行政院國家科學委員會專題研究計畫　期中進度報告

無線網路環境下語音處理技術之整合型研究(1/3)

計畫類別：個別型計畫
計畫編號：NSC92-2213-E-002-090-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：國立臺灣大學資訊工程學系暨研究所

計畫主持人：李琳山

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 6 月 16 日

行政院國家科學委員會專題研究計畫期中報告

中文口語處理技術之前瞻性研究課題(2/3)

計畫編號：NSC 92-2213-E-002-090
執行期限：92 年 8 月 1 日至 93 年 7 月 31 日

主持人：李琳山　國立台灣大學電信工程學研究所
E-mail: lslee@gate.sinica.edu.tw

ABSTRACT

In this project, a new framework of distributed speech recognition (DSR) over wireless fading channels is proposed. 2D-DCT based compression method and an iterative bit allocation algorithm are developed for source coding, while short BCH code integrated with interleaving proposed for channel coding. The high correlation among speech feature parameters in temporal domain is well exploited in the 2D-DCT based compression, and a carefully designed iterative bit allocation algorithm can very efficiently make use of every bit transmitted. The very low bit rate plenty planning space for strong error control. Short BCH code integrated with interleaving is initially proposed to efficiently handle the severe bursty errors usually encountered in wireless fading channel. Overall system simulation based on a GPRS wireless channel simulator indicated significant recognition performance improvements is achievable at a bit rate of 3.4 kbps as compared to the conventional SVQ compression approach (with bit rate 4.4 or 4.8 kbps). The low computation requirements for all processes involved make it easy to implement them in the mobile phone clients with existing technologics.

中文摘要

本計畫提出一個透過無線衰減通道（wireless fading channel）傳輸實現分散式語音辨識（DSR）的新架構。在訊源編碼（source coding）部份，我們發展以二維離散餘弦轉換（2D-DCT）為基礎的資料壓縮方法，以及遞迴式位元分配演算法(iterative bit allocation algorithm）；而在通道編碼（channel coding）方面，則是以短 BCH 碼(short BCH，Bose-Chaudhuri-Hochquenghem code）與交錯編排（interleaving）的整合性運用為基礎。語音信號相鄰音框間的高度相關性，使得 2D-DCT 壓縮方法能發揮其功用，並配合我們精心設計的遞迴式位元分配演算法，使得每一位元的傳輸更具效率；壓縮之後所得到的低位元率（very low bit rate），提供了選用有效位元錯誤控制機制（strong error control）的傳輸頻寬空間。BCH 碼與交錯編排（interleaving）的整合，能有效處理以無線衰減通道做為傳輸媒介時，所常遭遇的嚴重叢集錯誤（severe bursty errors）情形；我們並以 GPRS 無線通道模擬器（wireless channel simulator）建立完整的模擬環境；與傳統的 SVQ 壓縮法相比較（4.4 ~ 4.8 kbps），在我們的模擬中可觀察到辨識正確率的明顯進步，且本文所提出的壓縮方法可達到更低的傳輸位元率要求（3.4 kbps）。我們所提出的架構，計算量需求不高，因此能與現今各種行動通訊技術妥善地結合。

1. INTRODUCTION

DSR is getting more and more important because of its compatibility to the network environment. The client-server architecture is very attractive because the computation requirements for the hand-held devices can be very low. There are in general two main approaches to the client-server model. The first is to use an existing speech codec, such as that used by the GSM system, to extract speech features and transmit them over the network. The speech recognition is then performed at the server based on features received after the transmission. The second approach, on the other hand, is to extract recognition-oriented features locally at the client, then compress and transmit them to the channel to the server for recognition. The discussions of this paper are based on the second approach.

In a prior work [1], the Mel-frequency cepstrum coefficient (MFCC) vectors were first grouped into blocks, and 2-D discrete cosine transform (DCT) was applied to the blocks. The DCT coefficients with small energy were set to zero and zigzag sampling, scalar quantization, run-length algorithm and Huffman coding were then performed. Very good recognition accuracy was achieved on the TI digit task at 624 bps. However, with run-length algorithm & Huffman coding, the errors inevitably occurring during wireless transmission in the fading channels may seriously propagate at the receiving end and degrade the recognition performance. In this paper, we learn the concept of the prior work, but develop a whole set of source coding and channel coding techniques to handle the problems here. 2D-DCT and iterative bit allocation algorithm are specially designed for the source coding to make use of the special characteristics of the speech feature parameters, while short BCH code integrated with interleaving [2] are used for the channel coding to combat with the bursty errors in wireless fading channel. Very encouraging results were obtained in the preliminary experiments.

