Perceptual-MVDR based Analysis-Synthesis of Pitch Synchronous Frames for Pitch Modification

R. Muralishankar *Member, IEEE,* M. Ravi Shanker and A. G. Ramakrishnan *Senior Member, IEEE*

***Abstract*— In our earlier work [1], [2], we employed minimum variance distortionless response (MVDR) and MVDR Bauer respectively, as spectral estimation techniques in place of modified-linear prediction in Discrete cosine transform (DCT) based pitch modification [3]. We introduce psychoacoustic characteristics to [1], [2] resulting in Perceptual-MVDR (PMVDR) and PMVDR-Bauer algorithms utilized here for spectral estimation. Further, we employ Bauer method of spectral factorization in our later algorithm since it results in causal inverse filter. These are used to obtain residual signal from pitch synchronous speech frames. The residual signal is resampled using DCT/IDCT depending on the target pitch scale factor. Finally, forward filters realized from the above factorization are used to get pitch modified speech. The modified speech is evaluated subjectively by 10 listeners and mean opinion scores (MOS) are evaluated for pitch factors from 0.5 to 2. Modified bark spectral distortion (MBSD) measure is also employed to evaluate objective performance. We found that the proposed approach has been rated with higher MOS and has achieved lower MBSD than the time domain pitch synchronous overlap [4], modified-LP method [3] and MVDR based methods [1], [2]. Further, we modified the pitch contours of 20 affirmative sentences to sound like interrogative sentences, using the current as well as our earlier algorithms and compared their performance.**

I. Introduction

Pitch modification is the process of changing the pitch of a given speech signal without effecting its time scale, time-varying spectral envelope and speaker information. Many techniques exist in the literature that accomplish this in the time or frequency domain or both, of which time domain pitch synchronous overlap adding (TD-PSOLA, [4]) is the simplest. It requires the knowledge of the pitch pulses and exact pitch synchronicity between pitch marks. Frequency domain pitch synchronous overlap adding (FD-PSOLA, [5]) was the first technique proposed to achieve pitch modification. Here, each short-time analysis signal is modified by employing frequency domain resampling on the short-time Fourier transform signal. Techniques like residual PSOLA (LP-PSOLA, [4]) split the speech signal into an excitation component $E(z)$ and vocal tract component $A(z)$. Pitch modification is then carried out on the source signal also known as residual signal. The output is obtained by combining modified source, $\hat{E}(z)$ and $A(z)$ using linear prediction (LP) [6]. In [7], the pitch is modified by interpolating the residual signal, realized through either upsampling or downsampling to obtain new residual length corresponding to the target pitch modification factor. The spectral envelope responsible for the formant structure is superimposed by LP forward filtering of the modified residual.

In [3], LP and modified-LP spectral estimation were employed. The required pitch scaling was achieved by a transform domain resampling of the residual using DCT/IDCT. Recently, minimum variance distortionless response (MVDR, [8]) model has been employed in pitch modification schemes in [1] and [2]. In [2], we used Bauer method of MVDR spectral factorization to extract inverse filter [9] and showed its improved performance over [1] and [3]. In this article, we introduce psychoacoustic bark scale to the earlier schemes [1] and [2] for spectral estimation. This results in two pitch modification algorithms, hereby referred to as PMVDR and PMVDR-Bauer techniques.

In our approach, pitch synchronous speech frames are inverse filtered to obtain residual signals and we follow the procedure employed in [3] for residual resampling to achieve the targeted pitch scaling. In section II, we introduce MVDR spectral modeling and its computation using LP coefficients. Later, we generate PMVDR coefficients by replacing LP with warped-LP. Section III presents pitch modification using PMVDR-Bauer. Pitch contours of modified sentences are presented in Sec. IV for two different factors. Finally, we combine the subjective and objective performances of our technique with those of earlier pitch modification schemes.

II. Spectral Modeling using MVDR and PMVDR

Murthi et al [8] introduce MVDR based spectral model as an alternative to LP. They report that the MVDR follows input speech spectral envelope with a minimal distortion. It models unvoiced speech, and mixed speech spectra better than LP [8]. Further, it was noted in [9] that MVDR analysis would lead to better discrimination of vocal tract transfer function and excitation source. We utilize this property and devise a pitch modification scheme based on MVDR.

