Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 47 (2005) 322–335

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Frequency warped modeling of vowel spectra: Dependence on vowel quality ☆

Preeti Rao, Pushkar Patwardhan *

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

Received 17 July 2004; received in revised form 10 January 2005; accepted 14 February 2005

Abstract

The compact representation of harmonic amplitudes in the sinusoidal coding of speech is an important problem in low bit rate speech compression. A widely used method to achieve this is by the all-pole modeling of the spectral envelope. Often a perceptually warped frequency scale is applied in the all-pole modeling to improve perceived accuracy at low model orders. In this work, an attempt is made to obtain a suitable frequency scale warping function by an experimental study on synthetic and natural steady vowels. Subjective listening experiments indicate that the change in perceived quality brought about by frequency warping depends on the underlying signal spectrum or vowel quality. Objective distortion measures are computed to obtain insights into the subjective results. It is observed that an auditory distance measure based on partial loudness shows high correlation with the subjective test scores indicating that frequency masking plays an important role in spectrum envelope modeling.

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

1. Introduction

The problem of compactly representing a discrete harmonic spectrum is common to many applications in sound synthesis and coding wherever periodic signals arise. An example is the coding and synthesis of voiced speech based on sinusoidal coding models. Well-known sinusoidal models are the Sinusoidal Transform Coder (MacAulay and Quatieri, 1995) and Multiband Excitation (MBE) coder (Griffin and Lim, 1988). In the case of voiced speech, the parameters to be quantized are the pitch or, fundamental frequency, and the spectral amplitude of each harmonic. The number of discrete spectral amplitudes

☆ An early version of this work was presented at ICSLP, 2002, Denver.

* Corresponding author. Tel.: +91 022 25764670/9819073984; fax: +91 022 25723806.

E-mail addresses: prao@ee.iitb.ac.in (P. Rao), pushkar@ee.iitb.ac.in (P. Patwardhan).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.02.009

depends on the number of harmonics within the frequency bandwidth and therefore on the fundamental frequency. In the context of low bit rate speech coding, this set of spectral amplitudes must be quantized as efficiently as possible with minimal loss in perceptual accuracy. Over the years various techniques have been proposed to represent compactly the variable number of discrete spectral amplitudes. Non-parametric methods such as scalar quantization, variable dimension vector quantization (Das and Gersho, 1995) and the Discrete Cosine Transform (DCT) have been proposed. However, at very low bit rates parametric methods such as linear predictive (LP) modeling are far more efficient (Champion et al., 1994; Kondoz, 1991; MacAulay and Quatieri, 1995; Molyneux et al., 1998). A set of linear pre- diction coefficients approximates the spectral amplitudes by another set of spectral amplitudes which are samples of the all-pole modeled spectral envelope at the harmonic frequencies. Apart from the compactness of the LPC representation, several effective methods are available for their quantization at low bit rates.

The LP coefficients may be obtained by standard time-domain LP analysis of the input signal. For high-pitched speakers, the frequency response of the LP filter tends to follow the fine structure of the speech power spectrum. Formant frequencies are often biased toward pitch harmonics and formant bandwidth is underestimated (Hermansky et al., 1985; Jelinek and Adoul, 1999). Methods that have been proposed to alleviate these effects include Discrete All-Pole Modeling (DAP) (Jaroudi and Makhoul, 1991) and its variants (Wei and Gibson, 2000), spectral envelope-interpolated LP (Hermansky et al., 1985) and Minimum Variance Distortionless Response (MVDR) modeling (Murthi and Rao, 1997). A listening test by Molyneux et al. (2000), comparing the performance of these techniques in the context of a low bit-rate MBE vocoder revealed that while the speech quality was significantly improved over that obtained by standard time-domain LP modeling, differences among the three techniques were insignificant. Of these methods, spectral envelope interpolation based LP analysis has been particularly popular due to its computational simplicity, insensitivity to overestimated model orders and the ability to accommodate perceptual frequency warping. A smooth spectral envelope is fitted to the harmonic amplitudes and from the corresponding power spectral envelope, an autocorrelation function is obtained by the inverse DFT. Next, a standard algorithm is used to obtain the LP coefficients from the autocorrelation function. Typically, high model orders (16–22) are required to accurately represent the harmonic amplitudes of voiced sounds (MacAulay and Quatieri, 1995) and methods to minimise the perceptual degradation at moderate LP model orders have been the subject of recent research. A widely used method to improve the perceived quality at given model order is frequency-scale warping of the spectral envelope prior to modeling, according to a perceptual scale (Hermansky et al., 1985; Koljonen and Karjalainen, 1984; Makhoul, 1976). This can be viewed as a pre-distortion of the spectral envelope so that the final error after usual LP modeling (and then restoring the spectral amplitudes by the corresponding inverse mapping) is perceptually more acceptable.

In this work, we consider the all-pole modeling of narrowband speech vowel spectra in the context of a low bit-rate MBE (Multi-Band Excitation) speech codec. Vowels have been used also in previous studies on LP modeling of speech due to their easily observed sensitivity to spectrum modeling errors (Varho and Alku, 1998). For the present study, the MBE analysis–synthesis model offers a convenient framework for the evaluation of spectral envelope modeling (as also in (Molyneux et al., 1998)) although the results are applicable more generally. In the present work, we investigate specifically the performance of frequency-scale warping in obtaining improved quality at low LP model orders. A lower LP model order implies fewer parameters to be quantized and a lower achievable bit rate. The purpose of the present study is to investigate the relation, if any, between the performance of warping and the underlying signal spectrum in the context of low order LP modeling of steady vowels, such sounds being expected to be perceptually most sensitive to spectrum modeling errors. The goal is to develop a better understanding of how to select the warping parameter for a particular application and, if possible, to adapt it to the input signal.

In the following sections, we describe the frequency warped all-pole modeling of spectral amplitudes. We investigate the influence of vowel quality on the perceived modeling error by subjective experiments and suitable objective measures. An objective distance measure based on an auditory model is used to obtain an explanation of the experimental results.

