A NEW APPROACH TO MODELING EXCITATION IN VERY LOW-RATE SPEECH CODING

S. Ghaemmaghami and M. Deriche

Signal Processing Research Centre
Queensland University of Technology
2 George st., Brisbane, Q 4001, Australia
shahrokh@markov.eese.qut.edu.au m.deriche@qut.edu.au

ABSTRACT

A new method for two-band approximation of excitation signals in an LPC model, to improve speech naturalness in very low rate coding, is proposed. Based on a simplified model of Multi-Band Excitation, the method accurately determines the degree of periodicity, using the concept of Instantaneous Frequency (IF) estimation in frequency domain. The harmonic structure in the spectrum of LPC residual, within individual bands, is identified based on flatness of the IF as a criterion for pitch and voicing detection. On this basis, the excitation is modelled by combining a predefined periodic signal in the lower band and a random signal in the higher band. It is shown that this improves considerably the naturalness of reconstructed speech in very low rate coding in comparison with that obtained using traditional binary excitation [1]. The performance of the technique is also given in Temporal Decomposition (TD) based coding at 800 b/s.

1. INTRODUCTION

To improve reconstructed speech quality in very low rate coding, different methods have been proposed in the literature [4], mostly working at rates above 2 kb/s. However, it is found that simple versions of *Multi-Band Excitation* (MBE) outperforms *binary* model and improves the speech quality at low rates [1,3]. This is achieved through modeling speech spectrum within a number of individual bands in MBE coder, while it is roughly approximated for entire frequency band in the binary model [3].

Aiming at further simplifying MBE model, It was reported in [3] that only about 6% of speech frames could possess more than four different voicing bands while 70% could be described by just two bands, where the lower band was almost always voiced. These findings led to more saving in the number of bits with near natural voice quality at rates about 2.4 kb/s [1]. However, at rates below 2 kb/s, even typical two band models fail, as these need more than 800 b/s to encode excitation parameters [3].

In this paper, we extend our previous work in this area [6] and propose a new MBE-based model which significantly enhance the synthesized speech naturalness, with respect to that in binary model, at rates where typical MBE models are inapplicable. This is achieved through a novel approach to pitch and voicing detection using the concept of *Instantaneous Frequency* (IF) estimation. The performance of the model is given in *Temporal Decomposition* (TD) based coding, where about 300 b/s are used for the excitation signals.

2. PROPOSED METHOD

In the method we propose, the excitation applied to the LPC filter for each frame is modelled by a mixed signal characterized by a spectrum composed of two overlapping distinct bands, which is periodic in the lower band and random in the higher band. Accordingly, spectral characteristics of the signal change gradually at *transition* frequency specified by a time varying parameter. This parameter is computed from the spectrum of the residual signal in an LPC model using a novel *periodicity* measure, described in the following.

The *residual* signal, $e(n)$, is first calculated through inverse filtering in an LPC model, for each frame of speech, as:

$$E(k) = A(k)S(k), \qquad k = 1, 2, ..., N \tag{1}$$

where N is total number of samples of each frame of speech, $S(k)$ and $E(k)$ are DFT (Discrete Fourier Transform) coefficients of $s(n)$ and $e(n)$, $n = 1, 2, ..., N$, respectively, k is frequency index, and $A(k)$ represents the DFT coefficients of the inverse filter.

The DFT of residual signal, $E(k)$, $k = 1, 2, ..., N$, is then reduced to $E_1(k)$, which is given as:

$$E_1(k) = |E(k)|, \qquad k = 1, 2, ..., N/2 \qquad (N\ even) \tag{2}$$

As $e(n)$ is real, only the phase information is lost with such reduction.

Then, $E_1(k)$, $k = 1, 2, ..., N/2$, is *filtered* using a window function in time domain with low phase distortion. This is indeed a replica of band-pass filtering in frequency domain, which is expressed as:

$$E_2(k) = E_1(k) * W_t(k), \qquad k = 1, 2, ..., N/2 \tag{3}$$

where $W_t(k)$ represents the DFT coefficients of the window, $E_2(k)$ shows the windowed signal in frequency domain, and asterisk stands for convolution.

