SPECOM 1528 ARTICLE IN PRESS No. of Pages 16, Model 3+
31 January 2006 Disk Used

Available online at www.sciencedirect.com
SCIENCE DIRECT

Speech Communication xxx (2006) xxx–xxx

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Effect of voice quality on frequency-warped modeling of vowel spectra ☆

Pushkar Patwardhan *, Preeti Rao

Department of Electrical Engineering, ACRE, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India

Received 19 July 2005; received in revised form 7 December 2005; accepted 6 January 2006

Abstract

The perceptual accuracy of an all-pole representation of the spectral envelope of voiced sounds may be enhanced by the use of frequency-scale warping prior to LP modeling. For the representation of harmonic amplitudes in the sinusoidal coding of voiced sounds, the effectiveness of frequency warping was shown to depend on the underlying signal spectral shape as determined by phoneme quality. In this paper, the previous work is extended to the other important dimension of spectral shape variation, namely voice quality. The influence of voice quality attributes on the perceived modeling error in frequency-warped LP modeling of the spectral envelope is investigated through subjective and objective measures applied to synthetic and natural steady sounds. Experimental results are presented that demonstrate the feasibility and advantage of adapting the warping function to the signal spectral envelope in the context of a sinusoidal speech coding scheme.

Keywords: Voice quality; Spectral envelope modeling; Frequency warping; All-pole modeling; Partial loudness

1. Introduction

A successful low bit rate speech coding algorithm requires a good model for the speech signal together with an effective parameter quantization algorithm. A popular model for low bit rate speech coding has been the sinusoidal model, an important example of which is the multiband excitation (MBE) model (Griffin and Lim, 1988). In the case of voiced speech, the parameters of the model are the fundamental frequency (or pitch), and the amplitudes and phases of the harmonics. The harmonic amplitudes represent the product of the source excitation and vocal tract spectra. At low bit rates, estimated phases are usually dispensed with, and the accurate representation of the pitch and harmonic amplitudes becomes critical to the perceptual quality of decoded speech. Steady vowel sounds are particularly sensitive to harmonic amplitude representation errors.

The quantization of harmonic amplitudes is most demanding on the bit allotment in sinusoidal coding, and methods for efficient quantization have been an important topic of research. A widely used method for the quantization of the amplitudes of the harmonics is based on the modeling of a spectral

☆ A portion of this work was presented at the International Conference on Spoken Language Technology-04, New Delhi.

* Corresponding author. Tel.: +91 22 25346665; Mob: +91 9819073984; fax: +91 22 25723806.
E-mail address: pushkar@ee.iitb.ac.in (P. Patwardhan).

0167-6393/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.specom.2006.01.003

envelope fitted to the harmonic peaks (MacAulay and Quatieri, 1995). The spectral amplitudes are then reconstructed from samples of the modeled spectral envelope at harmonic frequencies. Representing the spectral envelope by the coefficients of an all-pole filter enables the use of one of the many efficient quantization methods available in the speech coding literature. The order of the all-pole model has a significant effect on the accuracy of the modeled spectral amplitudes. While the all-pole representation of the spectral envelope is expected to capture local resonances accurately, capturing additional features such as overall spectral tilt and spectral zeros due to nasality typically lead to an increase in the number of poles required for an adequate approximation. Further, for similar spectral envelope, low pitched sounds require higher LP model order for similar perceived quality levels (Champion et al., 1994; Rao and Patwardhan, 2005).

In the interest of achieving low bit rates, however, it is necessary to keep model order as low as possible. Frequency-scale warping before all-pole modeling of the spectral envelope is a widely used method to improve the perceptual accuracy of modeling for a given model order. Frequency-scale warping leads to a more accurate representation of the low frequency spectrum at the cost of increased errors in the high frequency region. Although perceptual scales such as the Bark-scale and its variants have been widely used in LP modeling of speech spectra, our recent work (Rao and Patwardhan, 2005) on synthetically generated steady vowel sounds (using fixed excitation source parameters) indicated that the performance of frequency warping depended to a great extent on the nature of the underlying sound spectrum. It was observed that front vowels such as [eɪ] and [i] in fact were better modeled without Bark-scale warping. This was explained by the low frequency first formant structure of these vowels which fails to mask high frequency distortion adequately. In a subjective experiment using natural speech, it was found that the comparative behaviour of modeling error under different warping conditions in the case of the non-front vowels was inconsistent across speakers indicating a further dependence on speaker voice quality as reflected in the overall spectral envelope.

In the present study we attempt to extend our previous work by exploring aspects of voice quality that could influence the spectrum modeling error. The influence of voice quality on perceived modeling error in frequency-warped all-pole modeling of the spectral envelope is studied via the framework of MBE model analysis–synthesis.

While the overall spectral envelope for voiced speech is determined by the glottal source waveform, vocal tract transfer function and lip radiation, it is the glottal source waveform that most directly affects the relative strengths of the low and high frequency harmonics. The scope of the present study is restricted to variations in laryngeal voice qualities for they are closely linked to gross differences in spectral envelope. An investigation based on subjective and objective experimental evaluation is presented. Finally, the applicability of the results to improving speech quality in a low bit rate sinusoidal coder is discussed.

2. Frequency-warped LP modeling of discrete spectra

A discrete spectrum, characterized by a fundamental frequency and the amplitudes of the components at harmonic frequencies, can be represented by the coefficients of an all-pole model. A smooth spectral envelope is first derived to fit the harmonic amplitudes using a suitable interpolation method such as linear interpolation of log amplitudes (Hermanksy et al., 1985; Rao and Patwardhan, 2005). The power spectrum obtained from the interpolated envelope is used to compute the autocorrelation function via the inverse DFT. The Levinson–Durbin algorithm is applied to obtain the LP coefficients. The spectral amplitudes can be recovered by sampling the reconstructed spectral envelope represented by the all-pole coefficients at the harmonic frequency locations. Frequency-scale warping may be incorporated in the spectral envelope modeling by mapping the input harmonic frequencies to corresponding warped frequency locations by means of a warping function based on a chosen perceptual scale. The log-linear interpolation of the discrete spectral amplitudes (now non-uniformly spaced in frequency) is carried out to obtain a densely sampled spectral envelope as detailed in Section 3 of (Rao and Patwardhan, 2005). Uniformly spaced samples at 20 Hz interval are found to be adequate for the LP modeling of narrowband speech spectra in the sinusoidal coding context.

In this work, we use the framework of MBE narrow band coding of speech to evaluate frequency-warped LP modeling of voiced sounds. The MBE analysis–synthesis model offers a convenient framework for the evaluation of spectral envelope modeling (Molyneux et al., 1998; Rao and Patwardhan, 2005) although the results are applicable more generally to other vocoders. Voiced regions are modeled by harmonics of a fundamental frequency in the MBE vocoder, and unvoiced bands by spectrally shaped random noise. Thus parameters of the MBE speech model consist of the fundamental frequency, band voicing decisions, and the harmonic amplitudes (Griffin and Lim, 1988).

MBE analysis involves the use of high resolution DFT in an analysis-by-synthesis loop for the accurate determination of pitch, voicing and spectral amplitudes for each input frame of speech (typically 20 ms duration). In the case of fully voiced speech, synthesis is achieved by summing of sinusoids each corresponding to one harmonic. Adjacent frames are combined using either overlap-add or the interpolation of phase depending on the extent of pitch variation (Griffin and Lim, 1988). With unquantised parameters, synthesized speech of very high quality is obtained, particularly in voiced regions. Spectral envelope-interpolated LP modeling is utilized to achieve the low bit rate coding of the harmonic spectral amplitudes in an MBE model based speech coder. In the present work, we investigate some aspects of the performance of frequency-scale warping in obtaining improved quality at low LP model order. Based on the considerations discussed in (Rao and Patwardhan, 2005) the LP model order selected for the speech quality evaluation experiments is 10. The test and evaluation framework is depicted Fig. 1 which shows the generation of reference and test signals used in the quality comparisons.

