MVDR Spectral Estimation for DCT based Pitch Modification

R. Muralishankar* M. Ravi Shanker†, A. G. Ramakrishnan†

*Department of Telecommunication Engineering, PES Institute of Technology
100 Feet Ring Road, Banashankari 3rd Stage
Bangalore-560085, India.
muralishankar@pes.edu
†Department of Electrical Engineering, Indian Institute of Science
Bangalore-560012, India.
{shanker, ramkiag}@ee.iisc.ernet.in

Abstract

This paper presents an improvisation to the work on Discrete cosine transform (DCT) based pitch modification in the residual domain (Muralishankar et al., 2004). The residual signal is obtained from pitch synchronous frames by inverse filtering the speech signal. The inverse filters are constructed using the spectral factorization of minimum variance distortionless response polynomial into a causal and a noncausal part. DCT/IDCT is used to resample the residual depending on the target pitch scale factor. 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 are tabulated. Further, modified bark spectral distortion measure is also computed for objective evaluation of the performance. We find that the proposed algorithm performs better compared to Time domain pitch synchronous overlap-add (Roucos and Wilgus, 1985) and modified linear prediction method (Muralishankar et al., 2004).

1. Introduction

Machine synthesis has been achieved by various procedures, of which concatenative synthesis is generally employed. It is known that concatenative synthesis gives the most natural sounding synthesized speech. This is due to concatenation of recorded utterances (in the form of basic units) directly. It requires a large database of basic units. However, the nature of the automated techniques for segmenting the waveforms sometimes results in audible glitches in the output, detracting from the naturalness of the synthesized speech. And even after unit selection, their concatenation may result in sudden pitch change from one unit to the other. Hence, we need a pitch modification scheme to employ on the units selected, to smoothen out the pitch contour at the place of concatenation. Pitch modification has been widely used in the applications that include voice delivery of text messages and email, voice response to database inquiries, mobile-environment communications that leave the hands and eyes free, adjusting the pitch in a singers voice to get the desired effect etc.

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 literature that accomplish this in the time or frequency domain or both, of which Time domain pitch synchronous overlap adding (TD-PSOLA, (Roucos and Wilgus, 1985; Moulines and Charpentier, 1990)) is the simplest method for pitch modification of speech signals. It requires a knowledge about the pitch pulses and an exact pitch synchronicity between pitch marks. Frequency domain overlap adding (FD-PSOLA, (Charpentier and Stella, 1986)) was the first technique proposed to achieve time and pitch scale modification. Here, each short-time analysis signal is modified by employing frequency domain resampling on the short-time Fourier transform signal. Further, techniques like residual PSOLA (LP-PSOLA, (Moulines and Charpentier, 1990)) split speech signal into an excitation (source) 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) (Kleijn and Paliwal, 1995). In (Gimenez de los Galanes et al., 1995), the pitch is modified by interpolating the residual signal, realized through either upsampling or downsampling to obtain new residual length corresponding to the given pitch modification factor. The spectral envelope responsible for the formant structure will be superimposed by LP forward filtering of the modified residual.

In this paper, we use Minimum variance distortionless response (MVDR) based spectral estimation technique (Murthi and Rao, 2000) to extract spectral envelope. However, LP and modified-LP were employed in the pitch modification scheme proposed in (Muralishankar et al., 2004). Their argument is based on the observation of LP spectral gain variations with the harmonic positions for medium and high pitched signals. Moreover, to compensate for this gain variations, modified-LP method was employed to obtain a smoothed spectral envelope (Muralishankar et al., 2004; Ansari, 1997). Depending on the pitch modification factor, the radius of the LP-pole is decreased (bandwidth increases) to accommodate the new harmonic positions. In contrast, we replace modified-LP with MVDR based spectral modeling and follow similar procedure in the case of residual resampling as in (Muralishankar et al., 2004). We employed causal and noncausal inverse and forward filters to extract residuals and to add spectral envelope to the modified residuals.

Section 2., introduces MVDR spectral modeling and its computation using LP coefficients. Section 3., presents pitch modification using MVDR based spectral modeling of the pitch synchronous speech frames. Pitch contours of

the pitch modified sentence are presented in sec. 4. for two different factors. Finally, we present subjective and objective performances of our technique and compare with an earlier pitch modification scheme (Muralishankar et al., 2004).