As in LP modeling of speech, MVDR spectrum for all frequencies can be conveniently represented in a parametric form. The MVDR spectrum can be simply computed as

$$P_{MV}(\omega) = \frac{1}{\mathbf{v^H}(\omega)\mathbf{R_{M+1}^{-1}}\mathbf{v}(\omega)}, \tag{1}$$

where $\mathbf{R_{M+1}}$ is the $(M+1)\times(M+1)$ Toeplitz autocorrelation matrix of the data and $\mathbf{v}(\omega) = [1, e^{j\omega}, e^{j2\omega}, ..., e^{jM\omega}]^T$. The above equation represents the power obtained by averaging several samples at the output of the optimum constrained filter. This averaging results in reduced variance [8]. The M^{th} order MVDR spectrum can be computed by the following fast algorithm [8].

$$P_{MVDR}(\omega) = \frac{1}{\sum_{k=-M}^{M}\mu(k)e^{-j\omega k}} = \frac{1}{|B(e^{j\omega})|^2} \tag{2}$$

where the MVDR coefficients, $\mu(k)$, are obtained as,

$$\mu(k) = \begin{cases} \frac{1}{P_e}\sum_{i=0}^{M-k} L a_i a_{i+k}^*, & k = 0, ..., M \\ \\ \mu^*(-k), & k = -M, ..., -1 \end{cases} \tag{3}$$

a_k are the LP coefficients, P_e is the prediction error and $L = (M+1-k-2i)$. For real input signal $\{\mu(k)\}$ is real and even (and so is $\frac{1}{|B(e^{j\omega})|^2}$). From (2), one can view MVDR power spectrum as an all-pole power spectrum. We use spectral factorization [10] to obtain a minimum phase filter, $\frac{1}{B(z)}$, whose power spectrum equals the one computed in (2). This can be written as

$$C(z) = \sum_{k=-M}^{M}\mu(k)z^{-k}. \tag{4}$$

A unique canonical factorization [10] of the form

$$C(z) = D(z)rD^*(1/z^*) \tag{5}$$

is possible with $D(z)$ being a minimum-phase M^{th}-order polynomial. The inverse filter is then

$$B(z) = \sqrt{r}D(z) \tag{6}$$

whose coefficients $b(n)$ are guaranteed to be real because $\mu(k)$ are real. We can factorize $C(z)$ directly for small model orders [9]

by extracting the polynomial roots that lie inside the unit circle. For higher orders, it is suggested in [9] to use iterative method to approximate exact coefficients $\mu(k)$'s. One can see that the former approach has been considered in [1], and later in [2].

A natural extension to the MVDR scheme is the incorporation of perceptually motivated mel-frequency into the otherwise linear frequency scale. Here, perceptual information can be incorporated directly into spectral estimation by using mel-scale filter banks. However, it can easily be seen that the filter bank structure is only a rough approximation to the perceptual scale, since it samples the perceptual spectrum at the center frequencies of the filter bank. Furthermore, the filter bank is less effective in completely removing the harmonic excitation information from the spectrum. Alternatively, warping can be incorporated directly into the discrete Fourier transform (DFT) power spectrum [11], or by using warped-LP coefficients in generating warped-MVDR spectrum [12]. In this paper, we generate Perceptual-MVDR (PMVDR) coefficients using warped-LP.

