HIGHER PRECISION PITCH MARKING FOR TD-PSOLA

Vincent Colotte and Yves Laprie
LORIA, Campus scientifique, BP 239, F-54506 Vandœuvre-lès-Nancy, FRANCE
Tel: +33 (0)3 83592074; fax: +33 (0)3 83413079
e-mail: Vincent.Colotte@loria.fr, Yves.Laprie@loria.fr

ABSTRACT

The paper describes techniques to improve the precision of prosodic modifications with TD-PSOLA. TD-PSOLA relies on the pitch synchronous decomposition of the signal into overlapping frames synchronised with pitch period. The main objective is thus to preserve the consistency of marks between neighbouring frames with respect to the temporal structure of pitch periods. First, we improve pitch marking by eliminating mismatch errors which appear during rapid formant transitions. This is achieved by pruning pitch mark candidates whose distance with other candidates is clearly not consistent with the current pitch period. From the synthesis point of view we exploit a fast re-sampling method which allows signal frames to be shifted finely where they should appear given both the initial pitch mark and the location of pitch mark for synthesis. Together with the pitch marking improvement, this fast re-sampling method enables very high quality transformations characterised by the absence of noise between harmonics.

1 INTRODUCTION

Marking pitch periods of speech signals is important for pitch synchronous methods used to perform time or pitch scale modifications, as well as to implement text-to-speech synthesis. Our work deals with the modification of prosodic parameters in the context of language learning. We accepted TD-PSOLA (*Time Domain - Pitch Synchronous OverLap and Add*) method because it is fast and allows speech rate and F0 to be modified simultaneously.

TD-PSOLA [6] is based on the decomposition of the speech signal into overlapping pitch synchronous frames. Pitch marks indicate centers of frames. Signal modifications consist in manipulating analysis marks to generate new synthesis marks. This corresponds to the duplication or the decimation of frames whose distance with neighbouring frames can be changed.

The main requirement is to preserve the consistency of mark location between frames in order to be able to preserve the original temporal structure of the signal under analysis. Therefore it is crucial to obtain an accurate marking of pith periods because it directly influences the quality of the resulting signal.

Various pitch marking methods have been described in he literature. Usually they are based on the seeking of precise events in the speech signal: glottal closure events [1], signal extrema, excitation instants of LPC models, last zero-crossing before a positive maximum...

As underlined by R. Veldhuis [8], these techniques suffer from the rigidity of the numerical criteria exploited. In particular, the numerical criterion may force the marking of samples which satisfy the numerical criterion but whose distance with neighbouring marks is far from the pitch period. In the framework of speech synthesis it is conceivable to correct some errors by hand, which is quite impossible in the context of automatically modifying sentences for language learning.

In [3] we have proposed a pitch marking algorithm which exploits the result of pitch determination and ensures the consistency of marks over the entire sentence. R. Veldhuis [8] has proposed an algorithm which slightly extends our approach by taking into count the correlation of signal in the vicinity of neighbouring pitch marks.

The improvement of the robustness, as well as the precision of the marking is the first aim of this paper.

Besides the quality of the pitch marking, another source of temporal discrepancies can originate in the location of synthesis marks themselves. As a first approach, synthesis marks can be chosen among sampling instants. For the same reasons as those mentioned above, synthesis marks have to be set carefully to prevent phase mismatches. The figure 1 shows the phenomenon: one period of sound /ɛ/ at 16 kHz is duplicated and the pitch scale of this stimulus is linearly modified. The analysis marks are placed exactly at the same place in each period. The first spectrogram shows the obtained result with the classical algorithm and the second with higher precision synthesis. In the first case, we can see the repercussion of phase mismatch on the quality of harmonics.

It is thus important to develop algorithms to achieve a higher precision for pitch marking as well as for accurate synthesis. First, we describe how we improve the pitch marking algorithm we proposed in [3]. Then we explain how this accurate pitch marking can be combined with a resampling technique to achieve a higher precision synthesis with TD-PSOLA.

2 PITCH MARKING

2.1 Principle

The idea of our algorithm is to select pitch marks among local extrema of speech signal.

Given a set of mark candidates which all are negative peaks (or all positive peaks) :

$$C = [c(i)] = c(1) \dots c(i) \dots c(N)$$

where $c(i)$ is the sample of the i^{th} peak, and N the number of peaks extracted ([3] explain how these candidates are found).

