Pitch-Synchronous Time-Scaling for Prosodic and Voice Quality Transformations

João P. Cabral and Luís C. Oliveira

L^2F Spoken Language Systems Lab., INESC-ID/IST,
Rua Alves Redol 9, 1000-029 Lisbon, Portugal
{jpcabral,lco}@l2f.inesc-id.pt

Abstract

Current time-domain pitch modification techniques have well known limitations for large variations of the original fundamental frequency. This paper proposes a technique for changing the pitch and duration of a speech signal based on time-scaling the linear prediction (LP) residual. The resulting speech signal achieves better quality than the traditional LP-PSOLA method for large fundamental frequency modifications. By using non-uniform time-scaling, this technique can also change the shape of the LP residual for each pitch period. This way we can simulate changes of the most relevant glottal source parameters like the open quotient, the spectral tilt and the asymmetry coefficient. Careful adjustments of these source parameters allows the transformation of the original speech signal so that it is perceived as if it was uttered with a different voice quality or emotion.

1. Introduction

The pitch and duration manipulation of speech signal have been a subject of great interest and several methods have been proposed to address this problem. The quality of concatenative speech synthesis can be improved by correcting pitch discontinuities across concatenated segments [1] and the unit inventory size can be reduced if broader prosodic modifications are allowed without perceived degradation of the synthesized speech. Current dialogue systems also require the ability to generate speech with variable emotional content. Indeed an important challenge of state of art speech synthesizers is to enhance the naturalness of speech using suprasegmental information about the emotional state and the voice characteristics of the speaker. However, the traditional prosodic parameters of rhythm and intonation are not enough to represent the emotional content of speech and additional acoustic parameters have been investigated such as the spectral energy distribution, harmonic-to-noise ratio and voice quality parameters [2].

Time domain approaches for pitch modification are known to be more advantageous than frequency domain techniques because they require much lower computational effort and provide good quality over a moderate transformation scale. Most time-domain techniques are based on a PSOLA (Pitch Synchronous Overlap and Add) approach [3] that produces both phase and spectral distortion. For large pitch transformations the degradation in the quality of the synthesized speech becomes noticeable and its use is normally limited to scaling factors between 0.5 and 2.

Time-scaling algorithms for pitch and duration modification have also been studied. A simple approach is to interpolate or decimate the speech signal but it often introduces audible artifacts and the distortion of the spectral envelope creates noisy-like chipmunk [4]. A possible approach to reduce spectral distortion is to use the linear prediction model for analysis-synthesis because the modification can be applied to the spectrally flat excitation signal without changing the formant structure [5]. In this paper we propose a new algorithm applied to the excitation signal for pitch and duration modification and an extension of this method for voice quality transformation.

2. Pitch and duration modification

We will start by presenting a modified approach of the LP-PSOLA method and a new algorithm for pitch and duration modification that try to reduce some of the problems encountered in the traditional LP-PSOLA algorithm.

2.1. Standard LP-PSOLA Algorithm

The PSOLA approach for modifying the speech signal duration and pitch was proposed in [3] using two methods. The TD-PSOLA (Time-Domain PSOLA) applies the pitch synchronous splitting into overlapped short-time signals directly on the speech signal while the LP-PSOLA (Linear Prediction PSOLA) splits the residual signal resulting from a linear prediction estimator. The results achieved by both methods are very similar but the TD-PSOLA is more commonly used given its lower computational requirements. In our case, we are interested in manipulating the features of the glottal signal for which the linear prediction residual is an approximate estimate of its derivative. For this reason the LP-PSOLA method was preferred.

In our implementation the pitch synchronization is achieved by using an estimate of the glottal closure instants for voiced speech, computed using a pitch marking algorithm. In the unvoiced regions, the pitch marks are equally spaced 5 ms apart. The LPC parameters are also computed pitch synchronously using a 20 ms Hanning window centered on each epoch estimate. The residual signal is obtained by inverse filtering the speech signal by a time-varying all-zeros filter using the LPC coefficients associated with each pitch mark.

The residual signal is then segmented into overlapping short-time signals using a weighting window centered in each analysis pitch mark $n_o(i)$ and with a length equal to two times the original pitch period $N_o(i) = n_o(i) - n_o(i-1)$:

$$x_i(n) = x(n)w(n - n_o(i)) \quad (1)$$

The synthesis epochs, $n_s(j)$, are computed in order to meet a specified duration and pitch contour.

