
PROSODY MODIFICATION OF SPEECH AND SINGING FOR TUTORING APPLICATIONS

Neehar Jathar and Preeti Rao

Department of Electrical Engineering, IIT Bombay, Mumbai - 400076

E-mail: neehar.j@iitb.ac.in

In this work, we discuss prosodic transformations in terms of syllable durations and pitch, in the context of speech and music tutoring applications. We address some specific issues that arise with the use of TD-PSOLA based time- and pitch-scaling in the context of the singing and speech transformation to pre-defined target prosody. Time alignment is performed by matching automatically detected syllable onsets of the source and target followed by time-scaling and pitch-shifting using TD-PSOLA with attention to the choice of pitch marks and analysis-synthesis windows. Experiments demonstrate that TD-PSOLA can provide artifact-free perceived quality without explicit pitch mark detection by using longer analysis-synthesis windows.

1. Introduction

The term *prosody* refers to the pitch, duration and loudness dynamics of speech. The pitch dynamics are also referred to as *intonation*. These are essentially supra-segmental properties of speech that do not affect the meaning of the utterance (unlike segmental properties that are specified by the phone sequences that form words, and further, sentences). On the other hand, the prosody enhances speech communication by helping a listener to detect phrase and sentence boundaries, the mode of the utterance (e.g. whether declarative or interrogative) and the associated emotions of the speaker, if any, such as anger or sadness. In languages such as English, with its prominent stress patterns, prosody in the form of syllable durations plays a role in correct word pronunciation too. Learning to speak a new language involves acquiring both, proper pronunciation of words and the prosody of the language. In fact, an important characteristic of the non-native sounding accents of language learners lies in the speaker's imputation of his own native language prosody in the new language. Spoken language learners usually acquire skills through the imitation of a teacher's 'model' utterances. Such learning is aided when the model voice is familiar or similar to the students' voice implying that the best such voice is the student's voice itself [1]. Hence in interactive language learning applications, using signal processing to modify a given speaker's utterance, thus achieving the desired prosody while retaining voice identity, can make for very effective learning

material. Very similar considerations would apply in the case of singing tutoring. Prosody has a special importance in music as it pertains to the melodic (pitch variation) and rhythmic (syllable onsets and durations) characteristics which must be accurately reproduced in the course of learning a song. Achieving natural sounding modified speech given varied extents of mismatch in the source and target prosodies is challenging [2].

The above applications clearly motivate the problem of automatic prosodic modification of speech and singing signals to achieve target prosody while retaining speaker voice characteristics. Various time- and frequency-domain signal transformations have been previously investigated by researchers [3]. The speaker voice characteristics and segmental (phone sequence) properties of the source utterance are retained while transforming the pitch and duration to match that of a given target utterance. The imputation of target prosody on a source utterance is based on the temporal alignment of the source and target syllables. The target prosody, however, is derived from the written score of the music for singing tutoring, and from the canonical prosodic form of the mode of the utterance in spoken language tutoring. Thus the target prosody is specified directly in terms of the model durations and pitches of the syllables without an acoustic realisation of the utterance.

In this paper, we present a system for prosody modification of input speech given the model prosody, using TD-PSOLA. TD-PSOLA is the algorithm of choice for both time scaling and pitch shifting in most cases because of its computational efficiency and relatively good performance in terms of naturalness of the synthesized sound [4] [5]. Temporal alignment is typically achieved using a dynamic time warping algorithm using a cost function based on similarity of frame-level LPC or MFCC vectors [4] [5]. However, this is not possible in the case where the original (target) song/utterance is either polyphonic or not available at all. For pitch shifting, traditional implementations of TD-PSOLA use analysis-synthesis windows whose length is twice the local pitch period and require high precision in pitch mark detection for good quality synthesis [6]. By eliminating the dependence on accurate pitch marking, the computational requirements of TD-PSOLA can be further reduced. The above mentioned improvements are the focus of the present work.