A. Extraction of Inverse Filter $B(z)$ using Bauer Method

This technique [10] is based on the Cholesky decomposition of Toeplitz matrices, whose first column consists of the PMVDR coefficients ($\mu(k)$'s, k positive). Let P_N be the $(N+1)\times(N+1)$ Toeplitz matrix; the sequence starts with

$$P_0 = \left(\mu(0)\right), P_1 = \begin{pmatrix}\mu(0) & \mu(1)\\ \mu(1) & \mu(0)\end{pmatrix} ... \tag{7}$$

Given a P_N matrix, we use Cholesky decomposition to get a $(N+1)\times(N+1)$ lower triangular matrix D_N with a unit diagonal and a $(N+1)\times(N+1)$diagonal matrix r_N, that satisfy the equation

$$P_N = D_N r_N D_N^T \tag{8}$$

It has been shown by Bauer that, as $k\to\infty$, the D_{N_j} elements on the last line of $\mathbf{D}_N$ in reversed order tend to the coefficients of the $D(z)$ polynomial in (5) and r_N, the $(N+1)^{th}$ element of r_N tends to r. Further, it can be written as

$$B(z) \simeq \sqrt{r_N}\sum_{k=0}^{M} D_{N(N-k)} z^{-k} \tag{9}$$

III. Pitch Modification Using PMVDR-Bauer

Our pitch modification algorithm uses DCT/IDCT based residual resampling procedure [3]. Further, we use here the Bauer spectral factorization of PMVDR. We note that the choice of MVDR model [8] in [1] has been driven by its interesting spectral estimation properties, namely minimum variance, low distortion and a better spectral match across wide range of pitch values. In our algorithm, shown in Fig. 1, we utilize these properties to capture vocal tract responses using Bauer method.

The residual resampling employed in [3] is repeated here for the clarity of presentation. Input speech is pitch-marked in voiced regions according to their pitch values and in unvoiced regions, pitch-marks are uniformly placed. LP coefficients are extracted from each pitch synchronous (PS) speech frame. PMVDR coefficients are then computed from the warped-LP coefficients using (3). Subsequently, we use (9) to get $B(z)$ from PMVDR coefficients [9]. Then, the residual signal is extracted by passing PS speech frames through the filter $B(z)$. The pitch is modified in the residual domain using DCT. 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 divided by the pitch modification factor. For pitch increase, $N_1 - N_2$ trailing DCT coefficients are removed; whereas,

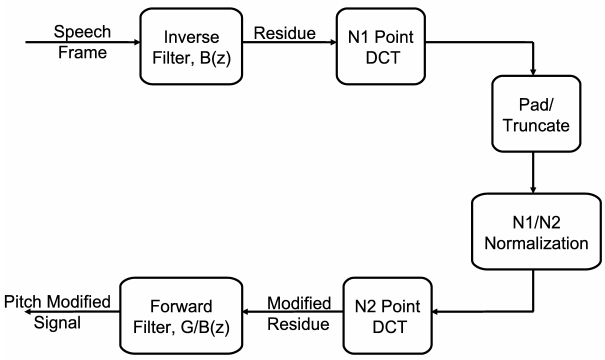

Fig. 1. Block diagram of pitch modification using DCT/IDCT via PMVDR spectral modelling.

for decreasing the pitch, $N_2 - N_1$ zeros are added to the DCT coefficients. Before taking IDCT, amplitude normalization must be carried out to compensate for the effect of change in length of the residual signal. The modified residue is used to re-synthesize the pitch modified speech using the forward filter, $\frac{1}{B(z)}$. The durational effects on the speech due to our pitch modification step are compensated by an appropriate time-scaling factor using known algorithms like TD-PSOLA [4] and WSOLA [13].

IV. Results and Discussion

All results of all our experiments are available at $http$: $//ragashri.ee.iisc.ernet.in/MILE/index\_files/content\_ra11.html$. To demonstrate the effectiveness of this technique, individual phonemes, words and sentences spoken by both male and female volunteers were extracted from the MILE Tamil TTS speech database, whose average SNR is about 40 dB and sampling frequency is 16 kHz. These utterances were analyzed and re-synthesized for different pitch factors. Figure 2(a) shows a speech segment /A/. Fig. 2(b) gives the corresponding residual signal extracted by inverse filtering the above signal using $B(z)$ coefficients (LP model order 16). Figs. 2(c) and (e) show the length-modified residual signals obtained via DCT/IDCT, the factor of increase (decrease) in pitch being 1.3 (0.7). Figs. 2(d) and (f) show the corresponding synthesized speech signals after forward filtering by $1/B(z)$ coefficients.

