

**MULTI-BAND EXCITATION CODING OF SPEECH AT 960 BPS USING  
 SPLIT RESIDUAL VQ AND V/UV DECISION REGENERATION**

**Cheung-Fat Chan**

**Department of Electronic Engineering  
 City Polytechnic of Hong Kong  
 Tat Chee Avenue, HONG KONG  
 E-Mail: EECFCHAN@CPHKVX.BITNET**

**ABSTRACT**

This paper describes a method to achieve high-quality coding of speech signals at 960 bps. The method employs the multiband excitation (MBE) model together with very efficient quantization schemes for coding the spectrum magnitudes and the voiced/unvoiced (V/UV) decisions. The spectrum magnitudes are coded using a novel vector quantization (VQ) scheme which has both the structural characteristics of multistage VQ and split VQ. The V/UV decisions for pitch harmonics are not transmitted to the decoder but are regenerated from the spectrum information available in the decoder. It was demonstrated that the proposed speech coder is capable of generating speech of good quality.

**1. INTRODUCTION**

Recently, multiband excitation (MBE) model has been shown capable of producing high-quality natural-sounding speech[1]. Instead of using a single voiced/unvoiced (V/UV) decision for the whole speech spectrum as in conventional LPC system, in multiband model, the excitation signal is represented by a series of bands centered at the harmonics of the fundamental frequency. Each band can be independently declared as voiced or unvoiced. This model allows voiced and unvoiced excitations co-exist within the signal band, and as a result of this improvement in modelling, the quality of the synthesized speech is increased. In the original MBE coding system, the parameters needed to be sent to the decoder compose of a pitch frequency, the band magnitudes and phases, and the V/UV decisions for all pitch harmonics. This requires a coding rate of 8 kbit/s[1]. In order to code speech signals at sub-kbit range more elaborate coding schemes are needed. This paper introduces an improved pitch detection scheme for MBE analysis, an efficient split VQ scheme for coding the band magnitudes, and a novel V/UV regeneration scheme so that only 1 bit is sufficient for coding the V/UV decisions.

**2. IMPROVED PITCH DETECTION FOR MBE ANALYSIS**

In MBE analysis, the input speech spectrum  $S(\omega)$  is matched to the synthetic spectrum by minimizing the error function  $\xi(\omega_p)$  with respect to the pitch frequency  $\omega_p$ , i.e.,

$$\xi(\omega_p) = \frac{\sum_{m=1}^M \sum_{\omega=\omega_m}^{b_m} [|S(\omega)| - A_m |E(\omega)|]^2}{(1 - \omega_p B) \sum_{m=1}^M \sum_{\omega=\omega_m}^{b_m} |S(\omega)|^2} \quad (1)$$

where  $B$  is a weighting factor for unbiasing the pitch dependent

error, and  $A_m$  is the band magnitude defined as

$$A_m = \frac{\sum_{\omega=\omega_m}^{b_m} |S(\omega)| |E(\omega)|}{\sum_{\omega=\omega_m}^{b_m} |E(\omega)|^2} \quad (2)$$

The number of harmonic bands is calculated as  $M = \left\lfloor \frac{\pi}{\omega_o} \right\rfloor$  with  $\omega_o$  being the pitch frequency such that  $\xi(\omega_o)$  is a minimum. If there are  $N$  frequency samples over the whole speech spectrum, then

$$a_m = \left\lfloor \frac{(m-0.5)N\omega_o}{\pi} \right\rfloor \text{ and } b_m = \left\lfloor \frac{(m+0.5)N\omega_o}{\pi} \right\rfloor \quad (3)$$

The excitation spectrum  $E(\omega)$  is constructed from shifted main lobes of a transformed Hamming window. The error in each individual band  $\xi_m \in [0, 1]$  is calculated as

$$\xi_m = \frac{\sum_{\omega=\omega_m}^{b_m} [|S(\omega)| - A_m |E(\omega)|]^2}{\sum_{\omega=\omega_m}^{b_m} |S(\omega)|^2} \quad (4)$$

which is an indication of the degree in mixture of voiced and unvoiced excitation in band  $m$  with center frequency  $m\omega_o$ . In the original MBE coding system proposed by Griffin,  $\xi_m$  is properly biased and compared to a threshold value to obtain a binary V/UV decision. The band is declared as voiced if  $\xi_m < 0$  and is declared as unvoiced if  $\xi_m > 1$ .