As evident from equation (3), the input-output relationship of the window appears in multiplication form in time domain:

$$e_2(n) = e_1(n)\, w_t(n), \qquad n = 1, 2, ..., N/2 \tag{4}$$

where $w_t(n)$ is treated as the time window which is a fourth order band-pass *Butterworth* function. This *windowing* procedure reduces the effect of formants on the pitch harmonics and smoothes the spectrum as well.

The lower and upper 3-dB attenuated points of the window function are obtained from the range for the fundamental frequencies expected. For a speaker independent system, this range is 70-450 Hz. Accordingly, m_1 and m_2, normalized 3-dB points of the window, are obtained from:

$$m_1 = \frac{F_s}{70N}, \qquad m_2 = \frac{F_s}{450N} \tag{5}$$

where F_s is the sampling frequency in Hz and m_1 and m_2 are expressed as normalized number of samples in time domain.

The technique could better be clarified if we take $E_1(k)$ as a function of time which is to be filtered. We need to inspect spectrum of $E_1(k)$, $DFT\{E_1(k)\}$, within a certain range where fundamental frequency is expected. The filter used here, picks those components of $DFT\{E_1(k)\}$ which are likely pitch harmonics. For the case of periodic speech, $E_1(k)$ shows also periodicity in frequency domain as the *fine structure* of speech spectrum. It is indeed this periodicity in frequency domain which is considered by the band-pass filter, and is checked for the number of variations in the space designated by $DFT\{E_1(k)\}$, in which m_1 and m_2 are also described. As $DFT\{E_1(k)\}$ is indeed a time function, the filter characteristics is to be specified in time domain.

$E_2(k)$ is then considered as the signal spectrum to search for pitch harmonics using the concept of IF estimation. To do this, a typical IF estimation technique is applied to $E_2(k)$. For a voiced frame, the fundamental frequency and a number of its harmonics appear as equally spaced peaks in $E_2(k)$ which provide the dominant frequency for the variations in $E_2(k)$. This dominant frequency, representing the pitch period, is detected through IF estimation.

For IF estimation, we use *spectrogram* technique which employs a segment-based analysis using an appropriate window [2]. Here, the windowing is performed in frequency domain, on a band analysis basis, using *Hanning* window:

$$S(k,l) = |\frac{1}{M_2} \sum_{r=1}^{M_1} E_2(k+r) e^{-j\frac{2\pi}{M_2} rl} w(r)|^2, \tag{6}$$

$$k = 1, 2, ..., N/2, \qquad l = 1, 2, ..., M_1,$$

where $M_1 = min\{N/2, k+M\} - k$, $M < M_2 < N/2$, $S(k,l)$ is the l-th spectrogram coefficient, M_2 is the number of DFT points, M is the pre-defined window length, and $w(r)$, $r = 1, 2, ..., M_1$, is the Hanning window in frequency domain. As evident, as long as $k + M < \frac{N}{2}$, M_1 equals M.

The peak of spectrogram, $S(k,l)$, $l = 1, 2, ..., M_1$, gives the IF of the spectrum, $E_2(k)$, across the frequency axis:

$$\xi(k) = max\{S(k,l)\}, \qquad k = 1, 2, ..., N/2 \tag{7}$$

$\xi(k)$ represents IF of the spectrum over frequencies from 0 to $F_s/2$.

The *transition frequency*, f_{trans}, which specifies a change in the spectrum characteristics from periodic to random, is obtained through measuring the *flatness* of $\xi(k)$ in a number of sub-bands, n_b. This is formulated as:

$$\zeta(j) = \frac{exp(\overline{log\, \underline{\xi_j}^2})}{\overline{\underline{\xi_j}^2}}, \qquad j = 1, 2, ..., \frac{F_s}{2n_b} \tag{8}$$

where j is the sub-band index, $\underline{\xi_j}^2 = \{\xi_{j1}^2\ \xi_{j2}^2 ...\}$, and vector $\underline{\xi_j} = \{\xi_{j1}\ \xi_{j2} ...\}$ is the jth part of $\xi(k)$, $k = 1, 2, ..., N/2$, located in jth band, whose flatness is represented by $\zeta(j)$.

