Modification of Pitch using DCT in the Source Domain

R. Muralishankar, A. G. Ramakrishnan and P. Prathibha

Department of Electrical Engineering, Indian Institute of Science, Bangalore-560012, INDIA.

Abstract

In this paper, we propose a novel algorithm for pitch modification. The linear prediction residual is obtained from pitch synchronous frames by inverse filtering the speech signal. Then the Discrete Cosine Transform (DCT) of these residual frames is taken. Based on the desired factor of pitch modification, the dimension of the DCT coefficients of the residual is modified by truncating or zero padding, and then the Inverse Discrete Cosine Transform is obtained. This period modified residual signal is then forward filtered to obtain the pitch modified speech. The mismatch in the positions of the harmonics between the pitch modified signal and the LP spectrum introduce gain variations, which is more pronounced in the case of female speech [16]. This is minimised by modifying the radii of the poles of the filter to smoothen the peaky linear predictive spectrum before forward filtering. This pitch modification scheme is used in our Concatenative Speech synthesis system for Kannada. The technique has also been successfully applied to creating interrogative sentences from affirmative sentences.

Keywords- Linear Prediction, Concatenative synthesis, residual signal, resampling, 3dB bandwidth.

1 Introduction

Machine synthesis of speech [1][2] facilitates convenient information transmission in a number of applications, including voice delivery of text messages and email, voice response to database enquires, reading aids for the blind and mobile communications. Speech synthesis presents a key challenge when it comes to improved quality [3], which is assessed by the attributes of intelligibility and naturalness. Of the various approaches to speech synthesis, concatenative synthesis has entailed speech with the highest quality to date. Concatenative synthesis involves selecting a class of basic acoustic units, creating an inventory of stored units by recording them from natural voice, and then generating utterances by concatenating appropriately modified segments from this inventory. A critical task in concatenative speech synthesis is that of modifying the prosody (pitch, amplitude and durations) of the voiced sections of the stored units and creating a concatenation of units that sounds seamless. Methods have been proposed in the literature for both time and pitch scale speech modification [4][5][6][7]. Pitch scale modification or pitch modification has applications such as adjusting the pitch in a singer's voice to get the desired effect, helping hearing impaired to understand speech better and modifying speech so that it is easier to code efficiently [8]. The objective of pitch modification is to alter the fundamental frequency of speech without affecting the time-varying spectral envelope. Techniques exist in the literature that accomplish this in the time or frequency domain.

1.1 Time domain pitch modification

Time domain pitch synchronous overlap adding (PSOLA[9]) is likely the simplest method that can be imagined for high quality pitch modification of speech signals. In practice, the implementation of pitch modification in time domain (TD-PSOLA) requires knowledge about the pitch pulse locations. Exact pitch pulse locations are not essential, but it is crucial to maintain an exact pitch synchronicity between successive pitch marks. The signal is windowed pitch synchronously using a Hamming window of length 2-4 pitch periods, centered around the current pitch pulse. A length of 2 periods is usually good for pure time-domain modification and a longer window (>2) is good for frequency domain PSOLA (FD-PSOLA). Because the intervals between the pitch pulses are altered, the total length of the signal is modified and

thus time scale modification of speech is also usually needed in order to maintain the original length of the signal. It is implemented in a simple way: If the pitch is increased, some frames are used twice and if it is lowered, some frames from the original signal are left out in the synthesized signal.

1.2 Frequency domain pitch modification

Historically, the FD-PSOLA was the first pitch synchronous time scale and pitch scale modification technique proposed in the literature [11]. FD-PSOLA and residual domain PSOLA (LP-PSOLA) are two methods that can be adapted almost directly from the TD-PSOLA paradigm. These two methods are more flexible than the TD-PSOLA technique because they provide a direct control over the spectral envelope at both the analysis and the synthesis stages. In FD-PSOLA, prior to overlap add synthesis, each short-time analysis signal is modified ; the modification is carried out in the frequency domain on the short-time Fourier transform signal. The algorithm used is basically a frequency domain resampling, which leads to some complex problems in the synthesis stage. It can be said that, if features such as speaker identity hiding are not needed, TD-PSOLA leads to the same results with a much simpler implementation. In practice, FD-PSOLA differs from TD-PSOLA only in the definition of the short-time synthesis signals for pitch scale modifications.

In LP-PSOLA, prior to PSOLA processing, the signal is split into an excitation component e(n) and the spectral envelope A(z). Pitch scale modification is then carried out on the source (residual) signal. The output is obtained by combining the modified source signal with the time-varying spectral envelope usually using linear prediction. Synthesis is again complex and the details can be found in the literature [12].

