Novel Speech Duration Modifier For Packet Based Communication System

Senthil Kumar Mani, Jitendra Kumar Dhiman and K. Sri Rama Murty

Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India
{ee13p1007, ee11m04, ksrm}@iith.ac.in

Abstract

In this paper, we propose a real-time method for duration modification of speech for packet based communication system. While there is rich literature available on duration modification, it fails to clearly address the issues in real-time implementation of the same. Most of the duration modification methods rely on accurate estimation of pitch marks, which is not feasible in a real-time scenario. The proposed method modifies the duration of Linear Prediction residual of individual frames without using any look-ahead delay and knowledge of pitch marks. In this method, multiples of pitch period is repeated or removed from a frame depending on a scheduling algorithm. The subjective quality of the proposed method was found to be better than waveform similarity overlap and add (WSOLA) technique as well as Linear Prediction Pitch Synchronous Overlap and Add (LP-PSOLA) technique.

Index Terms: Voice over IP, linear prediction, look-ahead delay, WSOLA.

1. Introduction

As network characteristics (such as jitter, out of order, packet loss, etc.,) of packet based communication system (PBCS) varies with time, adaptive jitter buffer (AJB) [1] is required to control the buffering delay dynamically. In a voice over IP (VoIP) scenario, sudden rise in jitter may delay the reception of the packets. Hence, appropriate duration expansion should be applied to the decoded speech so that listener will not perceive abrupt discontinuity. When the network jitter comes to normalcy, speech duration has to be scaled down to maintain optimum end-to-end delay. Hence, duration modification is mandatory feature for achieving optimal perceptual quality in a packet based speech communication system.

Methods have been proposed in the literature for compression and expansion of speech duration [2]-[10]. However, most of them operate directly on the speech samples using waveform similarity overlap and add technique (WSOLA) [8]. These methods compress/expand the duration by merging/repeating the pitch periods. Thus, any fractional mismatch at the overlapping/concatenating point leads to perceptual distortion. At the same time, waiting for a perfect similarity of two successive pitch periods will limit the maximum scaling rate and may impact adaptation rate of AJB.

Generally, the speech spectrum has more energy in the lower frequency range and accordingly frequency response of Linear Prediction (LP) synthesis filter has higher attenuation at high frequencies. The frequency response of LP synthesis filter closely resembles a low-pass filter and it tends to smoothen the output signal. So, even if there are any minor mismatches at the points of concatenation, the output is not affected as much as it usually would have been when operating with the waveform directly. Hence, doing manipulations in the LP domain may result in better perceptual quality compared to direct waveform manipulations [9]. In order to achieve better perceptual quality, the methods [2], [13], [14], [15] perform duration modification on the LP residual. These methods, typically, require the knowledge of the pitch marks which point to the instants of highest similarity in successive pitch periods.

The Linear Prediction Pitch Synchronous Overlap and Add (LP-PSOLA) method [2] divides the speech signal into overlapping frames by placing analysis window at pitch marks (analysis time instants). The analysis window is centered at a pitch mark and extended typically over two pitch periods during time scaling process of LP-PSOLA. It leads to usage of variable window size and introduces considerable algorithmic delay when used for PBCS. The epoch based technique presented by Sreenivasa et.al, [13] makes use of instants of significant excitations (Epochs) and LP residual. The accuracy of this method depends on prediction of exact epoch locations. This approach does not use frame based processing of speech instead it manipulates LP residual pitch synchronously using the epoch location. The determination of epoch locations and pitch synchronous processing of LP residual add more computational complexity compared to other algorithms. As this technique synchronizes across epochs, expansion or compression requires look-ahead delay of at least one pitch period when used for PBCS. Also, searching for pitch marks may not be feasible in real-time, particularly at the time instant AJB tries to reduce its buffering delay [11]. This paper addresses the above said issues by removing or expanding pitch periods in LP residual without using the knowledge of epoch locations/pitch marks. Hence, the proposed method gets rid of look-ahead delay introduced by other methods [2]-[15].

This paper is organized as follows: Section 2 provides detailed description about the proposed method. Section 3 provides test results. Section 4 finally concludes about pro- posed method and its advantages.

