes $\bar{\mathcal{E}} = \{...\mathbf{E}_{n_h}, \mathbf{E}_{n_i}, \mathbf{E}_{n_j}, ...\}$. However, in order to compute the interpolation of prototypes between $\mathbf{E}_{n_i}$ and $\mathbf{E}_{n_j}$, it is necessary that these two spectra have the same number of samples. This problem can be solved appending zeroes at the end of the prototype with fewer samples. This increasing of zeroes is associated with a compression in frequency domain and consequently an expansion (interpolation) in time domain. By appending zeroes, the duration of two consecutive prototypes $\mathbf{E}_{n_i}$ and $\mathbf{E}_{n_j}$ becames equal to $M_{ij} = Max\{T_{o_i}, T_{o_j}\}$.

After adjusting the duration of the prototypes, each spectrum $\mathbf{E}_n \in \mathcal{E}_{n_i}^{n_j} = \{\mathbf{E}_{n_i}, \mathbf{E}_{n_i+1}, ..., \mathbf{E}_{n_j-1}\}$, can be obtained by linear interpolation of the prototypes $\mathbf{E}_{n_i}$ and $\mathbf{E}_{n_j}$.

$$\mathbf{E}_n \approx \hat{\mathbf{E}}_n = \mathbf{I}_n\{\mathbf{E}_{n_i}, \mathbf{E}_{n_j}\} = \alpha(n) \cdot \mathbf{E}_{n_i} + \beta(n) \cdot \mathbf{E}_{n_j} \quad (6)$$

In order for Equation 6 to be considered as a linear interpolation, the real and imaginary parts of $\mathbf{E}_{n_i}$ and $\mathbf{E}_{n_j}$ are interpolated separately. However, in a linear interpolation $\alpha(n)$ and $\beta(n)$ do not need to be linear functions. In this work linear interpolation functions were used :

$$\alpha(n) = \frac{1 - (n - n_i)}{N_{ij}} \;\; and \;\; \beta(n) = 1 - \alpha(n) \quad (7)$$

With the spectra $\hat{\mathbf{E}}_n$ calculated using linear interpolation, an estimation of the vector $\mathbf{e}_{n_i}^{n_j} = \{e(n_i), e(n_i+1), ..., e(n_j-1)\}$ can be obtained using the following inverse transformation :

$$\hat{e}(n) = \begin{cases} \frac{\hat{\mathbf{E}}_n(0) + \hat{\mathbf{E}}_n(\frac{M_{ij}}{2}) \cdot (-1)^n}{M_{ij}} + \sum_{k=1}^{(\frac{M_{ij}}{2}-1)} C_k & M_{ij} \text{ even} \\ \frac{\hat{\mathbf{E}}_n(0)}{M_{ij}} + \sum_{k=1}^{(\frac{M_{ij}-1}{2})} C_k & M_{ij} \text{ odd} \end{cases} \quad (8)$$

where

$$C_k = \frac{2}{M_{ij}} \cdot \Re\{\hat{\mathbf{E}}_n(k) \cdot \exp^{j \cdot \varphi(n) \cdot k}\} \quad (9)$$

In this case $\varphi(n)$ is defined as :

$$\varphi(n) = [(\alpha(n) \cdot \frac{2 \cdot \pi}{T_{o_i}} + \beta(n) \cdot \frac{2 \cdot \pi}{T_{o_j}}) + \Phi_i] \cdot n \quad (10)$$

with $\alpha(n)$ and $\beta(n)$ given by Equation 7 and the phase function Φ given by :

$$\Phi_i = \Phi_h + \frac{2 \cdot \pi \cdot N_{ij}}{T_{o_j}} \quad (11)$$

where Φ_h is the phase associated to n_h pitch-mark (previous pitch-mark).

Equation 10 assures a linear transition, sample-by-sample, from the fundamental frequency of prototype $\mathbf{E}_{n_i}$ to the fundamental frequency of prototype $\mathbf{E}_{n_j}$. The phase function expressed in Equation 11 is necessary to assure the continuity of the speech signal.

Because the **LSF** coefficients were upsampled during the analysis procedure, then they also have to be upsampled before filtering the residue. Moreover, in order to be consistent with the interpolation of the prototypes $\mathbf{E}_{n_i}$ and $\mathbf{E}_{n_j}$, the corresponding sets of Line-Spectral Frequencies $\mathcal{LSF}_{n_i}$ and $\mathcal{LSF}_{n_j}$ (8 for each pitch-mark) have also to be linearly interpolated using the same linear functions $\alpha(n)$ and $\beta(n)$. The 8 new interpolated **LSFs** for each frame are transformed into **LP** coefficients and then used to synthesize the speech signal.