In this paper, we present a new method of modifying the residual obtained after inverse filtering with linear prediction coefficients. Gimenez [13] modified the pitch by interpolating the residual signal, realized by either upsampling or downsampling. Both upsampling and downsampling remap the 0 to π scale to the new residual length corresponding to the given pitch modification factor. Once the residual is modified, the spectral envelope responsible for the formant structure will be superimposed by forward filtering with the same LP coefficients. Our approach is similar to the one above, but differs in the interpolation of the residual signal. Interpolation is carried out using forward and inverse orthogonal transformation of the residual signal [14]. Traditionally, low-pass filters are used in sampling rate conversions for upsampling as well as downsampling to avoid spectral repetitions and aliasing. With the help of fast transforms, computational complexity involved in sampling rate conversion can be significantly reduced. Depending on the pitch modification factor, truncation or zero padding is performed on the forward transformed residual and the modified forward transformed residual is inverse tranformed. For a time-varying pitch modification using upsampling or downsampling, the low-pass filter must be redesigned every time because, the cutoff frequency varies according to the pitch modification factor. This could very well be avoided using an orthogonal transform, irrespective of whether the pitch modification factor is constant or time varying. This method preserves the formant structure and the speaker identity remains unchanged. We have also made some modifications to the above algorithm for handling female speech. In this method, the filter parameters are modified to produce a magnitude response that is significantly less peaky than the original linear predictive model used for inverse filtering. This reduces the filter sensitivity to pitch modification [16]. The discrete cosine transform (DCT) [15] has been used in our algorithm for resampling the residual. Energy loss is minimal in resampling process because DCT has high energy compaction.

2 Method

As an alternative to strictly time domain techniques, the ubiquitious source-filter model of speech can be invoked [17]. Prosody modification then becomes a task of separating the excitation and vocal tract components from speech, modifying the excitation, and then recombining with the vocal tract component. In principle, this allows retaining the vocal tract response without any modifications. Ideally, the analysis would separate the excitation signal, which could be modified independent of the vocal tract response. In practice, the system attempts to separate the speech signal into a spectral shaping component, and a

residual signal, ensuring in the process that the original signal's temporal detail is preserved. The LPC [18] residual (error, excitation) signal has a number of advantages over the speech signal in the context of pitch modification [19]. The former is spectrally flat and there is little correlation within each pitch period.

2.1 Pitch marking

The first step in our analysis is to pitch mark the speech signal. For this task, an algorithm based on the autocorrelation of the speech signal has been used. In the autocorrelation domain, finding the local maxima and the distance between successive local maxima gives the periodicity of the signal under consideration. After getting the pitch information, it is submitted to various periodicity constraint rules, and linked together in order to obtain a chain of marks. A nonlinear processing of these marks that includes deletion, delay, interpolation and extrapolation, results in the final pitch marks positioned at the peaks of the signal in the voiced segments. Figure 1 shows a segment of a pitch marked signal. Unvoiced segments are marked 10 msec apart. For the rest of this paper, the voiced and unvoiced marks are both called as pitch marks. Because the marks are positioned at specific samples of the speech signal, the resulting period is quantized to an integer number of samples. This is a common procedure in pitch synchronous TTS systems and is employed in our algorithm. For a 10 msec pitch period and a 16 kHz sampling rate, for example, the error in the pitch period due to quantization is lower than 0.63%.

2.2 Resampling using DCT

Let $\{e_n;\ 0 \leq n \leq N_1 - 1\}$ be the residual signal obtained after pitch synchronous inverse filtering with LP coefficients. Signal expansion in orthogonal functions can be written as

$$e_n = \sum_{k=0}^{N_1-1} \theta_k \phi_k(n), \quad 0 \leq n \leq N_1 - 1$$

where,

$$\theta_k = \sum_{n=0}^{N_1-1} e_n \phi_k^{\star}(n), \quad 0 \leq k \leq N_1 - 1$$

The set of coefficients $\{\theta_k; 0 \leq k \leq N_1 - 1\}$ constitute the spectral coefficients of $\{e_n\}$ relative to the given orthonormal family of basis functions. In our algorithm, we use IDCT after truncating or zero padding θ_k to obtain a different pitch frequency and corresponding harmonics. This operation can be explained as a linear transformation $A^{'} : R^{N_1} \rightarrow R^{N_2}$,where $A^{'}$ is the IDCT $N_2 \times N_2$ matrix . For $N_1 > N_2$, pitch frequency increases, and for $N_1 < N_2$, pitch frequency decreases. The forward transformation of the residual signal can be represented in matrix form as

$$\underline{\theta} = A\underline{e}$$

where, $\underline{e^T}$ is the residual signal, A is the DCT $N_1 \times N_1$ matrix and $\{\theta_k; 0 \leq k \leq N_1 - 1\}$ are the DCT coefficients. The linear transformation of θ_k to $\{e'_l; 0 \leq l \leq N_2 - 1\}$ can be performed by premultiplication of $\underline{\theta}$ by the IDCT matrix. For $N_1 > N_2$, we truncate $\theta_k; 0 \leq k \leq N_1 - 1$ up to $N_2 - 1$ and premultiply with $N_2 \times N_2$ IDCT matrix:

$$\begin{pmatrix} e'_0 \\ . \\ . \\ . \\ . \\ e'_{N_2-1} \end{pmatrix} = \begin{pmatrix} \phi_0(0) & . & \phi_0(N_2-1) \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ \phi_{N_2-1}(0) & . & \phi_{N_2-1}(N_2-1) \end{pmatrix} \begin{pmatrix} \theta_0 \\ . \\ . \\ . \\ . \\ \theta_{N_2-1} \end{pmatrix}$$

where ϕ_k's are the DCT basis vectors. For $N_1 < N_2$, we pad (N_2-N_1) zeros to $\theta_k;\ 0 \leq k \leq N_1 - 1$ to obtain $\theta_l;\ 0 \leq l \leq N_2 - 1$ and then premultiply with $N_2 \times N_2$ IDCT matrix:

$$\begin{pmatrix} e'_0 \\ . \\ . \\ . \\ . \\ e'_{N_2-1} \end{pmatrix} = \begin{pmatrix} \phi_0(0) & . & \phi_0(N_1-1) \\ . & . & . \\ . & . & . \\ . & . & . \\ . & . & . \\ \phi_{N_2-1}(0) & . & \phi_{N_2-1}(N_1-1) \end{pmatrix} \begin{pmatrix} \theta_0 \\ . \\ . \\ . \\ . \\ \theta_{N_1-1} \end{pmatrix}$$

