Proc. of the Int. Conf. on Spoken Language Technology (ICSLT-O-COCOSDA), 2004, New Delhi, India.

# Adaptive Frequency Warping for Improved Spectral Modeling

Pushkar Patwardhan and Preeti Rao
Dept. of Electrical Engineering, I.I.T. Bombay
*(e-mail: prao@ee.iitb.ac.in)*

## Abstract

*The compact representation of harmonic amplitudes in the sinusoidal coding of voiced speech is often achieved by the all-pole modeling of a spectral envelope. The perceptual accuracy of the representation may be enhanced by the use of frequency-scale warping prior to LP modeling. The effectiveness of a frequency warping function has been shown in previous work to depend on the underlying signal spectral shape as determined by phoneme quality as well as voice quality. This suggests the possibility of minimizing the perceived degradation due to LP modeling by adapting the warping function to the local signal spectrum. In this paper, an investigation is conducted to determine possible correlations between preferred warping condition and easily computable spectral cues. Experimental results are presented that demonstrate the advantage of adapting the warping function to the signal spectral envelope in the context of a sinusoidal speech coding scheme.*

## 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 speech coding has been the sinusoidal model, of which the Multiband Excitation (MBE) model is a prominent example. 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. 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. The quantization of harmonic amplitudes is most demanding on the bit allotment, and methods for efficient quantization have been an important topic of research. A widely used method for the efficient quantization of the amplitudes of the harmonics is based on the modeling of a spectral 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 many available quantization methods. 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, features such as overall spectral tilt and spectral zeros due to nasality typically require a relatively large number of poles for adequate approximation. Steady vowel sounds are especially sensitive to spectrum modeling errors. Also it has been observed that for similar spectral envelope, low-pitched sounds require higher LP model order for similar perceived quality levels (Champion et. al , 1994; Patwardhan and Rao , 2003). 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 perceptual accuracy for a given model order (Harma and Laine , 2001). It leads to the 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, recent experimental results on steady vowel sounds (Patwardhan and Rao , 2002) indicated that the performance of frequency warping depended to a great extent on the nature of the underlying sound spectrum. It was noted that front vowels such as [ε] and [i] were, in fact, best modeled without Bark-scale warping. This was explained by the low first formant structure of these vowels which fails to mask high frequency distortion adequately. Further studies (Patwardhan and Rao , 2003) revealed that these results were to an extent speaker dependent and were influenced by the voice quality as reflected in the spectral slope. Dark voices with their steep spectral slopes were found to benefit from frequency-scale warping but not so the relatively bright voices which actually degraded. Comparison with the predictions of perceived degradation by an auditory-model based objective measure suggested that the absence of dominating low frequency components could result in the inadequate masking of distortion in the coarsely modeled high frequency spectral region.

Based on the above observations, it is of interest to investigate the possibility of adapting the warping function to the spectral characteristics of the sound to be modeled. In the context of speech coding, this would necessitate transmission to the receiver of information regarding the warping function. In this work we consider the option of switching between Bark-scale warping and linear-scale warping (i.e. no frequency warping). This gives rise to only a single-bit overhead in coding. An important issue raised by this approach to the perceptual optimization of LP modeling is that of finding reliable spectral cues to guide the selection of the warping function. In this paper, we compute a measure of spectral slope and experimentally investigate its correlation with subjective judgement of preferred warping function across a set of synthetic and natural vowels.

In the next section we provide details of the implementation of frequency-warped LP modeling in MBE speech coding.

## 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 by the method depicted by Fig. 1. 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). 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 (Patwardhan and Rao , 2002). 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. Uniform 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 speech coding of narrowband speech (sampling frequency = 8 kHz) 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; Patwardhan and Rao , 2002) although the results are applicable more generally. 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 (Griffin and Lim , 1988). 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 phase of harmonics is not transmitted but predicted during synthesis in most low rate coders. 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 can be 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 specifically the performance of frequency-scale warping in obtaining improved quality at low LP model order. Based on the observation that speech quality improves only slowly as the LP model order is increased beyond 10, the LP model order selected for the experiments in this work is 10.