The paper is organized as follows. Section 2 discusses the temporal alignment and time-scaling method used while Section 3 presents the pitch shifting technique. The dataset and experimental evaluation of the presented methods appear in Section 4. Finally, a critical discussion of the achieved performance with suggestions for improvement and future work are presented.

2. Temporal Alignment and Time Scaling

Consonant-vowel transitions are detected in the input (I) utterance via spectral change detection [7]. Sub-band energy in the frequency range between 680-2000 Hz is calculated for each 10 ms frame of the signal and the positive peaks of the first difference function of the energies are denoted as syllable onsets. This method is found to achieve an accuracy of ~90-94%. The alignment of detected onsets with target (T) onsets is based on the temporal proximity after compensating for any overall fixed delays. Next, based on the aligned onsets, the corresponding syllables from I and T are identified, and the time scaling factor (α) to be applied to each I syllable is computed as below.

$$\alpha_k = \frac{T_{k+1} - T_k}{I_{k+1} - I_k} \quad (1)$$

T_k denotes the onset time of the k^{th} syllable in target clip T, and I_k denotes the onset time of the k^{th} syllable in input source clip I.

Time scaling is performed using TD-PSOLA. The input signal pitch is calculated for every 10ms frame using the YIN pitch detection algorithm [8]. We denote the first pitch mark as the first location such that windowing around the pitch mark does not require zero padding.

Every subsequent pitch mark is defined as the location which is one pitch period after the previous pitch mark location where the pitch period is the local pitch period at the previous pitch mark.

$$PM_i = PM_{i-1} + \text{pitchperiod}(PM_{i-1}) \quad (2)$$

Wherever there are unvoiced portions or silences, a uniform pitch of 150 Hz is assumed and pitch marks are defined accordingly. This allows a smooth transition from voiced to unvoiced regions.

The input signal is then windowed around every pitch mark with an analysis Hanning window of length 3.2 times the local pitch period (this is justified in Section 3.1). These windows are then placed at synthesis pitch marks (calculated in the same way as analysis pitch marks) and overlap-added. For a time scaled signal, t ms of input signal should correspond to αt ms of output signal where the time scaling factor (α_k) for every syllable has been calculated already. To achieve this, analysis windows are either discarded or repeated whenever the disparity between current synthesis pitch mark and the desired length of the synthesized signal so far exceeds a threshold. To calculate the disparity at any point, the cumulative scaling factor up to that point (α_{cum}) is calculated and disparity is denoted as

$$\text{disparity} = PM_{syn,curr} - \alpha_{cum} \cdot PM_{anal,curr} \quad (3)$$

If disparity is positive, an analysis window is discarded while if disparity is negative an analysis window is repeated. The threshold value used is 5 ms. The resultant signal is a time scaled version of the input signal with syllable onsets aligned with the onsets in the target signal and the same pitch as the original. This can then be passed on to the pitch shifter to match the input pitch to the target template.

3. Pitch Shifting

Once temporal alignment of the signals is achieved, the target pitch data in terms of sampled pitch values across the syllable duration can be applied to the input signal directly. The pitch is computed for the time scaled version of the input signal every 10 ms using [8]. The pitch shifting factor α_p is calculated at every 10 ms as follows:

- If the frame is voiced in both signals, α_p is, naturally, the ratio of the target signal pitch to the input signal pitch
- If the frame is unvoiced in both signals, α_p is 1 because pitch shifting has no meaning for unvoiced frames.
- If the frame is voiced in one signal and unvoiced in the other, the previous value of α_p is used for the frame.