2. Spectral Modeling using MVDR

Despite the popularity of LP as a method of spectral modeling, it has its own drawbacks. LP model is more suited for low pitch speech and its performance increases with the decrease in pitch frequency. It does not model well the spectral envelope for medium and high pitch voiced speech (Murthi and Rao, 2000). Further, if the model order of the LP filter is increased, then the corresponding envelope overestimates the original voiced speech power spectrum, resolving the harmonics and not the spectral envelope. However, the MVDR provides a smooth spectral envelope even when the model order is increased. Furthermore, the MVDR spectrum is capable of modeling unvoiced speech, and mixed speech spectra (Murthi and Rao, 2000).

As in LP representation of speech modeling, MVDR spectrum for all frequencies can be conveniently represented in a parametric form. The MVDR spectrum can be simply computed as (Haykin, 1991)

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

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 (Stoica and Moses, 1997). The M^{th} order MVDR spectrum can be computed by the following fast algorithm proposed by Musicus (Musicus, 1985).

$$P_{MV}(\omega) = \frac{1}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}} \quad (2)$$

where the MVDR coefficients, $\mu(k)$, are given by the following non-iterative computation, using the LP coefficients a_k and the prediction error P_e.

1. Compute the LP coefficients (LPCs), $a^M_{0..M}$ of the order M and the prediction error power, ϵ_M.

2. Correlate the LPCs, as

$$\mu(k) = \begin{cases} \frac{1}{P_e} \sum_{i=0}^{M-k} La_i a^*_{i+k}, & k = 0, ..., M \\ \mu^*(-k), & k = -M, ..., -1 \end{cases} \quad (3)$$

where $L = (M + 1 - k - 2i)$.

3. Compute the MVDR envelope

$$S_{MVDR}(e^{j\omega}) = \frac{\epsilon_M}{\sum_{k=-M}^{M} \mu(k) e^{-j\omega k}} \quad (4)$$

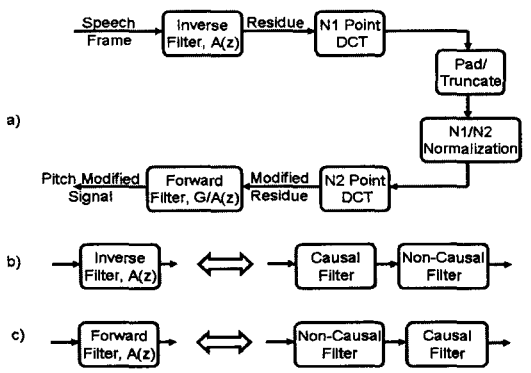

Figure 1: Block diagram of pitch modification using DCT/IDCT via MVDR spectral modelling.

3. Pitch Modification Method

We present our pitch modification algorithm that is similar to the one proposed in (Muralishankar et al., 2004), based on residual resampling using DCT/IDCT. Further, we select MVDR model in the place of LP and modified-LP used in (Muralishankar et al., 2004). In Murthi (Murthi and Rao, 2000), we note the interesting properties of MVDR based spectral estimation, namely minimum variance, low distortion and a better spectral match across a wide range of pitch values. We utilize these properties to capture the vocal tract responses in our algorithm and represent through a block diagram as shown in Fig. 1.

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 input pitch synchronous (PS) speech frame. MVDR coefficients are then computed from the LP coefficients using eq.(3). In order to use $A(z)$ in inverse filtering the PS speech frames, we decompose $A(z)$ as a cascade connection of a causal and a noncausal filters (Santarelli et al., 2003). This is achieved by finding the roots of $A(z)$ and segregating the roots depending on their radius $\gtrless 1$. We get a causal transfer function $A_c(z)$ by considering the roots having radii < 1. $A_c(z)$ is also used to get noncausal filter responses by using input/output flipping trick.