## 3. Voice quality and its spectral correlates

The overall spectral envelope for voiced speech is determined by three factors: (1) the spectrum of the glottal excitation pulse, (2) the vocal tract transfer function, (3) the radiation at the lips and nostrils. From the discussion of Sec. 1, we are interested in spectral cues that reflect the dominance of the lower harmonics in the spectrum or the roll-off of the spectrum. An important determiner of the slope of the spectrum is the shape of the glottal excitation pulse. An abrupt closure of the vocal folds 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 speech spectrum is the “brightness" of the voice. At a given fundamental

frequency, “dark" voices have relatively weak high harmonics while “bright" voices are characterized by flatter spectra.

A commonly used acoustic measurement to describe source spectral tilt is H1-A3 (Hanson , 1997). This is the ratio of amplitude of first harmonic (H1) relative to that of the third-formant spectral peak (A3) in the speech signal spectrum. An abrupt closure in the glottal cycle results in relatively strong middle and high frequency components leading to lower H1-A3 and a perceptually brighter voice.

## 4. Experiments

Experiments were designed to investigate the influence of voice quality on the modeling error from frequency-warped LP modeling of the spectral envelope. Listening tests were carried out to obtain subjective preferences on warping function for sets of isolated vowels as well as sentence utterances.

### 4.1 Isolated vowels

Natural and synthetic samples corresponding to eight distinct vowels as in the words {“guard” – (a), “cat”-(ae), “law”-(ow), “cut”-(uh), “boat”-(oe), “boot”-(oo), “beet”-(iy) and “wait”-(e)} were obtained for the experiment. Synthesis of vowels allowed the manipulation of the glottal timing parameters which in turn enabled control over the spectral slope and voice quality. The synthetic vowels were generated by articulatory synthesis with the articulation parameters estimated from 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. 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 the Table 1. These correspond to four different perceived voice qualities ranging from dark to bright. Four instances of each vowel corresponding to each set of glottal parameters were generated. The vowel sounds had 120 Hz pitch and a duration of 350 ms. The start and end of the vowel was tapered to avoid abrupt transitions at the boundaries. The set of natural vowels consisted of 4 instances of each vowel (Pitch 80 Hz to 120 Hz) taken from slowly uttered words of several speakers. The sounds were selected such that they corresponded to various degrees of “brightness” as judged by listening as well as visual examination of the spectral tilt. The duration of the sounds ranged between 350 ms to 600 ms. In all, the set of natural vowels consisted of 32 vowels (4 instances each of 8 distinct vowels).

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 $10^{th}$ order frequency-warped LP modeling with chosen frequency warping factor as detailed in Section 2. The synthesis was carried out using the spectral amplitudes obtained by frequency warped all-pole model approximation and compared with a reference sound synthesized using originally estimated spectral amplitudes. There were two test sounds for each reference sound: LP modeled sound without frequency warping (“U”) and LP modeled with Bark-scale warping (“B”).

A subjective listening experiment was set up to compare the perceived qualities of the LP modeling with different warping functions. Six normal hearing listeners participated in the test. The test material was presented to the subject at normal listening levels through high quality headphones connected to a PC sound card in a quiet room. The subjects were asked to rank the relative perceived degradations of the test sounds U and B with respect to the corresponding reference sound for each of the vowel sounds in test set. Subjects were allowed to listen to the reference and test sounds any number of times before making a decision. Each listener did the test using the same set of items in different order 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 test vowel by combining the numerical ranks across listeners and trials.

### 4.2 Sentence level test

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 2 lists the 15 sentences along with an indication of the dominant phonetic content as well as perceived brightness of the voice. 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 synthesised from $10^{th}$ order LP modeled spectral amplitudes with and without Bark-scale warping to obtain B and U versions respectively. 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.

## 5. Results and Discussion

### 5.1 Isolated vowels experiment

Fig. 2 (a) and Fig. 2 (b) show for each of the natural and synthetic sounds used in the experiments, the subjectively preferred warping condition (between non-warped and Bark-warped) in relation to the spectral slope characteristic of the sound. The latter is measured by H1-A3 for a representative frame picked from the centre of each steady vowel sound. (In all the samples under study, the fundamental frequency was low enough and the formants well separated so that estimating A3 was straightforward.) The bright sounding vowels fall at the lower end of the H1-A3 axis relative to the corresponding dark vowels. The vowel instances that have been ranked better with Bark-scale frequency warped all-pole modeling are depicted by triangles while the ones without warping before modeling are depicted by squares.