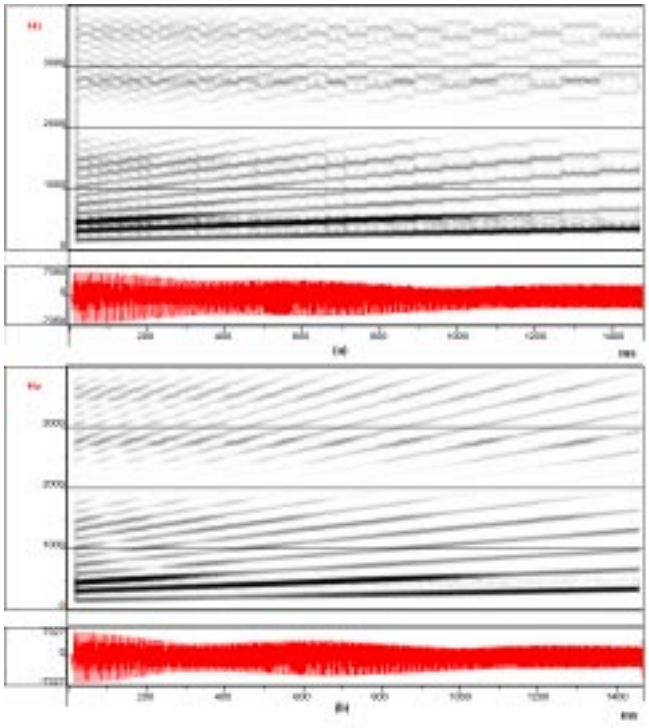

Figure 1: A stimulus has been constructed from one period of sound /ɛ/. Its pitch scale has been linearly modified: (a) with classic TD-PSOLA algorithm (b) with higher precision synthesis

Pitch marks are a subset of points out of C, which are spaced by periods of pitch given by the pitch extraction algorithm. The selection can be represented by a sequence of indices

$$J = [j(k)] = j(1) \dots j(k) \dots j(K)$$

with $K < N$. J has to preserve the chronological order which requires the monotony of j: $j(k) < j(k+1)$.

The sequence of indices along with the corresponding peaks is defined to be the set of pitch marks:

$$\overline{C} = [c(j(k))] = c(j(1)) \dots c(j(k)) \dots c(j(K))$$

The determination of j requires a criterion expressing the reliability of two consecutive pitch marks with respect to pitch values previously determined. The local criterion we chose is:

$$d(c(l), c(i)) = |(c(i) - c(l)) - pitchPeriod(c(l))| \qquad (1)$$

where $l < i$. It takes into account the time interval between two marks compared to the pitch period in samples. This criterion returns zero if the two peaks are exactly $pitchPeriod(c(l))$ samples away from one another and a positive value if the distance between these peaks is greater or less than the pitch period.

The overall criterion is:

$$D = \sum_{k=1}^{K-1} d(c(j(k)), c(j(k+1))) - B(c(j(k)) \qquad (2)$$

where B is the bonus of selecting an extremum as a pitch mark. In a first time, $B(c(j(k)) = \gamma |amplitude(c(j(k)))|$. The coefficient γ expresses the compromise between closeness to pitch values and strength of pitch marks. Minimising D is achieved by using dynamic programming.

We used the pitch determination algorithm proposed by Martin [4] to evaluate the local criterion defined in Eq.1. The signal was filtered with a low pass filter whose cutoff frequency is 2500 Hz.

2.2 Improving the pitch marking algorithm

The coefficient γ which controls the compromise between closeness to the pitch marks and strength of marks has been experimentally set to 1/400. When γ is too strong many peaks are kept as pitch marks as shown in Fig. 2a (at 1144 ms). The value, we accepted for γ, turned out to be sufficiently general to give good results with most of speech signals. However, it happens that this choice is not appropriate for some signals with substantial spectral transitions.

We investigated two solutions. The first consists of exploiting a similarity function to prevent phase mismatches between neighbouring extrema. We thus used the correlation as a bonus. Bonus of Eq.2 is replaced by:

$$\begin{aligned} B(c(j(k))) = \quad & \gamma \times (\delta |amplitude(c(j(k))| \\ & + \delta' corr_n(c(j(k)), c(j(k+1))) \end{aligned} \qquad (3)$$

where $corr_n(c(j(k), c(j(k+1)))$ is the correlation between segments of length n, centred at $c(j(k))$ and $c(j(k+1))$. The duration, over which the correlation is calculated, is set to the pitch period at instant $c(j(k))$. Coefficients δ and δ' allow amplitude and correlation to be weighted independently.

