CONTROLLING PERCEIVED DEGRADATION IN SPECTRUM ENVELOPE MODELING VIA PREDISTORTION

Pushkar Patwardhan and Preeti Rao

Department of Electrical Engineering,
Indian Institute of Technology, Bombay, India 400076
Email: prao@ee.iitb.ac.in

ABSTRACT

The compact representation of the discrete amplitude spectrum of voiced speech by an all-pole model of the spectral envelope is considered. Based on the properties of the all-pole modeling error, the use of spectrum predistortion for improving the perceptual fit at low model orders is motivated. Warping of the frequency scale before modeling of the spectral envelope of narrowband voiced sounds is investigated by subjective listening and objective measures. It is found that, contrary to what is generally accepted, the improvement in perceived quality brought about by frequency warping actually depends to a large extent on the underlying signal spectrum distribution. An objective distance measure based on partial noise loudness is found to show high correlation with subjective judgements of degradation, indicating that auditory frequency masking plays an important role in determining the perceptual accuracy of the spectrum envelope model.

1. INTRODUCTION

The problem of representing the envelope of a discrete spectrum is common to many applications in sound synthesis and coding. While the compactness of the representation is of primary importance in speech compression, concatenative speech synthesis systems such as [1] require, in addition, the flexibility for pitch modification. In the case of steady sounds with discrete harmonic spectra, which describes the large class of voiced speech sounds, the timbre or quality of the sound is a direct perceptual correlate of the harmonic spectral amplitudes [2]. The perceptually accurate representation of the discrete spectral amplitudes is therefore crucial to the performance of coding and synthesis systems. A popular approach to obtaining a compact spectral representation has been to fit a smooth "continuous" envelope to the set of harmonic spectral magnitudes, and then model and parameterise this envelope. The required amplitudes for synthesis are recovered by sampling the parameterised continuous spectral envelope at the harmonic frequencies. All-pole modeling has commonly been used to achieve compact and easily quantisable representations of the spectral envelope [3]. The spectral envelope modeling stage, based on LP (linear predictive) modeling, introduces errors in the spectral amplitudes. Typically, high all-pole model orders are required to represent sounds in a perceptually accurate way [4]. In speech compression applications, it is particularly important to control the perceived distortion in the reconstructed speech while keeping the model order as low as possible. Modifying the model fitting procedure is not straightforward. Rather, the alternative approach of introducing a controlled distortion in the spectral envelope *prior* to modeling, such that the model order required for the perceptually accurate encoding of the distorted speech spectrum is lowered, is attractive. At the synthesis stage, the spectrum envelope is restored by the inverse of the predistortion operation. Frequency and amplitude pre-warping of the spectrum have been used frequently in the past to shape the "noise" due to modeling as well as that due to quantisation. Strube [5] noted that warping the frequency scale according to the auditory Bark scale improved speech intelligibility and quality in vocoders for low model orders. In [6] speech phonemes were coded using regular LPC as well as Bark-scale warped LPC to demonstrate the superiority of the latter. More recently, the performance of frequency-warped LP modeling was compared with regular LP modeling for a variety of audio signals of various bandwidths and model orders, to conclude that Bark-scale warping led to improved perceived quality [7]. A reduction of the required LP model order by the use of frequency warping has also been reported in [4]. Similarly, spectral amplitude warping has been used obtain a perceptually improved fit of the LP model of a given order to the spectral envelope [8].

The focus of this paper is the in-depth investigation of the performance of frequency-scale warping for the representation of harmonic spectral amplitudes of narrowband voiced phonemes using subjective and objective measurement of perceptual degradation. In the next section, we briefly review LP modeling of the envelope of discrete amplitude spectra, and motivate the use of predistortion of the spectral envelope. The implementation of frequency-scale warping is described, followed by subjective and objective evaluation of different warping functions for a set of test sounds.