2. Frequency warped all-pole modeling

A property of the LP model spectral approximation is that it is equally accurate at all frequencies (Makhoul, 1976). However human auditory perception has a resolution that decreases with frequency. If the model order is low, it is possible that there is a preservation of spectral details at higher frequencies at the cost of the accurate modeling of the spectral envelope at the perceptually more important lower and middle frequencies. A suitable warping of the frequency scale so as to transform the spectral envelope into one in which the lower frequency regions now occupy a larger portion of the frequency range while the higher frequency regions are correspondingly compressed has the potential to result in perceptually more accurate LP spectrum matching. Frequency warped linear prediction has been discussed by Makhoul (1976) and Strube (1980), and applied to speech and audio compression by Koljonen and Karjalainen (1984) and Harma and Laine (2001) among others. Frequency warping can be implemented as a frequency to frequency mapping by the following transformation (Oppenheim et al., 1971):

$$\theta = f(\omega) = \arctan\left(\frac{(1-\alpha^2)\sin(\omega)}{(1+\alpha^2)\cos(\omega)+2\alpha}\right) \qquad (1)$$

where a frequency ω is mapped to the corresponding warped frequency θ. The parameter α controls the severity of warping.

A standard frequency warping scale is the Bark scale. It is based on critical band-rate as derived from auditory masking experiments. It is also closely related to perceptual scales based on other characteristics of hearing such as the Mel scale which is based on just noticeable differences in frequency (Zwicker and Fastl, 1974). By varying the warping parameter in (1) it is possible to approximate different perceptual scales for the case of 8 kHz sampling rate. Parameter $\alpha = 0.4$ approximates the auditory Bark scale at 8 kHz sampling frequency (Smith and Abel, 1999), while Mel scale is approximated by the warping parameter of $\alpha = 0.3$ in (1). Fig. 1 shows a few frequency scaling functions where 4 kHz represents the maximum frequency (equivalent to $\theta = \omega = \pi$ in (1)). Fig. 1 shows a close match between the originally experimentally derived Bark scale (Zwicker and Fastl, 1974), and bilinear transform (Smith and Abel, 1999) based Bark scale. We also illustrate two more mappings that correspond to the linear scale ($\alpha = 0$ resulting in conventional LP modeling) and a mild version of Bark scale warping ($\alpha = 0.2$). We observe that for a given warping parameter a fixed band of frequencies on the y-axis corresponds to a band of nearly the same width on the x-axis at low frequencies but to a wider band at high frequencies. This matches with the frequency-to-place

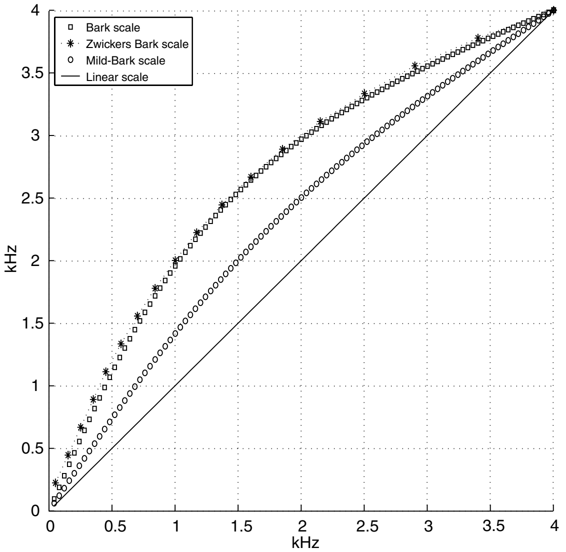

Fig. 1. Comparison of various frequency-scale warping functions. A mapping corresponding to $\alpha = 0.4$ in (1) closely approximates the Zwicker's Bark scale, while $\alpha = 0.2$ corresponds to the "mild-Bark" scale. An $\alpha = 0.0$ generates linear mapping.

transformation of the inner ear where with increasing frequency, a wider band of frequencies maps to a fixed width along the basilar membrane. While Harma and Laine (2001) reports significant gains in quality in narrow- and wide-band speech and audio coding from the use of Bark scale warping, he also suggests that the optimal warping function may depend on the test signal set as evident by the many distinct warping functions cited in the literature. However there are few studies comparing the performance of different warping functions in specific applications.

3. Implementation of MBE spectral envelope modeling

A widely known representative of sinusoidal modeling is the MBE speech model (Griffin and Lim, 1988). Voiced regions are modeled by harmonics of a fundamental frequency, and unvoiced regions by spectrally shaped random noise. The parameters of the MBE speech model consist of the fundamental frequency, band voicing decisions, and the harmonic amplitudes. The voicing information allows the mixing of the harmonic spectrum with a random noise spectrum in a frequency dependent manner. The spectral amplitudes represent the product of the excitation and vocal tract spectra. The phases of the harmonics are not transmitted but predicted during synthesis in most low rate coders. While the pitch and voicing decisions can be encoded efficiently, the quantization of spectral amplitudes presents a difficult trade-off between bit rate and quality. The perceptual quality of the decoded speech, particularly of steady voiced regions, depends strongly on the accuracy with which the harmonic amplitudes are quantized.

MBE analysis involves the use of a 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 unquantized parameters, synthesized speech of very high quality is obtained, particularly in voiced regions. The output of the analysis is a set of estimated amplitudes $\{S(\omega_1), S(\omega_2), \ldots, S(\omega_L)\}$ at the uniformly spaced L harmonic frequencies $\{\omega_1, \omega_2, \ldots, \omega_L\}$. The harmonic frequencies are then mapped to another set $\{\theta_1, \theta_2, \ldots, \theta_L\}$ of warped frequencies (now non-uniformly spaced) through (1).