If the pitch shifting is being done for the speech tutoring application, the pitch scaling factor α_p is calculated with respect to the mean pitch of the utterance, since we are interested in reproducing the pitch contour shape only (not absolute pitch), as follows where pitch is specified in Hz.

$$\alpha_p = \frac{\text{mean_pitch}_{input}}{\text{mean_pitch}_{target}} \times \frac{\text{pitch}_{target}}{\text{pitch}_{input}} \quad (4)$$

Pitch shifting is done using TD-PSOLA as summarized in Fig. 1. Analysis pitch marking is done in a similar way as described in Section 2. Synthesis pitch marks are placed using the target pitch period information instead of input pitch or by suitably scaling the pitch period by the local α_p . The signal is windowed around analysis pitch marks as in the temporal alignment case (details of windowing are discussed in Section 3.1). These windows are then placed at the synthesis pitch marks and overlap-added. Whenever α_p is <1 , pitch period is increased and some windows have to be discarded to maintain the length of the signal. Whenever α_p is >1 , pitch period is decreased and some windows have to be repeated. The decision to repeat or discard windows is done as in Section

2, with a disparity defined as the difference between the current synthesis and analysis pitch mark. The same threshold of 5 ms is used here as well.

$$disparity = PM_{syn,curr} - PM_{anal,curr} \quad (5)$$

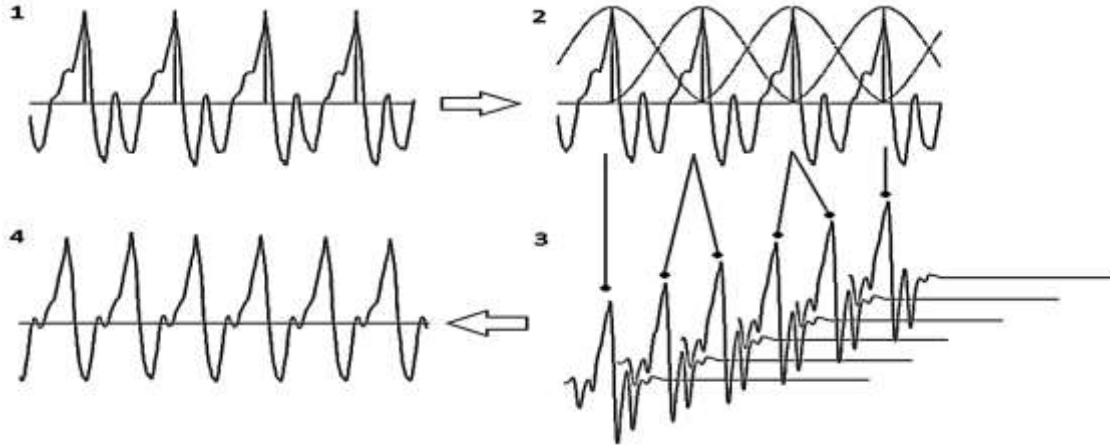


Figure 1: Overview of TD-PSOLA for pitch-shifting. (1) is the original signal - /a/ sound sung at 190 Hz (F# note). The vertical lines denote pitch marks. (2) denotes the signal windowed at every pitch mark with a Hanning window whose length L is proportional to the pitch period (here $L = 2 * \text{pitchperiod}$). (3) denotes the pitch shifting step with overlap-add. To increase the pitch, the windows are placed closer to each other than in the original. Further, some windows are repeated to maintain the time duration of the signal. (4) is the final synthesized signal which is pitch shifted and has the same duration as the original signal

In order to minimize round off errors when converting from pitch information to pitch period in terms of integer multiple of the sampling period, the sampling rate is increased to 48 kHz before pitch shifting and then restored after pitch shifting. Finally, the synthesized signal has the same temporal alignment of syllables and the same pitch contour as the target signal.

3.1 Choice of Window

Traditional PSOLA literature proposes the use of Hanning windows with length equal to twice the local pitch period around pitch marks when time-scaling or pitch-shifting [6]. However, when pitch marks are not explicitly marked at signal peaks, a window may be centred at a pitch mark lying between two peaks. If the window length is only twice the pitch period, the resultant rapid tapering of the window reduces most of the energy in the slightly offset peak leading to unnatural sound. Also when Hanning windows are overlapped with a hop equal to half the window length, the amplitude envelope across windows is constant. However, when the hop length is changed as in the case of pitch shifting, the amplitude envelope constancy is violated. This is depicted in Fig. 2, for pitch scaling factor (α) values of 0.8, 1 and 1.25. To alleviate both these concerns, we consider using a Blackman window instead of a Hanning window, with length equal to 4 times the local pitch period. The amplitude variation in this case is much lesser than in the case of the Hanning window, as is depicted in Fig. 2.