2. LP MODELING OF THE SPECTRAL ENVELOPE

The LP modeling of the spectral envelope is carried out by first deriving a continuous (or, more precisely, densely sampled) spectral envelope from the discrete spectral amplitudes by interpolation at frequency samples between the harmonics. Linear interpolation of the log magnitude spectrum as proposed in [8] has been found to provide good performance. An autocorrelation function is obtained from the resulting discrete power spectrum by an inverse Fourier transform [4],[8]. The LPCs for a desired model order are next obtained from the autocorrelation function using the Levinson-Durbin algorithm.

This work has been supported in part by a research grant from D.R.D.O., Govt. of India.

2.1. Properties of the modeling error

The error criterion, i.e. minimization of integral of original to approximated all pole spectrum, used in LP modeling [9] ensures that (i) the LP modeled spectrum envelope approximates the input spectrum uniformly across the linear frequency scale, and (ii) the local peaks are approximated more accurately than the valleys of the spectral envelope. It turns out that these properties of the LP spectrum modeling error may not achieve the perceptually most desirable spectral fit for a given model order. This can be justified from what is known about human auditory perception. Human hearing has a resolution that decreases with frequency. Also simultaneous frequency masking plays a role in the perception of distortion at various amplitude levels. These aspects of hearing typically lead to the spectrum modeling error in the lower frequency region being perceptually worse, and to the inherently larger errors in low amplitude regions (valleys) becoming even more prominent due to inadequate masking. Pre-distortion of the spectrum by explicitly incorporating transformations which resemble those going on in the auditory periphery can serve to redistribute the error in a perceptually more uniform manner. Frequency-scale warping based on the Bark-scale has been advocated by many to obtain a more appropriate shaping of the modeling error across frequency. Similarly, amplitude warping or compression has been used to achieve the more equal distribution of error at high and low amplitudes [8].

2.2. Frequency-scale warped LP modeling

The approximated spectrum, $\hat{S}(.)$, for LP order "p" is computed on a warped frequency scale, θ, and given by

$$\hat{S}(\theta) = \frac{G}{1 + \sum_{k=1}^{p} a_k e^{-j\theta k}} \tag{1}$$

where $\theta = f(\omega)$ for input frequency ω. The expression for the warping function is obtained from [10] as

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

The parameter α controls the extent of frequency warping: $\alpha = 0.47$ corresponds approximately to the auditory Bark scale frequency warping at 10 KHz sampling rate [5]. For the purpose of our investigation on narrowband speech sampled at 8 kHz, we choose 3 values for the warping parameter. These correspond to regular ($\alpha = 0.0$), a mild version of Bark-scale warping ($\alpha = 0.2$) and Bark-scale warping ($\alpha = 0.45$). The frequency mapping achieved by each of α's is shown in Figure 1. We see that increasing α leads to increasing non-linearity and a correspondingly higher degree of compression of the frequency scale in the high frequency region.

Spectral amplitudes obtained from analysis of the input speech signal are mapped to warped frequency locations by equation (2). The resulting set of non-uniformly spaced spectral amplitudes is log linearly interpolated to a fixed 20 Hz interval to get a good approximation to the continuous enveloipe. Next the autocorrelation function is obtained from the discrete power spectrum and used to compute the LPCs. Due to the greater number of interpolated spectral samples between the original low frequency harmonics compared with the high frequency harmonics, a correspondingly higher weighting of the low frequency region error is obtained in the subsequent LP model fitting.

Item ID	Vowel Sym.	Typ. word	Pitch (Hz)	Formant Frequencies F1 (Hz)	F2 (Hz)	F3 (Hz)	F4 (Hz)	LP ord.
1	E	bet	130	483	2043	2200	2700	10
2	OW	bought	117	558	985	2345	2928	10
3	UH	but	119	512	1200	2373	2877	10
4	A	hot	115	666	1206	2363	3057	10
5	AE	bat	110	624	1162	1777	3000	10
6	IY	beet	121	461	2090	2606	3361	12
7	IY	beet	100	378	1882	2436	3069	12
8	OO	boot	98	448	1622	2248	2952	12
9	OO	boot	113	375	1716	2544	3372	12
10	E	bet	155	544	1606	2075	2726	8
11	OW	bought	145	540	883	1875	2791	8
12	A	hot	153	613	1100	2019	2779	8
13	UH	but	149	600	1063	1821	2844	8
14	AE	bat	143	756	1285	1888	2719	8