3. Voice quality and its spectral correlates

Voice quality refers to the auditory impression a listener gets upon hearing the speech of another talker. Voice quality is determined by the articulators of the vocal tract as well as the characteristics of the vocal folds. For speech, the vocal folds have a predominant importance. We refer to this aspect of voice quality that results from differences in vocal cord vibratory patterns, or laryngeal voice quality (Childers and Lee, 1991). Periodic glottal excitation is a characteristic of "modal" voices. The perceptual and spectral variations within modal voices can be attributed to the differences in the timing parameters of the glottal pulse. Since we are concerned with the spectral representation of steady periodic sounds, we do not consider those voice types that are characterized by voicing aperiodicities such as aspiration noise and pitch/amplitude jitter.

3.1. Modal voice production parameters

The speech signal is generated by excitation of vocal tract caused by the modulated flow of air generated by the lungs. The volume of air passing through the vocal cords, also known as glottal volume flow is the excitation signal which is periodic in the case of voiced speech. The glottal pulse shape

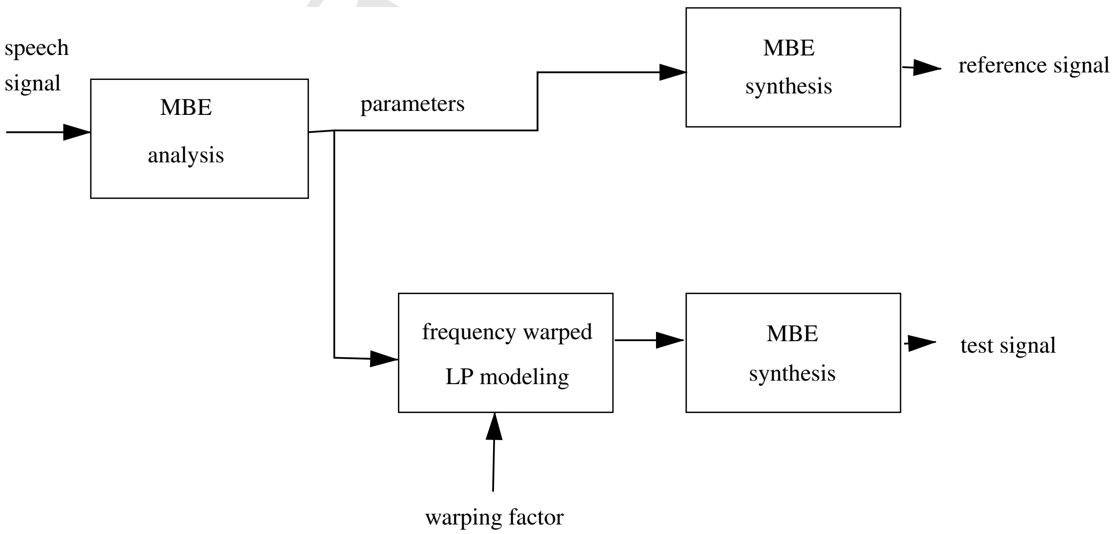

Fig. 1. Generation of MBE modeled test and reference signals.

has been extensively studied and modeled for different voice qualities. A typical glottal pulse period (glottal volume flow) and its derivative (Childers, 2000; Childers and Lee, 1991) are shown in Fig. 2(a) and (b).

The parameter t_p denotes the instant of the maximum in the glottal flow waveform. The parameter t_e is the instant of the maximum negative differentiated glottal flow. The t_a is the time constant of the exponential curve of the second segment of the LF model, and parameter t_c is the instant at which complete glottal closure is reached. The spectral shape of the glottal pulse derivative is dependent on the values of the timing parameters. Modal voices, common among male speakers, are characterized by the nearly complete closure of the vocal folds. Fig. 2(c) shows glottal pulse derivative waveforms corresponding to three different sets of timing parameters and Fig. 2(d) shows the corresponding spectra. An abrupt closure of the vocal folds (low t_a) creates strong high frequency harmonics while a relatively gradual closure of vocal folds results in spectra with strong low frequency harmonics and weak high frequency harmonics. The perceptual correlate of the relative strengths of the low and high frequency harmonics in the glottal source spectrum is the brightness of the voice. "Dark" voices have relatively weak high harmonics while "bright" voices are characterized by flatter spectra. Apart from the abruptness in the closure of the vocal folds, the symmetry of the open phase of the glottal pulse as captured by the "pulse skew" also affects the spectrum. It influences the relative amplitudes of the low frequency harmonics (especially the first and second harmonics) (Doval and d'Alessandro, 1997).

3.2. Spectral correlates

Spectral shape can provide useful cues to relevant aspects of voice quality. Some of the important cues that reflect the characteristics of glottal signal are as follows:

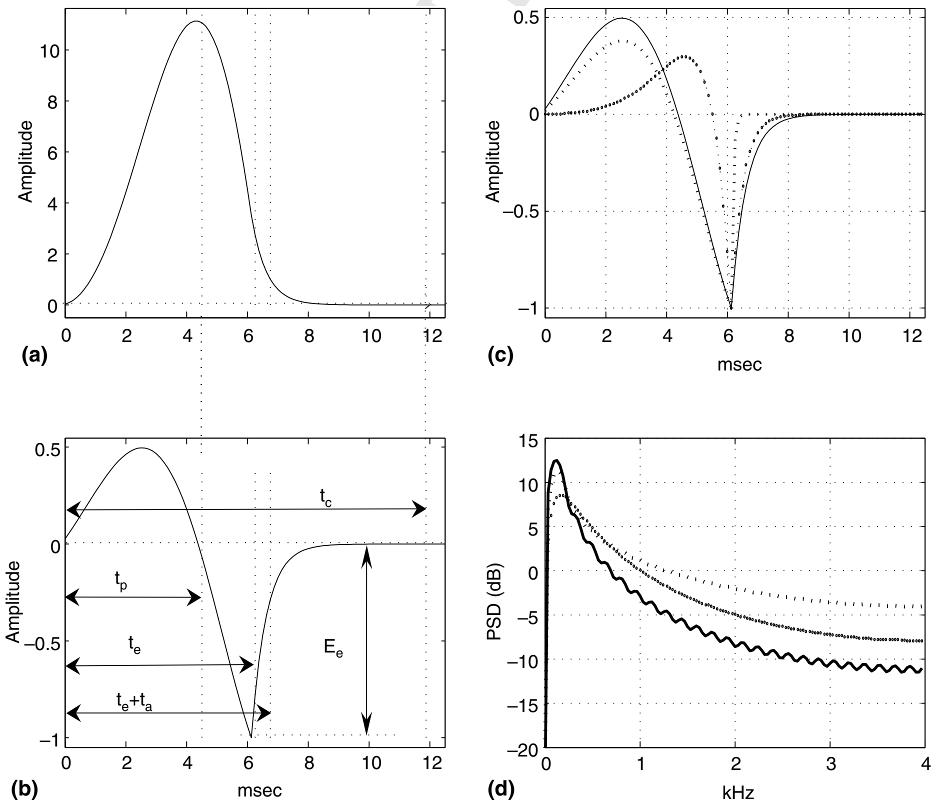

Fig. 2. Temporal and spectral structure of glottal pulse. (a) Glottal flow, (b) glottal flow derivative, (c) glottal flow derivative for three different sets of timing parameters, (d) spectra corresponding to (c).

H1–A3: This is the ratio of amplitude of first harmonic (H1) relative to that of the third-formant spectral peak (A3), and has been used by Hanson (1997) to characterize spectral tilt. The middle and high frequency components are directly affected by the duration of glottal pulse. An abrupt closure in the glottal cycle results in relatively strong middle and high frequency components, i.e. A3 is typically higher than what it would be with a gradual glottal closure. Fig. 3 shows the comparison of harmonic spectra for vowel [ɑ] for dark and bright sounds. It clearly shows the differences in slopes of overall spectral envelopes.