Interpolation of the spectral amplitudes to obtain a smooth spectral envelope can be carried out in different ways. Linear interpolation between log spectral amplitudes offers the advantages of modeling the assumed parabolic behaviour of the spectral envelope in the vicinity of a formant peak as well as computational simplicity (Hermansky et al., 1985). The log spectral amplitudes are linearly interpolated to a fixed frequency spacing (20 Hz is found to be sufficient) to get the interpolated spectrum as follows:

$$Q(\theta_j) = 10^{\log|S(\theta_k)| + \left(\frac{\theta_j - \theta_k}{\theta_{k+1} - \theta_k}\right)(\log|S(\theta_{k+1})| - \log|S(\theta_k)|)}$$
$$\text{for } \theta_k < \theta_j < \theta_{k+1} \tag{2}$$

where, $\{S(\theta_k)\}$'s are the set of estimated spectral amplitudes at the warped frequencies $\{\theta_k\}$ and the $Q(\theta_j)$'s are the interpolated amplitudes. We thus obtain 200 spectral samples in the 4 kHz speech bandwidth. The autocorrelation is computed from the power spectrum by the IDFT operation, given by

$$R_i = \frac{1}{200} \sum_{j=0}^{199} |Q(\theta_j)|^2 \cos(i\theta_j) \tag{3}$$

Finally, the warped LP coefficients $\{a_k, k = 1, \ldots, p\}$ are computed by solving the following simultaneous equations using the Levinson–Durbin algorithm.

$$\sum_{k=1}^{p} a_k R_{|i-k|} = -R_i \quad \text{for } 1 \leqslant k \leqslant p \tag{4}$$

where p is the chosen order of all-pole model. The all-pole model coefficients represent the spectral envelope. The harmonic spectral amplitudes are later recovered for speech synthesis by generating the envelope and sampling it at the warped frequency locations.

4. Experimental evaluation

To investigate the performance of frequency warping in improving the perceptual fit of the all-pole model at low model orders, a subjective listening experiment involving synthetic and natural vowels was carried out. A suitable model order is first determined, and then MBE synthesized reference and LP-modeled envelope test sounds based on three different warping functions were compared. Objective measures of degradation were computed to obtain insight into the results of the subjective testing.

4.1. Choice of model order

In the context of speech coding, it is of interest to use as low a model order as possible to minimize the bit allocation. Since the spectral envelope used for the LP modeling is obtained by smooth interpolation between the harmonic amplitudes, it is expected that it will reflect the spectral details of the actual underlying source-tract envelope only for low-pitched sounds. For high-pitched voices the underlying spectrum is only sparsely sampled and therefore is typically smoother due to the larger extent of interpolation. This characteristic is expected to impact the order of the LP model required to achieve a good spectral fit. It is observed that female speech is modeled better than male speech at a given LP model order. The superior quality of synthesized speech for female voices at a given all-pole model order is attributed to the relatively small number of harmonic amplitudes to be approximated as well as to the higher smoothness of the interpolated spectral envelope at higher fundamental frequencies (Champion et al., 1994).

MBE modeling with envelope-interpolated LP was implemented on a set of 34 predominantly voiced sentences spoken by 17 male and 17 female speakers. The utterances were drawn from the TIMIT database (1989) as well as a locally recorded database. The selected male speech lay in the approximate pitch range of 100–120 Hz, while the female speech lay in the pitch range of 240–260 Hz. The LP modeling order was varied in steps of 2 starting from 4 and until 16 and then 22. The ITU P.862 PESQ MOS algorithm (Rix et al., 2001) was used to quantify the distortion introduced by the all pole modeling. Fig. 2 shows the average PESQ MOS score separately for the low- and high-pitched speech. We observe that in both cases the PESQ MOS increases rapidly with LP order at low orders but increases only slowly beyond order = 10 and then almost flattens out beyond order = 16. It is likely that order = 10 serves to adequately capture the local resonances of the spectrum and therefore phoneme quality (at least for non-nasalised vowels). For a given order, the high-pitched speech shows better PESQ MOS compared with the low-pitched speech. This is in accordance with the discussion in the previous paragraph. It may be noted though that the sentence samples chosen were all clearly voiced i.e. female breathy voices with their steep spectral slopes were not included. In view of the results of Fig. 2, an LP order = 10 is chosen as the base model order for investigation of further possible improvements in quality by the incorporation of frequency-scale warping. Also since low-pitched voices tend to suffer relatively more degradation, we confine our

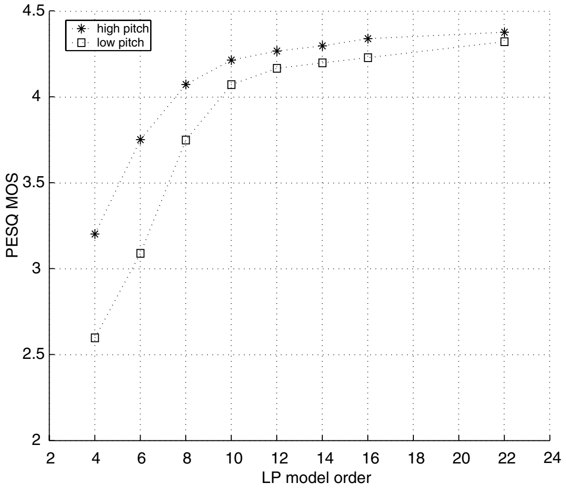

Fig. 2. Speech quality of the all-pole spectral envelope modeled sentences for various LP model orders as measured by PESQ MOS. □: Low pitch range, ∗: high pitch range.

experiments to synthetic and natural speech of low pitch.

4.2. Generation of test sounds

Our test set consisted of synthetic and natural steady vowels. 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). The articulatory synthesizer mimics a target vocal tract response based on an analysis-by-synthesis technique. For the synthesis of each of the vowels the corresponding articulatory parameters together with a source 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_p, t_e, t_a and t_0) (Childers, 2000). In this work, a set of parameters typical of a male modal voice as provided by Childers (2000) are chosen given by $t_p = 40\%$, $t_e = 55\%$ and $t_a = 1\%$ of the total period t_0. The synthetic vowel pitch was set at 120 Hz and duration at 350 ms. The start and end of the sound were tapered to avoid abrupt transitions at the boundaries. Table 1 contains a description of the 8 synthetic vowel sounds used in the experiment. The natural vowels were manually extracted from spoken words of several different speakers, including some from the TIMIT database (1989), in the pitch range 85–120 Hz. Two instances of each vowel spoken by different speakers were used in the final evaluation. The duration of steady vowels was tailored to lie in the range of 350–550 ms by periodic extension of the signal if necessary. The ends were tapered to eliminate abrupt transitions.

The sounds were analyzed to estimate the pitch and spectral amplitudes as described in Section 3. The spectral amplitudes thus obtained were modeled for each input speech frame using 10th order frequency-warped LP modeling with a chosen warping factor. Synthesis was carried out by standard sinusoidal synthesis methods (Griffin and Lim, 1988) using the spectral amplitudes obtained from the all-pole model approximation. A reference sound was synthesized using the original estimated spectral amplitudes. There were 3 test sounds for each reference sound: LP modeled without frequency warping (denoted "U"), LP modeled with mild-Bark scale warping (denoted "M") and LP modeled with Bark warping (denoted "B") corresponding to the three distinct frequency warping scales shown in Fig. 1.