Table 1. Description of vowel sounds used in the subjective listening experiment

Item ID	Subjective Ranking			Spectral Distortion			MFCC Distance			Partial Loudness		
	U	M	B	U	M	B	U	M	B	U	M	B
1	1	2	3	1	2	3	1	2	3	1	2	3
2	3	1	2	1	2	3	2	1	3	3	1	2
3	3	1	2	1	2	3	1	2	3	2	1	3
4	3	1	2	1	2	3	2	1	3	3	1	2
5	3	1	2	1	2	3	2	1	3	3	1	2
6	1	2	3	1	2	3	1	2	3	2	1	3
7	1	2	3	1	2	3	1	2	3	1	2	3
8	1	2	3	1	2	3	1	2	3	1	2	3
9	1	2	3	1	2	3	1	2	3	1	2	3
10	3	1	2	1	2	3	2	1	3	3	1	2
11	3	2	1	1	2	3	2	1	3	3	1	2
12	3	2	1	1	2	3	1	2	3	3	1	2
13	3	2	1	1	2	3	1	2	3	3	2	1
14	3	1	2	1	2	3	2	1	3	3	1	2

Table 2. Ranking of perceived similarity of the modeled sound to the reference sound for each warping condition. (1≡ best, 3≡worst)

3. SUBJECTIVE AND OBJECTIVE EVALUATION

Representative voiced phonemes were extracted from the vowel portions of single, slowly uttered words recorded from a number of speakers. The phoneme sound was extracted from each word manually and the ends were tapered to eliminate abrupt transitions. The duration of the sounds ranged between 200 and 700 ms. Table 1 describes the test set. The sounds were analysed to estimate the pitch and spectral amplitudes of frames of 20 ms duration by an analysis-by-synthesis method applied to the DFT spectrum [11]. LP coefficients of a chosen order (roughly increasing with decreasing fundamental frequency to account for the higher number of spectral amplitudes for low pitched voices) were computed for each frame using frequency warped LP modeling with the chosen warping factor. Synthesis was carried out by standard sinusoidal synthesis methods [11] using the spectral amplitudes obtained from the all-pole model approximation and compared with a reference sound synthesized using the originally 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 ("M") and LP modeled with Bark warping ("B").

3.1. Subjective listening test

A subjective listening test was conducted with three normal hearing subjects, where listeners were asked to rank the *relative* perceived degradations of the test sounds U, M and B with respect to the corresponding reference sound for each of the 14 phonemes of Table 1. 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. Each listener did the test using the same set of items in different ordering on three separate occasions. Although no instructions whatever were given to the listeners on the type of degradation to listen for, it was observed by them that the distortions due to modeling inaccuracies are characterized by changes in both, the intelligibility (clearness) and the "color" (brightness) of sound. It was noted that the subjects showed a high degree of consistency across trials, and also that there was largely agreement between subjects although the rankings themselves varied from item to item. Table 2 shows the result of the subjective test as rankings based on suitably averaged scores across trials and subjects for the phonemes of Table 1. We observe that several phonemes *degrade* under Bark-scale warped LP spectral modeling as compared with regular LP modeling.

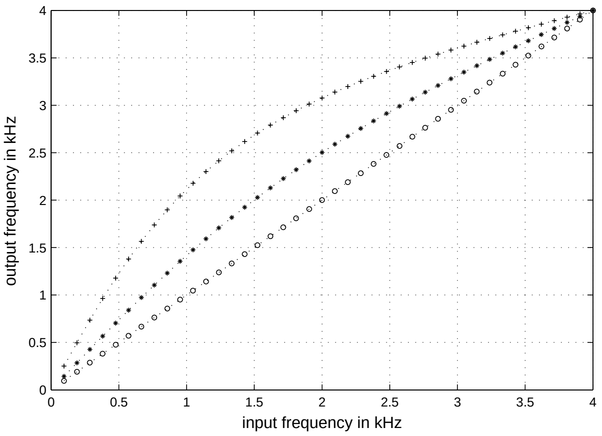

Fig. 1. Warping functions for different values of warping parameter. The symbols '$+$','$*$' and 'o' indicate warping functions for $\alpha = 0.45$, $\alpha = 0.2$ and $\alpha = 0.0$ respectively.