As evident, $0 < \zeta \leq 1$, which is used as an indication of flatness, where 1 is for an absolutely-flat vector ($\xi_{j1} = \xi_{j2} = ...$). f_{trans}, is then calculated through comparing $\zeta(j)$ with threshold th as:

$$f_{trans} = j_0 \frac{F_s}{2n_b} \tag{9}$$

where $j_0 = min\{j < th\}$.

The threshold is calculated based on the mean of the spectrum flatness within a certain band, averaged over a number of previous frames composed of voiced and unvoiced, given as:

$$th = .5[\frac{1}{j_2 - j_1 + 1} \overline{\sum_{j=j_1}^{j_2} \zeta(j)} + 1] \tag{10}$$

where j_1 and j_2 are pre-defined indices of the lowest and highest sub-bands considered.

As expressed by equation (10), the spectrum is assumed periodic at frequencies below f_{trans} while it is taken random at frequencies over f_{trans}, with a resolution specified by n_b. This assumption, although not always true, suffices to achieve a good naturalness, as will be shown later.

Figure 1 indicates three classes of speech, voiced, unvoiced, and mixed, along with the corresponding IF representations. As seen, a nearly flat IF is obtained when there are strong pitch harmonics in the residual signal spectra.

Pitch and Voicing Detection

Pitch is determined based on the average value of the IF within a pre-specified band below 1 kHz regardless of voicing status. Pitch period is then obtained from, $l_{max} \in \{1, 2, ..., M_1\}$, the time at which $\xi(k)$ (peak of spectrogram) is reached, as:

$$T = N l_{max} \tag{11}$$

where T is the pitch period in terms of number of samples.

Model→	A	B	C	Quantization
$d_s\,(dB)$	2.005	1.834	1.926	No
$d_s > 3\,dB(\%)$	8.25	1.23	1.36	No
$d_s\,(dB)$	4.891	3.667	4.053	Yes
$d_s > 5\,dB(\%)$	12.02	8.71	11.79	Yes

Table 1. Overall spectral distortion with different models for excitation signal.

The *degree* of voicing, or periodicity, is determined by the transition frequency. A low f_{trans} means that the periodic portion of the excitation spectrum is dominated by the random part and vise versa. For this reason, the accuracy in pitch detection during unvoiced, which is intrinsically ambiguous, is insignificant and non-effective in naturalness. Therefore, pitch values can be smoothed or simply made equal to zero, during unvoiced, based on the same criterion as used in pitch detection.

Figure 2 illustrates the pitch contour detected using the proposed method along with f_{trans} contour and the corresponding speech waveform. To make a comparison with pitch detection in binary model, the pitch contour detected by *cepstral* method is also indicated. As seen, the differences between two pitch contours occur mostly during unvoiced, where cepstral pitch is set to zero.

Reconstruction of Excitation

We consider three different models for excitation. Model A, is a binary model in which excitation signal takes one of two forms, periodic when frame is voiced, or random noise when frame is unvoiced. This is controlled by a voiced/unvoiced switch working typically based on a waveform periodicity measure applied to the residual signal in LPC model. Pitch and voicing are determined using cepstral method and are quantized by 2 bits/20 msec. Gain is also encoded by one bit/20 msec.

In models B and C, Excitation signal is reconstructed through combining a periodic spectrum at lower frequencies with a random spectrum at higher frequencies. Two portions of the spectrum are then joint together at the frequency specified by f_{trans}, for each frame of speech.

In model B, a fixed shape is used for excitation waveform which is a modified version of LPC10 excitation waveform. f_{trans} in model B is quantized using three bits. Pitch is detected using the proposed method described in previous section, quantized with 2 bits/20 msec, and gain is treated similar to the gain in model A.

Model C uses two pre-defined different waveforms for excitation which are selected as excitation signal based on spectral distance between original and reconstructed speech signals, on a frame by frame basis. In this model, two bits are allocated to f_{trans}, while pitch and gain are encoded in the same way as that in model B.