4.3. Subjective tests

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. 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. A rank of 1 would imply that

Table 1
Description of the synthetic vowel sounds used in the experiment

Vowel ID	IPA symbol	Typical word	$F1$ (Hz)	$F2$ (Hz)	$F3$ (Hz)	$F4$ (Hz)
1	ɑ	"harm"	541	1094	2309	3052
2	æ	"hat"	574	1482	2410	3172
3	ɛ	"ate"	479	1668	2446	3253
4	ɒ	"law"	468	856	2401	2987
5	ʌ	"hurt"	486	1086	2538	3416
6	œ	"boat"	377	744	2279	2967
7	i	"seat"	274	1689	2536	3091
8	u	"boot"	270	820	1589	2344

Table 2
Rankings of subjective and objective degradation due to spectral envelope modeling and the corresponding Spearman's correlation coefficient as obtained from the experiments on synthetic vowels

Vowel ID	IPA symbol	Subjective ranks			Objective ranks						Spearman's correlation	
					LSD based			PL based				
		U	M	B	U	M	B	U	M	B	LSD	PL
1	ɑ	3	2	1	1	2	3	3	2	1	−1	1
2	æ	3	2	1	3	2	1	3	2	1	1	1
3	ɛ	1	2	3	1	2	3	1	2	3	1	1
4	ɒ	3	2	1	2	1	3	3	1	2	−0.5	0.5
5	∧	3	2	1	3	1	2	3	2	1	0.5	1
6	œ	3	1	2	1	2	3	3	1	1	−0.5	0.75
7	i	1	2	3	1	2	3	2	1	3	1	0.5
8	u	3	2	1	1	2	3	3	2	1	−1	1

All sounds were modeled at 10th order LP with each of 3 frequency-scale warping functions: non-warped (U), mild-Bark (M) and Bark (B).

Table 3
The exact values of objectively measured degradation corresponding to the modeled sounds of Table 2

Vowel ID	IPA symbol	LSD (dB)			PL (sones)		
		U	M	B	U	M	B
1	ɑ	3.55	3.83	3.86	0.73	0.37	0.32
2	æ	1.87	1.85	1.82	0.45	0.06	0.05
3	ɛ	1.76	2.50	2.66	0.12	0.13	0.17
4	ɒ	3.59	3.56	3.60	0.27	0.13	0.14
5	∧	4.02	3.70	3.74	0.29	0.28	0.21
6	œ	6.98	7.60	8.44	1.32	0.26	0.26
7	i	2.56	2.94	3.57	0.22	0.21	0.41
8	u	5.97	6.16	6.60	0.53	0.15	0.14

the indicated warping function resulted in the closest reproduction of the reference sound. An undetectable difference would result in a suitable tied ranking. Subjects were allowed to listen to the reference and test sounds any number of times before making a decision. A particularly convenient method of listening was found to be the "reference–test–reference" sequence. Each listener did the test using the same set of vowel sounds in different orders on three separate occasions. Since it was found that the subjective ranks were highly consistent across listeners and trials, an overall ranking order was derived for each reference vowel by combining the numerical ranks across listeners and trials. Tables 2 and 4 show the results of overall subjective ranking for the synthetic and natural vowels respectively. Increasing score implies increasing perceived degradation for each of the items. Although no instructions whatever on the type of degradation to listen for were given to the listeners, it was observed by them that the distortions due to modeling inaccuracies are characterized by changes in voice "color" rather than the articulation of the phoneme.

4.4. Objective measurement

Log spectral distortion (LSD), given by the mean squared error between log magnitude spectra, is widely used to quantify spectral degradation in speech coding literature (Quackenbush et al., 1988). The LSD is known to not correlate well with subjective ratings of quality but is popular due to its simplicity and historic value. A more psychoacoustically sound distance measure is "partial loudness". The modeling error is treated as the signal whose audible significance (loudness) is to be estimated in the presence of a background masker (the reference sound). The concept of partial loudness, or masked loudness, was first applied to measure speech quality degradation due to quantization by Schroeder et al. (1979). Recently, a computational model of partial loudness (PL) was proposed by Moore et al. (1997) that accounts for a large body of subjective data from psychoacoustical experiments. This model is based on the approximate stages of auditory processing representing the conversion of an input sound

Table 4
Rankings of subjective and objective degradation due to spectral envelope modeling and the corresponding Spearman's correlation coefficient as obtained, from the experiments on natural vowels

Vowel ID	IPA symbol	Instance	Subjective ranks			Objective ranks						Spearman's correlation	
						LSD based			PL based				
			U	M	B	U	M	B	U	M	B	LSD	PL
1	ɑ	p	3	2	1	3	1	2	3	2	1	0.5	1
		q	1	2	3	1	2	3	1	3	2	1	0.5
2	æ	p	3	2	1	3	1	2	3	1	2	0.5	0.5
		q	1	2	3	1	2	3	1	2	2	1	0.75
3	ɛ	p	1	2	3	1	2	3	1	2	3	1	1
		q	1	2	3	1	2	3	1	2	3	1	1
4	ɒ	p	3	2	1	2	1	3	3	2	1	−0.5	1
		q	1	3	2	1	3	2	1	2	3	0.5	0.5
5	ʌ	p	3	2	1	1	2	3	3	1	1	−1	0.75
		q	3	1	2	3	2	1	3	2	1	0.5	0.5
6	œ	p	3	1	2	1	2	3	3	1	2	−0.5	1
		q	1	2	3	1	2	3	1	3	2	1	0.5
7	i	p	1	2	3	1	2	3	1	2	3	1	1
		q	1	2	3	1	2	3	2	1	3	1	0.5
8	u	p	3	1	2	1	2	3	3	1	2	−0.5	1
		q	1	2	3	1	2	3	1	3	2	1	0.5

All sounds were modeled at 10th order LP with each of 3 frequency-scale warping functions: non-warped (U), mild-Bark (M) and Bark(B). "p" and "q" indicate samples from different speakers.

spectrum to the excitation pattern on the basilar membrane as depicted by Fig. 3(a). In the case of a signal presented with a background masker, a partial loudness is derived for the signal based on the computed excitation patterns of the signal and the masker.