3.2. Objective measurement

In order to obtain an insight into the dependence of perceived quality on the spectral modifications obtained by frequency warped all-pole modeling, an objective estimation of the perceived distortion was attempted. Since there is no universally accepted objective measure for speech degradation, a number of different measures were computed. These included the traditional speech coding measure: log spectral distortion (SD)[3], and MFCC (Mel Frequency Cepstral Coefficients) distance which has been widely used by speech recognition and synthesis community over the years. A third measure used was the (psychoacoustically motivated) partially masked loudness. Partial loudness (PL) was first proposed to measure speech quality degradation due to quantization in [12]. The spectral distortion is treated as the signal whose audible significance is to be estimated in the presence of a background masker (the reference sound). Recently a computational model of partial loudness was developed that accounts for a large body of subjective data from psychoacoustical experiments [13]. The model was shown to perform well in the prediction of audible discrimination of spectral envelope distortions in vowel sounds as measured in psychoacoustical experiments [14]. The rankings of relative degradation based on the objective measures, were derived by computing the distortion measure for each of the reference-test(U/M/B) sound pairs. The ranks obtained by the objective measures are shown in Table 2. The performance of an objective distance 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 [15] is a suitable measure since it makes minimal assumptions about the data. Applying the Spearman's correlation coefficient to the results of Table 2, we find that PL shows a positive correlation (equal to either 1.0 or 0.5) for all the test items, while MFCC distance and SD show positive correlations in only 71% and 36% of the cases respectively.

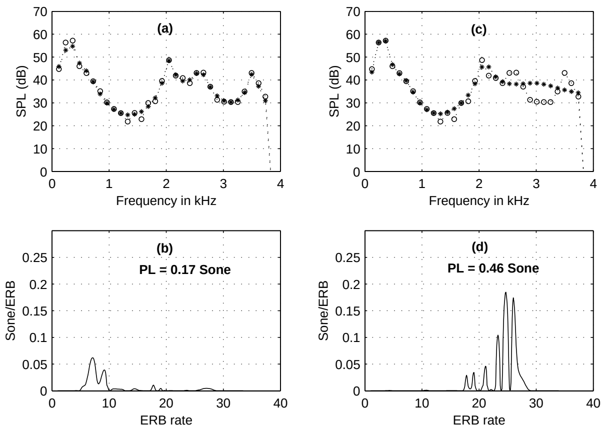

Fig. 2. Effects of regular and Bark-warped LP spectral modeling with model order = 12 on a frame of estimated spectral amplitudes of sound /iy/, pitch = 121 Hz. [**a**]: Original 'o' and LP modeled '$*$' spectral amplitudes; [**b**]: PL distribution for the spectral distortion of (a); [**c**]: Original 'o' and Bark scale warped LP modeled '$*$' spectral amplitudes; [**d**]: PL distribution for the spectral distortion of (c).

4. DISCUSSION AND CONCLUSIONS

The results of the subjective experiment clearly demonstrate that the effect of frequency warping on the perceptual accuracy of the modeled spectral envelope depends to a large extent on the characteristics of the underlying sound. Contrary to what is generally accepted, frequency warping according to the Bark scale does not universally improve perceived quality. We see that a single warping function cannot be universally applied for all types of sounds. Overall, mild-Bark warping (α=0.2) appears to perform the best.