H1–H2: This is the ratio of amplitudes of the first two harmonics. It is affected by the glottal pulse skew as well as the open quotient (OQ) of the glottal pulse. The glottal pulse skew is measured by the speed quotient (SQ) which is computed as ratio of opening interval to the closing interval. While the open quotient is ratio of open glottal interval to the pitch period. A higher SQ indicates lower H1–H2 (Doval and d'Alessandro, 1997), while higher OQ implies higher H1–H2.

Roll-off: Roll-off is an indicator of shape of the spectrum. This is defined as frequency within which 85% of the total accumulated magnitude is concentrated (Burred and Lerch, 2004). The roll-off is determined by the largest DFT bin "R", which satisfies

$$\sum_{k=N1}^{R-1} |X(k)| \leqslant 0.85 \sum_{k=N1}^{N2} |X(k)| \qquad (1)$$

where $X(\cdot)$ is the DFT spectrum of the input frame. $N1$ and $N2$ define the DFT bins corresponding to the range of frequencies over which the roll-off is computed. Based on the values of $N1$ and $N2$, R can represent the roll-off in any frequency region of interest. The roll-off computed over the full frequency spectrum (0, 4000 Hz) captures the overall slope of the spectrum. For right-skewed spectra the value of roll-off turns to be high while for left-skewed spectra the

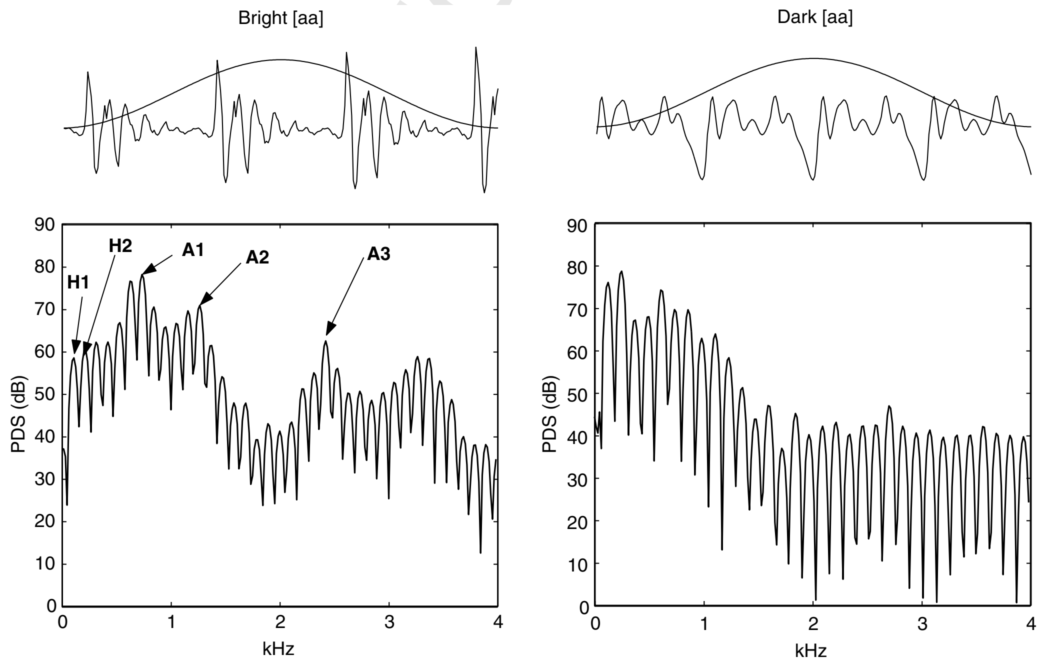

Fig. 3. Temporal waveforms and spectra of "dark" and "bright" vowel [ɑ]. Acoustic parameters labeled in "bright" [ɑ] are: H1 – amplitude of first harmonic, H2 – amplitude of second harmonic, A1 – amplitude of first formant, A2 – amplitude of second formant, A3 – amplitude of third formant.

roll-off is lower. For bright vowels we expect a high roll-off.

It must be remarked that perceived voice quality may also be influenced by recording conditions. The spectral measures described in this section would reflect both glottal waveform characteristics as well as the transfer function of the recording setup.

4. Experimental evaluation

Experiments were designed to investigate the influence of voice quality on the modeling error from frequency-warped LP modeling of the spectral envelope as presented in Section 2. Vowels uttered in low pitched modal voices of different brightness were obtained by synthesis as well as extraction from natural speech.

4.1. Test set generation

Natural and synthetic samples corresponding to eight distinct vowels, as shown in Table 1, were generated for the experiment. Synthesis of vowels allowed the manipulation of the glottal timing parameters which in turn enabled control over the voice quality. The synthetic vowels were generated using articulatory synthesis with the articulation parameters estimated from provided target speech based upon an analysis-by-synthesis approach (Childers, 2000). For the synthesis of each of the vowels the estimated articulatory parameters together with an excitation signal, separately generated using an LF model with fixed parameters, are used in a period-by-period synthesis. The synthesized voice quality is dependent on the LF model parameters (t_o, t_p, t_a, t_e, t_c and E_e) (Childers, 2000). By varying the timing parameters it is possible to generate variations in the spectral slope (and consequently in the perceived brightness) of the voice. The four sets of glottal pulse timing parameters used in the experiment are shown in Table 2. Four instances of each vowel corresponding to each set of glottal parameters (i.e. a total of 32 synthetic sounds) were generated. Based on the resulting overall spectral slope, we term these as "dark", "weak-dark", "weak-bright" and "bright". Each vowel was generated at 120 Hz pitch (i.e. $t_o = 8.3$ ms) for a duration of 350 ms. The start and end of the sound were tapered to avoid abrupt transitions at the boundaries.

The set of natural vowels consisted of four instances of each vowel (pitch ranging between 80 Hz and 120 Hz) taken from slowly uttered words with neutral intonation from a set of six male (Indian) speakers. The sounds were selected such that two instances of each vowel were of bright quality and two of dark quality as judged by listening as well as visual examination of the spectral slope. The ends were tapered to eliminate abrupt transitions. This resulted in a total set of 32 vowel sounds (four instances of eight distinct vowels selected in all across six speakers). In a few instances of the vowel sounds (i, ɔ, ʊ) it was found that the actual vowel durations were in between 250 ms and 280 ms and insufficient for judgment. The duration of such sounds were extended to reach the minimum duration of 350 ms by repeating the first two periods at the start of the sounds and final two periods at the end of file. This artificial extension was not detectable upon listening. The sounds were analyzed by the MBE model algorithm to estimate the pitch and spectral amplitudes. The spectral amplitudes thus obtained for each 20 ms speech frame were modeled with 10th order frequency-warped LP coefficients with a chosen frequency warping factor. The synthesis was carried out by standard sinusoidal synthesis method (MacAulay and Quatieri, 1995) using the spectral amplitudes obtained by frequency-warped all-pole model approximation. It

Table 1
Set of vowels used in the experiment

Vowel ID	IPA symbol	Typical word	ARPABET (symbol)
1	ɑ	"guard"	[aa]
2	æ	"cat"	[ae]
3	ɔ	"orchid"	[ao]
4	ʌ	"cut"	[uh]
5	oʊ	"lotus"	[ow]
6	ʊ	"boot"	[uw]
7	i	"beet"	[iy]
8	eɪ	"mate"	[ey]

Table 2
Glottal waveform timing parameters as a % of t_o, (Childers, 2000) used to generate the synthetic vowels with four voice qualities

Voice quality	Timing parameters		
	t_p (%)	t_e (%)	t_a (%)
Dark	35	50	15
Weak-dark	40	50	15
Weak-bright	40	50	2
Bright	45	50	2

$t_o = 8.3$ ms; $t_c = 0.9 t_o$; $E_e = 40$.

was then compared with a reference sound synthesized using originally estimated spectral amplitudes. There were three test sounds for each of the 32 reference sounds: LP modeled sound without frequency warping ('U'), LP modeled with a mild version of Bark-scale warping ('M') and LP modeled with Bark-scale warping ('B') based on the frequency warping scales described in (Rao and Patwardhan, 2005).