In the context of our experiment, excitation patterns are separately computed for the reference and modeled power spectra. Then based on the channel-wise comparison of these two excitation patterns, we redefine the signal ($|E_1 - E_2|$) and noise masker ($\min(E_1 - E_2)$) excitation patterns, as shown in Fig. 3(b), to be used in the partial loudness computational model of Moore et al. (1997). Next, the analytical expressions provided in (Moore et al., 1997) are applied to obtain the loudness of the distortion in the presence of the background masker. These computations represent a non-linear relation by which a fixed dB difference between the excitation distributions results in a specific loudness distribution that is higher at higher excitation levels. The overall partial loudness is obtained by integrating the resulting specific loudness distribution. The partial loudness model was shown to perform well in the prediction of audible discrimination of spectral envelope distortions in vowel sounds as measured in a psychoacoustical experiment (Rao et al., 2001). It was found that a threshold value of 0.01 sone corresponded well with the threshold of audibility of the degradation. In the present study, however, we are concerned with the ranking of degradations which are clearly audible ("supra-threshold" distortions).

The two objective measures (LSD and PL) were computed between each of the reference sounds and the corresponding test sound for each warping condition. The objective distance was computed between the estimated and modeled spectral amplitudes for each frame of speech and then averaged to obtain one value for the entire sample. To compute the PL, an average listening level of 65 dB SPL was assumed corresponding to normal headphone listening. The obtained LSD and PL values are shown in Tables 3 and 5. The numerical values in these tables were used to obtain the corresponding objective similarity rankings shown in Tables 2 and 4. The performance of an objective distance

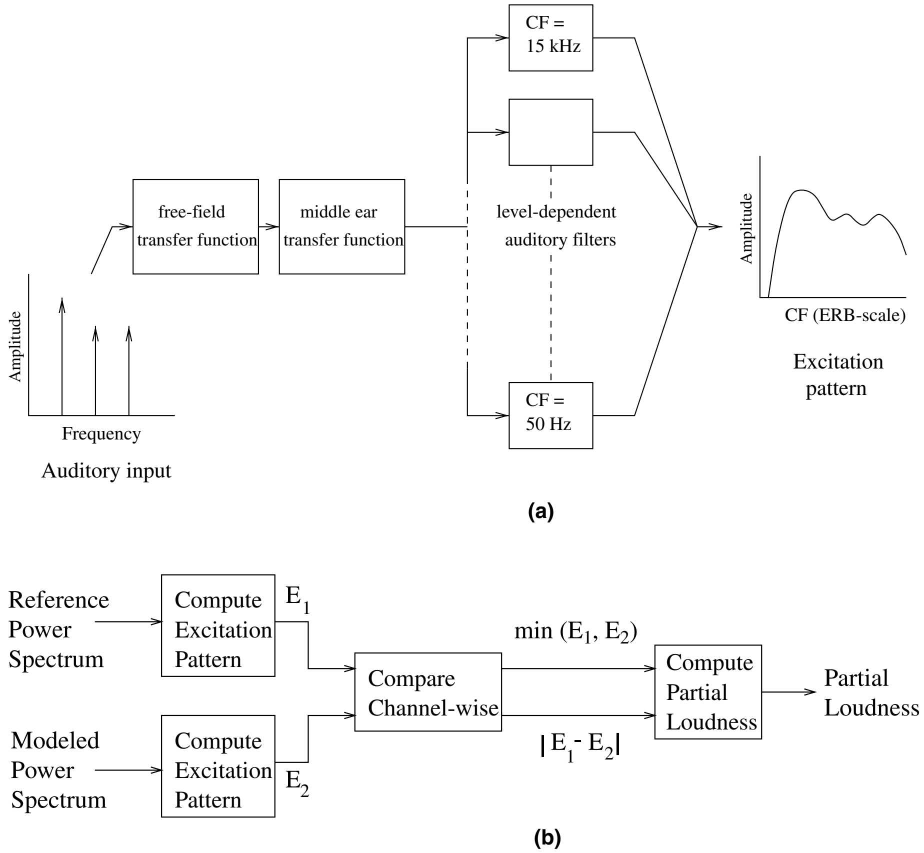

Fig. 3. Schematic diagrams for computation of excitation pattern and partial loudness. (a) Stages in obtaining excitation pattern from an input power spectrum. (b) Obtaining the partial loudness of the modeling error by appropriately defining the signal and masker excitation patterns.

measure in predicting subjective judgments may be evaluated by computing a measure of correlation between the objective rankings and subjective rankings. The Spearman's correlation coefficient (Miller, 1989) is a suitable measure since it makes minimal assumptions about the data. The Spearman's correlation coefficients between the subjective ranks and objective ranks are shown in the last columns of Tables 2 and 4. We note that the PL shows consistently high correlation with subjective judgment unlike the LSD.

5. Discussion

From the subjective experiment results on synthetic vowels shown in Table 2, we observe that the vowels [ɛ] and [i] have been marked as having degraded with warping (i.e. column B has been ranked as lower than column U), while the remaining vowels mentioned have been evaluated by the subjects as having improved with Bark scale warping compared with the non-warped condition. We note that, contrary to the generally accepted notion that Bark scale frequency warping improves the perceived quality of spectral modeling, certain vowels do not show this behaviour. The objective ranks based on PL for each of the vowels are highly correlated with the subjective ranks but not so for the LSD. The LSD is a relatively simple spectral distance measure that is sensitive to the relative change of intensity of spectral components but, of course, does not take into effect the auditory processes of critical band filtering and frequency masking.