2. Proposed Method

The architecture of proposed method is illustrated in Fig. 1. The LPC coefficients and pitch are extracted from every input frame using Levinson-Durbin algorithm and Auto Correlation Function (ACF) weighted inverse of an Average Magnitude Difference Function (AMDF) respectively. And, every input frame is filtered using the LPC coefficients extracted. Then, the LP residual is time scaled uniformly based on the targeted scaling rate. The time scaled LP residual is inverse-filtered to produce synthesized time scaled speech. Since LP filter whitens the input signal and its degree of whitening increases during tonal and carrier (pronouncing /a/,/e/,/i/,/o/,/u/) regions, the pitch period is estimated from input waveform instead LP residual for better accuracy. The latest pitch period is estimated from the current

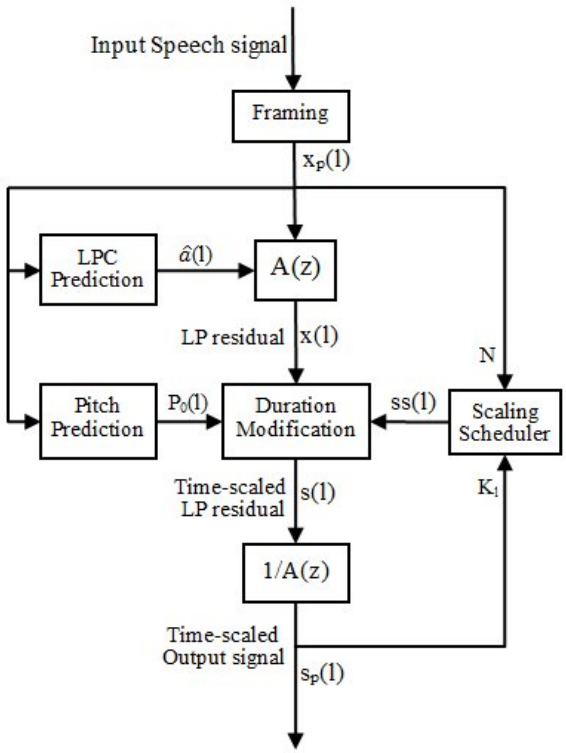

Figure 1: *Block diagram of proposed method*

frame by multiplying the inverse of an AMDF with ACF. The characteristics of the AMDF are very similar with that of the ACF. The AMDF produces a notch, while the ACF produces a peak. However, both functions essentially have the same periodicity. In noisy environment, the noise components included in the ACF and AMDF un-correlate each other. Hence, the peak of the ACF is emphasized in a noisy environment when the ACF is combined with the inversed AMDF [16]. As a result, it is expected that the accuracy of pitch extraction by the ACF will be improved when operating under shorter frame length or when detecting latest pitch period if it is different from previous pitch period. As proposed method removes or repeats latest pitch period or its multiple, extracting the latest pitch period using inverse AMDF weighted ACF will provide better quality than using ACF between current frame and previous frame. Thus, it helps in providing delay free scaling of speech signal when working on frame size of 10ms or multiple of 10ms in PBCS. The proposed method uses a scaling scheduler for triggering scaling uniformly. The scaling scheduler compares current scaling rate with target scaling rate and triggers time scaling accordingly.

Let us consider that l represents frame index, N represents frame size or number of samples from (LP residual) an input frame, $x(l, 1:N)$ represents input samples corresponding to frame index l and $s(l, 1:K_l)$ represents K_l number of output samples generated from the input frame l. Then, the total number of samples processed $t_i(l)$ till l^{th} frame from the starting (with $t_0(0) = 0$) is computed as given below

$$t_i(l) = t_i(l-1) + N \tag{1}$$

Similarly, the total synthesized output samples generated $t_0(l)$ till l^{th} frame from the starting (with $t_0(0)$ = 0) is computed as given below

$$t_0(l) = t_0(l-1) + K_l \tag{2}$$

where, K_l represents number of synthesized samples generated at l^{th} frame and its value may vary from frame to frame depending on the pitch period. Now, the current scaling rate $sr(l)$ at l^{th} frame can be computed as

$$sr(l) = \frac{t_0(l)}{t_i(l)} \tag{3}$$

If the current scaling rate $sr(l)$ is not meeting the targeted rate T after a frame processing, then scaling is triggered for next frame as given in equation (4). This process will be continued for subsequent frames.

