Pitch-Synchronous Time Alignment of Speech Signals for Prosody Transplantation

Vagner L. Latsch and Sergio L. Netto
PEE-DEL/COPPE-Poli, Federal University of Rio de Janeiro
POBox 68504, Rio de Janeiro, RJ, 21941-972, Brazil
Email: {latsch, sergioln}@lps.ufrj.br

Abstract— **Prosody transplantation is a speech signal modification procedure usually used to voice transformation or to evaluate the quality of speech synthesizers. In practice, the pitch contour is mapped onto a common segmental content and the target signal is modified adjusting position and length of speech frames to achieve the desired pitch contour and time duration from a speech reference. A new algorithm for prosody transplantation is presented based on a pitch-synchronous feature extraction of the speech signal, unifying the time-aligning and pitch-modification stages. The result is a computationally efficient algorithm for prosody transplantation that maximizes the spectral similarity between the target and reference signals.**

I. Introduction

Time alignment is the key stage on a prosody transplant system, in which the pitch contours from a reference signal is imposed onto a test signal by a time alignment curve. Such a system has been successfully implemented yielding the so-called MBROLIN algorithm, as described in [1], and used to prosodic evaluation of Text-To-Speech systems, as described in [2].

Since the 1980's, the associated literature [3] [4] has investigated the time alignment between a pre-labeled and a test speech samples, using a dynamic time warping (DTW) algorithm, for automatic speech segmentation or phonetic labeling. In [1], authors consider time alignment for automatic labeling and segmentation of speech sentences to bootstrap the intensive training processes of the hidden Markov models (HMM). Automatic segmentation can be used to add new voices into speech synthesizers [5] or to phonetically annotate large corpus employed by some synthesizers [6]. However, in this paper the automatic segmentation is used to evaluate the proposed method.

The present paper introduces a new algorithm for prosody transplantation. The proposed method performs time alignment using a DTW algorithm in a pitch synchronous fashion. In that manner, the time-warping determined by the DTW algorithm becomes completely compatible to the time-domain pitch synchronous overlap-and-add (TD-PSOLA) algorithm [7] that enforces the desired pitch contour and time scale. The resulting scheme, by unifying the speech partitioning for the DTW and TD-PSOLA algorithms, avoids pitch-mark interpolations in the DTW stage and yields a exact pitch transplant with high frame similarity.

The present paper is organized as follows: The standard DTW and TD-PSOLA algorithms are presented in Sections II and III, respectively, establishing the notation used in the remaining of the paper. Section IV discusses the combination of the DTW and TD-PSOLA algorithm in the context of a pitch-transplant application. Section V presents the proposed method, named PS-DTW-OLA, where the DTW uses the same time framework of the TD-PSOLA algorithm, simplifying the combination of these two techniques. Computer experiments illustrating practical utility of proposed algorithm are presented in Section VI, whereas Section VII concludes the paper emphasizing its main contributions.

II. DTW-Based Speech Alignment

Consider two sets $\mathbf{R} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_{N_r}\}$ and $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{N_t}\}$, of characteristic vectors $\mathbf{r}_i$ and $\mathbf{t}_j$ from a reference and test speech signals, respectively, where N_r and N_t are the number of speech frames in each sentence, and $\mathbf{r}_i = [r_{i,1}, r_{i,2}, \ldots, r_{i,N}]^T$, $\mathbf{t}_j = [t_{j,1}, t_{j,2}, \ldots, t_{j,N}]^T$ with N indicating the number of characteristics extracted in each frame. The time alignment between the two sets $\mathbf{R}$ and $\mathbf{T}$ determines the relationship between the corresponding time-indexes i and j, through proper parameterizations $i = \phi_i(k)$ and $j = \phi_j(k)$, for $k = 1, 2, \ldots, K$. The whole process is accomplished by minimizing some distance measure

$$d_\phi(R,T) = \sum_{k=1}^{K} d(\mathbf{r}_{\phi_i(k)}, \mathbf{t}_{\phi_j(k)}), \quad (1)$$

over all possible time parameterizations [8]. Such a procedure corresponds to finding the best path, defined as a sequence of points in a grid defined by the pointwise distance between vectors $\mathbf{r}_i$ and $\mathbf{t}_j$, for each $i = 1, 2, \ldots, N_r$ and $j = 1, 2, \ldots, N_t$.