2. PROPOSED APPROACH

The block diagram for the proposed system is shown in Fig. 1. At the client end, the feature vectors of the speech signal are first grouped into blocks. The source encoding includes 2D-DCT performed on these blocks and the quantization of the important components, thus obtained with carefully designed codebooks. The channel encoding includes short BCH encoding integrated with interleaving so are to protect the compressed speech information. These encoded data are then transmitted over wireless fading channels. At the server end, the received bit streams de-interleaved and BCH decoded first, and then the outputs are de-quantized and grouped into blocks for inverse 2D-DCT. The reconstructed feature vectors are finally used for the recognition. Only the static feature parameters were encoded,

quantized and transmitted; the delta and delta-delta features are computed at the server. The details of the source coding and channel coding are explained in sections 3 and 4 respectively, while the simulation results including those for a GPRS wireless channel is presented in sections 5 and 6.

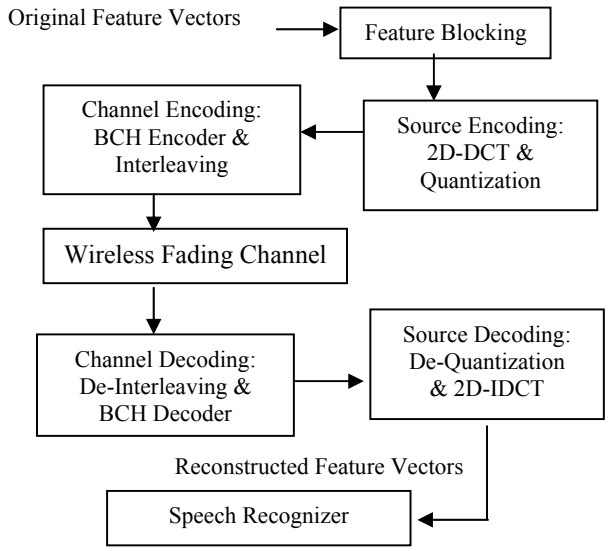


Fig.1 Block Diagram of the Proposed Approach

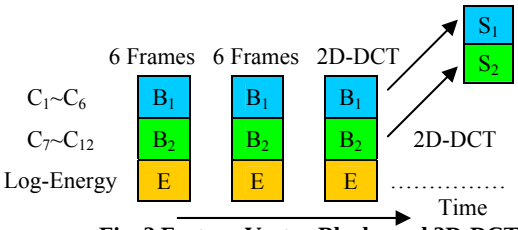


Fig. 2 Feature Vector Blocks and 2D-DCT

3. SOURCE CODING

3.1 Feature Vector Blocks & 2D-DCT

It has been reported that the first several MFCC parameters need more bits than other parameters in SVQ-based DSR approaches. This leads to the feature vector blocks and 2D-DCT scheme proposed here, as shown in Fig. 2. Every feature vectors includes 13 parameters, c_1~c_{12} plus the log-energy. Every 6 feature vectors form a segment, c_1-c_6 and c_7-c_{12} of such a segment are respectively grouped into matrices B_1 and B_2, which are then 2D-DCT transformed into matrices S_1 and S_2. The log-energy is handled separately using approaches for inter-frame SVQ [3]. As shown in Fig. 3 because the inter-frame correlation across time is much higher than intra-frame correlation among different parameters for the MFCCs, in S_1/S_2 most energy is concentrated in the first and second columns. Therefore instead of using zigzag sampling as was done before, we should only need to select important components from the first and second columns of S_1 and S_2 and quantize them. All other components can be set to zero.

3.2 Optimal Bit Allocation for Selected Components

Here, we try to perform optimal bit allocation to the most important 2D-DCT components by evaluating the degradation of recognition accuracy when deleting each component as well as losing each bit in each component, as given below.