The PMVDR Bauer spectra of phoneme /A/ and pitch modified signals are shown in Fig. 3 for pitch modification factors of 0.6, 0.8, 1.2 and 1.4, respectively. Phoneme /A/ is extracted from both the original and pitch modified sentences (/kAndaL poduwAka iLanyjiwaplu niRatlil amaendiruklum/). The figures illustrate the fact that noticeable deviations in the formant positions can be observed for the factors outside $0.7-1.3$. It is given in [14] that the speaker identity is not altered if the variation in the formant values is within $\pm 15\%$. To verify this, we evaluated the modified speech for speaker identity as reflected by the mean opinion score (MOS), in addition to other attributes. The MOS of the modified speech is found to be better than those of TD-PSOLA [15], modified-LP method [3], MVDR [1], MVDR Bauer [2] and PMVDR. Figure.4 shows the speech signal for a whole sentence /kAndaL poduwAka iLanyjiwaplu niRatlil amaendiruklum/, its original pitch contour and the contours after pitch change using the technique involving PMVDR Bauer coefficients for the factors 1.3 and 0.7.

We conducted subjective and objective tests to evaluate the performance of the proposed technique. Here, the modified bark spectral distortion (MBSD, [16]) is employed as an objective measure that is closely related to subjective evaluation. This estimates speech distortion in the loudness domain, taking into account the noise masking threshold in order to include only audible distortions in

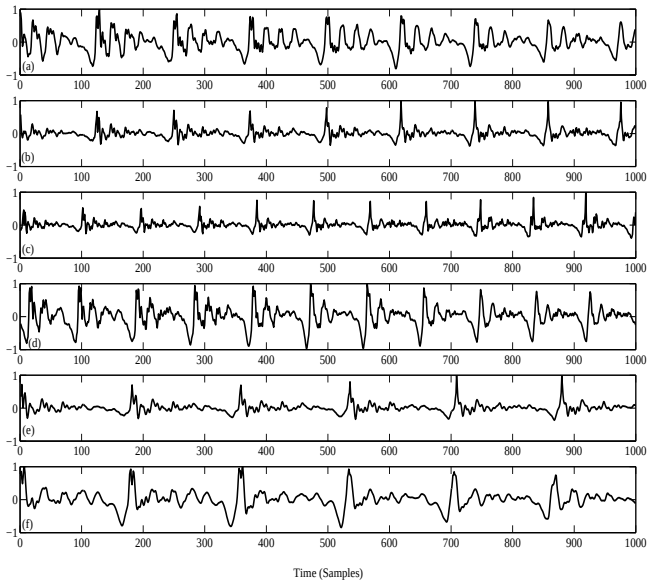

Fig. 2. (a) Few frames of an utterance /A/ and (b) its excitation.(c) Excitation in (b) modified for a pitch increase factor of 1.3. (d) Signal resynthesized by forward filtering the signal in (c) using $1/B(z)$ coefficients. (e) Excitation in (b) modified for a pitch decrease factor of 0.7. (f) Signal resynthesized by forward filtering the signal in (e).

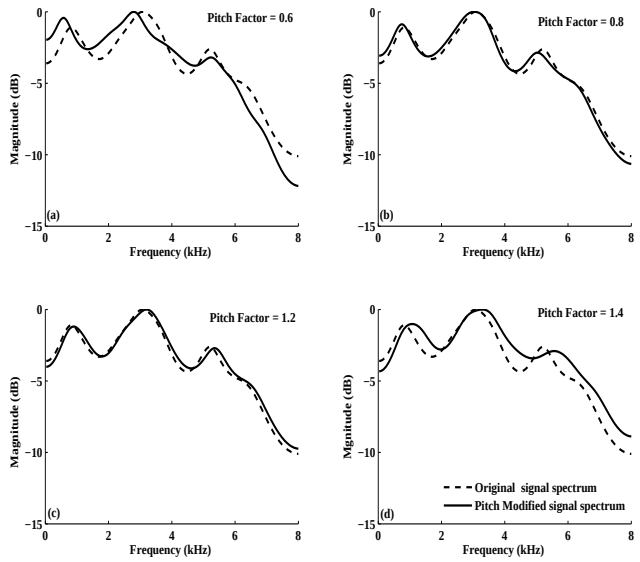

Fig. 3. PMVDR Bauer spectra of the original signal overlapped with those of the modified signals for pitch modification factors a) 0.6 (b) 0.8 (c) 1.2 (d) 1.4.