Components of the basis vectors after $\phi_k(N_1 - 1)$are not considered because, they multiply the padded zeros in θ_l and therefore, will not contribute to the output e'_l.

2.3 Modification of the residual signal

Figure 2 shows the details of the proposed method. The process starts with pitch synchronous extraction of the residual signal. The length of the residual signal of each frame is modified using DCT-IDCT. N_1 point DCT of each frame of the excitation signal is obtained, where N_1 corresponds to the actual number of samples in each extracted frame. An N_2 point IDCT is then obtained, where N_2 corresponds to N_1 multiplied by the ratio of modified to original pitches. Before taking IDCT, normalization must be carried out to compensate for the change in length of the residual signal after truncation or padding with zeros. The effect of taking a N_1 point DCT followed by an N_2 point IDCT has an effect almost amounting to resampling the excitation signal. This occurs because, while taking the IDCT, the new length of the transformed residual after truncation or padding is mapped to π. Each frame of the speech signal is then synthesized by forward filtering.

2.4 Modification of all-pole filter coefficients

It is known that linear prediction using a least sqares error criterion produces spectral estimates that are biased towards the pitch harmonics [18]. Bandwidth estimates are typically poor. One often observes signal degradation in LPC pitch modified speech, especially for female speech. This is because of the gain variations with the new harmonic positions. Observations on the difficulty of modeling the data from a single recording are discussed in [16]. In that work, the filter parameters are not chosen to model the data by minimizing the residual energy, but to have sensitivity to pitch modification. The system parameters are determined in a pitch synchronous manner. A 14^{th} order all-pole filter was used for representing the signal in each pitch period in the voiced portion. The magnitude response is chosen to have a significantly less peaky structure than that which is typically obtained in LPC. The covariance matrix of the data in each frame was modified so that it produces an all-pole filter with chosen lowpass response whenever the signal energy is reduced to zero. In our approach, the magnitude response of the LP spectrum is made less peaky (see Fig. 3), by adaptively decreasing the radius. The polynomial in z formed by the LP filter coefficients is solved for the roots (poles), which in turn give angle (Θ) and radius (roots represented in polar form). From theta, we get the information about the frequency of the corresponding peak in LP spectrum ($f_i = \Theta_i F_s/2\pi$) and from radius (r), we get the 3-dB bandwidth ($B_i = -ln(r_i)F_s/\pi$) of the peak, where, F_s is the sampling frequency. Depending on the pitch modification factor, the bandwidth is increased to accommodate the new harmonic positions. This in turn decreases the radius. The modified LP coefficients are used for forward filtering.

3 Results and discussions

To demonstrate the effectiveness of this technique, individual phonemes, words and sentences spoken by a male volunteer were analyzed and re-synthesized for different pitch change factors. Figure 4(a) shows a segment of a phoneme. Figure 4(b) gives the corresponding segment of the residual signal extracted by inverse filtering the phoneme (LP model order 14). Figure 4(c) shows the length-modified residual signal

obtained through DCT-IDCT, the factor of increase in pitch being 1.5. Figure 4(d) shows a segment of the resynthesized speech signal after forward filtering. Figure 5 shows the pitch contours for the phoneme shown in Fig. 4(a), and its pitch modified versions for factors 0.7 and 1.4. It can be seen that the pitch contour is maintained in the modified signal. Figure 6 shows the speech signal for a whole word, its original pitch contour and the contour after pitch modification using our technique. Incorporating modified version of the pitch modification algorithm discussed earlier, for a sentence from a female voice and its pitch contours are shown in Fig. 7. Time varying pitch modification using the above algorithm is shown in Fig. 8. The characteristics of interrogative sentences with an "yes or no" answer is that both the pitch contour and the amplitude rise sharply for the last syllable [20]. With a linearly increasing pitch modification (in addition to linear amplitude modification), we have raised the pitch of the last syllable of the affirmative sentence up to a factor of 1.3 to obtain interrrogative effect. Figure 9 shows energy loss due to the truncation of the DCT coefficents for pitch modification factor from 1 to 2. We can observe that the energy loss is a monotonically increasing function of the pitch modification factor. The loss is acceptably small at less than 13% of total residual energy up to a pitch modification factor of 1.6, as shown in Fig 9.