The computation of the correlation for each pair of candidates leads to a strong increase of the time required for pitch marking. Furthermore, the correlation is used to correct "obvious" errors which correspond to the matching of two local extrema whose distance significantly differ from the local pitch period.

For that reason, the second solution we investigated, relies on a pruning strategy. Without eliminating potential candidates, it is relevant to drastically increase the weight of the distance between two candidates compared against the local pitch period, when this distance is too far from the pitch period. In that way, the local criterion is given more importance than the bonus (see Eq.1).

2.3 Pitch marking results

In a first time, we tested the bonus based on the correlation alone. The ratio between correlation and local criterion was set to 40. The compromise between distance and correlation turned out to produce approximatively the same number of errors as the initial strategy. Indeed, this new strategy favours the selection of two peaks so close together that the correlation is close to 1.

Then we tested a bonus which incorporates both amplitude and correlation. These two contributions were given the same relative weighting. Results are satisfying for most of the signal but some errors cannot be eliminated. Furthermore, as mentioned above the remaining errors seem rather obvious because the distance between two pitch marks is very far from the pitch period. This stems from the dynamic programming algorithm which locally favours gross errors if this contributes to lower the global criterion over the entire sentence. On the other hand, as these errors can be easily located it appears that pruning can be exploited. Pruning is implemented by drastically increasing the local criterion for pairs of candidates whose distance is 20% greater or less than the pitch period expected. We set γ to 1/40000 instead of

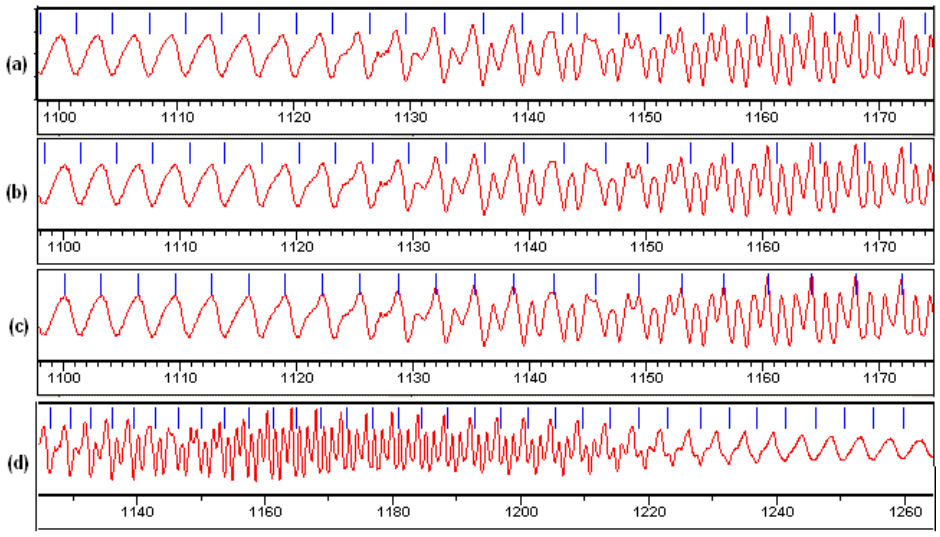

Figure 2: (a) Erroneous marking with the amplitude bonus but without pruning. (b) Marking with pruning and amplitude bonus. (c) Marking with pruning and with the correlation bonus. (d) Marking over a speech segment with formant transitions.

1/400 in Eq. 2. This arbitrary choice prevents dynamic programming from selecting two peaks inconsistent with each other. Furthermore, the 20% deviation allows gross errors to be eliminated but preserves the possibility of choosing two consecutive peaks, whose distance slightly differs from the pitch period calculated, but, which are consistent with the correlation criterion. This choice does not require fine tuning, and more importantly no tuning which would depend on the signal analysed.

Results obtained with the pruning strategy and with either the amplitude alone, correlation alone or both criteria turned out to be very good. These results are exhibited in Fig.2b and 2c. It can be seen that an accurate pitch marking can be achieved even in the case of transitions in the structure of the temporal signal (near ms 1145 and 1220 in Fig.2d).