The optimal path P, as described in equation (**??**), is commonly determined using a dynamic time-warping (DTW) algorithm based on Bellman's optimality principle [8].

III. TD-PSOLA Algorithm

A common tool for pitch modification and time scaling is the so-called TD-PSOLA algorithm [7], which has been widely applied in successful speech synthesizers based on unit-concatenation [5], allowing one to impose a wide variety of prosody patterns.

The TD-PSOLA algorithm modifies the time position of speech frames according to some desired pitch contour. The key aspect in the algorithm is that the each frame length

978-1-4244-9472-9/11/$26.00 ©2011 IEEE

varies according to the pitch of the original speech sample. These frames are separated or approximated individually in time and added together accordingly to obtain the desired pitch variation. When modifying the pitch contour along the voiced frames, the TD-PSOLA algorithm is forced to replicate or eliminate some frames to satisfy time-duration constraints. For instance, if one desires to decrease (or increase) pitch, consecutive frames are copied further (or closer) apart. Then, in order to keep time constraints, the algorithm is forced o eliminate (or replicate) some of these frames. Overall, the TD-PSOLA algorithm presents low computational complexity and yields high speech quality.

IV. Pitch-Transplant Application

In a pitch-transplant procedure, one is interested on imposing a pre-determined pitch period $P^s(t)$ of a reference signal onto a given test signal. To perform such a task, one first applies the DTW algorithm to determine a proper time alignment between the two speech signals. The TD-PSOLA algorithm is then used to modify the test pitch pattern according to the reference pitch contour.

This whole tandem operation is illustrated in Figure 1, where the pitch marks p_j^a, for $j = 0, 1, 2, 3, 4$, of the initial test signal are represented along the y axis with distinct color patterns. In this figure, the piecewise-linear curve W represents the time-alignment mapping yielded by the standard DTW scheme using a uniformly sampled grid. This mapping W is used to determine the new time position $p_j^{\prime a}$, along the x axis, for each test pitch mark p_j^a. Hence, in order to impose the desired pitch contour $P^s(t)$ onto the test signal, one replaces the reference frame associated to the pitch mark p_i^s by the nearest dislocated test frame associated to $p_j^{\prime a}$. In this process, as illustrated in the bottom part of Figure 1, some test frames must be eliminated or replicated several times by the TD-PSOLA algorithm to enforce proper time duration in the resulting signal. In this case, the reference pitch marks p_i^s, for $i = 0, 1, 2, 3, 4, 5, 6$, are respectively associated to the dislocated test marks $p_j^{\prime a}$ with $j = 0, 2, 3, 3, 4, 4, 4$, as indicated by the corresponding color pattern.

One important note on the DTW/TD-PSOLA joint operation is that in principle there is no direct correspondence between the pitch marks p_i^s and $p_j^{\prime a}$. This requires an interpolation of the path along W to locate the nearest mark $p_j^{\prime a}$ to p_i^s.

V. Proposed Pitch-Synchronous Time Alignment - Overlap and Add

In current DTW algorithm the test and reference signals are segmented into fixed-length frames, as illustrated in Figure 1 above. In this manner, the characteristic vectors are determined for each time instants $p_i^s = i\alpha$ and $p_j^a = j\alpha$, with $\alpha = \frac{f_s}{f_r}$, where f_s and f_r are the speech sampling frequency and frame rate, respectively.

By increasing the frame rate f_r, one may achieve better time resolution on the warping path at the cost of additional computational burden. However, all this additional resolution in the DTW stage may be in vain, since the subsequent TD-PSOLA scheme used to modify the pitch contour operates in a synchronous manner with the pitch marks, whatever the frame length. Additionally, the frame elimination/replication procedure performed by the TD-PSOLA does not follow in general any continuity pattern yielded by the DTW algorithm.

