PROSODIC MANIPULATION USING INSTANTS OF SIGNIFICANT EXCITATION

K. Sreenivasa Rao and B. Yegnanarayana

Speech and Vision Laboratory
Department of Computer Science and Engineering
Indian Institute of Technology Madras, Chennai - 600 036, INDIA.
E-mail: {ksr,yegna}@cs.iitm.ernet.in

ABSTRACT

This paper proposes a technique for prosodic (pitch and duration) manipulation using instants of significant excitation. Instants of significant excitation correspond to the instants of glottal closure (epochs) in voiced speech and to some random excitations like burst onset in the case of nonvoiced speech. Instants of significant excitation are computed from the average group delay of minimum phase signals. The manipulation of pitch and duration is achieved by modifying the Linear Prediction (LP) residual with the help of instants of significant excitation as pitch markers. The modified residual is used to excite the time-varying filter whose parameters are derived from the original speech signal. Perceptual quality of the synthesized speech is found to be natural, and is without any distortion. The original and corresponding synthesized speech signals from the proposed approach are available for listening at *http://speech.cs.iitm.ernet.in/Main/Results/Prosody.html*.

1. INTRODUCTION

The objective of prosodic manipulation is to alter the pitch and duration of a speech segment without affecting the characteristics of the spectral envelope [1]. Pitch and duration modification of speech is a subject of theoretical and practical interest. Applications include Text-to-Speech (TTS) synthesis, voice conversion and varying speech rate etc. [2]-[4]. For instance, in TTS, the basic-units and words must have their duration, and pitch is modified in order to incorporate the suprasegmental knowledge constraints of the utterance containing sequence of basic-units. This processing is necessary to avoid the production of monotonous sounding of synthesized speech, and is achieved by prosodic manipulation. In some other applications time-scale expansion is used to slow down rapid or degraded speech, enhancing its intelligibility. Conversely time-scale compression is used in message playback systems for fast scanning of recorded messages. To incorporate the prosodic knowledge many concatenation-based TTS systems employ the Time Domain Pitch Synchronous Overlap and Add (TD-PSOLA) approach [2],[4]. Some other systems perform prosody manipulation using sinusoidal-model, phase vocoders and Discrete Cosine Transform (DCT) based approaches, which are computationally intensive [1],[5],[6],[7]. Sinusoidal-model based approach represents the speech signal as a sum of sine-wave components with time varying amplitude, frequency and phase. For pitch modification, the model modifies the excitation signal frequency and phase. Problems arise when altering the pitch over a large scale, especially for increasing pitch, hoarseness is present in the reconstruction [1]. In TD-PSOLA approach, pitch marks are obtained using a pitch marking algorithm. Speech is synthesized by using superposition of Hamming windowed segments centered at the pitch marks, and extending between two adjacent pitch marks on either side. Duration modification is realized by deleting or replicating some of the windowed segments. Pitch period modification is realized by increasing or decreasing the superposition between windowed segments. This approach suffers from spectral and phase distortions. This is due to direct manipulation of the speech signal.The amount of distortion can be minimized to a large extent by operating on the Linear Prediction (LP) residual. In this paper an approach for prosodic manipulation is proposed, which operates on linear prediction residual using the knowledge of the instants of significant excitation as pitch markers.

The paper is organized as follows: In section 2 the proposed approach for prosodic manipulation is discussed. Manipulation of pitch is explained in section 3. Section 4 describes duration modification. The paper is concluded with the summary and possible extensions for future work.

2. PROPOSED APPROACH FOR PROSODIC MANIPULATION

The time-varying vocal tract system parameters and the corresponding LP residual signal are derived from the speech signal by LP analysis [8]. The instants of significant excitation are computed from the LP residual using group-

0-7803-7663-3/03/$17.00 ©2003 IEEE

This paper was originally published in the Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, & Signal Processing, April 6-10, 2003, Hong Kong (cancelled). Reprinted with permission.

delay analysis [9],[10]. The main step in extracting the instants of significant excitation is the computation of average slope of the unwrapped phase spectrum. This is also equivalent to computing the average group-delay. Accuracy of computation of the unwrapped phase depends on the (windowed) signal and the true phase values. If $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of the windowed signal *x(n)* and *nx(n)*, respectively, then the group-delay $(-\phi'(\omega))$ is given by [11],[9]

$$-\phi'(\omega) = \tau(\omega) = \frac{X_R Y_R + X_I Y_I}{X_R^2 + X_I^2}$$

where $X_R + jX_I = X(\omega)$ and $Y_R + jY_I = Y(\omega)$. Isolated peaks in $\tau(\omega)$ are removed by using a three point median filtering. The average value of smoothed $\tau(\omega)$ is computed. The resulting phase slope function is computed by moving the analysis window by one sample at a time. The positive zero-crossing instants of the phase slope function corresponds to the instants of significant excitation.