First, we develop a simple approach to rank the relative importance of the components out of the 24 coefficients (the first two columns of S_1, S_2). Initially, we assign enough bits (eg: 8 bits each) to train each component's non-uniform scalar quantization codebook, and evaluate the increase of error rates when deleting each single component, and rank the importance of these components according to the increase of error rates. Based on this ranking, a total of L coefficients in order of importance, denoted b_1~b_L below, can be selected, where b_1 is the most important and b_L the least important.

An iterative bit allocation algorithm is then developed here as described below. We first set a maximum acceptable recognition error rate, $R_0=\alpha$, and then the recursive search algorithm includes two steps as follows.

Step1: This step main include many small stages. In each stage, one bit is removed from all the components b_1~b_L, re-train the codebook for each component and evaluate the error rate. If the error rate is still lower than α, the same stage is repeated, i.e, one more bit is removed from all components, etc. This step terminates when the error rate exceedsα.

Step2: This step includes two stages. In the first stage, we add one bit to b_1, the most important component, then cheek if the error rate is reduced belowα. If no, we further add one bit to b_2, the second important component, and so on. When the most important m components, b_1, b_2…b_m, all obtain one more bit and the error rate eventually becomes lower than α, the stage stops. The bit number for b_1, b_2…b_m are now fixed. In the next stage, the importance of the rest of the components, b_{m+1}, b_{m+2}….b_L are

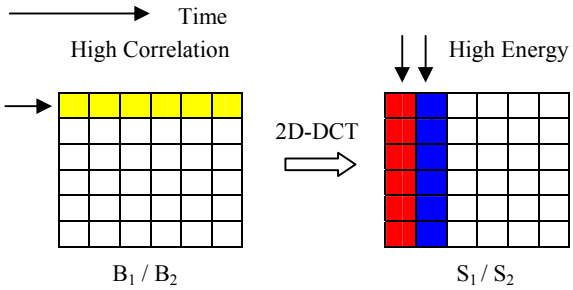
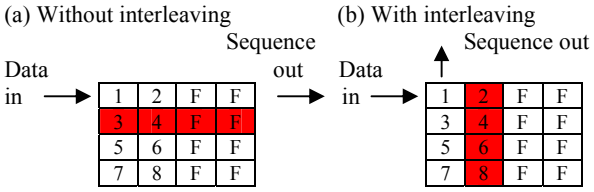


Fig. 3 Properties of B_1/B_2 and S_1/S_2



Transmitted Sequences:
(a) 1-2-F-F-3-4-F-F-5-6-F-F-7-8-F-F / 3-4-F-F error
(b) 1-3-5-7-2-4-6-8-F-F-F-F-F-F-F-F / 2-4-6-8 error

Fig. 4 BCH code, Interleaving & Bursty Errors

re-ranked using the approach mentioned above, i.e: evaluating the increases of error rate when deleting each single components. After this re-ranking, we have a new set of coefficients to be considered, b_1~b_{L-m}, where b_1 is the most important and b_{L-m} the least important. Now go back to step1 with this set of components b_1~b_{L-m}. The iterative bit allocation algorithm terminates when the bit number for all components are fixed. The above algorithm requires a large training/testing corpus to allocate right number of bits to the components. The test results will be given in section 5.

4. CHANNEL CODING

BCH code is a subclass of cyclic codes, which are originally designed for handling random errors. For an (N,k,t) BCH code, N-k cheek bits are used to protect k information bits, and the total number of transmitted bits is N. If there are less than t errors in the received N bits, the BCH decoder can correct all the errors and recover the original information bits [2]. However if there are more than t errors in the N bits, BCH code can't be able to handle these errors, so BCH code alone is not helpful for fading channels with bursty errors. Interleaving has

been used to disperse the bursty errors and make the bursty errors distributed like random errors. The degree which the bursty errors can be dispersed depends on the interleaving depth, which has to do with the time delay, an important issue in real time operation. In this research we integrate BCH code with interleaving to handle the bursty errors in wireless fading channels.