4.2. Subjective test

A subjective listening experiment was set up to compare the perceived qualities of the LP modeling with different warping factors. Six normal hearing listeners participated in the test. The test material was presented to the subject at normal listening levels through high quality head-phones connected to a PC sound card in a quiet room. For each of the test sounds the following "trio" presentation format was used: "reference-test-reference" where 250 ms silence separated the sounds. This format made the degradation of the test with respect to the reference sound relatively easy to detect. The reference sound was also separately available for listening.

The subjects were asked to rank (using ranks 1, 2 and 3) the relative perceived degradations of the test sounds U, M and B with respect to the corresponding reference sound for each of the vowel sounds in test set. Rank 1 would correspond to the least degradation. An undetectable difference would result in a suitable tied ranking. Subjects were allowed to listen to the test trios associated with a given reference any number of times before making a decision. Each listener did the test using the same set of items in different orders on three separate occasions. Since it was found that the subjective ranks were nearly constant across listeners and trials, an overall ranking order was derived for each test vowel by combining the numerical ranks across listeners and trials as follows.

A numerical value is assigned to each rank in a single trial as follows: rank $1 = 2$ points, rank $2 = 1$ point, rank $3 = 0$ point. In case of a tie, equal number of points are given to both (i.e. two points each for first position or one point each for second position). The points are summed for each item across listeners to get the "total score". The total scores are then used to derive an overall subjective ranking. While the data within each item (i.e. reference sound) can be compared, this is not so across items because the relative degradations were rated from best to worst (ranked) only within each item set. E.g. a high score in item 1 may indicate perceptually transparent modeling, while the same score in item 2 may indicate a small but clearly audible degradation.

4.3. Objective measurement of degradation

To quantify the perceived modeling error in each of the vowel sounds, we used partial loudness (PL), a psychoacoustical distance measure with demonstrated correlation with subjective judgement of spectral degradation in steady vowel sounds (Moore et al., 1997; Rao and Patwardhan, 2005). The spectrum modeling error is treated as the signal whose loudness is to be estimated in the presence of a background masker (the reference sound). The reference sound is intended to take the role of the background noise and the modeled sound that of the signal plus background noise (reference sound plus the modeling error). The linear spectral distortion is treated as additive noise with power spectrum given by the difference between reference and modeled power spectra (Rao et al., 2001). The PL of the spectrum modeling error is computed for each frame and averaged across frames to obtain the PL value for the entire sample.

The objective rankings of relative degradation across U/M/B warping conditions were derived based upon the computed distortion measure for each of the reference-test sound pairs. The performance of an objective distance measure in predicting subjective judgments may be evaluated by computing a measure of correlation between corresponding objective and subjective rankings. The Spearman's correlation coefficient (Miller, 1989) is a suitable measure since it makes minimal assumptions about the data.

4.4. Sentence level test

The experiments described so far involved quality judgments on isolated vowels. In order to obtain a perspective on the usefulness of the observations in the context of a speech coding application, a sentence-level listening test was designed. Phonetically balanced sentences as well as sentences which predominantly contained one or another vowel were collected for a subjective listening experiment. Low pitched voices from a set of male speakers were selected. Table 8 lists the 15 sentences along with an indication of the dominant phonetic content as well as perceived brightness of the voice. The sentences were uttered by speakers whose voices would be clearly classified as bright sounding or dark sounding in neutral speech.

Each of the sentences was subjected to MBE analysis and synthesis. The reference sentence was taken to be the one synthesized with the estimated spectral amplitudes while the test sentences were synthesized from 10th order LP modeled spectral amplitudes with and without Bark-scale warping to obtain "B" and "U" versions, respectively. Eight subjects were presented with the reference and the two corresponding test sentences, and asked to choose the one perceptually most similar to the reference sentence. Based on a spectral shape measure described in Section 5.3, test sentences constructed by switching the warping parameter between "U" and "B" (depending on the nature of the frame) were also included in the listening test.

5. Results and discussion

Tables 3 and 4 summarise the results of subjective and objective ranking of the experiment involving the natural vowels. Tables 5 and 6 are the corresponding results for the synthetic vowels.

The best rank (rank 1) implies the least perceived degradation among the three test sounds, while the worst (rank 3) implies perceived degradation was maximum among the three test sounds. Higher score, and lower PL value, implies lower perceived degradation. The last column contains Spearman's rank correlation (Miller, 1989) which is an indicator of how close to each other the subjective and objective ranks are. In this section, we comment on several aspects of the obtained experimental results.

5.1. Subjective ranking experiment

We first discuss the observations on frequency-warped all-pole modeling of natural vowels. From Table 3, which provides the subjective and objective evaluation of the reference-U, reference-M and reference-B pairs of the dark sounding vowels, we observe that the reference-B pair is ranked better than the reference-U for all the vowels except for the vowels [eɪ] and [i]. This is consistent with the experimental results of (Rao and Patwardhan, 2005). From Table 4 we observe that in the case of the "bright" vowels the reference-B pair is always degraded as compared to the reference-U pair. This is evident from the fact that it has always been ranked subjectively lower than the reference-U pair. This is contrary to what has been generally accepted in the literature. The reference-M pair is seen to be typically ranked between the other two warping scales. In the case of the synthetic vowels, we see from Table 5 that the dark vowels satisfy the same

Table 3
Objective and subjective results on degradation due to spectral envelope modeling of natural vowels with dark voice quality

ID	IPA sym.	Sample	Subjective results						Objective results						Correlation coefficient
			Scores			Ranks			PL			Ranks			
			U	M	B	U	M	B	U	M	B	U	M	B	
1	ɑ	1	2	14	19	3	2	1	0.77	0.28	0.25	3	2	1	1
		2	5	20	28	3	2	1	0.86	0.14	0.09	3	2	1	1
2	æ	1	3	22	24	3	2	1	1.34	0.45	0.57	3	1	2	0.5
		2	4	22	23	3	2	1	0.77	0.23	0.23	3	1	1	0.75
3	ɔ	1	8	25	28	3	2	1	0.51	0.25	0.24	3	2	1	1
		2	7	20	27	3	2	1	1.28	0.82	0.61	3	2	1	1
4	ʌ	1	7	15	25	3	2	1	0.44	0.33	0.33	3	1	1	0.75
		2	5	23	22	3	1	2	1.29	1.01	0.98	3	2	1	0.5
5	oʊ	1	9	26	18	3	1	2	0.59	0.13	0.14	3	1	2	1
		2	4	20	27	3	2	1	1.28	0.17	0.15	3	2	1	1
6	ʊ	1	1	26	18	3	1	2	0.85	0.11	0.13	3	1	2	1
		2	2	15	27	3	2	1	0.99	0.68	0.31	3	2	1	1
7	i	1	30	15	0	1	2	3	0.64	0.83	0.94	1	2	3	1
		2	24	21	4	1	2	3	0.38	0.32	0.42	2	1	3	0.5
8	eɪ	1	28	16	3	1	2	3	0.15	0.19	0.36	1	2	3	1
		2	24	20	3	1	2	3	0.09	0.16	0.2	1	2	3	1

All vowels were modeled with 10th LP model order with each of the three warping conditions: unwarped (U), mild-Bark-scale warped (M) and Bark-scale warped (B)."1", "2" indicate samples from different voice qualities. The last column shows the Spearman's correlation coefficient between subjectively and objectively measured ranks.