From Fig. 2 (a) for the natural vowels, we observe a clear correlation between the value of H1-A3 and the subjectively preferred warping function. Bark-scale warping improves the perceptual accuracy of the modeling in the case of the dark-sounding vowels with their relatively high H1-A3. Notable exceptions to this general behaviour are the vowels [ε] (/e/) and [i] (/iy/). These vowels are best modeled without frequency warping irrespective of individual spectral slope characteristics. This may be explained by the absence of significant components in the mid-frequency region due to the widely separated first and second formants of these vowels. The results from the synthetic vowels experiment in Fig. 2 (b) follow roughly the same pattern as that of the natural vowels but with apparently less consistency.

The isolated vowel experiment was also extended to nasalised vowels but the results were not as clear as in the case of non-nasalised natural vowels. Since a prominent feature of nasal vowels lies in the low frequency spectrum, it would be expected that Bark-scale warping before the LP modeling would help to reproduce more accurately the perceived nasality. However when asked to rank overall quality with and without Bark-scale warping before LP modeling, subjects often preferred modeling without warping over 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 compared with the results on the non-nasalised vowels. This may be attributed to the presence of conflicting cues on the perceived distortion arising from 2 different sources, namely, loss of nasality and high-frequency modeling error.

### 5.2 Sentence level test

Table 2 shows the preferred warping condition in terms of overall subjective quality for each of the 15 sentences in the listening experiment. We observe that in the case of the dark-voiced sentences, subjects preferred the Bark-warped condition except when the phonemes [ε] and [i] were predominant. In the case of the bright voices, the subjects preferred modeling without frequency warping in all cases. These observations are consistent with the isolated vowel results of Sec. 5.1.

An additional experiment (not listed in Table 2) was based on adapting the warping factor depending on the spectral characteristics of each 20 ms speech frame. For this experiment, we considered the sentences 6 and 9 of Table 2. Bark-scale warping was applied throughout but manually disabled in segments which corresponded to the phones [ε] or [i]. Subjective listening tests revealed that this "switched" warping resulted in better quality over that achieved by either of the fixed warping conditions. 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.

## 6. Conclusion

The eventual goal of the work reported in this paper is the design of a spectral envelope modeling scheme that adapts the frequency-warping function to the spectral characteristics of the speech signal frame in order to maximize the perceptual accuracy of the modeling. Steady vowel sounds are particularly sensitive to spectrum envelope modeling errors and, of these, low-pitched voices are the most difficult to model at moderately low orders in the context of sinusoidal coding. The important spectral characteristics for selecting the warping function, as they emerge from the reported experimental results, are the vowel quality and the voice quality as measured by the spectral slope. While precise methods for determining vowel quality

(or at least identifying the front vowels [ε] and [i]) still need to be formulated, H1-A3 has been shown to be a suitable indicator of the spectral slope. Worthy of further investigation are simple and reliable alternatives to H1-A3 that do not involve estimating formants but still represent the extent of prominence of low frequency harmonics.

Finally, as is well known, the incorporation of pre-emphasis would serve to further improve the perceptual accuracy of LP-modeling. Our own studies have shown that while pre-emphasis alone (with a high pre-emphasis factor of > 0.9) greatly improves the modeling of voices with high spectral tilt (such as breathy female voices), a high pre-emphasis factor typically results in degraded modeling of bright voices as well as voices characterized by low H1-H2 (corresponding to a glottal excitation waveform with a low open quotient (Doval and d'Alessandro , 1997)). On the other hand, a mild pre-emphasis factor in combination with appropriate frequency-scale warping could serve well in achieving perceptually more accurate all-pole modeling of the spectral envelope.

## 7. References


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

Childers, D., 2000. Speech processing and synthesis toolboxes. John Wiley and Sons Inc.