It had been noticed that using (1) alone to search for the pitch frequency is insufficient because, if the energy of band  $m$  is small,  $A_m$  is small and the error due to band  $m$  is small comparing to the total error, then the estimated pitch value may be multiples of the true pitch value. In order to overcome this problem, a corrective error function is defined as

$$\xi_c(\omega_p) = \frac{1}{M(1 - \omega_p B)} \sum_{m=1}^M \xi_m \quad (5)$$

In this error function, the error energy in each band is normalized so that equal weighting is applied to each band. However, if the number of voiced bands in the analysis frame is small, then, using (5) alone to search for the pitch value is also not accurate. We have found that a modified error function using the sum of  $\xi(\omega)$  and  $\xi_c(\omega)$  would serve better. Therefore, the pitch frequency is

10.21437/ICSLP.1994-523

determined by

$$\omega_o = \underset{\forall \omega_p}{\operatorname{argmin}} [\xi(\omega_p) + \xi_c(\omega_p)] \quad (6)$$

Fig. 1 shows the original, corrective, and the summed error functions. It can be seen that it is very effective to remove double-pitch errors using the summed error function.

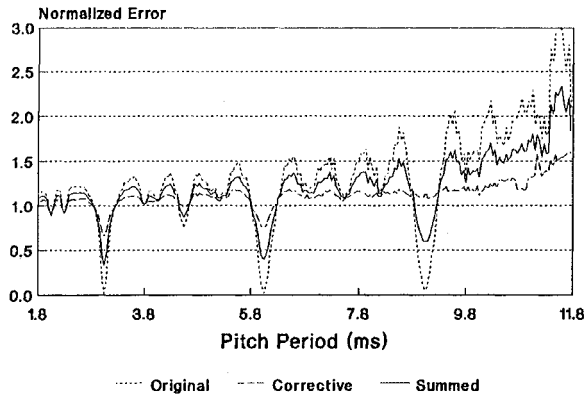


Fig. 1 Error Functions for Pitch Estimation

### 3. CODING OF BAND MAGNITUDES

Our approach to code the band magnitudes is to firstly convert the magnitudes into a fixed number of autocorrelation coefficients for LPC analysis. The LPC parameters are then quantized using a split residual vector quantization (SRVQ) scheme[2]. The autocorrelation are calculated from the band magnitudes using Fourier transform relation;

$$r(k) = \frac{1}{M} \sum_{m=1}^M A_m^2 \cos(2\pi mk\omega_o) \quad 0 \leq k \leq L \quad (7)$$

where  $L$  is the LPC order. This transformation will introduce distortion into the spectrum magnitudes as LPC is an all-pole model, however, experimentally, we found that the lost in subjective quality is quite small. A block diagram of the split residual vector quantizer is shown in Fig. 2. The structure of the quantizer resembles a segmented lattice filter. The PARCOR coefficients in each lattice segment are grouped as vector for quantization. Therefore, this structure is a split VQ structure as PARCOR coefficient vector is split into subvectors for quantization. However, as the residuals (both forward and backward) are coupled across VQ stages, this structure can also be classified as residual (multistage) VQ.

The decoding process of the SRVQ scheme is relatively simple, however, the encoding process is more complicated. A stage-by-stage search strategy must be used. In the first stage, the autocorrelation sequences are fed to the first lattice segment. An optimum codeword is then selected such that the sum of forward and backward residual energies at the output of the lattice segment is minimized[2]. After the codeword is found, lattice analysis is performed using the optimum codeword obtained from the first stage search, and the autocorrelation sequences are then transformed to two new sets of correlation sequences. These correlation sequences are actually the autocorrelation of the forward and backward residuals, and the cross-correlation of the forward and backward residuals. These two sets of correlation sequences are then fed to the second stage for VQ encoding. The search strategy in the second stage is exactly the same as in the first stage. Codebook generation is based on Generalized Lloyd

algorithm with likelihood ratio as distortion measure[3]. Details about the encoding method and the codebook generation procedure are given in Reference [2]. It has been shown that the SRVQ scheme outperforms, in terms of averaged spectral distortion and number of outlier frames, other split VQ schemes which use line spectrum frequency (LSF) representation[2, 4].