Fig. 4 illustrates an example of how interleaving can help to disperse the bursty errors. Assume a simple (4,2,1) BCH code is used and F is used to denote the check bit. Fig 4(a) is the case of without interleaving. Every row in Fig 4(a) is the four bits, where the first two are information bits while the next two F's are check bits for them. So in the 4 rows of Fig 4(a), a total of eight information bits (1,2,3,4,5,6,7,8) are transmitted together with eight check bits. The transmitted sequence is listed below in the figure. If a bursty error occurred, for example, the bits in the second row are all erroneous as shown in the figure,

Coefficients	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	$S_{1,5}$	$S_{1,6}$
Ranking	1	4	8	7	3	6
Coefficients	$S_{1,7}$	$S_{1,8}$	$S_{1,9}$	$S_{1,10}$	$S_{1,11}$	$S_{1,12}$
Ranking	11	9	15	18	19	21
Coefficients	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	$S_{2,5}$	$S_{2,6}$
Ranking	5	12	10	14	13	2
Coefficients	$S_{2,7}$	$S_{2,8}$	$S_{2,9}$	$S_{2,10}$	$S_{2,11}$	$S_{2,12}$
Ranking	16	22	20	23	24	17

Table I: Importance Ranking for S_1/S_2 Components

Coefficients	$S_{1,1}$	$S_{1,2}$	$S_{1,3}$	$S_{1,4}$	$S_{1,5}$	$S_{1,6}$
Bit Allocation	5	5	4	4	5	5
Coefficients	$S_{1,7}$	$S_{1,8}$	$S_{1,9}$	$S_{1,10}$	$S_{1,11}$	$S_{1,12}$
Bit Allocation	3	3	3	3	0	0
Coefficients	$S_{2,1}$	$S_{2,2}$	$S_{2,3}$	$S_{2,4}$	$S_{2,5}$	$S_{2,6}$
Bit Allocation	5	4	3	3	4	5
Coefficients	$S_{2,7}$	$S_{2,8}$	$S_{2,9}$	$S_{2,10}$	$S_{2,11}$	$S_{2,12}$
Bit Allocation	3	0	0	0	0	2
Inter-Frame Log-Energy SVQ: 6 bits						

Table II: Bit Allocation for S_1/S_2 Components at 1.45 Kbps

there is no way to recover the information bits(3,4) because the associate check bits are also erroneous. When interleaving is used, on the other hand, as shown in Fig 4(b) the bit sequence to be transmitted becomes column-wise instead, also listed below in the figure. In this case it is clear the error burst is actually dispersed across the four rows, and they can all be recovered because in each row there is only one bit error. This is why integrating interleaving with BCH code can protect the transmitted data very well against the bursty errors. The interleaving depth is in fact the number of rows in Fig 4, with which the interleaving to be performed as a group.

5. EXPERIMENTAL RESULTS FOR SOUCE CODING

5.1 Iterative Bit Allocation

The iterative bit allocation algorithm was carried out with a large-vocabulary Mandarin speech corpus. The speech signal was originally recorded in normal laboratory environment and down-sampled from 16 kHz to 8 kHz. The hamming window length for feature extraction was 32 ms, with window shift 10 ms. 40,000 sentences were used in training, by 100 male and 100 female speakers, 200 sentences from each. The acoustic models for recognition tests were trained from uncompressed data. The test data included 1635 sentences, produced by 2 male and 3 female speakers. The error rate used was the free-decoded syllable error rates considering the monosyllable nature of Mandarin Chinese, i.e, the syllable error rates without knowledge of lexicon and language model. The baseline syllable error rate was 39.34%. In order to simulate all the operations in an existing mobile phone, we truncated all the original feature parameters from 32 bits to 16 bits, and used 16 bits multiplication & addition in all following experiment [4].

The initial ranking obtained with the approach mentioned above for the relative importance for the 24 components in the 2D-DCT matrices S_1,S_2 are listed in Table I, in which $S_{i,k}$, are the k-th components in S_i (i=1, 2), and the ranking is from 1 to 24, where 1 indicates the most important and so on. We can see that the components of the first columns of either S_1 or S_2 are always more important than those of the second columns, and in general the components of S_1 was more important than those of S_2. Based on the results in Table I, we selected the 10 most important components of S_1 and 8 most important components of S_2 to perform the iterative bit allocation algorithm. We set the maximum acceptable error rate α=39.34%, the baseline error rate without compression. We found that the data rate can be compressed to as low as 1.45 kbps without any error rate increase as compared to the baseline (39.34%). The bit allocation for this method is given in Table II. If we allowed α=1% higher than the baseline, i.e.α=40%, then we could even compress the data rate down to 1.2 kbps, which gave an error rate of 39.89%.