We observed that a Blackman window with length 4 times the pitch period gives better pitch shifting results than a Hanning window with length 2 times the pitch period. However, the Hanning window with length 3.2 times the pitch period (PP) has a shape very similar to the Blackman window of length $4 * PP$. The amplitude fluctuation of the longer Hanning window, while still greater than the Blackman window, is much smaller than before (depicted in Fig. 2) and pitch shifting results are indistinguishable from the Blackman window case.

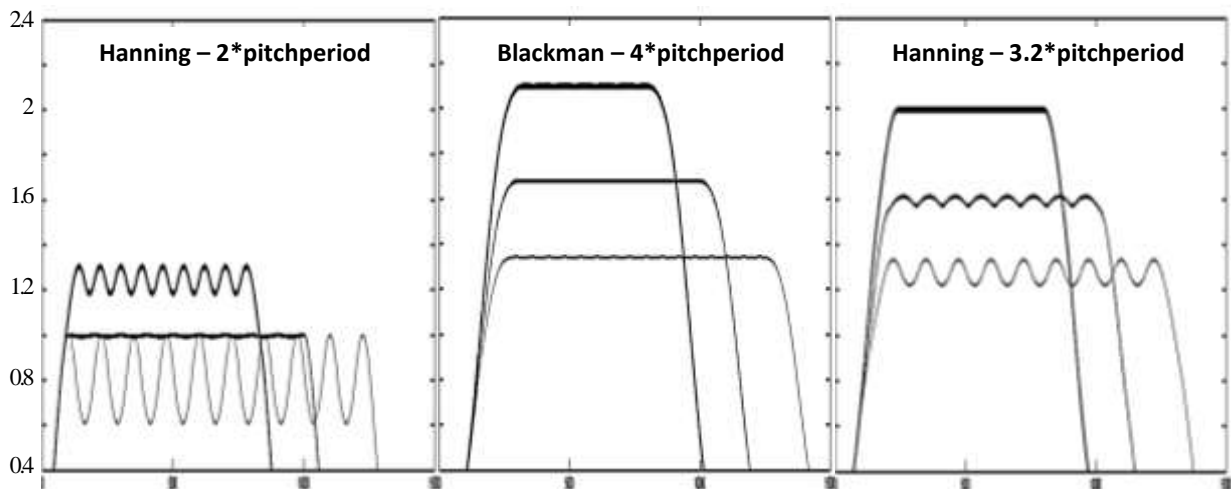


Figure 2: Amplitude versus time over several overlap-added synthesis windows for various pitch-scaling factors. From top to bottom in each figure: $\alpha = 1.25$, $\alpha = 1$ and $\alpha = 0.8$

4. Experimental Evaluation

The dataset used for listening evaluation consisted of excerpts from 8 Hindi movie songs sung by amateur singers (recorded in karaoke singing mode). The 8 singers comprised 4 male and 4 female singers. Each clip, sampled at 8 kHz, was approximately 100s long with silences during the instrumental part of the song. From this audio data, we selected 8 sung sections (1 per singer) of duration 15s. 4 of these 8 clips were time-scaled and pitch shifted to the final template and the remaining 4 were only time-scaled to align syllable onsets. Thus, a total of 8 synthesized and 8 natural (original) clips were a part of the listening test. Natural and time-scaled clips were selected such that the singer is largely in tune, to avoid the listener bias that an off-key performance implies a natural clip.