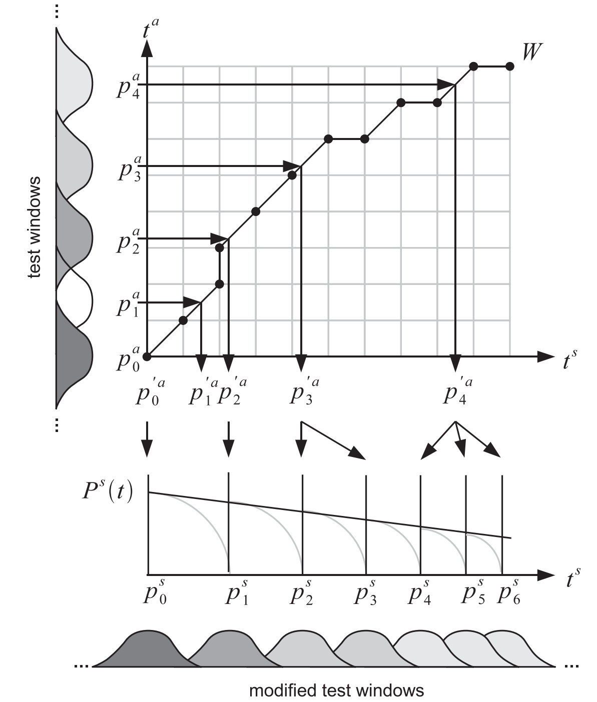

Fig. 1. Example of DTW mapping W that repositions test marks p_j^a along $p_j^{\prime a}$. The new positions identify the test frames nearest to reference pitch marks p_i^s. The corresponding test frames are combined by the TD-PSOLA algorithm to obtain desired pitch pattern $P^s(t)$.

To compensate for all these aspects, a new joint time-alignment and pitch-modification algorithm is proposed. The new scheme incorporates the pitch-synchronous (PS) characteristic of the TD-PSOLA algorithm onto the DTW alignment stage, allowing one to perform proper pitch modification and time alignment simultaneously, as illustrated in Figure 2.

The PS-DTW-OLA symbiotic combination optimizes the time resolution for the two processes, as indicated by the non-uniform grid shown in Figure 2. This strategy results in a computationally efficient time-alignment process in which the optimal pitch-synchronous path W_{ps} provides direct correspondence between test p_j^a and reference p_i^s pitch marks, thus avoiding additional interpolations within pitch-modification stage. Moreover, since the frame elimination/replication is performed in a guided manner within the PS-DTW-OLA algorithm, the modified test signal inherits a prosody pattern with nice similarity spectral features.

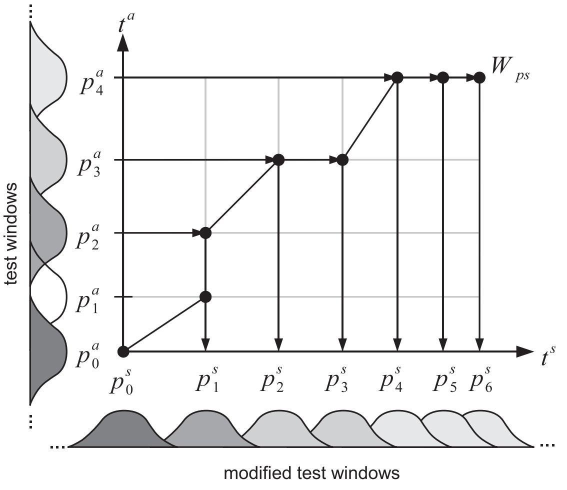

Fig. 2. Example of proposed PS-DTW mapping W_{ps} in which reference marks p_i^s are precisely associated to test marks p_j^a, simplifying subsequent TD-PSOLA stage.

A. Implementation Aspects

In the PS-DTW scheme, the characteristic extraction for voiced speech was performed in frames centered around a pitch mark with lengths determined by the neighboring marks. In unvoiced intervals, the frame length was set to 20 ms with 50% superposition to allow proper reconstruction by the TD-PSOLA algorithm.

The time-alignment stage employs local constraints, to expedite the search, while enforcing continuity on the resulting mapping between the two speech signals. Figure 3 includes, for instance, a local constraint that only considers paths with at most one (horizontal) step forward at a time.

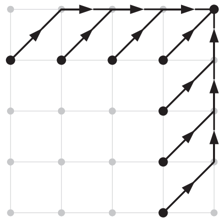

Fig. 3. Example of local constraints for time-warping stage.

This constraint yields nice pitch-transplant results with the proposed PS-DTW-OLA structure, as illustrated in Section VI.

In addition, since the pitch marks are avaiable, the time-alignment process may be performed in parts using explicit speech features such as voiced/unvoiced (V/UV) classification. By doing so, one forces the resulting time mappings to pass by V/UV frontiers determined in advance in the two signals. This modification is illustrated in Figure 4, still for the standard DTW method, where the three circled points denote the path restrictions determined by the V/UV marking. A proper V/UV classification improve the time-alignment as illustrated in Section VI below.