To evaluate the performance of the proposed technique, perceptive evaluation tests were carried out. The pitch contours of many phonemes, words and sentences were modified by different factors. Nine people were asked to rate the intelligibility, speaker identity and distortion after the modification. The result of the evaluation test is given in Table 1. From this table, we can see that quality of the pitch modified speech is better for the modification range from 0.8 to 1.3. This is a reasonably sufficient range for speech synthesis. Even for the case of modification to obtain interrogative effect (shown in Fig. 8), the maximum factor we needed to use was only 1.3. Table 1 shows that the perceptual evaluation for higher pitch modification factors results in fairly acceptable reconstructed speech. Currently, we are using our algorithm to convert an emotional utterance to a non emotional one, and vice versa. Thus, a sentence spoken in surprise is converted to a normal one by reducing the pitch by a suitable factor. Since the sentence spoken in surprise is naturally of shorter duration than a normal one, there is no need for duration modification. Similarly, normal utterances have been modified to generate emotions such as surprise and anger. Obtaining time varying pitch modification with TD-PSOLA is a very difficult task, because it involves shifting of the overlapped segment by different number of samples for different pitch synchronous frames. This is cumbersome, since the windowing effect needs to be compensated for differing lengths of overlap. However, in our case, since the DCT-IDCT operation is individually performed for each pitch synchronous frame, no additional complexity is introduced when the pitch change factor is time varying. Further, as already explained, the method proposed by Gimenez [13] requires redesigning the interpolation and decimation filters for varying factors of pitch change.

4 Conclusions

The proposed algorithm is simple and elegant. It directly follows from the basic source-filter model of speech. Perceptive evaluation shows that this performs well for the range of pitch change factors sufficient for a TTS system. The algorithm uses DCT-IDCT, and thus is not computationally intensive. The proposed scheme maintains the relative pitch contour of the original signal, without any additional processing or precautions to be taken. The same basic scheme is valid for both constant and time-varying pitch modification factors. In the case of female speech, when the pitch is modified even by small factors, gain difference occurs due to the peaky nature of the LPC spectrum. In such cases, the 3dB-bandwidth of the peak in LPC spectrum is adaptively increased to position the new harmonic peak so as to minimize the gain variation. This whole process smoothens the LPC spectrum and the modified algorithm is not sensitive to pitch marking errors[16].

Acknowledgements

We thank the Ministry of Information Technology, Government of India for funding part of this research under the project titled "Algorithms for Kannada speech synthesis".

References

[1] D. B. Roe and J. G. Wilpon (Eds), *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[2] A. Syrdal, R. Bennet, and S. Greenspan, *Applied Speech Technology*, CRC Press, Boca Raton, FL, 1995.

[3] M. Liberman, Computer speech synthesis: its status and prospects, *Voice communication between humans and machines*, National Academy of Sciences, 1994.

[4] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/overlap -add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, 40(6), pp. 497-516, June 1992.

[5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 34(4), pp. 744-754, 1986.

[6] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 30, pp. 374-390, June 1981.

[7] T. F. Quatieri and R. J. McAulay, "Speech transformation based on a sinusoidal representation," *IEEE Trans. Acoustics, Speech, Sig. Proc.*, 34(6), pp. 1449-1464, 1986.

[8] Ramo Anssi, "Pitch modification and quantization for offline speech coding," *M.S. Thesis*, Tampere University of Technology, May 1999.

[9] E. Moulines and F. Charpentier. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, 9, pp. 453-468, 1990.

[10] R. Vergin et al. "Time domain technique for pitch modification and robust voice transformation," *Proc. ICASSP 97, Vol. II of V , Speech Processing* pp. 947-950.

[11] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proc. Int. Conf. on Acoust. Speech and Sig. Proc.*, pp 2015-2018, 1986.

[12] W. Baastian Kleijn, Kuldip K. Paliwal, "*Speech coding and synthesis*", Elsevier Science B. V. 1995.

[13] F. M. Gimenez de los Galanes et al. "Speech Synthesis System Based on a Variable Decimation/Interpolation Factor,"in *Proc. Int. Conf. on Acoust. Speech and Sig. Proc.*, pp 636-639, 1995.

[14] K. R. Rao., and P. Yip, "*Discrete Cosine Transform: Algorithms, Advantages, Applications*", New York: Academic Press, 1990.

[15] N. Ahmed and K. R. Rao, "*Orthogonal Transforms for Digital Signal Processing*". New York: Springer, 1975.

[16] Rashid Ansari. "Inverse filter approach to pitch modification: Application to concatenative synthesis of female speech," *Proc. Int. Conf. on Acoust. Speech and Sig. Proc.*, pp 1623-1626, 1997.

[17] L. R. Rabiner, and R. W. Schafer. "*Digital Processing of Speech Signals*". Prentice-Hall, Inc Englewood Cliffs, New Jersy 07632.

[18] J. Makhoul, "Linear prediction: a review", *Proc. IEEE*, Vol. 63, pp. 561-580, April 1975.

[19] M. Edgington and A. Lowry. "Residual-based speech modification algorithms for TTS synthesis". BT laboratories, Martlesham Health IPSWICH, IP5 7RE, U. K.

[20] Masanobu Abe, Speaking Styles: Statistical Analysis and Synthesis by a Text-to-Speech System, *Progress in Speech Synthesis*, New York: Springer, 1996.

Table 1: Perceptual evaluation of pitch modified utterances as a percentage of the 9 evaluators.