Doval, B., d'Alessandro, C., 1997. Spectral correlates of glottal waveform models: An analytic study. In: Proc. of the IEEE Int. Conf. on Acoust, Speech, Signal Processing, 1295-1298.

Griffin, D., Lim, J., 1988. Multiband excitation vocoder. IEEE Trans. Acoustics, Speech and Signal Processing 36 (8), 1223-1235.

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

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

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

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

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

Patwardhan, P., Rao, P., 2002. Controlling distortion in spectral envelope modeling. In: Proc. of the Int. Conf. on Spoken Language Processing, 2002, Denver.

Patwardhan, P., Rao, P., 2003. Frequency warped all-pole modeling of vowel spectra: Dependence on voice and vowel quality. In: Workshop on Spoken Langauge Processing, 2003, Bombay.


| | Timing parameters as % of $t_o$ (83ms) | | |
|---|---|---|---|
| Voice Quality | $t_p$ | $t_a$ | $t_e$ |
| 1 | 35% | 50% | 15% |
| 2 | 40% | 50% | 15% |
| 3 | 40% | 50% | 2% |
| 4 | 45% | 50% | 2% |

**Table 1. Timing parameters used to generate synthetic vowels with 4 voice qualities.**

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

**Table 2. Description of sentences used in the subjective ranking of frequency warping.**

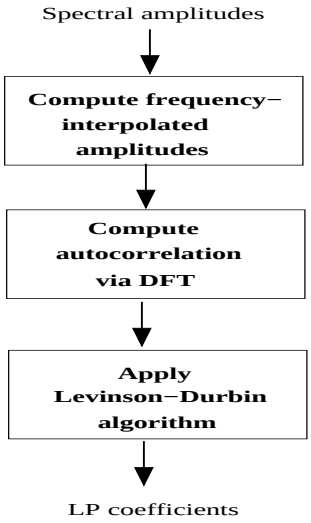


**Fig. 1 Steps in the frequency domain approach to all-pole modeling**

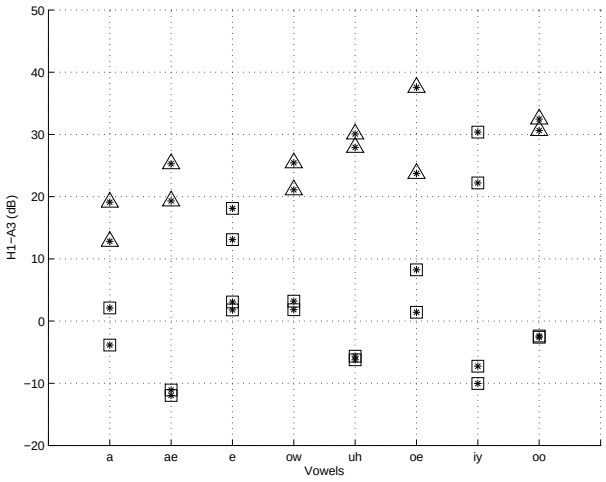


**Fig. 2(a) Spectral cue H1-A3 and subjectively preferred warping function for various natural vowels. Δ-Bark-scale warping, □- no warping.**

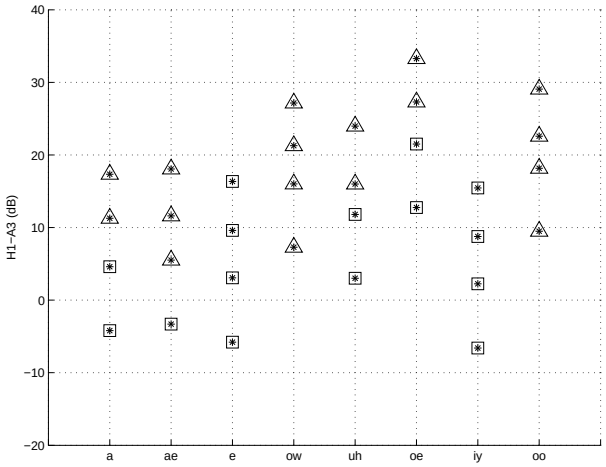


**Fig. 2(b) Spectral cue H1-A3 and subjectively preferred warping function for various synthetic vowels. Δ-Bark-scale warping, □- no warping.**