Voiced and nonvoiced portions of the residual are identified using energy thresholds. Fig.1 shows a segment of voiced speech, the LP residual, the phase slope function and the instants of significant excitation.

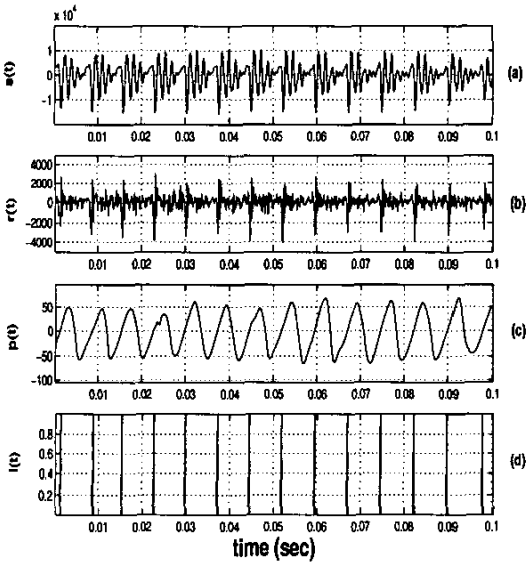

Fig. 1. (a) A segment of voiced speech (s(t)), (b) LP residual (r(t)), (c) Phase slope function (p(t)) and (d) Instants of significant excitation (i(t)).

The proposed approach for prosodic manipulation is to derive a new excitation signal incorporating the desired modification in the duration and pitch period. This is done by first taking the information about the epochs, and creating new epochs according to the desired pitch and duration modifications. Each epoch is associated with the time, pitch period, LP residual and LP coefficients (LPCs). For duration manipulation, new scaled time instants are obtained using the time scale manipulation factor. Likewise, for pitch manipulation, the pitch period associated with each instant is scaled appropriately. After obtaining the modified residual and LPCs for the desired prosodic manipulation, speech is synthesized by passing the residual through a time-varying filter defined by the LPCs. The synthesized speech quality is intelligible and distortionless.

3. PITCH MODIFICATION

For pitch manipulation, only voiced portions of the signal are considered. The new epoch sequence is generated from the epochs of the original signal, as per the required pitch modification factor. The residual is modified according to new instants, which represents the desired pitch modification.

Starting with the first instant in a voiced region, obtain the residual from the first pitch period. Depending on the modification factor, change the residual by padding with weighted random noise in the case of decreasing pitch, or by truncating the portion of the residual in the case of increasing pitch. In the case of decreasing pitch, after padding the residual with weighted random noise, the end point of the residual is checked against the following two instants. If it is close to first instant, then process the residual of next pitch period. Otherwise skip the next pitch period. While appending the weighted random noise, proper windowing is to be performed for smooth transition between the tail portion of the residual and the appended signal. In the case of increasing pitch, after truncating the portion of the residual, check the end point of the residual with the present and next instants. If it is close to the present instant, then replicate the residual until it is close to the next instant. Otherwise process the residual of the next pitch period. Truncating a portion of residual creates an abrupt discontinuity with the following instant, leading to audible distortions. This can be minimized by using proper window to smooth the abrupt truncation, and to provide a smooth roleoff to the next instant. In the case of pitch modification, the length of the residual is same before and after modification. Therefore the LPCs are not modified.

To demonstrate the effectiveness of the proposed pitch modification algorithm, sentences spoken by male and female speakers were analyzed and resynthesized by different factors. Fig.2 shows the LP residual of a voiced segment and modified residuals by modification factors of 1.5 and 0.75. As shown in the figure, the duration is kept constant. The modification in pitch values can be seen in the corresponding pitch contours shown in Fig.3. Fig.4 shows the

speech signals for the residuals shown in Fig.2. Fig.5 shows narroband (NB) spectrograms of the original and pitch modified (1.5 and 0.75) speech signals for the utterance *"This is a test recording for pitch manipulation and duration manipulation"*.