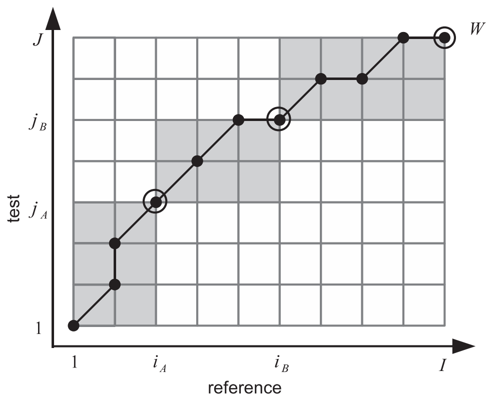

Fig. 4. Example of time-mapping using U/UV classification, indicated by shaded areas, defining path restrictions represented by circled dots.

VI. Computer Experiments

The proposed PS-DTW-OLA algorithm was implemented in C++ as part of a support system for text-to-speech development tool. Following the standard literature [8], the characteristic vectors included 12 mel-cepstral, 12 delta mel-cepstral, 1 normalized energy, and 1 delta-energy coefficients.

Experiment 1: To demonstrate the proposed method, a 2s-sentence was recorded and the corresponding pitch marks were determined. This sentence, referred to as the natural sentence, was then time-aligned to a synthetic sentence formed by the concatenation of proper diphone units in Brazilian Portuguese. The result is illustrated in Figure 5, which includes the spectograms for the synthetic, natural, and modified signals, respectively. These plots show how the PS-DTW-OLA scheme was able to simultaneously map the label marks from the synthetic to the natural sentence (performing automatic speech labeling) and perform the frame insertion/deletion of the synthetic frames to satisfy the alignment path.

In particular, two critical marks are annotated as A and B in Figure 5. Mark A indicates the first part of the /t/ plosive sound, which was expanded in the modified signal, as determined by the natural signal. In mark B, the /u/ vowel sound was compressed significantly, preserving its final portion, which is more similar to the sound present in the natural speech as detected by the PS-DTW algorithm.

Experiment 2: The PS-DTW method was applied to an automatic labeling system, in which segment marks of an annotated signal are automatically transferred to a test signal via a time-alignment operation. These transferred time marks are then compared to reference marks obtained manually by an expert. Performance of standard and PS-DTW methods are compared with and without V/UV partial classification. The mean squared error (MSE) between transferred and manual marks for each time-alignment method is shown in Table I. This table also includes the percentage of absolute errors be-

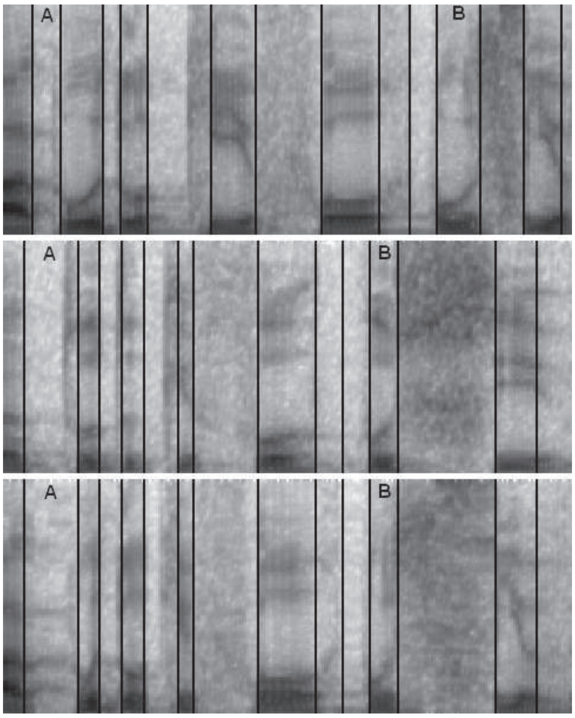

Fig. 5. Spectograms for synthetic, natural, and modified speech signals in Experiment 1 using PS-DTW-OLA algorithm, yielding automatic segmentation and labeling of natural speech or prosody transplant onto synthetic speech.

low 5, 10, 15, 20, and 30 ms. Results shown in Table I indicate great efficiency of PS-DTW methodology when combined to U/UV method, improving MSE and error percentages below a given threshold.