$$ss(l+1) = \begin{cases} 1 & if(sr(l) < T) \& (T > 1) \\ 1 & if(sr(l) > T) \& (T < 1) \\ 0 \end{cases} \tag{4}$$

where, $ss(l+1)$ represents scaling indicator flag, $ss(l+1) = 1$ indicates that scaling is to be done at next frame, $ss(l+1) = 0$ indicates that scaling should not be done at next frame, and T represents targeted scaling factor.

For achieving higher rate of duration expansion (>50%), the same pitch period is repeated more than once. But, the maximum rate of duration compression will depend on number of pitch periods in the input frame. On average, approximately two pitch periods per frame can be removed.

2.1. Compression of Speech Duration

Whenever scaling scheduler triggers compression, a pitch period P_0 or multiple of pitch periods are removed in backward direction from latest sample in the current frame. If the sum of two successive pitch periods is $\leq 3N/8$, latest two pitch periods are removed by taking P_0 as sum of two successive pitch periods. In this case, P_0 is estimated from third formant of the inverse AMDF weighted by ACF. After removing pitch period(s), an overlap and add (OLA) is performed at the edge of balance data in the current frame for smooth continuity with the next frame as illustrated in Fig.2. The first segment used for OLA contains latest (P_0/4) samples as given below

$$cs_1(l, 1:P_0/4) = x(l, N - (P_0/4) + 1 : N) \tag{5}$$

The (P_0/4) samples present just before the segment (of duration P_0) that has to be removed is used as second segment for OLA. The OLA uses a triangular window with the rising edge for first segment and the falling edge for second segment as given below.

$$cs_2(l, 1:P_0/4) = x(l, N - P_0 - (P_0/4) + 1 : N - P_0) \tag{6}$$

$$ola(l,i) = cs_1(l,i)(4i/P_0) + cs_2(l,i)(1 - 4i/P_0) \quad \forall i = 1, ..., P_0/4 \tag{7}$$

Now, the last $P_0 + (P_0/4)$ samples are removed and the balance samples if any, in the current frame along with overlapped samples are copied to the output buffer as given in equation (8).

$$\begin{aligned} K_l &= N - P_0 \\ s(l, 1:K_l - (P_0/4)) &= x(l, 1:K_l - (P_0/4)) \\ s(l, K_l - (P_0/4) + 1 : K_l) &= ola(l) \end{aligned} \tag{8}$$

From the equation (8), it can be observed that the output samples generated from any input frame does not depend on samples from previous input frames. Hence, the proposed method

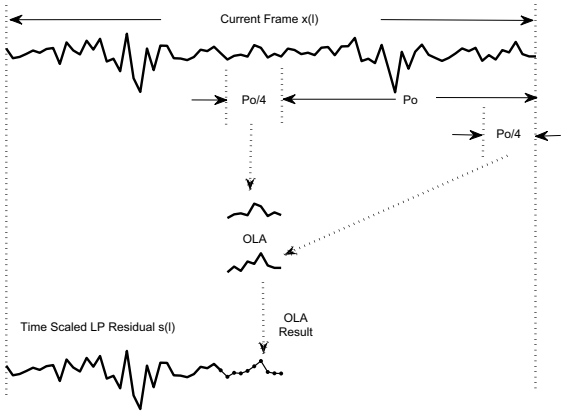

Figure 2: Illustration of speech duration compression

does not require any look-ahead delay unlike the conventional methods [8] - [13]. Also, the proposed method removes a pitch period completely without merging two successive pitch periods. Hence, there is no necessity to search for a segment with similar pitch periods to compress the duration unlike the conventional methods [8] - [13].

The compressed LP residual $s(l)$ is filtered with $\frac{1}{A(z)}$ to obtain time-scaled linear PCM samples.

2.2. Expansion of Speech Duration

The duration expansion is similar to the process of duration compression, but instead of removal, an extra pitch period is repeated. Here, unlike compression, the pitch period being larger than frame size can be handled with history of the past frame samples. If the pitch period is greater than 3/4 of frame size, the past frame is used.