3. PERFORMANCE EVALUATION

The method is evaluated through a perceptually based

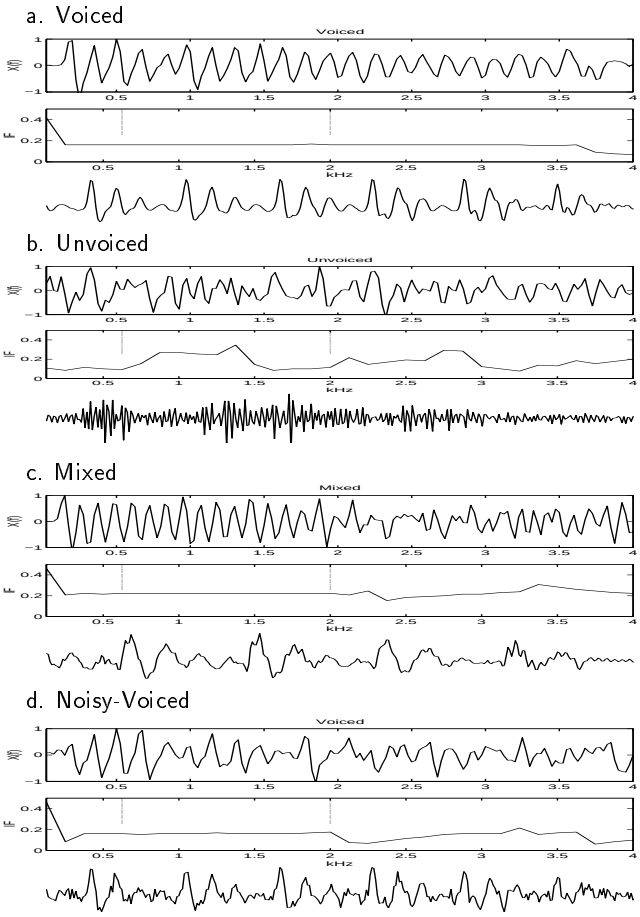

Figure 1. DFT of residual signal, IF estimation across frequency axis, and corresponding speech waveform. Dotted lines mark the boundaries of the band considered for voicing detection.

spectral distance measurement of speech reconstructed using the proposed method in an LPC model. We developed and used *log-Bark spectral* distance measure, defined as:

$$d_s\,(dB) = \frac{1}{N_f}\sum_{i=1}^{N}\Big(\frac{1}{15}\sum_{k=1}^{15}[10\log P_2^i(k) - 10\log P_1^i(k)]^2\Big)^{1/2} \tag{12}$$

where i is the frame index, N_f is the total number of frames, $P_1^i(k)$ and $P_2^i(k)$ are the normalized powers of original and encoded speech signals for the ith frame at the kth filter of the *Bark-scale* filter-bank [7].

4. EXPERIMENTAL RESULTS

Based on our previous work on Temporal Decomposition (TD) based coding [5], we used the proposed method for approximating the excitation in a TD based coder working at rate 800 b/s. For coding, ten LAR parameters were used as spectral parameters where *Cepstral* coefficients of the LPC filter impulse response were used for *event* detection (see [5]). We used the proposed method in the LPC synthesizer, with non-quantized LPC parameters, to

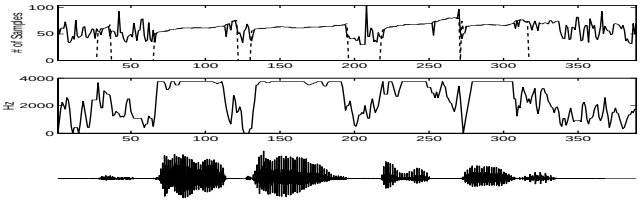

Figure 2. Top to bottom: pitch contour detected using proposed method (solid) along with cepstral pitch (dashed), time-varying f_{trans} over the utterance, and corresponding speech waveform (*/He stole a dime from a beggar/*).

eliminate the effect of error in quantization of spectral parameters on the quality of synthesized speech.

For the excitation signal, all three models described earlier were tested, where: $N = 320$, $M = 32$, $M_2 = 64$, $F_s = 8000$ Hz. The frame rate in the coder was 200 frame/sec while gain and pitch could change every 20 msec.

The results for the overall spectral distortion, d_s using a large number of different speech utterances from TIMIT database, are shown in Table 1. In this table, percentages of frames in each scheme, for which d_s axceeds a certain amount, are also indicated. These results were closedly confirmed by our informal listening tests.