Residual signal is extracted by passing PS speech frames through the filters (see Fig. 1). Here, pitch is modified in the residual (or source) 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. In the DCT domain, for pitch increase, $N_1 - N_2$ trailing DCT coefficients are removed; whereas, 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, $\frac{1}{A(z)}$. Here, the forward filtering is carried out by using the coefficients of $A_c(z)$ to realize $\frac{1}{A_c(z)}$. Sim-

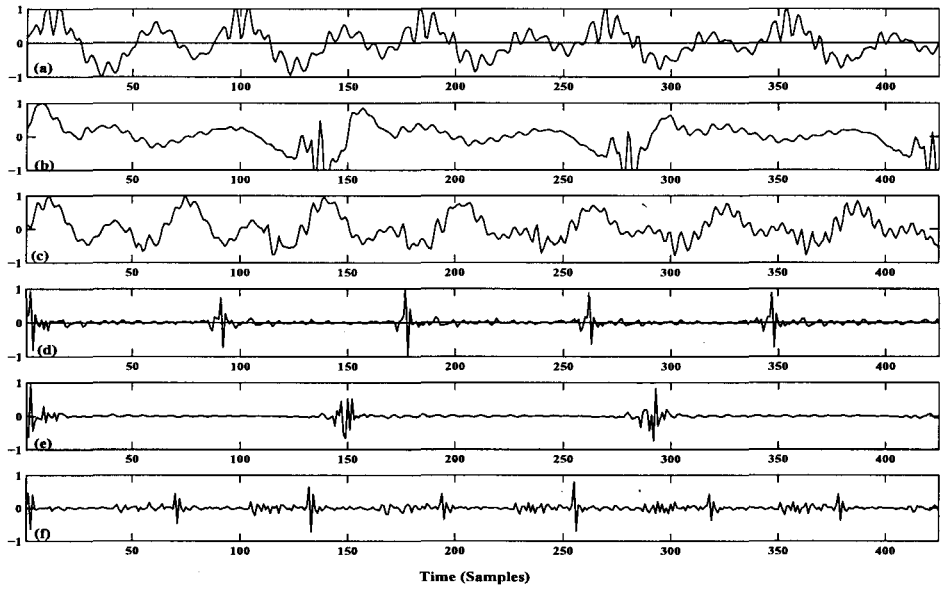

Figure 2: (a) Few frames of the original signal /pI/. (b) Few frames of the signal reconstructed by forward filtering the signal in (e) using MVDR coefficients. (c) Few frames of the signal reconstructed by forward filtering the signal in (f) using MVDR coefficients. (d) Few frames of the original excitation. (e) Few frames of the modified excitation for a pitch decrease factor of 0.7. (f) Few frames of the modified excitation for a pitch increase factor of 1.3.

ilarly, we can realize cascade connection of noncausal and causal forward filters, respectively (see Fig. 1). The durational effects due to our pitch modification setup on the modified speech are compensated by an appropriate time-scaling factor using the well known algorithms like TD-PSOLA (Roucos and Wilgus, 1985) and WSOLA (Verhelst and Roelands, 1993).

4. Results and Discussion

To demonstrate the effectiveness of this technique, individual phonemes, words and sentences spoken by both male and female volunteers were extracted from Tamil speech database with an average SNR of about 40 dB with a sampling frequency 16 kHz. These utterances were analyzed and re-synthesized for different pitch change factors. Figure.2(a) shows a speech segment /pI/. Fig.2(d) gives the corresponding residual signal extracted by inverse filtering the above signal using MVDR coefficients (LP model order 16). Fig. 2(e) shows the length-modified residual signal obtained through DCT/IDCT, the factor of decrease in pitch being 0.7. Fig. 2(b) shows the corresponding synthesized speech signal after forward filtering by the same MVDR coefficients. Fig. 2(f) shows the length-modified residual signal for a pitch modification factor of 1.3. Fig. 2(c) shows the corresponding synthesized speech signal after forward filtering.