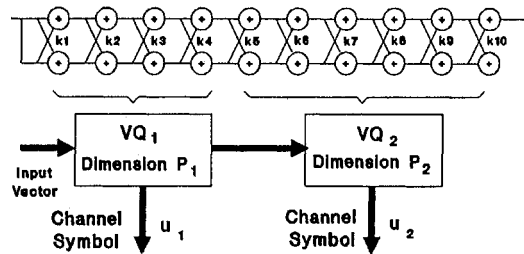


Fig. 2 Split Residual Vector Quantization Based on Segmented Lattice Filter

### 4. CODING OF V/UV DECISIONS

In the original MBE coding system, each V/UV decision is coded using 1 bit. Because the number of harmonic bands is data dependent, in order to achieve a fixed coding rate, V/UV decisions are usually grouped for encoding. A simple question to ask is how many bits are absolutely necessary for coding the V/UV information? Before answering this question, let us examine some typical spectra of voiced and unvoiced frames as shown in Fig. 3(a) and Fig. 3(b). For voiced frames, there are spectrum regions with strong formants and the excitations near the formant regions are most likely to be voiced. However, for unvoiced frames, there are unlikely to have strong formants in the spectrum and the chance for excitations to be classified as unvoiced is high. Also, it is most likely that speech signals at the high frequency regions are unvoiced. From this observation, hypothesis can be made that the short-term speech spectrum and the excitation are highly correlated. Since the spectrum information are available in the decoder, then, if we can extract these correlation and save them in a knowledge base, we can regenerate the V/UV decisions in the decoder for MBE synthesis without explicitly transmitting the V/UV decisions.

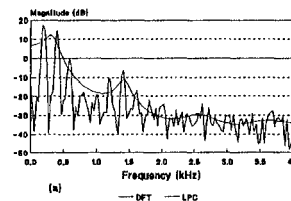


Fig. 3(a) Voiced Speech Spectrum

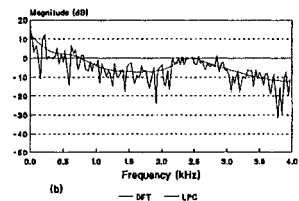


Fig. 3(b) Unvoiced Speech Spectrum

A long-term statistical approach based on training on real speech signals can be used to extract the V/UV excitation information. Our target is to use a single bit to identify the entire speech frame to be as voiced or unvoiced and utilize the spectrum information available in the decoder to estimate all V/UV decisions.

In the first step of training, VQ is applied to classify speech spectrum vectors into a finite set of prototype vectors. We use a similar VQ procedure described in previous section for training the spectrum codebook. Note that a large speech data base comprised of many speakers must be used for training and the speech data should be carefully chosen so that they contain a right

proportion of voiced and unvoiced sounds. In the second step, an analysis/classification procedure is performed as follows: For each input speech segment, an MBE analysis is performed to obtain a pitch frequency and a set of V/UV mixture values over the pitch harmonics. By assuming that the band error  $\xi_m$  calculated from (4) is constant within the harmonic band, we can define a V/UV mixture function as

$$\alpha(\omega) = \begin{cases} 1 - \xi_1 & a_1 \leq \omega \leq b_1 \\ 1 - \xi_2 & a_2 \leq \omega \leq b_2 \\ \dots\dots\dots \\ 1 - \xi_M & a_M \leq \omega \leq b_M \end{cases} \quad (8)$$

Therefore, one  $\alpha(\omega)$  can be generated for each input speech segment. After MBE analysis, the same speech segment is then mapped to one of the prototype vectors in the codebook by using a nearest neighbor rule similar to VQ encoding. This analysis/classification procedure continues until all input speech frames are analyzed and mapped. Once all training vectors are mapped to the codebook, each codevector in the codebook should own a cluster of training vectors and each training vector would have an associated  $\alpha(\omega)$ . In the next step, all  $\alpha(\omega)$ 's within the cluster are then averaged to achieve a single V/UV mixture function  $\bar{\alpha}(\omega)$  for each codevector. Afterward,  $\bar{\alpha}(\omega)$  is biased and clipped to a threshold level to obtain a binary number which is then stored alongside with the corresponding codevector. The threshold level is optimally determined such that the total number of wrongly classified V/UV decision after averaging is minimized.

Because there is a possibility that speech segments having voiced excitation and speech segments having unvoiced excitation are all mapped to the same codevector. Averaging  $\alpha(\omega)$ 's would thus degrade the discriminative power of the V/UV mixture function. Therefore, extra information to indicate the overwhelming nature of excitation is needed. A voiced/unvoiced classification (for the whole speech spectrum) will be performed in MBE analysis during the training. This single-bit V/UV information will be sent to the decoder for improving the performance of V/UV regeneration. By doing this, a codevector now owns two clusters; one contains speech vectors which are classified as voiced (actually, it may be partially voiced) and the other contains unvoiced speech vectors (it may also be partially voiced). Two averaged V/UV mixture functions for each spectrum codeword are then derived.