	Stationary Noise			Non-stationary Noise		
Accuracy	Exhibition	Car	Test A	Restaurant	Airport	Test B
Baseline	71.03	66.99	67.62	58.07	60.82	62.94
2D-DCT	68.66	63.74	65.97	59.12	60.43	62.19

Table III: Experimental Results for Aurora II Corpus

5.2 Tests with Aurora II Corpus and Noisy Environment

Because the above bit allocation was obtained with a Mandarin speech corpus, it is reasonable to wonder whether it is over-fitted to that database. It is also interesting to investigate how it works with noise environments. This is why in this section we tried to apply the 2D-DCT compression with the bit allocation obtained above on a completely different corpus, the Aurora-II English digit corpus. We used the Aurora Front-End to extract 13-dimentional feature vectors, C_1~C_{12} & log-energy. Only the quantization of the log-energy was based on Aurora's specification, while all the other coefficients C_1~C_{12} were compressed with the 2D-DCT and bit allocation discussed above and trained with the Mandarin corpus. The baseline experiment for feature parameters without compression gave an accuracy of 99.01%, the average clean sentence accuracy of test A, B & C. With the compression based on 2D-DCT and bit allocation, the accuracy was 98.42% at bit rate of 1.45 kbps. We can conclude that the difference is practically negligible, and this may verify that the compression approach proposed here should be equally applicable to different corpus and different tasks.

The experimental results for Aurora II database with noisy environment for clean speech training condition are summarized in Table III. The results for exhibition hall noise and car noise are separately listed in the left half of the table, while Test A also listed is the average of four noise conditions, where three of them are primary stationary. Similar on the right half of the table, the results for Restaurant noise and Airport noise are separately listed. Also listed is Test B, which is the average of four noise conditions, where three of them primarily non-stationary. In each case the number listed here is the average of accuracies for six noise levels: clean speech, SNR of 20, 15, 10, 5, 0 dB. The bit rate for the compression is 1.45 kbps. From table III, we can see that when the additive noise was non-stationary (right half of the table) the proposed 2D-DCT compression and bit allocation only very slightly degrades the accuracy as compared to the uncompressed features. In some cases the compressed features even perform slightly better. On the other hand, when the additive noise is primarily stationary, a degradation of about 2.0~3.5% in accuracy is observable. This phenomenon may be

explained as follows. The 2D-DCT compression very much depends on the correlation of the cepstral coefficients in temporal domain. When the noise is primarily stationary, some characteristics of the noise may be somehow absorbed into the feature components during the 2D-DCT compression, thus disturbs the coefficients and produces the mismatch. When the additive noise was non-stationary, on the other hand, the disturbance on feature parameters are lowly correlated in temporal domain, thus more or less rejected by the 2D-DCT compression.

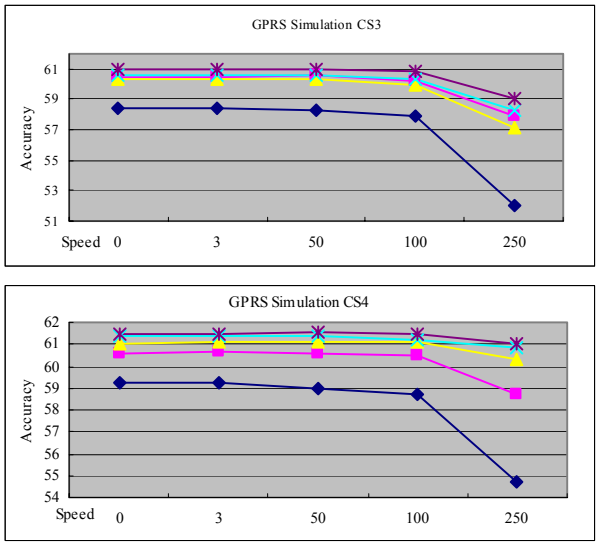
6. SIMULATION RESULTS OVER GPRS WIRELESS CHANNEL

The source coding and compression approaches proposed here in this paper seem to work satisfactorily as presented in the above experimental results. It is then important to check whether the channel coding approaches proposed here can properly complement the source coding approaches on the one hand, and can handle the problems for a wireless fading channel on the other hand. This is why in this section we try to perform the simulation for an overall DSR system over a wireless channel. This includes the source coding, the channel coding, the fading channel simulation, and speech recognition. The General Packet Radio Service (GPRS) was chose simply as an example for wireless channels. GPRS was been developed by ETSI to enhance the GSM system based on the packet switching framework. GPRS shares the GSM frequency bands and uses several properties of the physical layer of the GSM system. It includes four different error control code schemes, CS1~CS4, each with a different code rate.