The modified residual is the sum of the overlap-add short-time signals centered in the corresponding synthesis pitch mark:

$$y(n) = \sum_j (x_j(n - n_s(j))) \quad (2)$$

The shift between successive synthesis pitch marks is equal to the desired pitch period $N_s(j) = n_s(j) - n_s(j-1)$.

The quality of this method depends on the overlap factor F_R, defined as a function of the weighting window length (L) and the modified pitch period ($N_s(j)$):

$$F_R(j) = \max\left(0, 100\left(1 - \frac{N_s(j)}{L}\right)\right) \quad (3)$$

Distortion is perceived when F_R is too different from 50%. Since we are using $L = 2N_o(j)$, an increase in pitch period by a factor of 2 corresponds to an overlapping factor of 0%. Halving the pitch period gives an overlapping factor of 75%.

2.2. Modified LP-PSOLA

A simple modification can be made to the standard LP-PSOLA method in order to prevent the overlapping factor to increase above 50%. When reducing the pitch period, instead of using a weighting window length of twice the original period, we use a length of twice the desired (smaller) period, $L = 2N_s(j)$. Since the residual signal energy is concentrated close to the center of the window, the shorter window maintains the relevant features of the short-term signal.

When F_R is too small, an energy correction factor could also be applied but it does not prevent the energy fluctuation of the longer pitch periods. A more general solution is proposed to overcome both the period increase and decrease cases.

2.3. Pitch Synchronous Time-Scaling (PSTS) Algorithm

To preserve the shape of the residual waveform within the period a time-scaling transformation is proposed. By time-scaling the short-term signals, the overlapping interval can be changed maintaining the energy balance of the modified signal. Assuming that the prediction residual is an estimate of the glottal flow derivative, the time-scaling operation also preserves the most import features of the glottal pulse: open coefficient, abruptness of closure and spectral tilt.

The original and modified pitch marks ($n_o(i)$ and $n_s(j)$) and the linear prediction residual signal ($x(n)$) are determined the same way as for the standard LP-PSOLA method described in section 2.1. The short-time signals are computed by shifting $x(n)$ to start in the previous pitch mark and by multiplying it with a rectangular window $r(n)$ with length equal to the analysis pitch period $N_o(i)$:

$$x_i(n) = x(n + n_o(i-1))r(n) \quad (4)$$

where

$$r(n) = \begin{cases} 1, & 0 \le n < N_o(i) \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

and the original pitch period is $N_o(i) = n_o(i) - n_o(i-1)$. Thus, the short-time signal $x_i(n)$ has the length $N_o(i)$ and contains the ith pitch period of the original signal.

The short-time signals $x_j(n)$ are obtained by mapping the synthesis pitch marks $n_s(j)$ on the estimated analysis epochs $n_o(i)$ using the pitch and duration modification contours as in LP-PSOLA. Each short-time signal $x_j(n)$ is time-scaled by the factor $L = N_s(j)/N_o(i)$ producing the modified short-time signal $z_j(n)$, where $N_s(j)$ is the desired pitch period:

$$N_s(j) = n_s(j) - n_s(j-1) \quad (6)$$

Although L may be a non-integer factor, it is always rational, and the time-scaling can be implemented by an interpolator followed by a decimator. The scaling operation is performed with four times oversampling in order to minimize aliasing due to non-ideal low-pass filtering. When increasing the pitch period ($L > 1$) the time scaling operation compresses the spectrum and an "energy hole" appears at higher frequencies. To overcome this problem we use a simple and effective excitation regeneration method fully described in [9].

The pitch and duration modifications can require the skipping or duplication of short-time signals. To produce smooth spectral transitions and to prevent the discontinuities on the energy envelope, the concatenation of non-adjacent short-time signals is performed by overlapping and adding these signals. For example, suppose that the jth period of the modified signal is obtained by overlapping the lth and rth periods of the original signal. First, we need to time-scale both short-time signals to the same length $N_s(j)$ using the procedure stated above. The resulting short-time signals $z_l(n)$ and $z_n(n)$ are weighted by and summed:

$$y_j(n) = w(n)z_l(n) + w(N_s(j) - n)z_r(n) \quad (7)$$

where $w(n)$ is a decaying weighting window with length $N_s(j)$ such that $w(n) + w(N_s(j) - n) = 1$ to produce perfect reconstruction in the case of no modification. The last half of a Hanning window meets this criteria.