Table 4
Same as Table 3 but for natural vowels with bright voice quality

ID	IPA sym.	Sample	Subjective results						Objective results						Correlation coefficient
			Scores			Ranks			PL			Ranks			
			U	M	B	U	M	B	U	M	B	U	M	B	
1	ɑ	1	30	11	5	1	2	3	0.34	0.93	0.6	1	3	2	0.5
		2	30	15	0	1	2	3	0.11	0.35	0.49	1	2	3	1
2	æ	1	30	17	2	1	2	3	0.31	0.57	0.57	1	2	2	0.75
		2	25	23	8	1	2	3	0.43	0.24	0.48	2	1	3	0.5
3	ɔ	1	26	8	12	1	3	2	0.59	0.64	0.94	1	2	3	0.5
		2	14	7	5	1	2	3	0.04	0.06	0.14	1	2	3	1
4	ʌ	1	28	13	4	1	2	3	0.33	0.31	0.38	2	1	3	0.5
		2	29	16	0	1	2	3	0.43	0.49	0.78	1	2	3	1
5	oʊ	1	30	9	6	1	2	3	0.15	0.39	0.34	1	3	2	0.5
		2	28	16	1	1	2	3	0.29	0.16	0.39	2	1	3	0.5
6	ʊ	1	30	14	1	1	2	3	0.43	1.04	1.02	1	3	2	0.5
		2	30	15	0	1	2	3	0.21	0.92	0.77	1	3	2	0.5
7	i	1	29	19	3	1	2	3	0.11	0.07	0.49	2	1	3	0.5
		2	30	15	0	1	2	3	0.1	0.37	0.83	1	2	3	1
8	eɪ	1	28	17	2	1	2	3	0.3	0.3	0.37	1	1	3	0.75
		2	27	10	8	1	2	3	0.4	0.39	0.52	2	1	3	0.5

Table 5
Objective and subjective results on degradation due to spectral envelope modeling of synthetic vowels with dark voice quality

ID	IPA sym.	Sample	Subjective results						Objective results						Correlation coefficient
			Scores			Ranks			PL			Ranks			
			U	M	B	U	M	B	U	M	B	U	M	B	
1	ɑ	1	5	19	11	3	1	2	0.46	0.09	0.11	3	1	2	1
		2	6	14	13	3	1	2	0.11	0.10	0.09	3	2	1	0.5
2	æ	1	0	19	11	3	1	2	0.70	0.08	0.06	3	2	1	0.5
		2	3	19	11	3	1	2	0.29	0.06	0.04	3	2	1	0.5
3	ɔ	1	3	19	14	3	1	2	0.82	0.11	0.12	3	1	2	1
		2	5	18	15	3	1	2	0.32	0.10	0.13	3	1	2	1
4	ʌ	1	9	13	11	3	1	2	0.67	0.29	0.44	3	1	2	1
		2	5	13	14	3	2	1	0.39	0.32	0.28	3	2	1	1
5	oʊ	1	4	11	16	3	2	1	0.42	0.13	0.12	3	2	1	1
		2	2	13	15	3	2	1	0.25	0.17	0.13	3	2	1	1
6	ʊ	1	7	12	17	3	2	1	0.38	0.10	0.12	3	1	2	0.5
		2	0	16	14	3	1	2	0.36	0.10	0.12	3	1	2	1
7	i	1	16	16	4	1	1	3	0.42	0.21	0.27	3	1	2	−0.25
		2	16	15	2	1	2	3	0.22	0.18	0.35	2	1	3	0.5
8	eɪ	1	18	14	2	1	2	3	0.19	0.15	0.24	2	1	3	0.5
		2	20	10	0	1	2	3	0.06	0.12	0.15	1	2	3	1

All vowels were modeled with 10th LP model order with each of the three warping conditions: unwarped (U), mild-Bark-scale warped (M) and Bark-scale warped (B).“1”, “2” indicate samples from different voice qualities. The last column shows the Spearman’s correlation coefficient between subjectively and objectively measured ranks.

trend as the dark natural vowels (that is reference-B pair has been ranked better except for [i] and [eɪ]). Table 6 for the bright vowels, shows some inconsistencies in [ɔ] and [ʊ] while all other vowels show the expected superiority of reference-U. An explanation for the inconsistencies is provided in Section 5.3.

5.2. Correlation with partial loudness

Tables 3–6 indicate an overall high positive correlation between subjective ranks and partial loudness based ranks. In fact, the relative positions of non-warped and Bark-warped in the subjective preference are correctly captured by the objective dis-

Table 6
Same as Table 5 but for synthetic vowels with bright voice quality

ID	IPA sym.	Sample	Subjective results						Objective results						Correlation coefficient
			Scores			Ranks			PL			Ranks			
			U	M	B	U	M	B	U	M	B	U	M	B	
1	ɑ	1	15	14	5	1	2	3	0.06	0.07	0.15	1	2	3	1
		2	15	12	6	1	2	3	0.05	0.06	0.15	1	2	3	1
2	æ	1	2	18	10	3	1	2	0.13	0.04	0.09	3	1	2	1
		2	15	13	3	1	2	3	0.08	0.03	0.09	2	1	3	0.5
3	ɔ	1	1	19	12	3	1	2	0.43	0.05	0.20	3	1	2	1
		2	8	18	12	3	1	2	0.28	0.06	0.23	3	1	2	1
4	ʌ	1	13	14	5	2	1	3	0.25	0.34	0.5	1	2	3	0.5
		2	14	13	6	2	1	3	0.31	0.52	0.43	1	3	2	0.5
5	oʊ	1	12	12	8	2	1	3	0.15	0.17	0.22	1	2	3	0.5
		2	14	13	6	1	2	3	0.19	0.16	0.27	2	1	3	0.5
6	ʊ	1	3	14	19	3	2	1	0.70	0.16	0.17	3	1	2	0.5
		2	2	15	13	3	1	2	0.55	0.20	0.27	3	1	2	1
7	i	1	18	14	2	1	2	3	0.19	0.09	0.56	2	1	3	0.5
		2	12	17	2	2	1	3	0.37	0.32	0.75	2	1	3	1
8	eɪ	1	20	10	2	1	2	3	0.05	0.13	0.22	1	2	3	1
		2	20	10	7	1	2	3	0.05	0.14	0.26	1	2	3	1

tance. By and large, the only inconsistencies are with respect to the relative position of mild-Bark-warped test sounds.

The partial loudness is an indication of how audible is the spectral distortion. It is the sum of the partial specific loudness contributions of the distortion distributed across the signal spectrum. In frequency-warped LP modeling, there is an improved spectral match in the low frequency region at the cost of increased errors in the high frequency region. Whether frequency warping improves the overall perceptual accuracy of the fit will depend on whether the increase in the partial loudness of the high frequency distortion is exceeded by the decrease in the partial loudness of the low frequency distortion. Fig. 4 shows, for a particular frame of a dark vowel sound [ɑ] the comparison between Bark-warped and non-warped LP modeling in terms of spectral amplitude distortion and partial loudness distribution. On comparing Fig. 4(a) and (b), we note the reduced low frequency distortion and increased high frequency distortion introduced by the Bark-scale frequency warping. Further, a comparison of the partial loudness distributions of Fig. 4(c) and (d) indicates that the reduction in the low frequency distortion contributes more to decreasing audible error than the effect of the increase in the high frequency distortion. The correct prediction of this perceptual effect by the partial loudness model indicates that the high frequency spectral distortion is masked to a great extent by the relatively strong low frequency signal spectral components. This explanation is consistent with the opposite behaviour of the front vowels [i] and [eɪ] since their low first formant causes reduced spread of masking to the high frequency region. In the corresponding plots of Fig. 5 for a frame of a bright vowel [ɑ], we note the similar effect. That is, the reduction in the low frequency distortion due to Bark-scale warping does not compensate adequately for the increased high frequency distortion as observed in the partial loudness distributions of Fig. 5(c) and (d). The higher "loudness" of the high frequency error can be attributed to the insufficient masking from the low frequency components of the reference signal which are only comparable in strength to the high frequency components.