Bark-scale frequency warping leads to largely accurate modeling in the low frequency region at the cost of modeling errors in the high frequency region. Whether the modeling inaccuracies will be audibly significant or not is expected to depend upon the spectral distribution or formant locations of the phoneme. It was found that the PL measure correlates well with subjective judgements of degradation. This may be used to obtain a better understanding of the perceived distortion. For example, Figure 2 illustrates how the estimated and modeled amplitudes for the phoneme /iy/ translate into a partial loudness distribution on the auditory ERB-rate scale [13]. We see that Bark-scale warped LP provides a superior match to the original spectral amplitudes in the region of the first formant (indicated by insignificant contribution to the PL at low ERB-rate) but a very poor fit in the upper half of the frequency spectrum. This results in an overall degradation (high PL) due to the fact that the phoneme /iy/ is characterized by a very low frequency first formant and therefore displays an upward spread of masking that is not adequate to mask the gross errors due to the poor spectral fit of the Bark-warped LP model at high frequencies. We have found that the "high" vowels /oo/ and /iy/ typically degrade under Bark-scale warped modeling; this is also amply borne out at sentence-level listening (e.g. "She seemed irritated"). The results thus indicate that auditory frequency masking plays an important role in the perceptual accuracy of the spectrum envelope model. Sound samples demonstrating some of the presented results may be found at [16].

Work is currently underway to evaluate the performance of amplitude warping in a similar manner. As in the case of frequency warping, it is expected that whether spectral modeling will be perceptually improved, will depend to a great extent on the actual spectrum and to what extent modeling errors are masked by existing spectral features. Preliminary results indicate that for spectral envelopes that show a steep roll-off of amplitude level (e.g. for many female voices), a low degree of amplitude warping can lead to a reduction of the spectral modeling error at weaker harmonics leading to improved perceived quality.

5. REFERENCES

[1] Dutoit T. and Leich H., "MBR-PSOLA: Text-to-speech synthesis based on a MBE re-synthesis of the segments database," *Speech Communication*, vol. 13, no. 3-4, Nov 1993.

[2] Plomp R., *Aspects of Tone Sensation*, Academic London, 1976.

[3] Smith A.M., Rambadran T., and McLaughlin M.J., "Modeling and quantization of speech magnitude spectra at low data rates-evaluating design trade-offs," in *Proc. IEEE Workshop on Speech Coding for Telecommunications*, Sep 1997.

[4] MacAulay R.J. and Quatieri T.F., "Sinusoidal coding," in Speech Coding and Synthesis, Elsevier, Amsterdam, 1995.

[5] Strube H. W., "Linear prediction on a warped frequency scale," *J. Acoust. Soc. of Am*, vol. 68, no. 4, pp. 1071–1076, Oct 1980.

[6] Koljonen J. and Kajalainen M., "Use of computational psychoacoustical models in speech processing: Coding and objective performance evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1984.

[7] Harma Aki and Laine U.K., "A comparison of warped and conventional linear predictive coding," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 5, pp. 579–588, Jul 2001.

[8] Hermansky H., Hanson B.A., Wakita H., and Fujisakim H., "Linear predictive modeling of speech in modified spectral domain," in *Digital Processing of Signals in Communications*, Apr 1985, pp. 55–63.

[9] Makhoul J., "Spectral linear prediction: Properties and applications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-23, no. 3, pp. 283–296, Jun 1976.

[10] Oppenheim A.V., Johnson D.H., and Steiglitz K., "Computation of spectra with unequal resolution using fast Fourier transform," *Proc. of IEEE*, vol. 59, no. 2, pp. 299–301, Feb 1971.

[11] Griffin D.W.and Lim J.S, "Multiband excitation vocoder," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, Aug 1988.

[12] Schroeder M.R., Atal B.S., and Hall J.L, "Objective measure of certain speech signal degradations based on masking properties of human auditory perception," in Frontiers of Speech Communication Research, Academic NewYork, 1979.

[13] Moore B.C.J., Glasberg B.R., and Baer T., "Model for prediction of thresholds, loudness and partial loudness," *J. Audio Eng. Soc.*, vol. 45, no. 4, pp. 224–240, 1997.

[14] Rao P., van Dinther R., Veldhius R., and Kohlrausch A, "A measure for predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds," *J. Acoust. Soc. Am*, vol. 109, no. 4, pp. 2085 –2097, Apr 2001.

[15] Jane Miller, "Correlation," in Statistics for Advanced Level , Cambridge University Press, 1989.

[16] Patwardhan P. and Rao P., "ICSLP-02 demo. samples," http://www.ee.iitb.ac.in/˜prao/speechcoding/icslp02 /icslp02res.htm, Mar 2002.