Pitch change factor	Intelligibility			Distortion			Speaker identity		
	Good	Fair	Bad	Low	Medium	High	Good	Fair	Bad
0.6	55	45		22	34	44		78	22
0.8	77	23		34	66		55	45	
1.2	100			100			55	45	
1.3	89	11		100			55	45	
1.4	55	45		33	67		22	78	
1.6		100		22	56	22		55	45
1.9		78	22		67	33		22	78

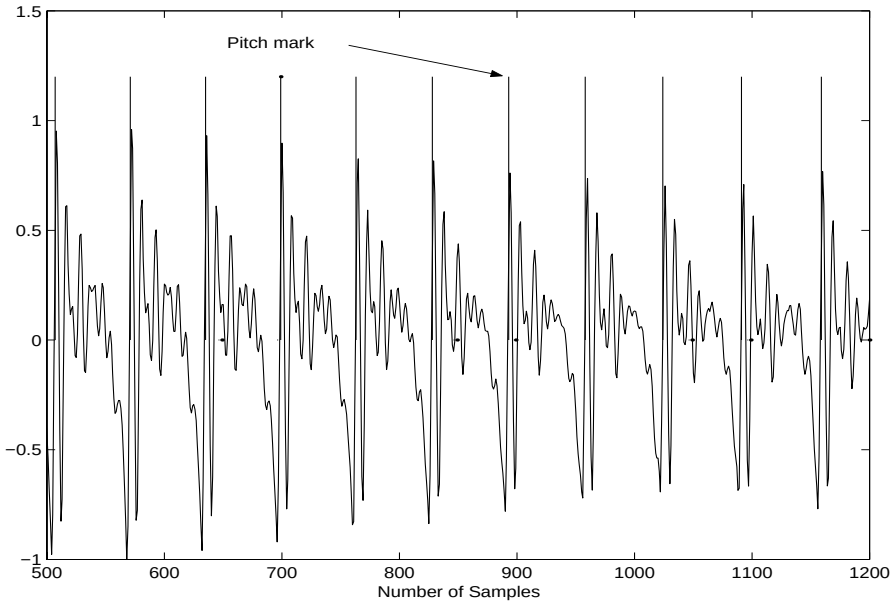

Figure 1: Pitch marked speech '/a/'