TABLE I
MEAN SQUARED ERROR (MSE) AND PERCENTAGE OF ABSOLUTE TIME ERRORS BELOW $T = 5, 10, 15, 20, 30$ MS IN EXPERIMENT 2.

Method	MSE	< 5	< 10	< 15	< 20	< 30
DTW	11.3	49%	75%	89%	92%	96%
DTW U/UV	10.3	52%	78%	91%	93%	97%
PS-DTW	11.1	55%	76%	89%	92%	95%
PS-DTW U/UV	9.5	58%	81%	92%	95%	97%

In pitch-transplant situations, PS-DTW advantages become even clearer, since its pitch resolution is perfectly suited for the TD-PSOLA stage, as illustrated next.

Experiment 3: Figure 6 illustrates the process of modifying the pitch contour from a male-speaker test signal (a) according to the pitch pattern from a female-speaker reference signal (b). The results from standard DTW/TD-PSOLA and PS-DTW-OLA are shown in subplots (c) and (d), respectively, in this same figure.

Figure 6 shows the exact correspondence of the PS-DTW-OLA pitch marks in subplot (d) with the ones from the reference signal in subplot (b). Meanwhile, the interpolations required by standard DTW/TD-PSOLA yield only an approximated correspondence, as observed in subplot (c).

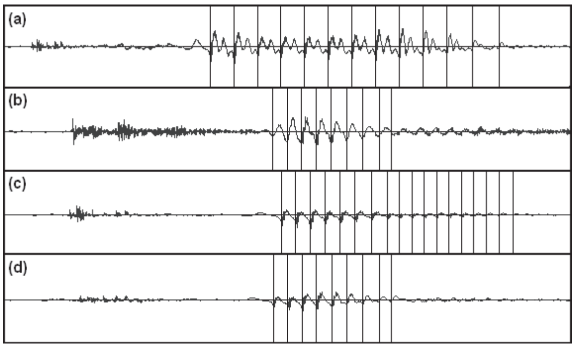

Fig. 6. Pitch transplant Experiment 3. Vertical bars indicate pitch marks in each signal: (a) test signal; (b) reference signal; (c) modified signal with standard DTW/TD-PSOLA; (d) modified signal with PS-DTW-OLA.

Informal subjective tests show the same perceived quality for the results of both standard DTW/TD-PSOLA and proposed PS-DTW-OLA transplantation methods. However, the proposed method requires less time resolution in the feature extraction, no interpolation and approximation in the OLA stage, and consequently less processing time: in our system development the proposed method was showwn to be 10 times faster than the standard one.

VII. Conclusion

A new algorithm for prosody enforcement was presented based on the pitch-synchronous (PS) time-alignment of speech signals. The proposed PS-DTW-OLA algorithm unifies the speech time-alignment (DTW) and pitch-adjusting (TD-PSOLA) stages. This simplifies the resulting alignment procedure, avoiding pitch-mark interpolations and subsequent time/spectral discontinuities in the modified signal, as illustrated in several computer experiments.

References

[1] F. Malfrere, O. Deroo, T. Dutoit, and C. Ris, "Phonetic alignment: Speech synthesis-based vs. Viterbi-based," *Speech Communications*, vol. 40, no. 4, pp. 503–515, 2003.

[2] M.-N. Garcia, C. d'Alessandro, G. Bailly, P. B. de Mareüil, and M. Morel, "A joint prosody evaluation of french text-to-speech synthesis systems," in *Proceedings of International Conference on Language Resources and Evaluation*, 2006, pp. 307–310.

[3] L. R. Rabiner, A. E. Rosenberg, J. G. Wilpon, and T. M. Zampin, "A bootstrapping training technique for obtaining demisyllable reference patterns," *Journal of the Acoustical Society of America*, vol. 6, no. 71, pp. 453–467, June 1982.

[4] H. Hohne, C. Coker, S. Levinson, and L. Rabiner, "On temporal alignment of sentences of natural and synthetic speech," in *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 31, 1983, pp. 807–813.

[5] A. W. Black and K. A. Lenzo, *Building Synthetic Voices*. Festvox, Jan. 2003. [Online]. Available: http://festvox.org/bsv/

[6] A. J. Hunt and A. Black, "Unit selection in a concatenative speech synthesis system using large speech database," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 373–376.

[7] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communications*, vol. 9, pp. 453–467, 1990.

[8] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, USA: Prentice-Hall, 1993.