The modified LP residual waveform is the pitch-synchronous sum of all the short-time signals:

$$y(n) = \sum_j (y_j(n - n_s(j) + N_s(j))) \quad (8)$$

Figure 1 (a) shows the LP residual where the pitch mark correspondent to the short-time signal i is located in n_o^i. Figure 1 (b) and (c) represent the modified residual signal after pitch period increasing with a factor of 2.3 using the PSTS algorithm (b) and the standard LP-PSOLA method (c), where n_s^j is the location of the synthesis pitch mark corresponding to the analysis pitch mark n_o^j.

Since the asymmetric overlap-add windows are centered in the pitch marks, the region of major excitation around the glottal closure is less altered. However, we rely on the assumption that the synthesis pitch period is approximately constant for adjacent glottal cycles since the pitch marks do not coincide with the beginning and ending of the glottal cycle. In practice this is not a problem because the pitch perception is mostly dependent on the distance between the instants of significant excitation.

This method gives a better representation of the residual after prosodic modification and overcomes the problem of energy fluctuation when the pitch modification factor is large. The overlap-add operation reduces the phase and frequency discontinuities when joining non-contiguous signals. Also, unlike LP-PSOLA, this approach tries to preserve the most relevant glottal source parameters.

3. Voice quality transformations

Having a robust technique to change the intonation and rhythm, our goal is now to try to extend these modifications to other characteristics of the speech related with the glottal source parameters known to be related with voice quality. Previous studies proposed several parametric models for the glottal source that allows the control of several source-related acoustic features and contributed to improve the flexibility of speech syn-

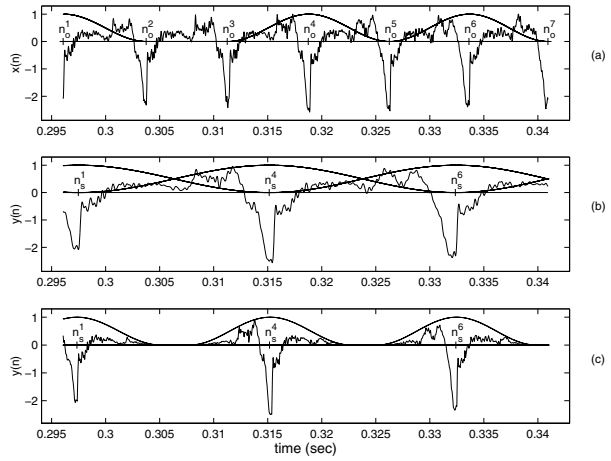

Figure 1: Pitch modification with a factor of 2.3. (a) A segment of the LP residual, (b) PSTS algorithm, (c) LP-PSOLA method.

thesizers [6]. Also, great attention has been dedicated to parameter estimation of the glottal flow waveform in the time-domain and frequency-domain and the effect of these parameters in the voice quality [7]. In this section we propose to extend our time-scaling method to allow modifications of the LP residual which try to mimic the effects produced by changing parameters of glottal source models. The advantage of this approach is that it does not require a detailed modeling of all the parameters that affect the glottal flow because most of the resulting effects are already in the LP residual signal. Thus, the complexity is reduced and the resulting speech sounds more natural.

3.1. Extraction of time-instants from the glottal flow derivative

The pitch marks and the LP residual signal are computed the same way as for the LP-PSOLA method described previously with the exception that before computing the LPC model for each analysis frame the original speech waveform is high-pass filtered with a pre-emphasis filter ($\alpha = 0.97$) to attenuate low-frequency rumble and obtain a flatter residual.

Each short-time signal corresponds to a pitch cycle, as given by equation (4). In the voiced regions, we estimate the three relevant instants represented in Figure 2:

- *Glottal closure* n_{cl}: is estimated as the instant of the first peak after the first zero crossing in the short-time signal.
- *Glottal opening* n_{op}: is obtained using the threshold based method described in [8].
- *Maximum of the glottal flow* n_p: first, it is computed the DC value between the glottal opening (n_{op}) and every zero crossing to the end of the short-time signal. Then, it is chosen the zero crossing correspondent to the maximum DC value.