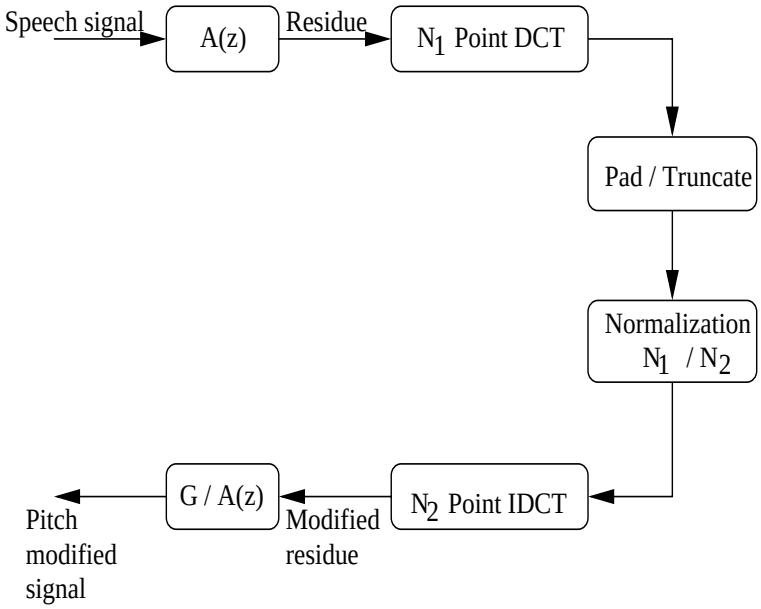

Figure 2: Block diagram of DCT based pitch modification

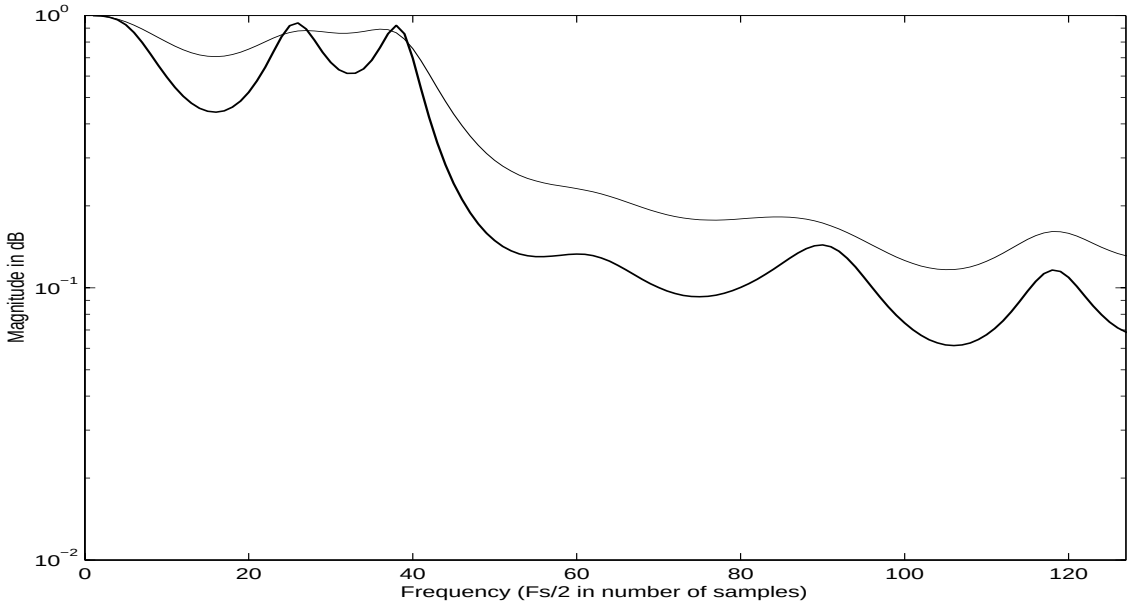

Figure 3: LPC spectrum and its smoothened version

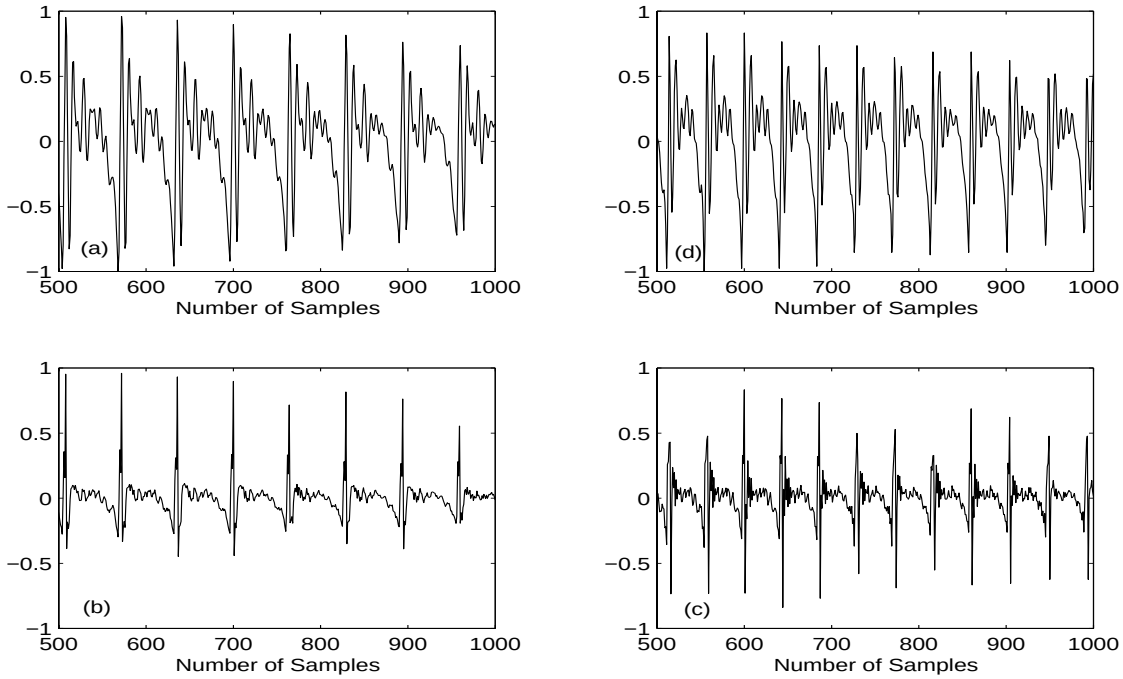

Figure 4: (a) Few frames of the original signal '/a/'. (b) Few frames of the original excitation. (c) Few frames of the modified excitation. (d) Few frames of the reconstructed signal.

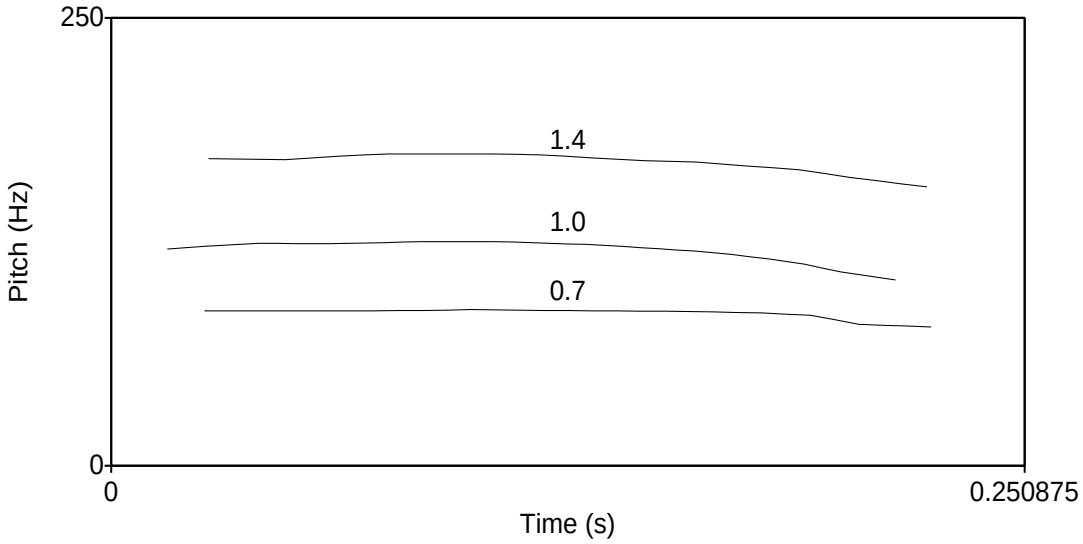

Figure 5: Pitch contours of the original and modified phoneme '/a/'.