The MVDR spectra of phoneme /A/ and pitch modified signals are shown in Fig. 3. Phoneme /A/ is extracted from the original and pitch modified sentence /nAyanaklAran mella nAyanatlae udaTlil waetlu pI pI enRu satlam pArtlAn/ with factors 0.6, 0.8, 1.2 and 1.4 respectively. The figures illustrate the fact that while there is no appreciable shift in the formants for factors 0.8, 0.6 and 1.2, there is a minimal shift in the third formant for a pitch increase factor of 1.4. It is known that the speaker identity is not disturbed if the variation in the formant values is within $\pm 15\%$ (Abe, 1996) of the original values. To verify this, we evaluated the resultant speech for speaker identity as reflected by the MOS (mean opinion score), in addition to other attributes. The MOS of the modified signals is found to be better than the modified-LP method (Muralishankar et al., 2004). Figure.4 shows the pitch contours for the segment shown in Fig. 2, and its pitch modified versions for factors 0.7 and 1.3. It can be seen that the shape of the pitch contour is maintained in the modified signals. Figure.5 shows the speech signal for a whole sentence */nAyanaklAran mella nAyanatlae udaTlil waetlu pI pI enRu satlam pArtlAn/*, its original pitch contour and the contours after pitch change using the technique involving MVDR coefficients for two factors 1.3 and 0.7.

To evaluate the performance of the proposed technique, we conducted subjective and objective tests. We employed an objective measure, Modified bark spectral distortion (MBSD, (Yang et al., 1998)) that is closely related to subjective performance. This estimates speech distortion in the loudness domain, taking into the account the noise masking threshold in order to include only audible distortions in the calculation of the distortion measure. This new addition of the noise masking threshold replaces the empirically derived distortion threshold-value used in the conventional bark spectral distortion (Yang et al., 1998). Since MBSD compares the distorted speech to the original speech, its performance would be sensitive to the temporal misalignment (Benbouchta, 1998). 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 distortions, 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 having a duration of about 1 min. We pitch modify these sentences using the proposed algorithm and compare with the TD-PSOLA (Roucos and Wilgus, 1985), modified-LP method (Muralishankar et al., 2004), for a range of factors from 0.5 to 1.5 with a step of 0.1, along with factors

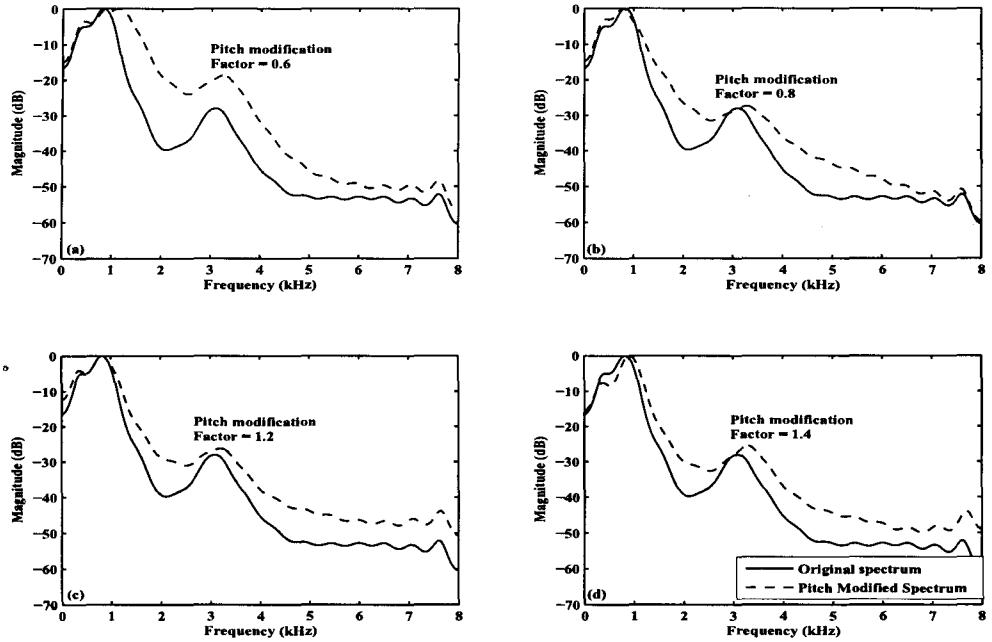

Figure 3: MVDR spectra of the original signal overlapped with the MVDR spectra of the modified signals. (a) Pitch modification factor = 0.6. (b) Pitch modification factor = 0.8. (c) Pitch modification factor = 1.2. (d) Pitch modification factor = 1.4.