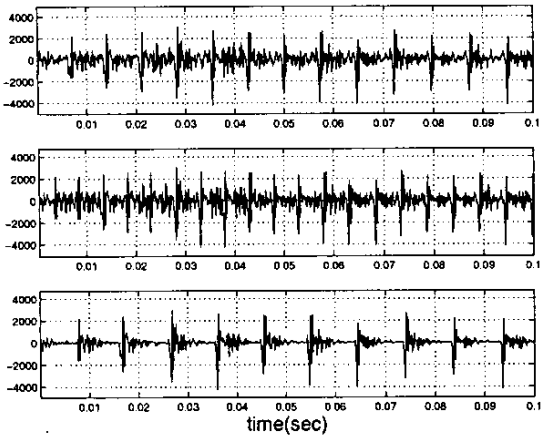

Fig. 2. Original LP-residual and modified residuals for modification factors 1.5 and 0.75 (respectively from top to bottom).

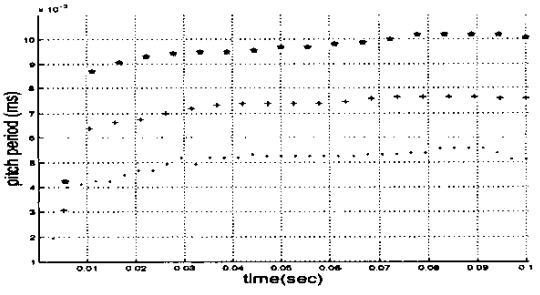

Fig. 3. Pitch contours for speech segments of the original (with /+/ symbols), pitch modified by 1.5 (with /./ symbols) and 0.75 (with /*/ symbols) factors.

4. MODIFICATION OF DURATION

In the case of duration modification, pitch contour should not alter, but the number of pitch cycles will vary according to the desired modification factor. In time-scale modification, the residual signal is modified uniformly from the starting instant to the last instant. Derive the new epoch sequence according to the required time scale modification. Access the residual using original instants and modify according to the new instant sequence. To increase the duration, replicate some portions of the residual at specified locations given by the algorithm. Similarly for reducing the

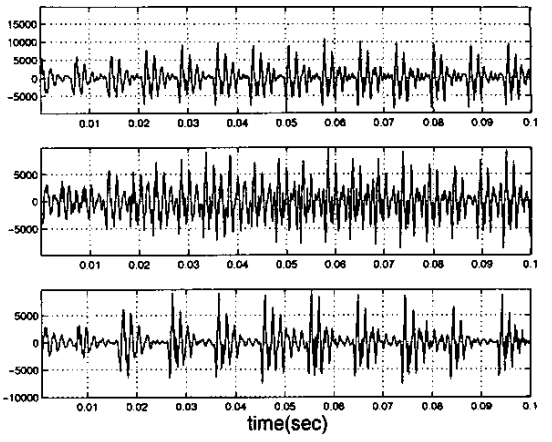

Fig. 4. Original speech signal and synthesized versions for the pitch modification factors of 1.5 and 0.75 (respectively from top to bottom).

duration, some portions of the residual are to be omitted at specific locations given by the algorithm, keeping track of the LPCs for the modified residual.

Fig.6 shows the speech utterance *"This is a test recording for pitch manipulation and duration manipulation"* , along with its synthesized versions and wideband (WB) spectrograms for the modified duration factors 1.5 and 0.75 .

5. CONCLUSIONS

The proposed method for time-scale and pitch-scale modification permits creation of high quality synthesized speech. The method is based on processing the LP residual using the knowledge of the instants of significant excitation. This approach provides flexibility in prosody manipulation for large range of the modification factor. The effectiveness of the proposed approach depends mainly on the accuracy in detecting glottal closure instants, because the entire residual manipulation is performed using these instants as anchor points. As we are performing the prosody manipulation in the residual domain without altering the vocal-tract characteristics, phase, spectral and audible distortions are not present in the synthesized speech. Along with prosodic manipulation it is also possible to vary the pole positions in z-plane, which gives variable vocal tract shapes suitable for applications like voice conversion.

6. REFERENCES

[1] Thomas F. Quatieri and Robert J. McAulay, "Shape invarient time-scale and pitch modification of speech," *IEEE Trans. Signal Processing*, vol. 40, pp. 497–510, Mar. 1992.