4. PROSODIC MODIFICATIONS

4.1. TSM - Time Scale Modifications

Time-Scale Modifications (TSM) can be performed simply by changing the original $time - position$ of the prototypes. No modification has to be done through Frequency-Domain. For example, the original distance between the prototypes $\mathbf{E}_{n_i}$ and $\mathbf{E}_{n_j}$ was defined as $N_{ij} = T_{o_i}$. However, after a TSM with a factor ν, the new distance N_{ij} will be equal to $\nu \cdot T_{o_i}$. Therefore, in order to preserve the continuity of the phase function, the new value of N_{ij} has to be used in the Equation 11.

This procedure can be performed for both voiced and unvoiced segments. However, a TSM bigger than 2 can produce some metalic-like sounding in the unvoiced segments. A good solution to minimize this problem is the REW (Rapidly-Evolving Waveform) and SEW (Slowly-Evolving Waveform) decomposition proposed in [2]. Using this approach the REW component can be upsampled through time-domain before interpolation. This new REW presents a slower evolution through time and can be better interpolated using the TFI Technique.

4.2. PSM - Pitch Scale Modifications

During Pitch-Scale Modifications (PSM) the original $time - positions$ of the prototypes must be preserved. However, the prototypes must be changed by resampling their envelopes at the positions of the new Pitch-Frequencies. For example, the original Pitch-Frequency of the prototype $\mathbf{E}_{n_i}$ is $Fo_i = \frac{1}{To_i}$. However, after a PSM with a factor of $\frac{1}{\rho}$, the new Pitch-Frequency will be equal to $\frac{Fo_i}{\rho} = \frac{1}{\rho \cdot To_i}$. Therefore, the new prototype must have samples taking place at multiples of this new fundamental frequency. Moreover, after a PSM with a factor of $\frac{1}{\rho}$, the Equation 11 must be changed to :

$$\Phi_i = \Phi_h + \frac{2 \cdot \pi \cdot N_{ij}}{\rho \cdot T_{o_j}} \quad (12)$$

The PSM can only be applied for voiced segments. In this work the resampling of the prototype envelopes has been done by means of a simple linear interpolation of the real and imaginary components of the spectra.

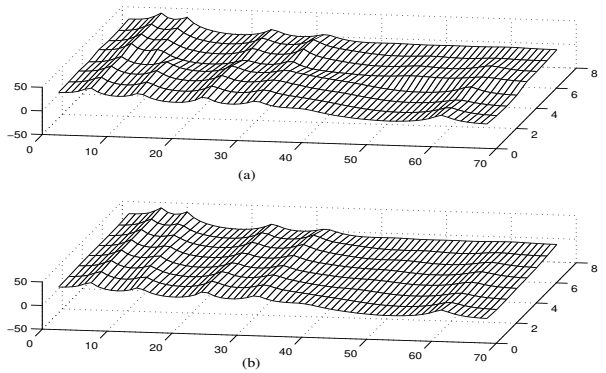

Figure 4: Frequency-response of the synthesis filter at the Boundary of two SUs. (a) Before smothing. (b) After smothing ($\mathbf{N}_p = \mathbf{N}_c = 4$).

5. SPECTRAL SMOOTHING

Using the flexibility of the TFI technique to perform spectral smothing, several methods can be developed to minimize the spectral mismatch at the boundaries of the SUs. The method described here is applied exactly in the same way to both prototypes and **LSF** coefficients. Defining N_p as the p last prototypes in the previous SU and N_c as the c first prototypes in the current SU, we will describe here the proposed procedure for a particular case where $N_p = N_c = 2$.