Table 5
The exact values of objectively measured degradation corresponding to the modeled sounds of Table 4

Vowel ID	IPA symbol	Instance	LSD (dB)			PL (sones)		
			U	M	B	U	M	B
1	ɑ	p	2.35	1.79	1.86	0.86	0.14	0.09
		q	2.86	5.05	5.35	0.34	0.93	0.60
2	æ	p	2.94	2.43	2.60	1.34	0.45	0.57
		q	2.47	3.44	4.24	0.31	0.57	0.57
3	ɛ	p	2.53	3.46	4.02	0.15	0.19	0.36
		q	2.15	3.45	3.68	0.09	0.16	0.20
4	ɒ	p	3.94	4.41	4.60	0.51	0.25	0.24
		q	2.98	4.40	5.30	0.59	0.64	0.94
5	ʌ	p	3.52	5.38	5.84	0.44	0.33	0.33
		q	5.10	4.51	4.35	1.29	1.01	0.98
6	œ	p	2.30	2.94	2.95	0.59	0.13	0.14
		q	3.68	5.20	5.44	0.15	0.39	0.34
7	i	p	4.00	6.34	6.81	0.64	0.83	0.94
		q	5.35	6.26	6.92	0.38	0.32	0.42
8	u	p	2.10	2.62	3.08	0.85	0.11	0.13
		q	2.38	4.56	5.43	0.43	1.04	1.02

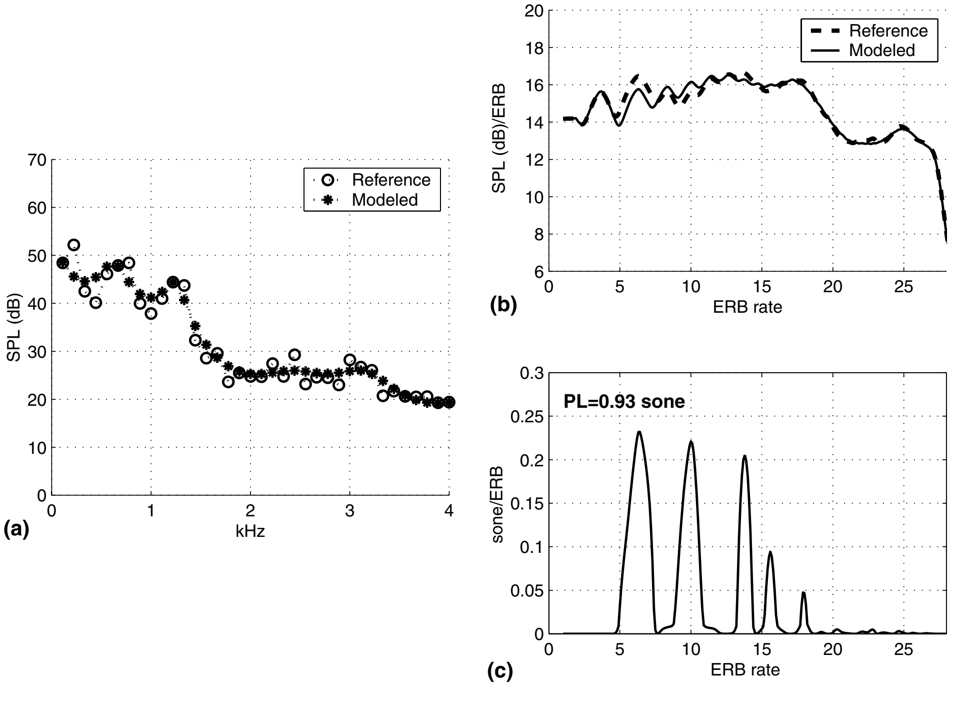

Fig. 4. Modeling of vowel [ɑ], having pitch = 105 Hz, without warping at LP model order = 10. (a) Spectral amplitudes of reference (○) and modeled (∗) sounds are connected to show the corresponding spectral envelopes. (b) Reference (- -) and modeled (—) auditory excitation patterns. (c) Partial loudness distribution of the spectral distortion.

Bark scale warping leads to more accurate modeling of low frequency spectral details at the cost of greater errors in the high frequency region. It is reasonable to suppose that whether or not the inaccuracies introduced in the high frequency region are audibly significant depends on the

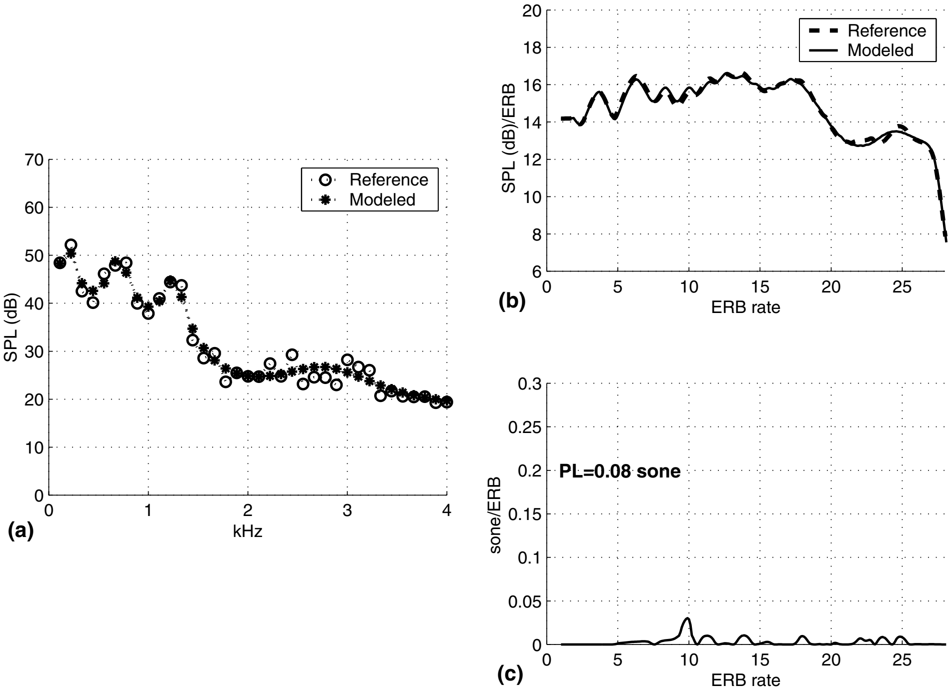

Fig. 5. Modeling of vowel [ɑ], having pitch = 105 Hz, with Bark scale frequency warping at LP model order = 10. (a) Spectral amplitudes of reference (○) and modeled (∗) sounds are connected to show the corresponding spectral envelopes. (b) Reference (- -) and modeled (—) auditory excitation patterns. (c) Partial loudness distribution of the spectral distortion.

spectral distribution or formant locations of the phoneme. To obtain an understanding of this behavior, we turn to the partial loudness measure which shows good correlation with subjective data. Shown in Figs. 4–7 are single frame-level comparisons of the original and modeled spectral envelopes for the non-warped and Bark-warped conditions for two vowel sounds, along with the auditory excitation patterns and the specific loudness distributions. The excitation pattern represents the output level of the successive auditory filters as a function of their center frequencies on an ERB-rate scale. Due to the decreasing frequency selectivity of the auditory filters with the increasing center frequency the excitation patterns have a relatively high resolution in the lower frequency region with the harmonics in this region clearly resolved. The asymmetric shape of the auditory filters results in the upward spread of the signal frequency components and therefore the “upward spread of masking”. The difference in the original and modeled spectral envelopes translates to a difference in the excitation patterns. The dB difference in the excitation distributions of the estimated and modeled amplitudes is converted to the masked specific loudness (in sones) by a non-linear relation in which a fixed dB difference corresponds to higher loudness at higher excitation levels (Moore et al., 1997). The integral of the specific loudness is the partial loudness which captures the overall audible significance of the spectrum difference.