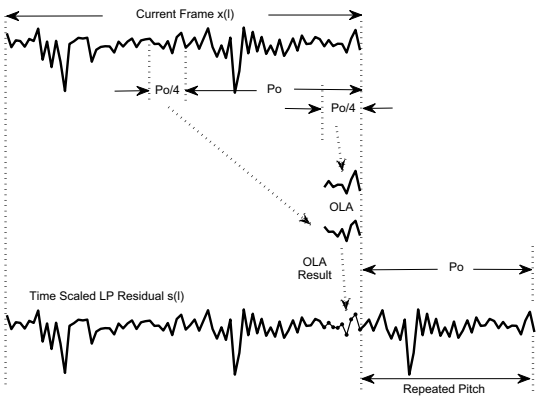

Figure 3: Illustration of speech duration expansion

Whenever duration expansion is triggered by the scheduler, latest pitch period or its multiple is repeated with OLA at the edge as illustrated in Fig.3. Similar to duration compression, if the sum of two successive pitch periods is $\leq 3N/8$, latest two pitch periods are removed by taking P_0 as sum of two successive pitch periods. Also, the segment of latest (P_0/4) samples and (P_0/4) samples present just before the P_0 samples are overlapped and added as given below

$$\begin{aligned} &es_1(l,1:P_0/4) = x(l,N-(P_0/4)+1:N) \\ &es_2(l,1:P_0/4) = x(l,N-P_0-(P_0/4)+1:N-P_0) \\ &ola(l,i) = es_1(l,i)(4i/P_0) + es_2(l,i)(1-4i/P_0) \\ &\forall i = 1,....,P_0/4 \end{aligned} \tag{9}$$

The current frame samples and OLA samples are copied into the output buffer as given below

$$\begin{aligned} &s(l,1:N-(P_0/4)) = x(l,1:N-(P_0/4)) \\ &s(l,N-(P_0/4)+1:N) = eola(l) \end{aligned} \tag{10}$$

Now, latest pitch period(s) is appended to output buffer as given below.

$$\begin{aligned} &K_l = N + P_0 \\ &s(l,N+1:K_l) = x(l,N-P_0+1:N) \end{aligned} \tag{11}$$

From the equation (11), it can be observed that the output samples from any input frame processing contain two parts. The first part contains all the samples of corresponding input frame except few samples used for OLA at the edge. The second part follows first part and contains samples of latest pitch period P_0 that was repeated. As this process does not modify any samples present at the edge of previous frame, duration expansion does not require look-ahead delay unlike the conventional methods [8]-[13]. Also, the conventional methods [8]-[13] insert a pitch period within the input frame (for duration expansion) using waveform similarity or pitch marks. But, the equations (10) and (11) illustrate that the pitch repetition is followed at the end of the input frame. Hence, only one OLA is required for every pitch period(s) repetition unlike the conventional methods that require two OLAs.

After repeating a pitch period, the scaling rate difference between current scaling rate and target rate is computed as given below

$$\bigtriangledown(l) = |sr(l) - T| \tag{12}$$

When $\bigtriangledown(l)$ is greater than 0.05 (i.e current scaling rate is 5% slower than target rate), multiple pitch periods are repeated to achieve the target rate. The number of pitch periods$\rho(l)$ repetition depends on the value of the $\bigtriangledown(l)$ as given below

$$\rho(l) = \begin{cases} 1 \;\; if(\bigtriangledown(l) < 0.05) \\ 2 \;\; if(\bigtriangledown(l) \geq 0.05) \& (\bigtriangledown(l) < 0.10) \\ 3 \;\; if(\bigtriangledown(l) \geq 0.10) \& (\bigtriangledown(l) < 0.15) \\ 4 \;\; if(\bigtriangledown(l) \geq 0.15) \end{cases} \tag{13}$$

The latest pitch period can be repeated at most 4 times (experimentally found) to avoid robotic sound as well as to avoid perceptual distortion. Every further repetition overwrites the output samples present at the edge of previous pitch period(s) repetition with OLA samples obtained in equation (10).

The expanded LP residual $s(l)$ is filtered with $\frac{1}{A(z)}$ to obtain time-scaled linear PCM samples.