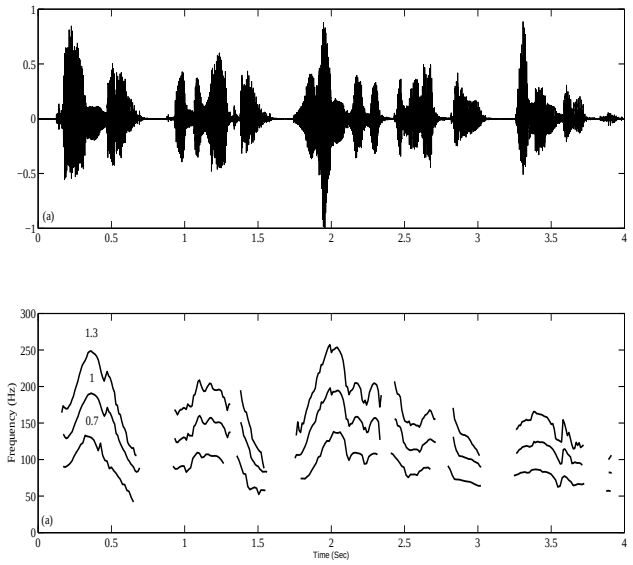

Fig. 4. Pitch contours of an utterance before and after pitch modification. (a) Waveform of the original utterance /kAndaL poduwAka iLanyjiwaplu niRatlil amaendiruklum/. (b) Comparison of pitch contours for factors 0.7 and 1.3.

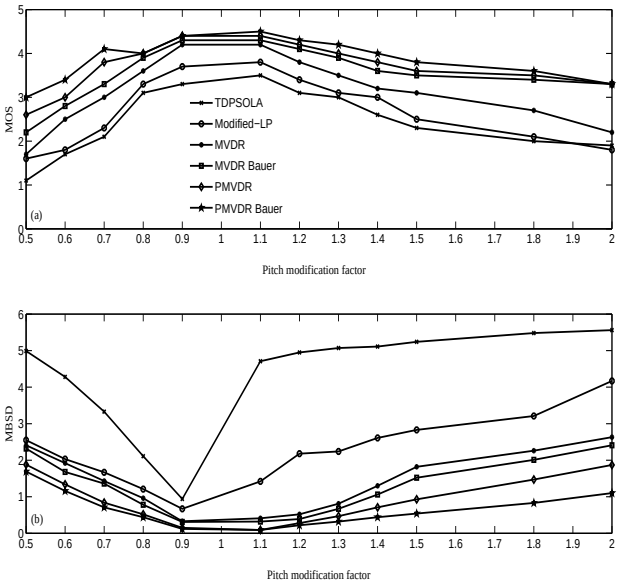

Fig. 5. Comparison of a) Subjective and b) Objective measures for different pitch modification schemes.

the calculation of the distortion measure. Since MBSD compares the distorted speech to the original, its performance would be sensitive to the temporal misalignment [16]. So a synchronization algorithm based on loudness domain is applied prior to performing the MBSD. Higher distortion in modified speech results in MBSD score away from 0 and for lower, it is close to 0.