Figs. 4 and 5 show the modeling of sound [ɑ] with non-warped and Bark warped frequency scaling. Fig. 4(a) shows the mismatch in spectral envelope in the low frequency region in absence of frequency warping, while in Fig. 5(a) the mismatch is significantly reduced. We observe that the audible significance of the distortion as predicted by the PL is very high in non-warped condition i.e. PL = 0.93 sone, while for Bark warped condition, PL = 0.08 sone. Figs. 6 and 7 show the modeling of the synthetic vowel [i] with non-warped and Bark warped frequency scaling. The vowel [i] is a

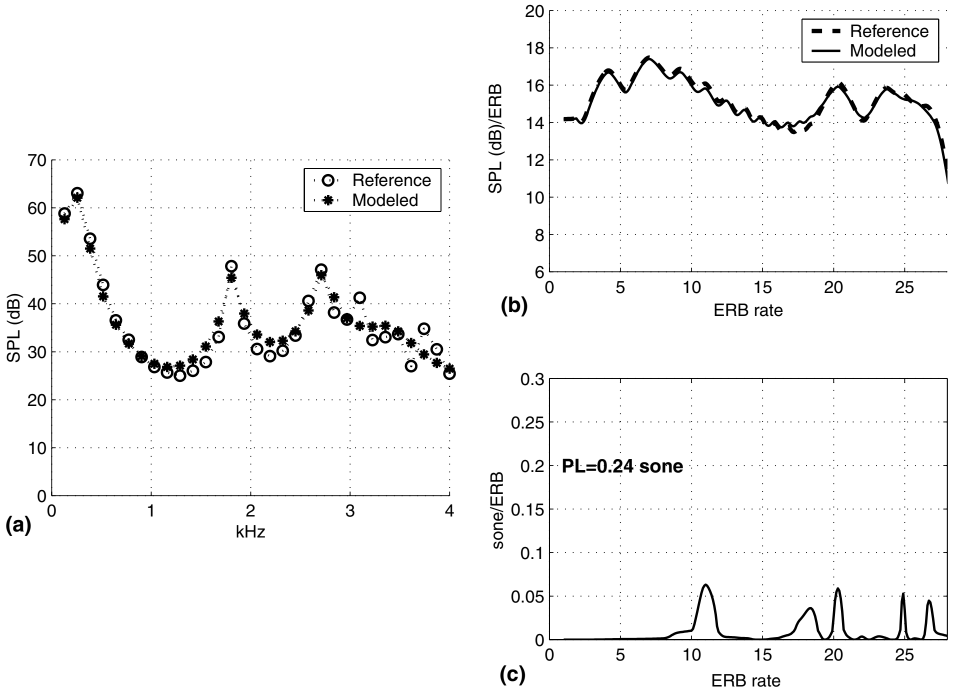

Fig. 6. Modeling of vowel [i], having pitch = 120 Hz, without frequency warping at LP model order = 10. (a) Spectral amplitudes of reference (○) and modeled (∗) sounds are connected to show the corresponding spectral envelopes. (b) Reference (- -) and modeled (—) auditory excitation patterns. (c) Partial loudness distribution of the spectral distortion.

front-high vowel characterized by a low first formant (274 Hz) and widely placed second formant ($F2$ = 1689 Hz). The spectral envelope of [i] shows a deep valley in the region between 400 Hz and 1500 Hz. The harmonics in this region are not strong enough to mask distortion in the high frequency region. Fig. 7(a) shows that large distortion is introduced in the modeled spectral envelope after Bark scale warping. The high frequency spectral distortion translates to significant perceived audible degradation in absence of significant masking components in its vicinity. The behaviour of [ɛ] (front-mid vowel) is similar to [i] due to its shared characteristic of widely spaced first and second formants. The “center of gravity” of the spectrum, which is known to play a role in the perception of vowels (Stevens, 2004), is located in the high frequency range for [ɛ] and [i], due to their prominent formant structure in the high frequency region. Any errors in the high frequency region directly affect the perceived qualities of these vowels. The remaining vowels (mainly back and mid) have strong enough low frequency components to mask the high frequency distortion arising from Bark scale warping. That frequency masking plays a role in the perception of spectral modeling errors has been recognized also by Lukasiak et al. (2000). In this work, only those spectral regions above a computed masking threshold are selected for LP modeling.

Turning to the subjective ranking results on natural vowels in Table 4, we see that the results match those of the synthetic vowels in the case of [i] and [ɛ] where Bark scale warping leads to increased perceived distortion. However the results are not so consistent in the case of the vowels [ɑ], [æ], [œ] and [u]. Each instance of a given vowel belongs to a different speaker. This suggests that apart from the nature of the vowel, speaker variability plays a role in influencing the performance of frequency-warped modeling. Speaker differences can influence the overall spectral tilt as well as exact formant locations and bandwidths. The studies in this paper were restricted to vowel