3.2. Transformation of the glottal flow parameters

In order to modify the glottal flow parameters we perform a time-scale transformation over the segments associated with the glottal flow phases. To avoid aliasing modifications are performed as previously, using oversampling with a factor of four.

The duration of the glottal cycle phases are computed from the extracted time instants (represented in Figure 2):

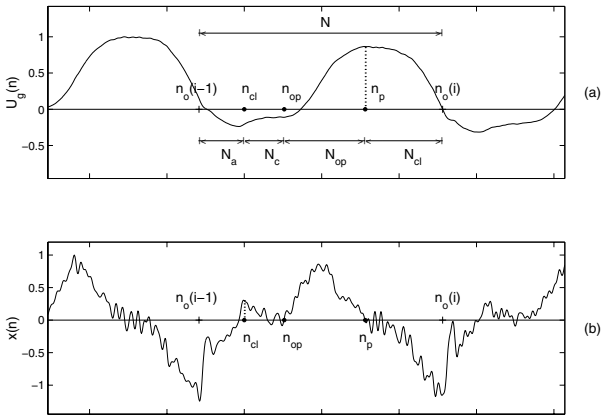

Figure 2: Representation of the extracted time instants in the glottal flow waveform (a) and its time-derivative (b).

Return phase: $N_a = n_{cl}$
Peak flow duration: $N_e = N - n_{op}$
Closed phase: $N_c = N - N_a - N_e$
Opening phase: $N_{op} = n_p - n_{op}$
Closing phase: $N_{cl} = N - n_p$

Our goal is to modify the three most important glottal parameters: the open quotient, the speed quotient and the length of the returning phase.

3.2.1. Open quotient

The open quotient is related to the duration of the open phase and can be expressed as:

$$OQ = (N_a + N_e)/N \tag{9}$$

To increase the OQ both the length of the return phase N_a and the length of the peak flow duration N_e must be increased. To decrease the OQ both lengths must be shortened. Thus, the time-scale factor is equal to the pretended variation rate for the OQ.

Due to the time-scale transformation it is necessary to adjust the duration of the closed phase to preserve the pitch period of the glottal waveform. If the modified short-time signal is shorter than the pitch period then additional samples are inserted in the middle point of the closed phase by repeating two neighborhood parts to the left and right side of that instant so that the continuity of the signal is preserved at the end points of the padded portion. To smooth the transition between the two parts first-order linear windowing is performed to the added portion. When the modified short-time signal is shorter than the pitch the adequate number of samples are truncated in the center of the closed phase. To avoid the discontinuity due to the truncation a proper windowing is also performed. In the closed phase of the glottal flow derivative waveform, the energy is normally lower than in the rest of the short-time signal so it is expected less abrupt discontinuity due to the truncation or padding of samples. Figure 3 shows a cycle of the glottal flow derivative waveform (a), the modified signal for a decrease of the OQ (b), and the modified signal for an increase of the OQ (c). In this figure n_m represents the middle point of the closed phase, N_w represents the windowed part of the modified short-time signal, n'_{cl} is the new instant of glottal closure and n'_{op} is the new instant of glottal opening.

We also adjust the DC component of the modified short-time signal to the original DC value by multiplying the new opening phase (with duration N'_{op}) with the appropriate scale factor.

3.2.2. Speed quotient

This parameter is related to the asymmetry coefficient and accounts for variations in the shape of the open phase of the glottal-flow. The speed quotient is:

$$SQ = N_{op}/N_{cl} \tag{10}$$

The speed quotient can be increased with a time-scale expansion of the opening phase and a time-scale compression of the closing phase so that the peak flow duration $N_e = N_{op} + N_{cl}$ remains constant. SQ can be decreased by the opposite transformation.

3.2.3. Return quotient

This parameter is related with the duration of the closing phase and determines the cut-off frequency of the spectral tilt. The return quotient is expressed by

$$RQ = N_a/N \tag{11}$$

The return quotient can be increased or decreased by a time-scale expansion or compression of the return phase. To maintain the pitch period and the open quotient, the peak flow is also time-scaled by an adequate factor.