Subjective and objective tests are conducted on 20 sentences spoken by both male and female volunteers, each of which is of duration about 1 min. We pitch modify these sentences using the proposed algorithm and compare with TD-PSOLA [15], modified-LP [3], MVDR [1], MVDR-Bauer [2] and PMVDR methods, for a range of factors from 0.5 to 1.5 with a step of 0.1, along with factors 1.8 and 2.0. Ten people rated the quality of the pitch modified sentences in terms of MOS. A MOS of 5 indicates 'excellent' and 1 indicates 'bad' with respect to naturalness, intelligibility and speaker identity. The performance comparison between our algorithm and the other methods is presented in Fig. 5. The figure displays significant improvements in subjective and objective performances for our algorithm over all the other methods for pitch factors between 0.7 and 1.3. Here, we know that the factors between 0.7 to 1.3 are useful in concatenative speech synthesis [3]. Better performance of our algorithm can also be observed for factors outside 0.7 and 1.3. One can also see a meagre improvement in objective performances and a good MOS score over other approaches. It was noted in [9] that MVDR analysis could better discriminate vocal tract transfer function and excitation source. Correspondingly, MVDR-Bauer obtained through spectral factorization of MVDR, has most of the spectral estimation properties of [9]. Further, its causal structure minimizes the

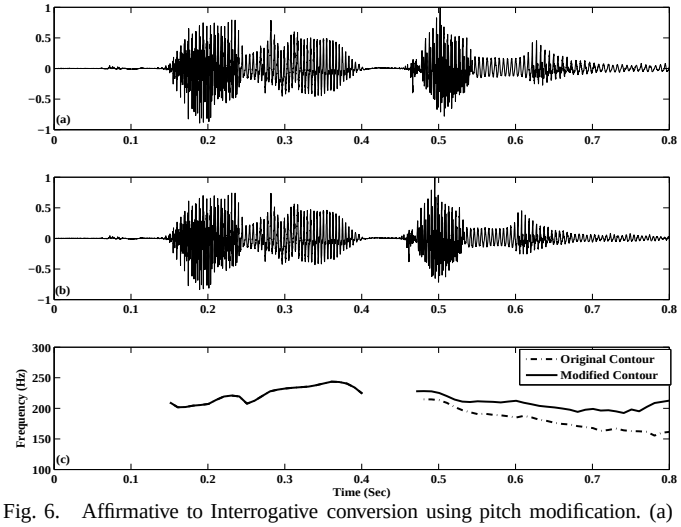

Fig. 6. Affirmative to Interrogative conversion using pitch modification. (a) Waveform of an affirmative sentence. (b) Synthesized Interrogative sentence (c) Pitch contours of (a) and (b).

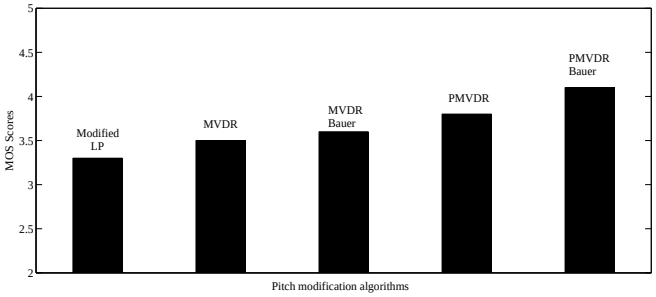

Fig. 7. MOS performance of various pitch modification algorithms in conversion of affirmative to interrogative sentences.

number of filters required to achieve pitch modification. In summary, one can observe similar properties by PMVDR-Bauer employed in our algorithm. These inputs suffice us to believe that the improved performance of our algorithm is attributed to good envelope match with low variance and minimal distortion of PMVDR-Bauer spectral factorization. Here, we use the Cholesky decomposition of PMVDR coefficients to obtain $B(z)$, a representation of PMVDR Bauer where PMVDR coefficients are obtained using (3) with the Warped-LP model of order 16. Finally, problems regarding bandwidth loss due to pitch lowering using residual resampling can be compensated by having a high bandwidth original speech [3].

To test the usefulness of our technique, we modified the pitch contours of 20 affirmative sentences from the MILE Tamil TTS database to make them sound like interrogative ones. For example, a characteristic of any interrogative sentence with an "yes or no" answer is that both the contour and the amplitude rise sharply for the last syllable [14]. We modified the pitch contours of different sentences appropriately to realize the objective. The result of pitch modification by a time varying factor using our algorithm is shown in Fig. 6. Figures 6(a) and (b) show the waveforms of an affirmative sentence and that of the interrogative sentence derived from that by pitch contour modification. Figure 6 (c) shows the pitch contours of (a) and (b). Finally, ten subjects were asked to rate the quality of the interrogative sentences synthesized using the different algorithms. It can be seen from Fig. 7 that PMVDR Bauer has higher MOS than Modified-LP, MVDR, MVDR-Bauer and PMVDR.