3. Test Results

The proposed algorithm is implemented in Matlab and its performance is evaluated subjectively with 20 research scholars in

the age group of 25 to 35 years. The test results of two utterances (male and female) are demonstrated here. For each utterance, the duration is modified by factors of 10%, 20%, 30%, 40%, 50%, 60% and 70% using proposed, Epoch and WSOLA methods. Then, the output wave files are played to the subjects in a laboratory environment and subjects are asked to judge quality of the speech on a regular five point scale (i.e. 1-unsatisfactory, 2-poor, 3- fair, 4-good and 5-excellent). The performance comparison of speech duration expansion and compression using average mean opinion score (MOS) of both male and female utterances is depicted in Fig. 4 and Fig. 5 respectively. It can be observed that the performance of Epoch based [13] method is slightly better than WSOLA [8]. But, the proposed method out performs to both Epoch based and WSOLA methods. A set of duration modified utterances and corresponding results discussed in this work are available for listening at http://www.iith.ac.in/~ksrm/duration/Speech_Duration.html.

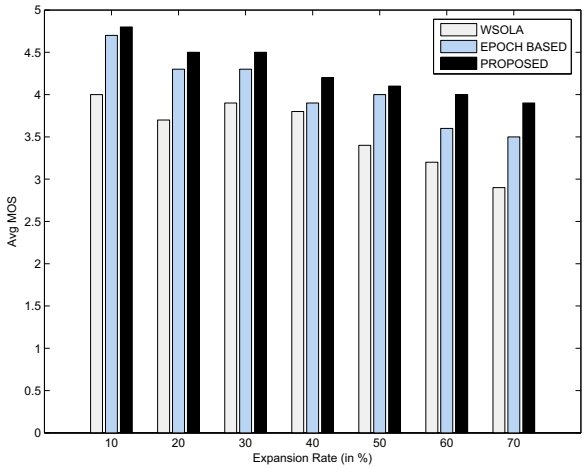

Figure 4: Performance comparison of speech duration expansion

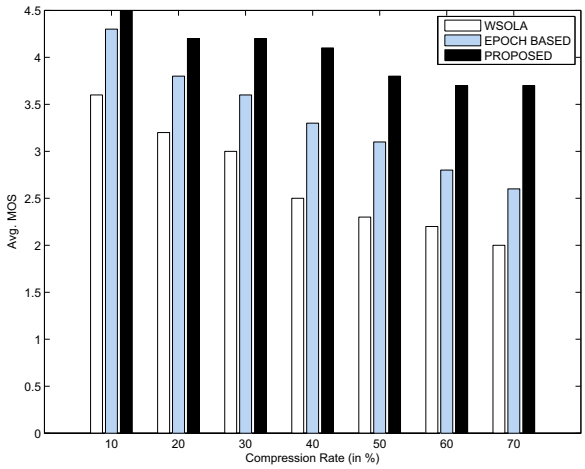

Figure 5: Performance comparison of speech duration compression

Whenever scheduler triggers time scaling, spotting similar pitch periods may not be possible. Hence, the phase discontinuity can be observed with WSOLA and Epoch based methods at some regions of duration modification where neighbouring pitch periods are different. However, this issue is resolved with the proposed method and can be used at any point of speech region for duration modification.

4. Conclusions

This paper proposed a look-ahead delay free, spotting location independent and high quality speech duration modification method that operates on LP residual. The performance results demonstrate that the proposed method is better than other conventional methods. The proposed speech duration expansion works with frame size of 10ms or its multiples. But, the duration compression works with frame size of 20ms or higher. However, it can be extended to work with 10ms frame size using 20ms memory buffer. If the pitch period is greater than 10ms, the algorithm will output zero samples and processes the current frame along with the next frame. During the next frame processing, the portion of the pitch period that extends to previous frame will be removed and rest of the procedure is similar to 20ms case.

5. References

[1] Yi J. Liang, Nikolaus Frber, and Bernd Girod, “Adaptive Playout Scheduling and Loss Concealment for Voice Communication Over IP Networks”, IEEE transaction on Multimedia, vol. 5, No. 4, December 2003.