The number of bits allocated to different parameters used in modeling the excitation are indicated in Table 2.

Model→ Parameters↓	A	B	C
V/UV	50	-	-
Waveform	-	-	50
f_{trans}	-	150	100
Total	50	150	150

Table 2. Number of b/s allocated to excitation modeling parameters used in TD-based coding for the three excitation models.

5. DISCUSSION

As shown in Table 1, spectral distortion in reconstructed speech, confirmed by our informal listening tests, significantly decreases when voicing diversity in the signal spectra is considered as a time-varying parameter. Particularly, the method highly reduces the maximum spectral error associated with the synthesized speech (see Table 1, percentage of frames with distortion larger than a certain amount). The slight difference between two models B and C arises from the fact that perceptual distortion is more sensitive to the error in the transition frequency, f_{trans}, compared to the error in temporal features of excitation signal. A sudden change in the spectral characteristics happens when f_{trans} is displaced in frequency domain for more than one fundamental frequency due to quantization error. This causes a change in the number of pitch harmonics in the periodic band which could generate perceivable distortion during voiced. For this reason, female speech could be more sensitive to this error and needs more bits than male speech for quantizing f_{trans}, in general.

An important feature of the proposed method is robustness to noise. This stems mainly from searching for pitch harmonics in the frequency domain within independent bands, provided by windowing the residual signal spectra. In addition, the *integration* process associated with the spectrogram technique for IF estimation, reduces the effect of random noise on the signal spectra. This can be seen in Figure 1-d representing the IF vector extracted from a frame of noisy speech (the same voiced frame as indicated in Figure 1-a) with 0-dB Signal-to-Noise Ratio.

The proposed method, can be used similarly at higher rates, specifically in MBE based coding. The main advantage at higher rates could be the possibility of sinusoidal approximation of periodic part of excitation, as considered in typical MBE coding.

6. CONCLUSION

We have proposed, in this paper, a new method for modeling excitation and pitch and voicing detection in very low rate coding to improve naturalness in synthesized speech. The method was evaluated using a perceptually-based log-Bark spectral distance measure and informal subjective tests which resulted in a considerable improvement in speech quality in a two-band MBE based coding scheme, particularly when the rate was highly restricted in TD-based coding. This is achieved through a novel approach to measurement the periodicity in speech based on the flatness of instantaneous frequency of the LPC-residual spectra within individual bands.

The method is also applicable at higher rates for accurate excitation modeling and pitch and voicing determination in MBE based speech coding, combined with traditional sinusoidal approximation technique.

7. REFERENCES

[1] F. Beritelli, S. Casale, F. Spataro, "A Simple and Efficient Two-Band Speech Coder at 2.4 Kbits/s for Real-Time Implementation on a Single Low-Cost DSP", *Proc. MELECON'96*, pp. 1674-1677, 1996.

[2] B. Boashash, "Estimation and Interpreting The Instantaneous Frequency of a Signal", *Proc. IEEE*, Vol. 80, No. 4, pp. 520-568, Apr. 1992.

[3] K. M. Chiu, P. C. Ching, "Quad-band excitation for low bit rate speech coding", *J. Acoust. Soc. Am.*, Vol. 99(4), pt.1, pp. 2365-2369, Apr. 1996.

[4] A. Gersho, "Advances in Speech and Audio Coding", *Proc. IEEE*, Vol. 82, No. 6, June 1994.

[5] S. Ghaemmaghami, M. Deriche, "A New Approach to Very Low-Rate Speech Coding Using Temporal Decomposition", *Proc. ICASSP'96*, Vol. 1, pp. 224-227, May 1996.

[6] S. Ghaemmaghami, M. Deriche, B. Boashash, "A New Approach to Pitch and Voicing Detection through Spectrum Periodicity Measurement", *Proc. TENCON'97*, 1997.

[7] S. Wang, A. Sekey, A. Gersho, "An Objective Measure for Predicting Subjective Quality of Speech Coders", *IEEE J. Select. Areas Comm.*, Vol. 10, No. 5, pp. 819-829, June 1992.