3 HIGHER PRECISION SYNTHESIS

In order to achieve higher precision synthesis we could imagine to oversample the signal for pitch marking and re-synthesis. Indeed we oversampled the original signal for pitch marking. However, we did not oversample the signal for synthesis because oversampling would require a much higher rate to enable accurate synthesis marks corresponding to any time-scale and F0-scale objective (i.e. to arrange it so that the synthesis mark corresponds to a sample).

Therefore, our algorithm works as follows:

1. **Pitch marking**
   - Oversampling and low-pass filtering (cutoff frequency 2500 Hz).
   - Marking pitch periods of the oversampled signal.
2. **Re-synthesis**
   - Applying TD-PSOLA algorithm with these marks to obtain the exact position of synthesis marks and to associate an analysis frame.
   - Shifting the frame by re-sampling to obtain the true synthesis frame (see Fig.3).

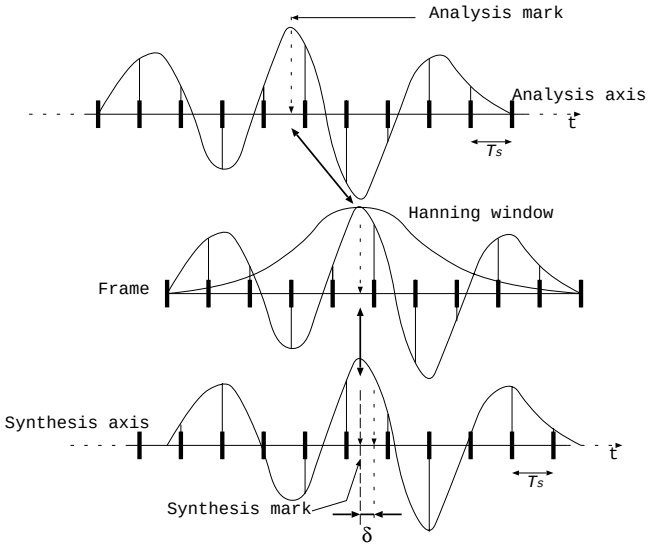

Figure 3: Matching of an analysis frame on synthesis time axis.

   - Reconstruction of the signal.

Therefore, given the pitch mark and the synthesis mark of a given frame we use a fast re-sampling method described below to shift the frame precisely where it will appear in the new signal.

Let $x[n]$ the original frame, the re-sampled signal is given by A. Oppenheim [7]:

$$x(t) = \sum_{n=-\infty}^{\infty} x[n] sinc\left(\frac{\pi(t - nT_s)}{T_s}\right) \qquad (4)$$

where T_s is the sampling period.

Calculating the result frame $y[m]$ corresponding to the frame $x[n]$ shifted by a small delay δ amounts to evaluate $x(mT_s - \delta)$ (see Fig. 3). Therefore, $y[m] = x(mT_s - \delta)$ i.e:

$$\begin{aligned} y[m] &= \sum_{n=-\infty}^{\infty} x[n] sinc\left(\pi f_s[(mTs - \delta) - nT_s]\right) \\ &= \sum_{n=-\infty}^{\infty} x[n] sinc\left(\pi f_s[(m-n)Ts - \delta]\right) \end{aligned} \quad (5)$$

where f_s is the sampling frequency $(1/T_s)$.

Now, by rewriting $sinc$ as $sin(x)/x$ and by using the following formula:

$$sin\left(\pi f_s[(m-n)T_s - \delta]\right) =$$

$$cos\left(\pi f_s \delta\right) sin\left(\pi(m-n)\right) - sin\left(\pi f_s \delta\right) cos\left(\pi(m-n)\right)$$

but $cos\pi(m-n) = \pm 1$, and $sin\pi(m-n) = 0$ so

$$y[m] = \sum_{n=-\infty}^{\infty} x[n] \frac{(-1)^{(m-n+1)} sin\left(\pi f_s \delta\right)}{\pi f_s[(m-n)T_s - \delta]} \quad (6)$$

As $0 < \delta < T_s$ (resp. $-T_s < \delta < 0$), we define $\delta = \alpha T_s$, where $0 < \alpha < 1$ (resp. $-1 < \alpha < 0$). Then the equation becomes:

$$y[m] = \sum_{n=-\infty}^{\infty} (-1)^{(m-n+1)} x[n] \left(\frac{\sin \alpha\pi}{\pi}\right) \frac{1}{(m-n) - \alpha} \quad (7)$$

The limit of the summation can not be infinite. We have used a short window (1-2 ms $\simeq$ 50 samples).