1.8 and 2.0. Ten people were asked to rate the pitch modified sentences in terms of MOS by taking into account naturality, intelligibility and speaker identity. A MOS of 5 indicates 'excellent' and 1 indicates 'bad' with respect to naturality, intelligibility and speaker identity. The performance comparison between our algorithm and both TD-PSOLA and the modified-LP method are presented in Table 1. From the table, we can see significant improvements in subjective and objective performances for our algorithm over TD-PSOLA and modified-LP method for pitch factors between 0.8 to 1.3. Better performances of our algorithm can also be noted for factors outside 0.8 and 1.3. We believe that this improvement in the performance is due to useful spectral estimation properties of MVDR. Further, (Muralishankar et al., 2004) reports that the factors between 0.8 to 1.3 are useful in concatenative speech synthesis.

It was noted in (Santarelli et al., 2003) that MVDR analysis could lead to better results in fine discrimination of vocal tract transfer function and excitation source. Hence, we believe that the improved performance of our algorithm is attributed to good envelope match with low variance and minimal distortion of MVDR. Further, we factorize a non-causal filter $A(z)$ in our algorithm by extracting their polynomial roots that lie inside the unit circle. However, it was stated in (Santarelli et al., 2003) that the factorization procedure can be used for small model orders and for higher orders, one must go for iterative methods. Here, we choose LP order equal to 16 to compute MVDR coefficients using eq.(3).

Presently, we are exploring the benefits of iterative method in the proposed pitch modification scheme. Problems regarding bandwidth loss due to pitch lowering using residual resampling can be compensated by having a high bandwidth original speech (Muralishankar et al., 2004).

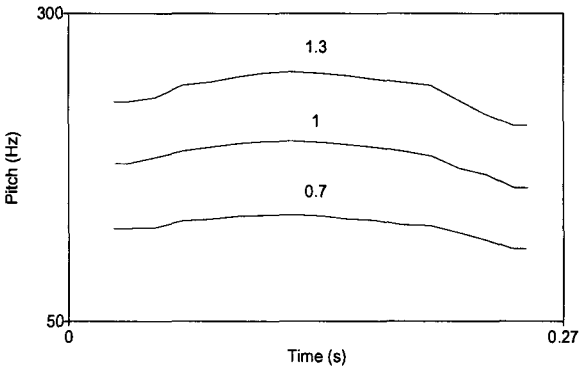

Figure 4: Pitch contours of the original and modified speech segment /pI/. The numbers shown are the pitch modification factors.

MVDR spectral envelope match although smooth with low variance and minimal distortion in its representation of true spectrum does not have enough resolution to capture all formants. In particular, higher formants are sometimes missed in MVDR spectral representation for small orders. This can be compensated by choosing sufficiently higher LP order to start with. Further, the loss of higher formants representation in MVDR affect mostly speaker information than speech information and is observable as a drop in MOS scores for pitch change factors away from 1 (see Table 1).

5. Conclusion

MVDR based spectral estimation is employed in our pitch modification algorithm. Residual signal is obtained by inverse filtering the pitch synchronous speech frames with MVDR coefficients. We extracted causal and non-

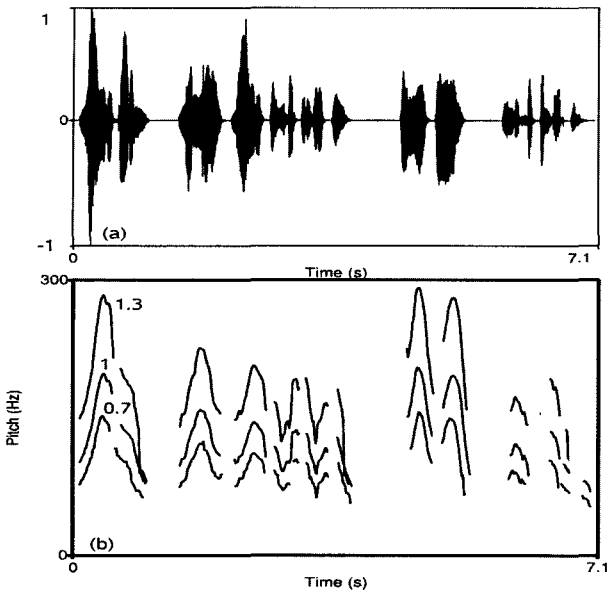

Figure 5: Pitch contours of original utterance and after pitch modification. (a) Waveform of the original utterance */nAyanaklAran mella nAyanatlae udaTlil waetlu pI pI enRu satlam pArtlAn/.* (b) Comparison of pitch contours (factors 0.7 and 1.3).