5.3. Relation to spectral cues

Although the partial loudness is able to predict the subjectively preferred warping condition accurately in most cases, it is desirable to have a less computationally complex measure for use in practice. The simple spectral cues of Section 3.2 capture the overall spectral balance in some way or other and therefore merit investigation.

Fig. 6 plots the H1–A3 of each of the set of 32 natural vowels plus 32 synthetic vowels used in the experiments. The H1–A3 is measured for a repre-

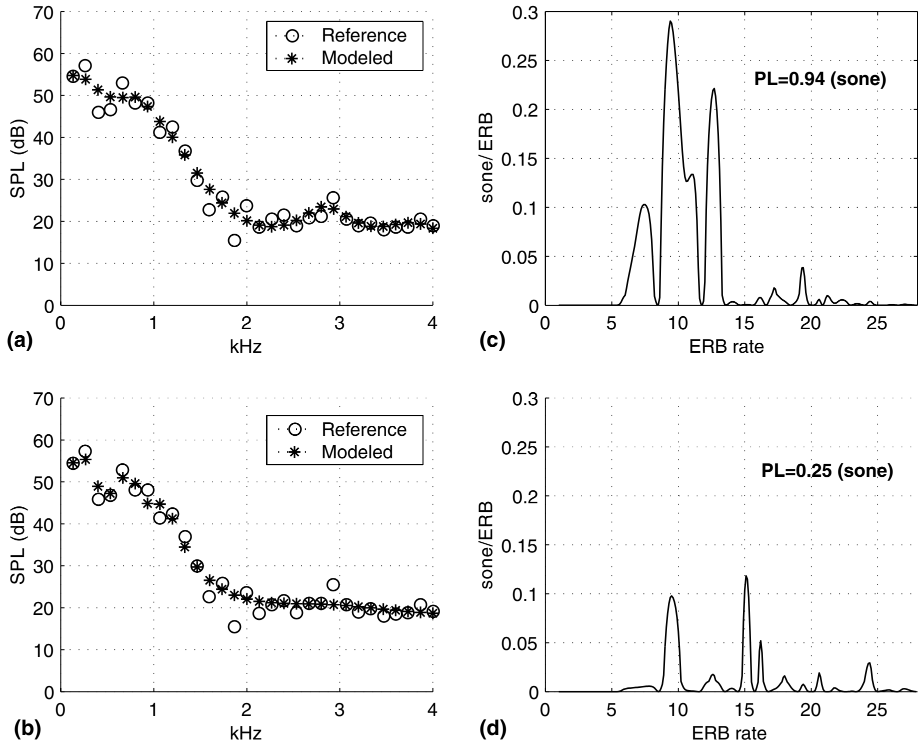

Fig. 4. Modeling of "dark" natural [ɑ] with pitch = 105 Hz, without and with Bark-scale frequency warping at LP model order = 10. (a) and (b) Spectral harmonic amplitudes of reference (○) and modeled (*) sounds are connected with a dotted line to show the spectral envelopes without warping and with Bark-scale warping respectively. (c) and (d) Partial loudness distribution of spectral distortion in unwarped and Bark-warped modeling, respectively.

sentative frame picked from center of steady vowel sound. The bright sounding vowels fall at the lower end of the *y*-axis relative to the corresponding vowels with dark voice quality.

Based on the results of Tables 3 and 4, the vowel instances that have been ranked better with Bark-scale frequency-warped all-pole modeling are shown encapsulated by a triangle while the ones without warping before modeling are shown encapsulated by a square. Fig. 6 clearly shows that the sounds with higher H1–A3 benefit more from Bark-scale warping whereas the ones with lower H1–A3 benefit less (but for the expected cases of [i] and [eɪ]). In fact, from the data on natural vowels, it appears that H1–A3 = 10 dB may be considered a rough threshold for prediction on whether Bark-warping before LP modeling would help to improve perceived quality. From Fig. 6(b) we note that the synthetic vowels [ʊ] and [ɔ] in our set had H1–A3 values clustered at the higher end of the scale which may explain why they do not show any variation in the subjectively preferred warping condition. It is seen that while H1–A3 captures the overall spectral tilt, it is insensitive to the relative distribution of spectral energy in the low frequency region, and hence cannot distinguish the low first formant characteristic of the vowels [i] and [eɪ].

In the view of these above observations, it is desirable to have a measure that describes not only the overall frequency concentration but also that in a specific frequency region. The spectral roll-off defined in Section 3.2 was adapted for the purpose. Experiments revealed that the spectral roll-off value of (1) evaluated in the region (0, 1500 Hz) was a good indicator of the lowness of the first formant as well as of the largeness of the gap between the first and second formant. That is, this measure is low only in cases of front high and front-mid vowels where there is large gap between F1 and F2. Further the spectral roll-off computed in the reverse direction in the same interval was a good indicator of the low frequency skew as represented by H1–H2. Based on these considerations and experimentally determined thresholds, the following roll-off based rule was applied to determine whether warping should be enabled:

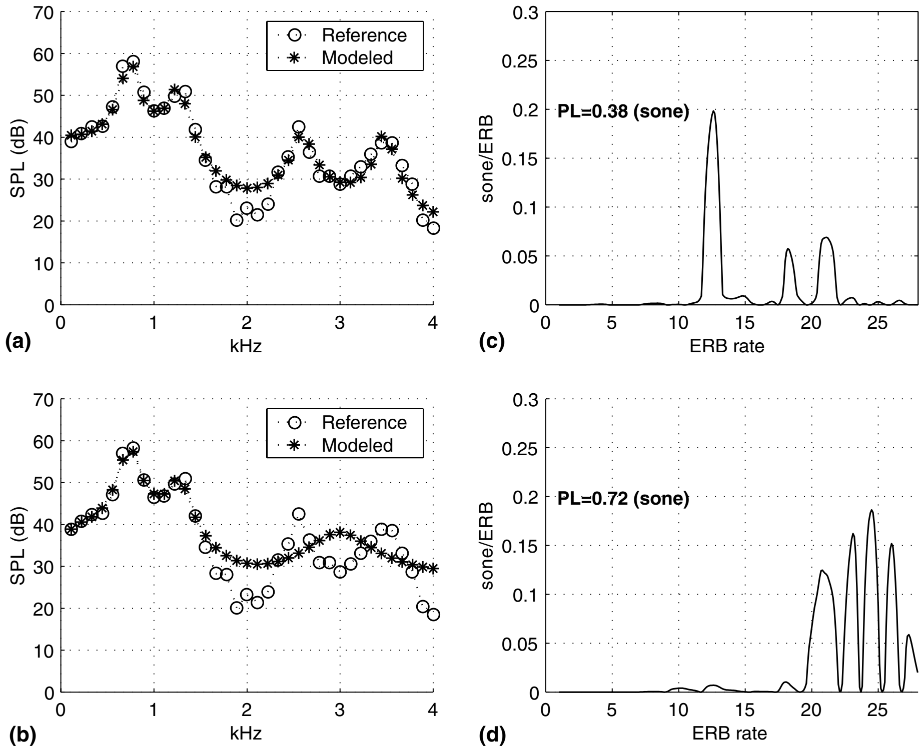

Fig. 5. Modeling of "bright" natural [ɑ] with pitch = 105 Hz, without and with Bark-scale frequency warping at LP model order = 10. (a) and (b) Spectral harmonic amplitudes of reference (○) and modeled (*) sounds are connected with a dotted line to show the spectral envelopes without warping and with Bark-scale warping respectively. (c) and (d) Partial loudness distribution of spectral distortion in unwarped and Bark-warped modeling, respectively.