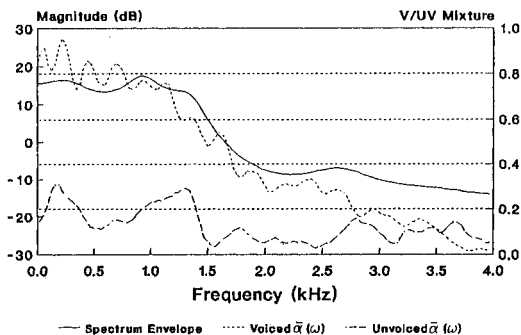


Fig. 4 Plots of V/UV Mixture Functions for a Prototype Spectrum

Fig. 4 shows plots of  $\bar{\alpha}(\omega)$  for both voiced and unvoiced speech and the spectrum envelop of the corresponding codevector. The

results were achieved by using a spectrum codebook having only 64 codevectors. The codebook was generated using over 15000 autocorrelation vectors obtained after MBE analysis and magnitude to autocorrelation conversion. We can see from these plots that each averaged V/UV mixture value falls naturally either in the voiced or unvoiced level; this means that the V/UV levels in similar speech frames are strongly correlated. Note that, for voiced speech, the V/UV mixture function tends to follow the trend of spectrum envelop. This actually confirmed our observations as stated early. Also, it is interesting to point out that the V/UV mixture functions for the voiced and unvoiced speech are distinctively different.

In order to evaluate the performance of this classification scheme objectively, we calculate the percentage error in V/UV classification by counting the number of wrongly classified V/UV decisions as

$$\zeta_{vuv} = \frac{1}{T} \sum_{i=1}^T \left[ \frac{n_i}{M_i} \right] \times 100\% \quad (9)$$

where  $n_i$  and  $M_i$  are the number of wrongly classified decisions and the total number of harmonic bands in each speech frame, respectively, and  $T$  is the total number of the training vectors. Table I shows the percentage errors for spectrum codebooks of various sizes.

Codebook Size	32	64	128	256
$\zeta_{vuv}$	18.2	13.2	12.1	11.4

Table I Percentage Error of V/UV Classification

In fact, from our observation, most errors occur in the regions where  $\alpha(\omega)$  is in the vicinity of the V/UV threshold level. From this result, we see that a 6-bit spectrum codebook is sufficient to capture the V/UV excitation information in speech signals. Obviously, because spectrum codebooks have already been used by the SRVQ system, in order to save the storage for an additional codebook for V/UV regeneration, we utilize the first stage VQ codebook (which is a 4-dimension, 7-bit codebook) from the SRVQ system introduced in previous section for analysis/classification of V/UV excitation during training and for V/UV regeneration during decoding. We found that the lost in accuracy in V/UV classification is very small as compared to a 10-dimension spectrum codebook of the same size.

## 5. MBE SYNTHESIS

A block diagram of the proposed speech coding system is shown in Fig. 5. By using the decoded index of the first spectrum codebook and the single-bit V/UV flag to select the voiced or unvoiced table, a binary V/UV function associated with the optimal codevector is then extracted. With an aid of a decoded pitch value, a set of V/UV decisions over the harmonic bands is derived by counting the majority of voiced and unvoiced components within the band. These V/UV decisions together with the decoded pitch value and the sampled spectrum magnitudes are then fed to a MBE synthesizer for speech generation. Practically, voiced speech is synthesized in time domain by summing the outputs of sine wave oscillators with the oscillating frequencies set at pitch harmonics declared voiced. The amplitude of the sine wave is set to the decoded band magnitude. The oscillating frequencies and amplitudes are linearly interpolated between

adjacent frames. Since phases of pitch harmonics are not available, a phase interpolation scheme based on a quadratic polynomial function is used to achieve smooth evolution of phases[5]. Unvoiced speech is synthesized in frequency domain. A segment of Gaussian white noise is firstly generated and windowed by a Hamming window. The windowed noise are then transformed to frequency domain and normalized to have unit energy. Noise in the unvoiced bands are scaled by the band magnitudes while noise in the voiced bands are set to zero. The filtered noise spectrum is then inverse transformed and added to the voiced speech using a weighted overlap-add procedure.