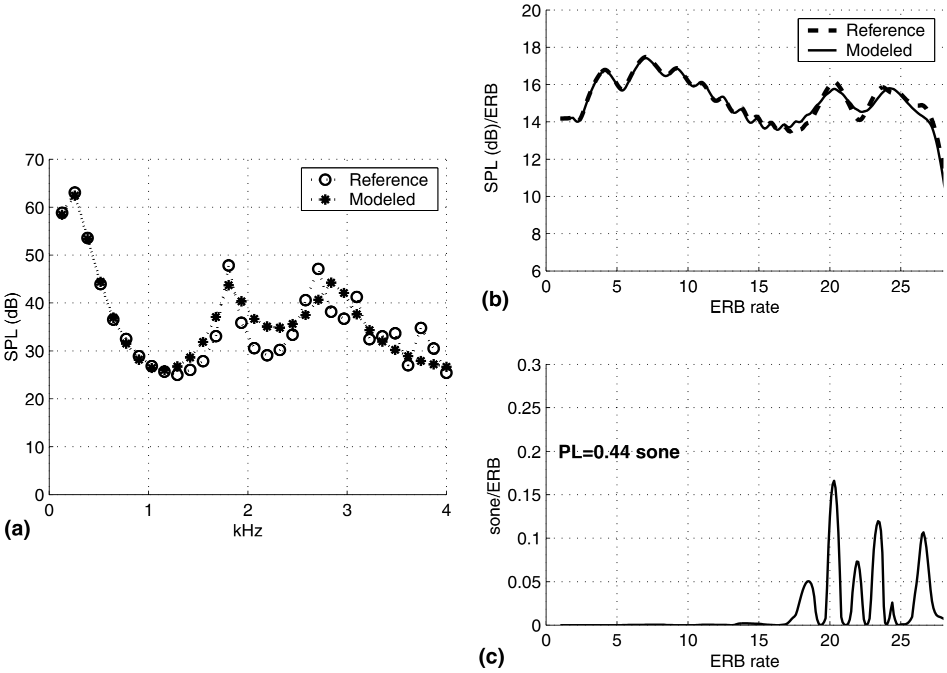

Fig. 7. Modeling of vowel [i], having pitch = 120 Hz, with Bark scale frequency warping at LP model order = 10. (a) Spectral amplitudes of reference (○) and modeled (⁎) sounds are connected to show the corresponding spectral envelopes. (b) Reference (--) and modeled (—) auditory excitation patterns. (c) Partial loudness distribution of the spectral distortion.

spectra synthesized with a single fixed set of source parameters, and will be extended to include speaker dependent effects in a subsequent work.

Summarising the contributions of this paper, the relative improvement in perceived quality due to perceptual frequency warping, prior to low order LP modeling of vowel sound spectral envelopes, was found to depend on phoneme quality. A consistent degradation in quality with Bark scale warping was noted in the case of vowels with widely spaced first and second formants. An auditory model based distance measure was found to closely predict the subjective judgements indicating that frequency masking plays an important role in determining the perceptual accuracy of spectral envelope modeling. Further, the experimental results presented in this paper support the validity of the partial loudness model of Moore et al. (1997), in predicting the perception of supra-threshold distortions of the type considered in the present study.

References

Champion, T., MacAulay, R., Quatieri, J., 1994. High-order all-pole modeling of the spectral envelope. In: Proc. IEEE Int. Conf. on Acoust, Speech, Signal Process., pp. 529–532.

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

Das, A., Gersho, A., 1995. Variable dimension spectral coding of speech at 2400 bps and below with phonetic classification. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 492–495.

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

Harma, A., Laine, U., 2001. A comparison of warped and conventional linear predictive coding. IEEE Trans. Speech Audio Process. 9 (5), 579–588.

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

Jaroudi, E., Makhoul, J., 1991. Discrete all-pole modeling. IEEE Trans. Signal Process. 39 (2), 411–423.

Jelinek, M., Adoul, J., 1999. Frequency domain spectral envelope estimation for low rate speech coding. In: Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process., pp. 253–256.

Koljonen, J., Karjalainen, M., 1984. Use of computational psychoacoustical models in speech processing: Coding and objective performance evaluation. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 191–194.

Kondoz, A., 1991. Chapter 8, Digital Speech: Coding for Low Bit Rate Communication Systems. John Wiley, New York.

Lukasiak, J., Burnett, I., Chicharo, J., Thomson, M., 2000. Linear prediction incorporating simultaneous masking. In: Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process., pp. 1471–1474.

MacAulay, R., Quatieri, T., 1995. Sinusoidal coding. In: Speech Coding and Synthesis. Elsevier, Amsterdam.

Makhoul, J., 1976. Spectral linear prediction: Properties and applications. IEEE Trans. Acoust., Speech, Signal Process. ASSP-23 (3), 283–296.

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

Molyneux, D., Ho, M., Cheetham, B., 2000. Robust application of discrete all pole modelling of sinusoidal transform coding. In: Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process., pp. 1455–1458.

Molyneux, D., Parris, C., Sun, X., Cheetham, B., 1998. Comparison of spectral estimation techniques for low bit-rate speech coding. In: Proc. Int. Conf. on Spoken Language Process., 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.

Murthi, M., Rao, B., 1997. Minimum variance distortion less response modeling of voiced speech. In: Proc. IEEE Int. Conf. on Acoust, Speech, Signal Process., pp. 1687–1690.

National Institute of Standards, 1989. The TIMIT cdrom.

Oppenheim, A., Johnson, D., Steiglitz, K., 1971. Computation of spectra with unequal resolution using fast Fourier transform. Proc. IEEE 59 (2), 299–301.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective Measure of Speech Quality. Prentice-Hall Inc.

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.

Rix, W., Beerends, J., Hollier, M., Hekstra, A., 2001. A new method for speech quality assessment of telephone networks and codecs. In: Proc. IEEE Int. Conf. on Acoust., Speech, Signal Process., pp. 749–752.

Schroeder, M., Atal, B., Hall, J., 1979. Objective measure of certain speech signal degradations based on masking properties of human auditory perception. In: Frontiers of Speech Communication Research. Academic, New York.

Smith, J., Abel, J., 1999. Bark and erb bilinear transform. IEEE Trans. Speech Audio Process. 7 (6), 697–708.

Stevens, K., 2004. Acoustic Phonetics. MIT Press.

Strube, H., 1980. Linear prediction on a warped frequency scale. J. Acoust. Soc. Am. 68 (4), 1071–1076.

Varho, S., Alku, P., 1998. Separated linear prediction—A new all-pole modelling technique for speech analysis. Speech Commun. 24, 111–121.

Wei, B., Gibson, J., 2000. Comparison of distance measures in discrete spectral modeling. In: Proc. IEEE Digital Signal Process. Workshop.

Zwicker, E., Fastl, H., 1974. Facts and Models in Hearing. Springer-Verlag, Berlin.