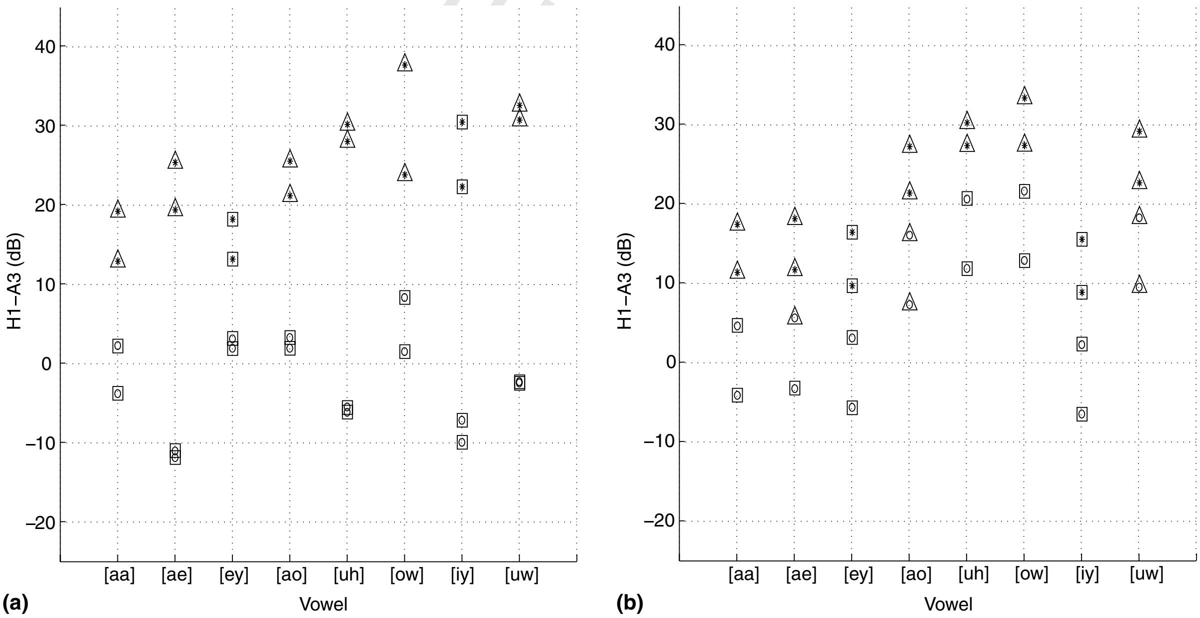

Fig. 6. Spectral cue H1–A3 along with subjective preference of quality between Bark-warped and unwarped modeling (a) natural vowels and (b) synthetic vowels. (△) Bark-scale warping preferred. (□) Unwarped preferred, "o": Vowels with "bright" voice quality, "*": Vowels with "dark" voice quality.

if (SR $<$ 1700 Hz) and (PSR $>$ 400 Hz) and (RPSR $<$ 350 Hz), use Bark warping.

SR is the roll-off computed (0, 4000 Hz), PSR is the roll-off computed in (0, 1500 Hz), RPSR is the roll-off computed in (1500, 0 Hz). Table 7 shows the normalized correlation of the objectively predicted preferred warping condition with the subjectively preferred condition for each vowel for the three objective measures: PL, H1–A3 and roll-off based. We see that the roll-off based rule does better than H1–A3 and is comparable to PL while being significantly less complex computationally.

5.4. Some comments on nasalised vowels

Nasalisation of vowels is known to be associated with a change in the first formant region of the spectrum that may be modeled with an added pole and zero, and therefore expected to be difficult to model with a low order LP model. Since a prominent feature of nasal vowels lies in the low frequency spectrum, it is expected that Bark-scale warping before the LP modeling would help to retain the perception of nasality. Informal listening experiments with artificially nasalized natural vowels support the above observation. To nasalize the vowels, a filter was applied that had a zero and a pole located above the first formant (Feng and Castelli, 1996). A separation of 100 Hz between F1 (frequency of first formant), FNZ (frequency of zero) and FNP (frequency of pole) is sufficient to introduce perceptible nasality in the vowel. In a listening experiment with nasalised vowels, when asked to rank overall quality, however, with and without Bark-scale warping before LP modeling, subjects often judged that modeling without warping was less degraded compared with modeling with Bark-scale warping, especially in the case of bright vowels. In the case of dark vowels too, high frequency distortion from warping was sometimes more significant than in the case of the corresponding non-nasalised vowel. This may be attributed to lowered masking from reduced low frequency components due to the presence of a spectral zero. Overall there was less consistency between subjects and also less correlation between subjective and objective rankings (based on the partial loudness measure) compared with the results on the non-nasalised vowels. This may be explained by the existence of conflicting perceptual effects from the two different spectral cues namely of loss of nasality, and high frequency spectrum distortion. In such cases, it is expected that higher level cognition would come into play in a subject's judgement of rank, a situation that cannot be predicted from a low level distance measure based on a model of the peripheral auditory system.

Table 7
Normalised correlation values between subjectively observed and objectively predicted preferred warping condition (between non-warped and Bark-warped only) for each vowel set

Vowel	Natural			Synthetic		
	PL	H1–A3	Roll-off	PL	H1–A3	Roll-off
ɑ	1	1	1	1	1	1
æ	1	1	1	1	0.75	0.5
ɔ	1	1	1	1	0.75	1
ʌ	1	1	1	1	0.5	0.5
oʊ	1	1	1	1	0.5	0.5
ʊ	1	1	1	1	0.75	1
i	1	0.5	1	0.75	0.75	1
eɪ	1	0.5	1	1	0.75	1

5.5. Sentence level test

Table 8 shows the preferred warping condition (as selected between non-warped, "U" and Bark-warped, "B") in terms of overall subjective quality on sentences. We note that in the case of the dark-voiced sentences, subjects preferred the Bark-warped condition except when the phonemes [eɪ] and [i] dominated. In the case of the bright voices, the subjects preferred modeling without warping in all cases. These observations are consistent with the isolated phoneme results discussed earlier. For the speech coding application, this suggests the possibility of improving overall perceived quality by making available a limited set of warping factors for dynamic selection based on frame-level spectral characteristics. Information on the selected warping factor can be conveyed to the decoder via one or two bits depending on the number of distinct warping factors available.

In the second part of the experiment instead of switching the frames manually, the prediction rule of Section 5.3 was implemented and warping condition was switched automatically frame by frame. A subjective listening test comparing the dynamically switched warped sentences with the preferred fixed warping condition of Column 5 of Table 8 revealed that in all cases the subjects rated the former either better or indistinguishable from the latter. This is shown in Column 6 of Table 8. The percentage of frames that were selectively switched to Bark-scale

Table 8
Sentences used in the subjective evaluation of preferred frequency warping at the sentence level

Sr. no.	Sentences	Voice quality	Predominant phoneme	Subjective preference		% Switched to B
1	All the balls were brought from shopping mall	Dark	[ɔ]	B	S	49
2	By and large he was un-harmed	Dark	[ɑ]	B	S	42
3	Lord Paul was tall	Dark	[ɔ]	B	N	71
4	Early bird earns a worm	Dark	[ʌ]	B	N	57
5	Can that man carry those pans	Dark	[æ]	B	S	39
6	Draw each graph on new axis	Dark	Balanced	B	S	35
7	They may say the same in Spain	Dark	[eɪ]	U	N	15
8	His meal is eel meat	bright	[i]	U	N	0
9	We were away to walla walla	Dark	[eɪ], [i], [ɑ]	U	S	55
10	They all agree that the essay is barely intelligible	Bright	Balanced	U	N	17
11	Thick glue oozed out of the tube	Bright	[i], [ʊ]	U	N	4
12	Dont ask me to carry an oily rag like that	Bright	Balanced	U	N	4
13	A muscular abdomen is good for your back	Bright, nasal	Balanced	U	N	7
14	Withdraw only as much money as you need	Dark, nasal	Balanced	B	N	48
15	Withdraw only as much money as you need	Bright, nasal	Balanced	U	N	13

The first sub-column of the "subjective preference" indicates the warping condition preferred by listeners between two fixed warping conditions ("U" and "B"). The second sub-column indicates the subjective preference when the automatic warping condition prediction rule was used. "S": automatically switched preferred, "N": unable to differentiate and "O": fixed warping preferred.

warped modeling is shown in the last column of Table 8. We find that, for sentences no. 7, 8, 10, 11, 12, 13 and 15, the automatic algorithm keeps most frames unwarped. These sentences are characterized by either bright voice or presence of [eɪ] and [i], which are better modeled without warping.