Defining $\mathcal{E}_p = \{\mathbf{E}_{n_l}, \mathbf{E}_{n_m}\}$ as the last two prototypes of the previous SU and $\mathcal{E}_c = \{\mathbf{E}_{n_n}, \mathbf{E}_{n_o}\}$ as the first two prototypes of the current SU, then at first the prototypes in $\mathcal{E}_p$ and $\mathcal{E}_c$ are normalized to have exactly the same size (putting zeroes at the end of the smaller ones) and after that they are submitted to the following linear combinations : $\mathbf{E}_{n_l} = \{\alpha_0 \cdot \mathbf{E}_{n_l} + \beta_0 \cdot \mathbf{E}_{n_o}\}$, $\mathbf{E}_{n_m} = \{\alpha_1 \cdot \mathbf{E}_{n_m} + \beta_1 \cdot \mathbf{E}_{n_n}\}$, $\mathbf{E}_{n_n} = \{\alpha_2 \cdot \mathbf{E}_{n_n} + \beta_2 \cdot \mathbf{E}_{n_m}\}$ and $\mathbf{E}_{n_o} = \{\alpha_3 \cdot \mathbf{E}_{n_o} + \beta_3 \cdot \mathbf{E}_{n_l}\}$.

The goal of this linear combination is to perform a sharing of information, across the boundary. The coefficients β_i have been defined as a normalized Blackman window with maximum amplitude equal to 0.5, and $\alpha_i = 1 - \beta_i$. This linear combination has been performed only for $Voiced$ segments.

Now using these combined prototypes, a double interpolation through time-domain is performed. The first one is a Spline interpolation at the following time-instants : n_l, $\frac{n_l+n_m}{2}$, $\frac{n_m+n_n}{2}$, $\frac{n_n+n_o}{2}$ and n_o. The second one is a linear interpolation that returns the Splined prototypes to the original positions. The aim of these interpolations is to minimize spectral discontinuities close to the boundary.

The Figure (4.a) shows the Frequency-response of the synthesis filter at the boundary of two SUs. Figure (4.b) shows the Frequency-response of the same synthesis filter after applying the proposed procedure of smoothing, where $\mathbf{N}_p = \mathbf{N}_c = 4$.

6. EVALUATION

An informal subjective test was performed between the proposed system and the standard TD-PSOLA system. The sequence of SUs (**Diphones**) and the Prosodic-Modifications were generated using the Festival System [1]. The test was composed of 14 sentences. These sentences were listened to by 12 speech-experts who had to express their preference. The results highlighted a superior quality of the proposed system in comparasion with the standard TD-PSOLA. 71.4% of the listeners said that the TFI was better than the TD-PSOLA and only 5.71% said that the TD-PSOLA was better than the TFI (22.89% of the listeners did not express a preference). These sentences can be heard on the following HomePage : http://www.cstr.ed.ac.uk/ $\sim$ emorais/pwitfi.html.

7. FINAL CONSIDERATIONS

In this paper some preliminary methods to apply Time-Frequency Interplolation [3] to Concatenative TTS Systems were presented. It was shown that the TFI technique allows PSM and TSM as well as Spectral Smothing at the boundaries of the SUs in a quite straightforward manner and with high quality.

Comparing TD-PSOLA and TFI regarding computational cost, it is clear that TFI has much higher complexity than TD-PSOLA. However, TFI presents the advantage of allowing very easy and efficient low-bit rate coding. Moreover, expecting the computer power to increase in the future, TFI complextiy will not be at all a problem.

In order to improve the quality of TFI techinque (in terms of Prosodic Modifications and the Spectral Smothing at the boundaries of the SUs) the authors suggest investigations on the REW and SEW decomposition [2] as well as on the adaptation of the technique to operate directly with the speech signal, instead of using the Linear Prediction model.

8. REFERENCES

1. A. Black, P. Taylor, R. Caley. The Festival Speech Synthesis. *Avaliable at htt://www.cstr.ed.ac.uk/projects/festival.html*, 4(5), Sept. 1996.

2. B. Kleijn, K. Paliwal, eds. *Speech Coding and Synthesis. Elsevier, Amsterdam*, 1998.

3. Y. Shoham. High-quality Speech Coding at 2.4 to 4.0 kbps Based on Time-Frequency Intepolation. *IEEE Proc. ICASSP '93*, II.167–II.170, April,1993.

4. F. Violaro, O. Boeffard. A Hybrid Model for Text-to-Speech Synthesis. *IEEE Trans. on Speech and Audio Processing*, 6(5), Sept. 1998.

5. E.S. Morais, F. Violaro, P. Barbosa. Prosodic Speech Modification Using Pitch-Synchronous Time-Frequency Interpolation. *Proc. of the International Telecommunication Simposium, Sao Paulo, SP, Brazil*, June, 1998.