10 listeners judged each clip as either synthetic or natural. It was observed that listeners rated 30% of natural clips as synthetic implying that listener judgments are not perfect. Time scaled clips were virtually indistinguishable from natural clips and were rated as synthetic with similar frequency. Pitch shifted clips were rated as synthetic more often (~70%) due to artifacts in localised regions (detailed in Section 5) and not due to any overall quality degradation.

5. Conclusions and Future Work

Prosodic transformations are discussed in the context of speech and music tutoring applications. For given speech utterances or songs, target prosody is specified in terms of the desired syllable durations and pitch. We addressed some specific issues that arise with the use of TD-PSOLA based time- and pitch-scaling in the context of the singing and speech transformation to pre-defined target prosody. Since an acoustic target utterance is not necessarily always available, time-alignment is achieved by matching detected consonant-vowel transitions in the input utterance to syllable onsets in the available target prosody. The resulting alignment is observed to provide acceptable quality, free of perceptible artifacts, in the transformed signal. A second issue in the TD-PSOLA implementation is the choice of analysis-synthesis window shape and duration. The traditionally used Hanning window of duration twice the local pitch period works well for time scaling but is observed to give rise to waveform envelope modulation for moderate pitch scaling factors due to the imperfect overlap-add that occurs here and tapering due to the non-coincidence of window centre with a peak. Experimental and theoretical arguments are presented to show the superiority of a longer Hanning or Blackman window analysis-synthesis with pitch-scaling factors in the range of 0.8 to 1.25.

TD-PSOLA has been widely recognised to provide for high-quality transformations with computationally efficient implementations. Therefore, it is of continuing interest to address any signal-dependent issues that lead to artifacts or unnatural sounding speech. Syllable onset alignment is critical and dynamic programming based onset matching may be necessary when there are phonetic mismatches between the target and speaker. Further, intra-phone articulation variation leading to unnatural time-stretching merits attention. Pitch detection, used here without explicit pitch marker detection, may not be accurate enough in regions of rapidly changing pitch as can occur during singing. In such cases, the resulting small errors in analysis window placement can lead to signal cancellation especially when the waveform is locally fast varying due to the presence of strong high-frequency formants. The problem can be partly assuaged by resetting pitch markers to start at the first strong waveform peak after each syllable onset.

REFERENCES

- [1] Daniel Felps, Heather Bortfeld, and Ricardo Gutierrez-Osuna, 2009. Foreign accent conversion in computer assisted pronunciation, *Speech Communication*, **51**(10).
- [2] Werner Verhelst and Henk Brouckxon, 2003. Rejection Phenomena in InterSignal Voice Transplantations, *In IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003, New Paltz, New York, October (2003)
- [3] E. Moulines and W. Verhelst, 1995. Time-domain and frequency-domain techniques for prosodic modification of speech, *in Speech Coding and Synthesis*, pp. 519-555, Amsterdam(The Netherlands): Elsevier, 1995.
- [4] Werner Verhelst and Henk Brouckxon, 2002. Voice Modification for Lip Synchronization, Voice Dubbing and Karaoke, *Proc. 1st IEEE Benelux Workshop on Model based Processing and Coding of Audio*, Leuven, Belgium, (2002).
- [5] K.H.Lau, 2002. A System for Hybridizing Vocal Performance, *Audio Engineering Society Convention Paper 5625, presented at the 112th AES Convention*, Munich, Germany (2002).
- [6] Thomas Ewender and Beat Pfister, 2010. Accurate Pitch Marking for Prosodic Modification of Speech Segments, *Interspeech (ISCA) 2010*, Makuhari, Chiba, Japan, September 26-30 (2010).
- [7] Pradeep Kumar P, Preeti Rao, Sumantra Dutta Roy, 2007. Note Onset Detection in Natural Humming, *International Conference on Computational Intelligence and Multimedia Applications*, (2007).
- [8] A. de Cheveigne and H. Kawahara, 2002. YIN, a Fundamental Frequency Estimator for Speech and Music, *J. Acoust. Soc. Am.* 111, 1917–1930, 2002.