At this point, it is relevant to comment on the use of high frequency pre-emphasis commonly used to improve LP modeling of speech spectra (Markel and Gray, 1976). Pre-emphasis serves to suppress the spectral tilt to an extent and thus improve the LP modeling of the formants. However it is suggested that pre-emphasis be avoided for unvoiced speech, while in the case of voiced speech, the pre-emphasis factor should ideally be signal dependent (Markel and Gray, 1976). This is also borne out by our own observations where we found that while pre-emphasis improves the modeling of voices with high spectral tilt (such as breathy, female voices), it degrades sounds with low energy at low frequencies [also noted by (Wong et al., 1980)] including sound spectra characterised by low H1–H2. Pre-emphasis was not used in the present study. It is anticipated that the selective use of pre-emphasis combined with frequency-scale warping based on the considerations presented in this paper can help to achieve further improvements in the perceptual accuracy of low order LP modeling of the speech spectral envelope.

6. Conclusion

Frequency-warping according to a perceptual scale is often applied to improve the perceptual accuracy of low order LP modeling of the speech spectral envelope in sinusoidal speech coding. Understanding the factors that influence the subjective perception of spectral envelope modeling errors can be useful in improving the attained speech quality. Steady vowel sounds are particularly sensitive to spectrum envelope modeling errors. Experimental investigations of the relative improvement in perceived quality from the use of different warping functions on natural and synthetic steady vowels have been presented. Contrary to what is generally accepted, it is found that whether the widely used Bark-scale frequency warping is effective depends on the vowel and voice quality attributes of the signal spectrum. Dark voices with their steep spectral slopes were found to benefit from Bark-scale warping but not so the relatively bright voices. The exceptions to this are the front-mid and front-low vowels [eɪ] and [i]. These vowels are observed to typically degrade when modeled with Bark-scale frequency warping relative to the reproduced quality without warping. In summary, our studies based on observations presented in this paper and a previous paper (Rao and Patwardhan, 2005) suggest that the vowels [eɪ] and [i] dominate over the dark/bright voice quality differences.

The subjective ratings of relative perceived degradation were closely predicted by an objective distance measure based on the partial loudness of the spectrum distortion. This suggested that frequency masking plays a role in determining whether the improved low frequency spectrum match from Bark-scale warping is sufficient to compensate for the accompanying high frequency spectrum distortion. Only when the low frequency components of the spectrum are relatively strong and sufficiently spread is there a clear benefit from Bark-scale warping. This condition was shown to be closely linked to the spectral slope as quantified by measures such as H1–A3 and roll-off which can therefore act as useful cues in predicting the suitability of a warping function for a particular sound. A rule has been proposed to predict the subjectively preferred warping condition. The rule is based on a combination of spectral roll-off values computed in different regions of the spectrum.

A sentence-level listening test confirmed the results of the isolated vowel experiments and also demonstrated the effectiveness of the warping prediction rule. For the speech coding application, this suggests the possibility of improving overall perceived quality by making available a limited set of warping factors for dynamic selection based on frame-level spectral characteristics. Information on the selected warping factor can be conveyed to the decoder via one or two bits depending on the number of distinct warping factors available. Alternatively, a simple but slightly less effective solution is to use fixed warping according to a mild version of the Bark-scale.

Audio demonstrations may be accessed at http://www.ee.iitb.ac.in/~prao/speech_comm_1/index.html.

7. Uncited references

Klatt and Klatt (1990) and Hanson and Chuang (1999).

Appendix A. Results on synthetic vowels

A subjective ranking experiment similar to that carried out for natural set of vowels was carried out on the synthetic vowels. The results appear in Tables 5 and 6. The trend nearly matches with the natural vowels, except for the dark vowel [IY], which may be attributed to the difficulty of carefully controlling the output of the articulatory synthesizer for this particular phoneme. (We found that the articulatory synthesizer produced a harmonic envelope which became less smooth on the higher frequency end for this phoneme. The partial loudness based ranking was not consistent with the subjective test in such cases probably due to the deviation from the model assumptions.)

References

Burred, J., Lerch, A., 2004. Hierarchical automatic audio signal classification. J. Audio Eng. Soc. 52 (8), 724–739.

Champion, T., MacAulay, R., Quatieri, J., 1994. High-order all-pole modeling of the spectral envelope. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, pp. 529–532.

Childers, D., 2000. Speech Processing and Synthesis Toolboxes. John Wiley and Sons.

Childers, D., Lee, C., 1991. Vocal quality factors: analysis, synthesis and perception. J. Acoust. Soc. Am. 90 (5), 2394–2411.

Doval, B., d'Alessandro, C., 1997. Spectral correlates of glottal waveform models: an analytic study. In: Proceedings of the IEEE International Conference on Acoustics, Speech, Signal Processing, pp. 1295–1298.

Feng, G., Castelli, E., 1996. Some acoustic features of nasal and nasalized vowels: a target for vowel nasalization. J. Acoust. Soc. Am. 99 (6), 3694–3706.

Griffin, D., Lim, J., 1988. Multiband excitation vocoder. IEEE Trans. Acoust. Speech Signal Process. 36 (8), 1223–1235.

Hanson, H., 1997. Glottal characteristics of female speakers: acoustic correlates. J. Acoust. Soc. Am. 101 (1), 466–481.

Hanson, H., Chuang, E., 1999. Glottal characteristics of male speakers: acoustic correlates and comparison with female data. J. Acoust. Soc. Am. 106 (2), 1064–1077.

Hermanksy, H., Hanson, B., Wakita, B., Fujisaki, H., 1985. Linear predictive modeling of speech in modified spectral domain. In: Digital Processing of Signals in Communications, pp. 55–63.

Klatt, D., Klatt, L., 1990. Analysis, synthesis and perception of voice quality variations among female and male talker. J. Acoust. Soc. Am. 87 (2), 820–855.

MacAulay, R., Quatieri, T., 1995. Sinusoidal Coding, in Speech Coding and Synthesis. Elsevier, Amsterdam.

Markel, J., Gray, A., 1976. Linear Prediction of Speech. Springer Verlag, Berlin.

Miller, J., 1989. Correlation, in Statistics for Advanced Level. Cambridge University Press.

Molyneux, D., Parris, C., Sun, X., Cheetham, B., 1998. Comparison of spectral estimation techniques for low bit-rate speech coding. In: Proceedings of the International Conference on Spoken Language Processing, pp. 946–949.

Moore, B., Glasberg, B., Baer, T., 1997. Model for prediction of thresholds, loudness and partial loudness. J. Audio Eng. Soc. 45 (4), 224–240.

Rao, P., Patwardhan, P., 2005. Frequency warped modeling of vowel spectra: dependence on vowel quality. Speech Commun. 47, 322–335.

Rao, P., van Dinther, R., Veldhius, R., Kohlrausch, A., 2001. A measure for predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds. J. Acoust. Soc. Am 109 (4), 2085–2097.

Wong, R., Hsiao, C., Markel, J., 1980. Spectral mismatch due to preemphasis in LPC analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. ASSP 28 (2), 263–264.