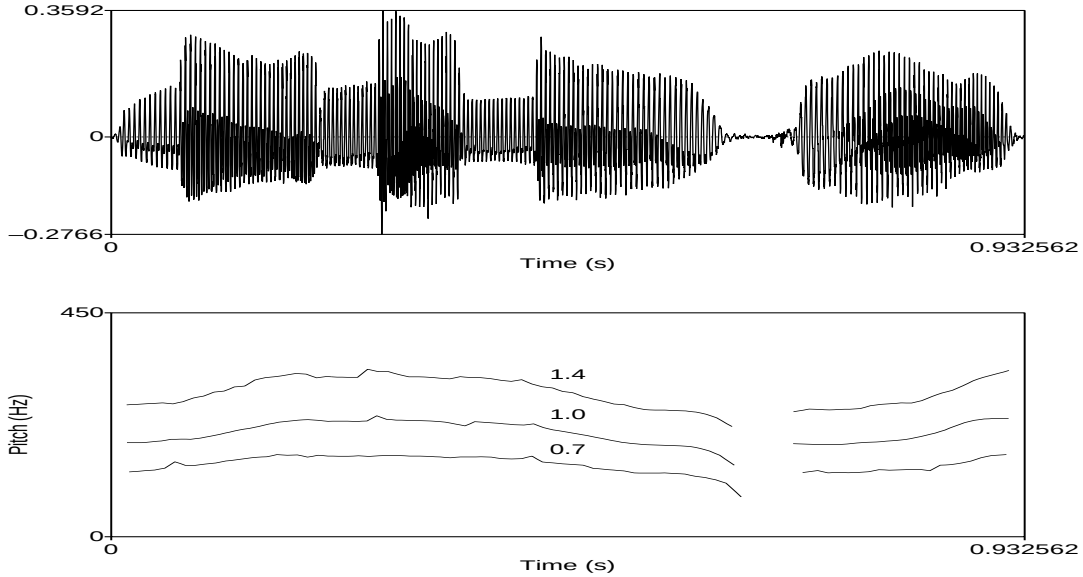

Figure 6: Speech signal for the word /niilamegha/ spoken by a female volunteer and its modified pitch contours.

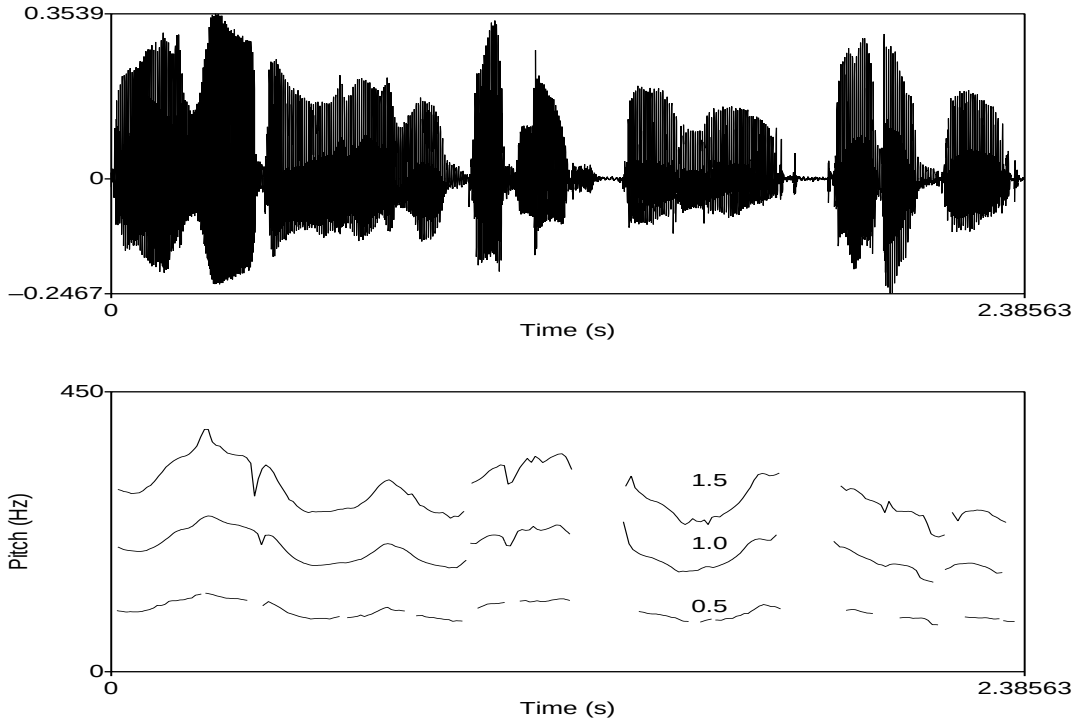

Figure 7: Speech signal of a sentence / kaaveeriya ugama sthana kodagu/ spoken by a female volunteer and its modified pitch contours.