The GPRS simulation software was developed by NTU Wireless-Communication-Lab, in which all complicated transmission phenomena have been carefully taken care of, such as the propagation model, the multi-path fading, the Doppler spread, etc. The example results presented below are based on the following simulation configuration: 15dB additive white noise, typical urban (TU) environment (a more frequently encountered with more severe fading problem), client traveling speed of 0/3/50/100/250 km/hr, one antenna and hard decision at receiver. We chose CS3, CS4 (with relatively weak protection) as testing code schemes. We used the source coding mentioned above including 2D-DCT compression and bit allocation with bit rate at 1.45 kbps, and (7,3,2) BCH code plus interleaving as the channel coding, so that the total transmission rate was 3.4 kbps. The tested task was Mandarin syllable recognition and the performance measure was syllable accuracy.

Three interleaving depths were tested, 1, 2, 3 block delay, where each block includes 6 frames of feature vectors with a length of 60 ms. The conventional intra frame SVQ compression scheme with or without error concealment based on interpolation were also performed here for comparison. The SVQ scheme compressed each feature vectors into a 44-bit symbol, and for error concealment purposes the symbol errors were detected by cyclic redundancy code [5], for which 4 extra bits are added to each to each symbol. So the total transmission rate for SVQ scheme was 4.4 or 4.8 (without/with error concealment) kbps.

The training data was the same as mentioned in section 5.1 and the test data included 1635 sentences, produces by 3 male and 2 female speakers, which was different from the test data in section 5.1. The simulation results for the overall system are shown in Fig.5. It can be seen from the figure that the SVQ scheme gave reasonable results (e.g. 58.75% at 100 km/hr, taking CS4 as an example), the error concealment offered very good improvement (e.g. 60.5% at 100 km/hr), while the proposed DCT approach achieved much better performance (e.g. 61.08% at 100 km/hr) at a bit rate of 3.4 kbps even with interleaving depth of only 1 block (60ms delay). Deeper interleaving certainly provided even better results. The degradation was moderate even if the client speed went up to 250 km/hr (e.g. 61% for Delay-3, CS4). Similar situation can be found for the case of CS3.



SVQ: ------: Without Interpolation ------: With Interpolation
DCT: ------: 1 block delay ------: 2 block delay ------: 3 block

Fig. 5 Overall System Simulations

7. CONCLUSION

An overall source/channel coding compression and error control framework for DSR is proposed in this paper. Iterative bit allocation algorithm offers very low bit rate, which in turn provides enough space for strong error control by BCH code and interleaving. Very significant important in recognition accuracy were obtained in the simulation. Although it seems that the 2D-DCT compression may suffer some degradation for stationary noise, substantial works have been reported that many new approaches are emerging which may be able to handle the problem of stationary additive noise. It should be pointed out that the computation requirements for all the processes mentioned here, including 2D-DCT, scalar quantization, BCH code and interleaving are all reasonably low, so it will be easy to realize the system in mobile phone clients with existing technologic.

8. REFERENCES

[1] Oifeng, Zhu, and Abeer. Alwan , “An efficient and scalable 2D DCT-based Feature Coding Scheme for Remote Speech Recognition,” in Proc. ICASSP, vol. I, May 2001, pp.113-116.

[2] Lijun. Zhang, Victor O.K.Li, and Zhigang Cao, “Short BCH Codes for Wireless Multimedia,” in wireless communications and networking conference, vol. 4. May 2001, pp 2613-2616.

[3] Constantinos Boulis, Mari Ostemdorf, “Graceful Degradation of Speech Recognition Performance Over Packet-Erasure Networks,” in IEEE Transactions on speech and audio processing, vol. 10, No, 8, November 2002, pp 580-590.

[4] Wireless Subscriber System Group, Motorola, “Integrated Dual-Core Baseband Processor, “1999.

[5] John B. Anderson, and Seshadri Mohan, *Source And Channel Coding*, Kluwer Academic , USA, 1991.