Pitch Scale Factor	TD-PSOLA		Modified-LP		MVDR	
	MOS Score	MBSD Score	MOS Score	MBSD Score	MOS Score	MBSD Score
0.5	1.1	4.99	1.6	2.55	1.7	2.41
0.6	1.7	4.28	1.8	2.03	2.5	1.92
0.7	2.1	3.33	2.3	1.67	3.0	1.43
0.8	3.1	2.11	3.3	1.21	3.6	0.96
0.9	3.3	0.94	3.7	0.67	4.2	0.33
1.1	3.5	4.71	3.8	1.42	4.2	0.41
1.2	3.1	4.95	3.4	2.18	3.8	0.52
1.3	3.0	5.07	3.1	2.24	3.5	0.81
1.4	2.6	5.11	3.0	2.61	3.2	1.3
1.5	2.3	5.24	2.5	2.83	3.1	1.82
1.8	2.0	5.48	2.1	3.21	2.7	2.26
2	1.9	5.56	1.8	4.17	2.2	2.63

Table 1: Comparison of subjective and objective measures for different pitch modification schemes.

causal filter coefficients from MVDR and employed in inverse and forward filtering. Pitch modification is achieved in the source domain using DCT/IDCT based resampling (Muralishankar et al., 2004). Forward filtering is carried out to obtain pitch modified speech. Its MVDR envelope is shown to have minimal deviation in formant positions compared with the original envelope. We observe that the present algorithm outperforms TD-PSOLA and the modified-LP method in both objective and subjective analysis and significant differences in performance can be seen for the factors between 0.8 and 1.3. It would be more worthwhile to see the impact of our algorithm on the synthesized speech. Preliminary results of our algorithm used in the Tamil speech synthesizer (Jayavardhana Rama et al., 2002) indicates an improved quality of synthesized speech. Currently, we are conducting further tests to ascertain usefulness of our approach in text-to-Speech systems. Furthermore, we believe that by introducing psychoacoustic scale in MVDR based envelope extraction would enhance the overall pitch modification performance.

6. References

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

Ansari, R., 1997. Inverse filter approach to pitch modification:application to concatenative synthesis of female speech. *Proc.ICASSP*:1623–1626.

Benbouchta, M., 1998. A waveform synchronization algorithm in the context of objective measure of speech quality. *M.S.Thesis, Temple University, Philadelphia, PA.*

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

Haykin, S., 1991. *Adaptive Filter Theory.* Englewood Cliffs, NJ:Prentice-Hall.

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

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

Jayavardhana Rama, G. L., A. G. Ramakrishnan, R. Muralishankar, and P. Prathibha, 2002. A complete text-to-speech synthesis system in tamil. *Proc. IEEE 2002 Workshop on Speech Synthesis*:191–194.

Moulines, E. and F. Charpentier, 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5):453–467.

Muralishankar, R., A. G. Ramakrishnan, and P. Prathibha, 2004. Modification of pitch using DCT in the source domain. *Speech Communication*, 42:143–154.

Murthi, M. N. and B. D. Rao, 2000. All-pole modelling of speech based on the minimum variance distortionless response spectrum. *IEEE Trans. Speech and Audio Pro.*, 8(3):221–239.

Musicus, B. R., 1985. Fast MLM power spectrum estimation from uniformly spaced correlations. *Proc. IEEE Trans. Acou. Speech Sig. Pro.*, (4):1333–1335.

Roucos, S. and A. Wilgus, 1985. High quality time-scale modification of speech. *Proc. ICASSP*:493–496.

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

Stoica, P. and R. Moses, 1997. *Spectral Analysis.* Englewood Cliffs, NJ: Prentice-Hall.

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

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