At last, the obtained frame is weighted with a Hanning window. Gimenez and Talkin [2] use an asymmetric window to reduce the phenomena of distortion and reverberation which are introduced by the windowing.

We can observe the improvement of the quality of the resulting signal in the figure 4. The first spectrogram comes from a signal modified with the classical TD-PSOLA method. The second is the spectrogram of a signal modified with the high resolution method explained above. In particularly, we observe the clearer structure of harmonics (for instance about 1000, 1500 and 1700 ms).

4 CONCLUSION

In this paper we have proposed a higher precision algorithm for pitch marking at two levels: analysis and synthesis marks. At first, our algorithm of pitch marking overcomes errors which may appear with other algorithms. In addition, the algorithm is very fast in computation, that is very suitable for TD-PSOLA method.

Secondly, the combination of our pitch marking with a fast re-sampling method during the synthesis step increases the signal quality. This gain in accuracy avoids the reduction of quality between original and synthetic signal observed with the classical TD-PSOLA method. This can be clearly observed in the quality of harmonics (in Fig.4) where the level of noise between harmonics is reduced with our method.

In future works, we will investigate how hyper-resolution F0 computation algorithms [5] could be exploited to achieve further improvement in the determination of pitch marks.

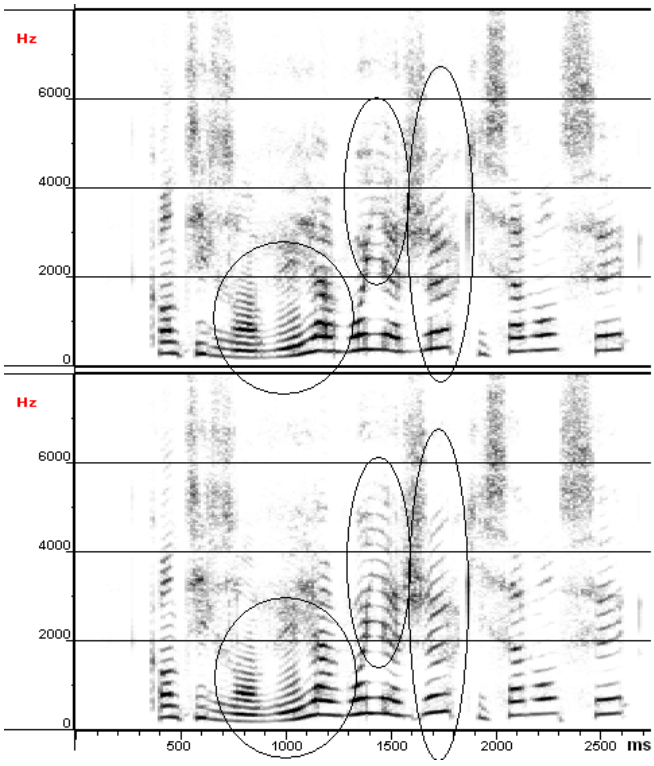

Figure 4: Top: Pitch scale modified without re-sampling. Bottom: with re-sampling

REFERENCES

[1] Y. M. Cheng and D. O'Shaughnessy. Automatic and reliable estimation of glottal closure instant and period. *IEEE Trans. Acoust., Speech, Signal Processing*, ASSP-37(12):1805–1815, December 1989.

[2] F. Gimenez de los Galanes and D. Talkin. High resolution prosody modification for speech synthesis. In *Eurospeech*, pages 557–560, Rhodes, Greece, 1997.

[3] Y. Laprie and V. Colotte. Automatic pitch marking for speech transformations via TD-PSOLA. In *IX European Signal Processing Conference*, Rhodes, Greece, 1998.

[4] Ph. Martin. Comparison of pitch detection by cepstrum and spectral comb analysis. In *Proc. of Int. Conf. Acoust., Speech, Signal Processing 1982*, pages 180–183, 1982.

[5] Y. Medan, E. Yair, and D. Chazan. Super resolution pitch determination of speech signals. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, ASSP-39(1):40–48, January 1991.

[6] E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for a text-to-speech synthesis using diphones. *Speech Communication*, 9(5,6):453–467, 1990.

[7] A. V. Oppenheim and R. W. Schafer. *Digital Signal Processing*. Prentice-Hall, Inc, 1975.

[8] R. Veldhuis. Consistent pitch marking. In *International Conference on Speech Language Processing*, Beijin, 2000.