[2] H. Valbret, E. Moulines, and J. P. Tubach,“Voice trans- formation using PSOLA techniques”,Speech Communication, vol. 11, pp. 175187, 1992.

[3] Brett Ninness and Soren John Henriksen, “Time-Scale Modification of Speech Signals”, IEEE transaction on Signal Processing, Vol. 56, NO. 4, APRIL 2008.

[4] Grofit, S. Lavner, Y. “Time-Scale Modification of Audio Signals Using Enhanced WSOLA With Management of Transients”, IEEE Transactions on Audio, Speech, and Language Processing, Page(s): 106 - 115, 2008.

[5] Werner VERHELST and Marc ROELANDS, “An Overlap-Add Technique Based on Waveform Similarity (WSOLA) for high quality Time scale Modification of Speech”, ICASSP, 1993.

[6] Dorran, David, Lawlor, R. , Coyle, E , “High quality time-scale modification of speech using a peak alignment overlap-add algorithm (PAOLA)”, ICASSP, 2003.

[7] Kim, J. Clements, M., “Time-scale modification of audio signals using multi-relative onset time estimations in sinusoidal transform coding”, Signals, Systems and Computers (ASILOMAR), 2010.

[8] Wang Lizhong, Wu Muqing, Li Mojia, “Waveform sim- ilarity over-and-add technique with gain control”, 2nd IEEE International Conference on Broadband Network and Multimedia Technology, 2009.

[9] S. Grofit and Y. Lavner, “Time-scale modification of audio signals using enhanced wsola with management of transients”, IEEE Transactions on Audio, Speech, and Language, vol. 16, no. 1, pp. 106-115, Jan. 2008.

[10] E. Moulines and J. Laroche, “Non-parametric techniques for pitch-scale and time-scale modification of speech”, Speech Communication, vol. 16, pp.175205, Feb.1995.

[11] AtriMukhopadhyay, TamalChakraborty, SumanBhunia, ItiSahaMisra, Salil Kumar Sanyal, “An Adaptive Jitter Buffer Playout Algorithm for Enhanced VoIP Performance”, Advances in Computing and Information Technology Communications in Computer and Information Science, Volume 198, 2011, pp 219-230.

[12] RabulHussainLaskar, KalyanBanerjee, Fazal Ahmed Talukdar K. Sreenivasa Rao, “A pitch synchronous approach to design voice conversion system using source-filter correlation”, International

Journal of Speech Technology, Springer Science+Business Media, LLC 2012.

[13] K. Sreenivasa Rao and B. Yegnanarayana, "Prosody Modification Using Instants of Significant Excitation", IEEE Transections on Audio, Speech and Language Processing, vol. 14, no. 3, may 2006.

[14] E. Gunduzhan and K. Momtahan, "A Linear Prediction Based Packet Loss Concealment Algorithm for PCM Coded Speech", IEEE Transactions on Speech and Audio Processing, vol. 9, no. 8, pp. 778-785, 2001.

[15] McAulay, R.J., "Sine-wave based PSOLA pitch scaling with real-time pitch marking", IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Oct. 2013.

[16] Senthil Kumar M., "Enhanced ITU-T G.711 Packet Loss Concealment (PLC) Algorithm", International Conference on Emerging Technologies and Applications in Engineering, Technology and Sciences (ICETAETS),13-14 January 2008.

[17] R. Crochiere, "A weighted overlap-add method of short time Fourier analysis/synthesis", IEEE Trans. Acoust., Speech, Signal Process., vol.ASSP-28, no. 1, pp. 99102, Feb. 1980.

[18] A. Jayan , P. Pandey and P. Lehana, "Automated detection of transition segments for intensity and time-scale modification for speech intelligibility enhancement", Proc. IEEE Int. Conf. Signal Process. Commun. Netw., pp.63 -68 2008.

[19] F. Nsabimana and U. Zolzer "Audio signal decomposition for pitch and time scaling", Proc. 3rd Int. Symp. Commun., Control, Signal Process., 2008.

[20] J.Wayman, R. E. Reinke, and D.Wilson, "High quality speech expansion, compression, and noise filtering using the SOLA method of time scale modification", in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 1989, pp. 714717.