V. Conclusion

PMVDR Bauer based spectral estimation is employed in our pitch modification algorithm. Residual signal is obtained by inverse filtering the pitch synchronous speech frames with PMVDR Bauer coefficients. Pitch modification is achieved in the source domain using DCT/IDCT based resampling [3]. Forward filtering is carried out to obtain pitch modified speech. We have shown that the resulting speech has minimal deviations in formant positions for factors from 0.7 to 1.3. We observe that the present algorithm outperforms TD-PSOLA, modified-LP, MVDR, and MVDR-Bauer methods in both objective and subjective analyses. Significant differences in performance can be seen for factors between 0.7 and 1.3. Moreover, we can see a minor improvement in objective performance over PMVDR approach. Considerable improvement can be observed in subjective scores over other algorithms for most of the factors. We have also shown its usefulness in transforming affirmative sentences to sound like interrogative sentences. The next logical step is to explore the utility of our approach for prosody modification in our Tamil text-to-speech synthesis system. We are currently working in that direction.

References

[1] R. Muralishankar, M. Ravi Shanker, and A. G. Ramakrishnan, "MVDR spectral estimation for DCT based pitch modification," *accepted, 3rd Language & Technology Conference*, Oct. 5-7, Poznan, Poland 2007.

[2] M. Ravi Shanker, R. Muralishankar, and A. G. Ramakrishnan, "Bauer method of MVDR spectral factorization for pitch modification in the source domain," *accepted, IEEE Workshop on ASPAA'07*, Oct. 21-24, Mohonk Mountain House New Paltz, New York, 2007.

[3] R. Muralishankar, A. G. Ramakrishnan, and P. Prathibha, "Modification of pitch using DCT in the source domain," *Speech Communication*, vol. 42, pp. 143–154, 2004.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467, 1990.

[5] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," *Proc ICASSP*, pp. 2015–2018, 1986.

[6] W. B. Kleijn and K. L. Paliwal, *Speech Coding and Synthesis*. Elsevier B.V, New York, 1995.

[7] F. M. Gimenez de los Galanes, M. Savoji, and J. M. Pardo, "Speech synthesis system based on a variable decimation interpolation factor," *Proc. ICASSP*, pp. 636–639, 1995.

[8] M. N. Murthi and B. D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech and Audio Pro.*, vol. 8, no. 3, pp. 221–239, 2000.

[9] A. Santarelli, M. Omologo, and L. Armani, "Separation of excitation source and vocal tract transfer function via an MVDR analysis of speech," *Proc. IEEE workshop on ASPAA*, pp. 115–118, Oct. 2003.

[10] A. H. Sayed and T. Kailath, "A survey of spectral factorization methods," *Numerical linear algebra with applications*, vol. 08, pp. 467–496, 2001.

[11] U. H. Yapanel and J. H. L. Hansen, "A new perspective on feature extraction for robust in-vehicle speech recognition," in *EUROSPEECH*, 2003, pp. 1281–1284.

[12] M. Wolfel, J. McDonough, and A. Waibel, "Warping and scaling of the minimum variance distortionless response," in *ASRU*, 2003, pp. 387–392.

[13] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high-quality time-scale modification of speech," *Proc.ICASSP*, pp. 554–557, 1993.

[14] M. Abe, "Speaking styles: Statistical analysis and synthesis by a text-to-speech system," *Progress in Speech Synthesis,Springer, New York*, 1996.

[15] S. Roucos and A. Wilgus, "High quality time-scale modification of speech," *Proc. ICASSP*, pp. 493–496, 1985.

[16] W. Yang, M. Benbouchta, and R. Yantorno, "Performance of a modified bark spectral distortion measure as an objective speech quality measure," *Proc. ICASSP*, pp. 541–544, 1998.