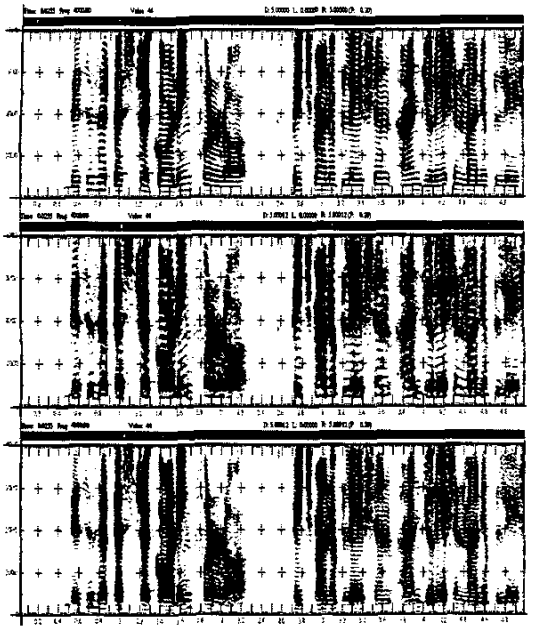

Fig. 5. NB spectrograms of the original and pitch modified (1.5 and 0.75) speech signals for the utterance *"This is a test recording for pitch manipulation and duration manipulation"*

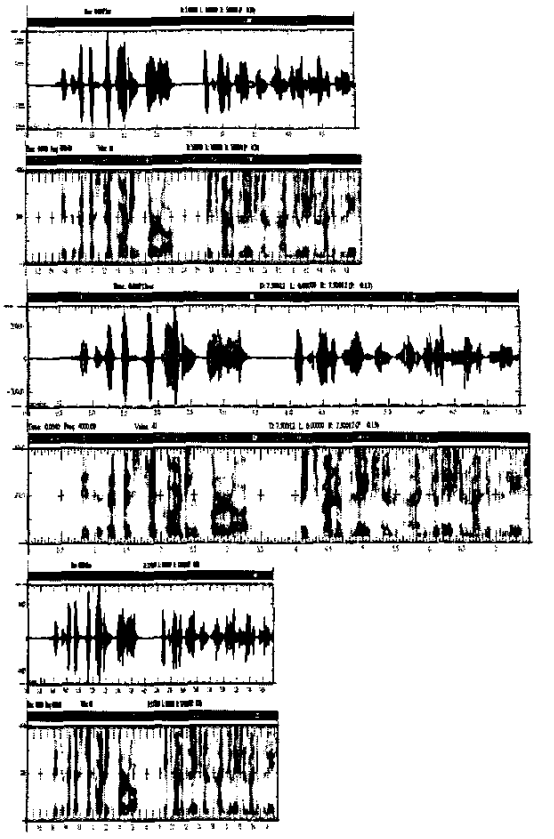

Fig. 6. Speech utterance *"This is a test recording for pitch manipulation and duration manipulation"*, and its synthesized versions with WB spectrograms for modified duration factors of 1.5 and 0.75

[2] Eric Moulines and F. Charpentier, "Pitch-synchronous wave form processing techniques for text to speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, Dec. 1990.

[3] D. G. Childers, Ke Wu, D. M. Hicks, and B. Yegnanarayana, "Voice conversion," *Speech Communication*, pp. 147–158, June 1989.

[4] Eric Moulines and Jean Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Communication*, vol. 16, no. 2, pp. 175–205, Feb. 1995.

[5] Joseph di Marino and Yves Laprie, "Supression of phasiness for time-scale modifications of speech signals based on a shape invarience property," *IEEE Proc. Int. Conf. Acoust., Speech Signal Processing*, May 2001.

[6] Michael R. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. ASSP-29, no. 3, pp. 374–390, June. 1981a.

[7] R. MuraliSankar, A. G. Ramakrishnan, A. K. Rohitprasad, and M. Anoop, "Dct baced pitch modification," *Proc. SPCOM 2001 6th Biennial Conference*, pp. 114–117, July 2001.

[8] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[9] P. Sathyanarayana Murthy and B. Yegnanarayana, "Robustness of group-delay-based method for extraction of significant excitation from speech signals," *IEEE Trans. Speech, Audio Processing*, pp. 609–619, 1999.

[10] Roel Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. Speech, Audio Processing*, vol. 3, no. 5, pp. 325–333, Sept. 1995.

[11] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck, *Discrete-Time Signal Processing*, Prentice-Hall, Upper Saddle River, NJ., 1999.