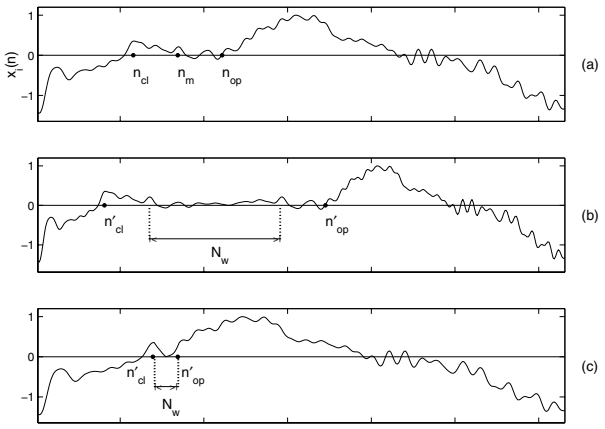

Figure 3: (a) One cycle of the glottal flow derivative waveform (b) Decrease of the open quotient by a factor of 0.7 (c) Increase of the open quotient by a factor of 1.2.

4. Results

Informal listening tests showed that the PSTS algorithm generates speech with higher quality than the traditional LP-PSOLA for large pitch scale factors, specially for increasing pitch period (above a factor of 1.5). However, for large scale factors, the modified speech does not achieve the quality of natural speech because other parameters of the glottal source are not properly modified. We obtained an improvement in naturalness of the output speech when we transformed both the pitch and the open quotient. For increasing pitch period we reduced the open quotient with an adequate factor while for decreasing pitch we increased this parameter. We are conducting a formal listening test to evaluate the performance of this algorithm.

Sample speech files are available at http://www.l2f.inesc-id.pt/~jpcabral/psts_prosodic_vq.

5. Conclusions

This paper proposes a method for pitch and duration modification based on a pitch-synchronous time-scale modification (PSTS) of the linear prediction residual. This method overcomes the problem of the energy envelope fluctuation characteristic of the PSOLA based methods. We also used this approach to modify the three most important parameters of the glottal source waveform. Thus, this method provides higher flexibility for prosody manipulation and voice quality transformation of speech with small increase in complexity comparing with existing time-domain techniques. Informal listening tests indicate that, for larger pitch modification factors, this method generates synthetic speech with higher-quality and an increased naturalness when compared with the traditional LP-PSOLA method. Further evaluations and experiments are being conducted.

6. Acknowledgements

This work was partially founded by the Portuguese Foundation for Science and Technology (FCT) within the project POSI/SRI/41071/2001.

7. References

[1] Quazza, S., Donetti, L., Moisa, L. and Salza, P.L., "ACTOR: a Multilingual Unit-Selection Speech Synthesis System", 4th ISCA Workshop on Speech Synthesis, Perthshire, Scotland, September 2001.

[2] Rank, E. and Pirker, H., "Generating emotional speech with a concatenative synthesizer", Proc. ICSLP 98, Sydney, Australia, Vol.3, pp. 671-674, November 1998.

[3] Moulines, E. and Charpentier, F., "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones", Speech Communications, Vol. 9, pp. 453-476, December 1990.

[4] Lee, F., "Time Compression and Extraction of Speech by the Sampling Method", Journal of the Audio Engineering Society, 20(9):738–742.

[5] Edgington, M. and Lowry, A., "Residual-based Speech Modification Algorithms for Text-to-Speech Synthesis", ICSLP 96, Philadelphia, USA, pp. 1425-1428, 1996.

[6] Childers, D.G., "Glottal source modeling for voice conversion", Speech Communication, 16(2):127–138, 1995.

[7] Alku, P., "Parameterisation methods of the glottal flow esimated by inverse filtering", Proc. ITRW VOQUAL'03, Geneva, Switzerland, pp. 81-88, August 2003.

[8] Arroabarren, I. and Carlosena, A., "Glottal source parameterization: a comparative study", Proc. ITRW VOQUAL'03, pp. 29-34, Geneva, Switzerland, August 2003.

[9] Cabral, J. P. and Oliveira, L. C., "Pitch-Synchronous Time-Scaling for High-Frequency Excitation Regeneration", Proc. Interspeech'2005, Lisbon, Portugal, September 2005.