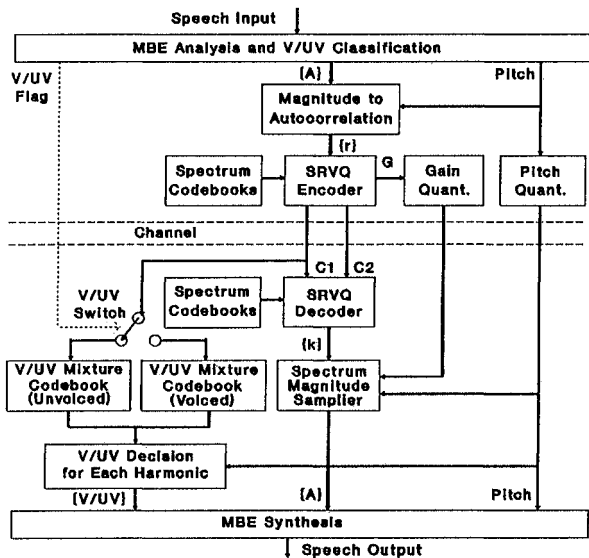


Fig. 5 Block Diagram of the Proposed Speech Coder

### 5.1 Adaptive Post Filtering

It is well known that an adaptive post filter could be used to improve the perceptual quality of CELP coders. Same post filtering technique can also be applied to the MBE speech. We found a post filter of the form shown below provides the best perceptual quality.

$$W(z) = \frac{1 + \sum_{i=1}^P a_i \beta^i z^{-i}}{1 + \sum_{i=1}^P a_i \gamma^i z^{-i}} (1 - \mu k_1 z^{-1}) \quad (10)$$

where  $a_i$  are the linear predictive coefficients. The controlling parameters are  $\beta=0.6$ ,  $\gamma=0.8$  and  $\mu=0.5$ .

### 6. SIMULATION AND RESULTS

In this simulation, speech signal is sampled at 8 kHz and windowed by a 256-point Hamming window. The frame update rate is 40 frames per second. The windowed signal are transformed to frequency domain using FFT. All 128 useful frequency samples are then up-sampled 3 times and linearly interpolated to obtain 384 samples for analysis. Therefore, fractional pitch of one third of the integer pitch can be achieved and  $\alpha(\omega)$  consists of 384 points. We must, however, emphasize that although fractional pitch must be used in order to obtain accurate estimates of harmonic amplitudes, the pitch value can be

coded fairly coarsely as human listeners are not so sensitive to small pitch error. We found that a 5-bit codebook with non-uniformly distributed codewords over the pitch range from 1.75 to 12 ms is sufficient for coding the pitch period with no noticeable loss of speech quality. The band magnitudes are converted to 11 autocorrelation coefficients. A two-stage SRVQ scheme is used. Both spectrum codebooks are 7-bit codebooks but the first codebook has a dimension of 4 (i.e.,  $k_1-k_4$ ) while the second codebook has a dimension of 6 (i.e.,  $k_5-k_{10}$ ). All spectrum codebooks were generated from a speech database contributed by 4 male and 3 female speakers. The speech was recorded from FM radio broadcast. The LPC gain is calculated as the square root of the forward residual energy in the SRVQ system and is coded to 4 bits using ADPCM. Table II lists the bit allocation scheme for the proposed coder operating at 960 bps.

Gain	VQ1	VQ2	Pitch	V/UV	Total
4	7	7	5	1	24

Table II Bit Allocation Scheme

Several English speech sentences, which were not included in the training set for generating the spectrum codebooks and the V/UV tables, were used to test the performance of the proposed coder. For the time being, we haven't performed any formal subjective test on the reproduction speech, but informal listening tests show that the coder is capable of generating speech of natural-sounding quality with high intelligibility. Only a slight reverberant quality was noticed, but more importantly, we didn't find any artifacts as a result of regeneration of V/UV decisions in all the speech sentences tested. The quality is judged to be significantly better than that of a 2.4 kbit/s standard LPC-10 coder.

### 7. CONCLUSION

This paper introduces an efficient split residual VQ scheme to quantize the spectrum magnitudes obtained from MBE analysis. A method to regenerate the V/UV excitation information from the coded spectrum envelop is also described. This allows efficient coding of V/UV decisions using only 1 bit per frame. Simulation results show that the proposed speech coder is capable of generating speech of good quality at 960 bps.

### References

- [1]. D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," IEEE Trans. on Acoustics, Speech, and Signal Processing, Vol. 36, No. 8, pp.1223-1235, 1988.
- [2]. K. W. Law and C. F. Chan, "A Novel Split Residual Vector Quantization Scheme for Low Bit Rate Speech Coding," IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Adelaide, Australia, pp.1.493-1.496, April, 1994.
- [3]. R. M. Gray, "Vector Quantization," IEEE ASSP Magazine, pp.4-27, April 1984.
- [4]. K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing, pp.661-664, 1991.
- [5]. INMARSAT-M Voice Coding System Description, 1991.