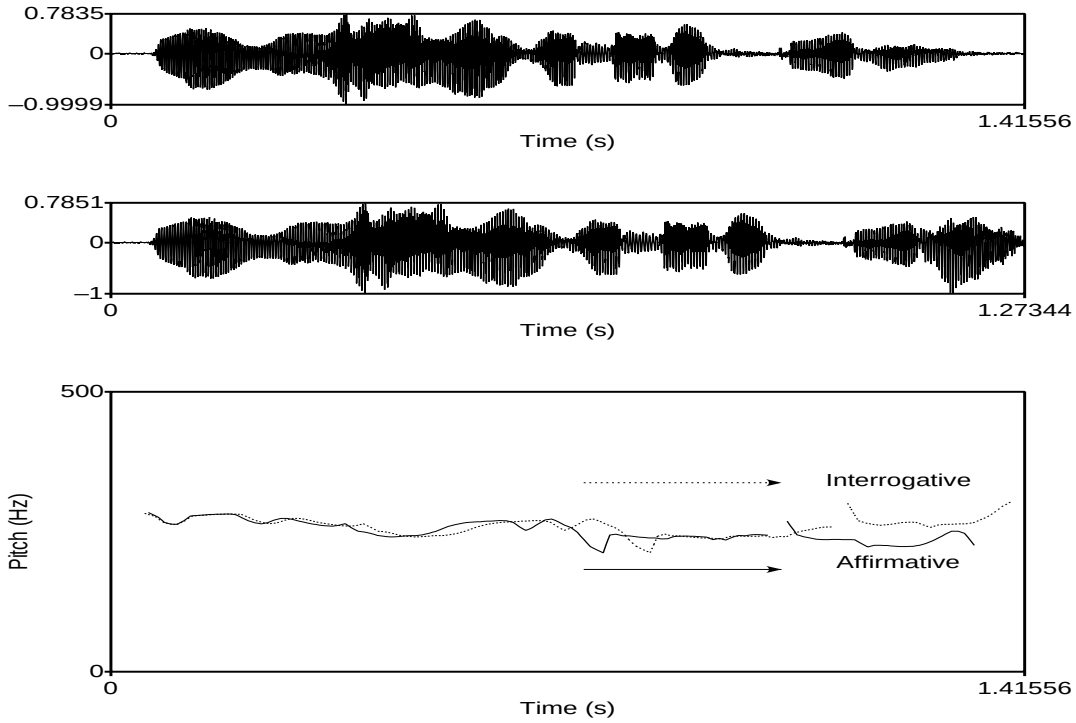

Figure 8: Modification of affirmative sentence /niivu yaavaga baruthira/ to interrogative one (female voice).

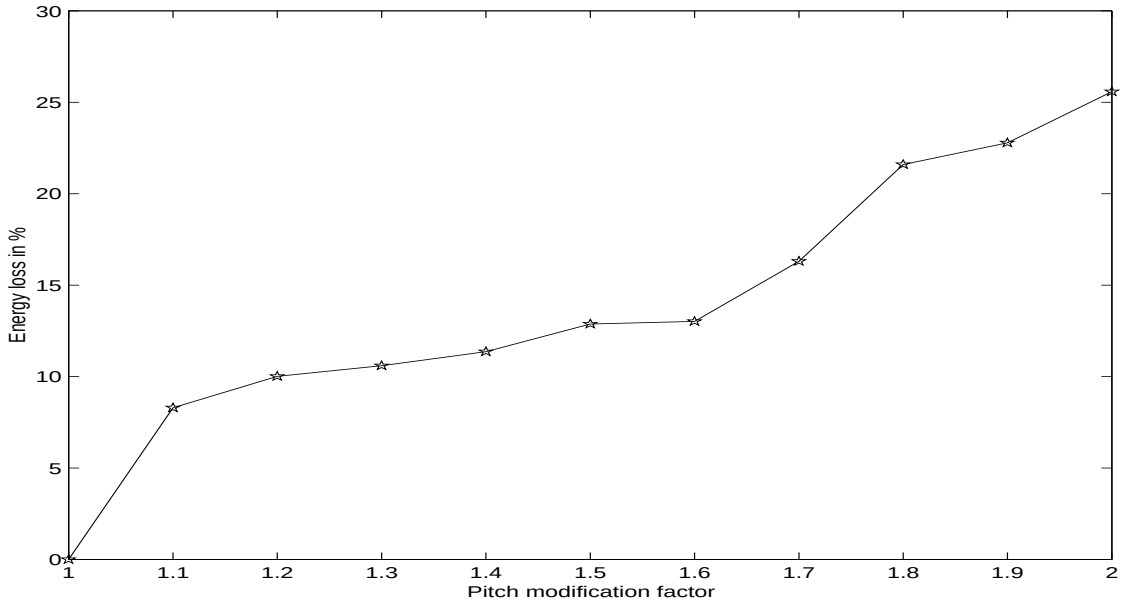

Figure 9: Energy loss w.r.t to the total residual energy against pitch modification factor