

# Variable-rate Coding Techniques for Mandarin Speech Transmission over Packet Network

Ding Zhong-qiang Ian McLoughlin

School of Applied Science, Nanyang Technological University  
Nanyang Avenue, S639798, Singapore  
Email: [p147253273@ntu.edu.sg](mailto:p147253273@ntu.edu.sg), [ivm@computer.org](mailto:ivm@computer.org)

Abstract-- Following the increasing popularity of Internet, the needs of speech transmission over packet network increase. The 13kb/s GSM RPE\_LTP has been accepted as an international speech coding standard, which is used to code speech over GSM digital cellular telephony networks. In this paper, the phonetic characteristics of Mandarin speech are analyzed and incorporated into GSM RPE\_LTP to make it suitable for Mandarin speech transmission over packet networks.

## I. INTRODUCTION

The main problems of speech transmission over traditional telephone networks are loss, noise and echo. When speech is transmitted over packet networks, these problems are minimized. However, some new problems in packet networks appear, for example: transmission delay, packet loss, packet disorder, which are caused by the physical transmission media and designs of network protocols [1]. These problems seriously affect the speech intelligibility. Because the packet networks may consist of several different media and network protocols, the error conditions are very complex when speech passes through these networks. The conventional speech coders generally do not consider the channel conditions. However, facing the new conditions in packet networks, the improvement of processing abilities of speech coders on channel errors may be more efficient for speech transmission than just performance improvement on one or several network protocols.

This paper discusses the characteristics of Mandarin speech and utilizes these characteristics to construct new coding schemes. A new structure of coder, which is based on the LTP\_RPE technique, is presented.

## II. ANALYSIS OF CHARACTERISTICS OF MANDARIN FOR SPEECH CODERS

Zhang [4] and Lee [7] discussed the phonetic and linguistic features of Mandarin speech, and compared them with those of English speech. According to Zhang, the intelligibility of Mandarin syllables increases as the intelligibility of phoneme increases. Every Mandarin

character consists of consonant, vowel and tone. The structure is CV. In order to improve the intelligibility of syllables, we must improve the intelligibility of consonants, vowels and tones. Syllables with the same consonant and vowel, but with different tone stand for different meanings. The tone is represented with the change of the pitch of the vowel, meaning that the performance of algorithms for pitch detection and preservation is significant for intelligibility of Mandarin. Most consonants (81%) are unvoiced. Most unvoiced consonants can be combined with the same vowel and tone. For example: A-set: {[a],[ba],[pa],[ma],[fa],[da],[ta],[na],[ma],[la],[ga],[ka],[ha],[ja],[cha],[sha],[zha],[sa],[ca],[za]} and An-set: {[an],[ban],[pan],[man],[fan],[dan],[tan],[nan],[man],[lan],[gan],[kan],[han],[jan],[chan],[shan],[zhan],[shan],[chan],[zhan],[ran]}. This makes the intelligibility of unvoiced consonants very important in Mandarin.

The number of syllables in Mandarin is limited to only 415, and there are four basic pitch frequency contours of tones. These characteristics are helpful for performance improvements of speech coding algorithms.

## III. CHINESE RPE\_LTP (CRPE\_LTP)

To improve the intelligibility of consonants in Mandarin speech transmission, we propose a new scheme called Chinese RPE\_LTP (CRPE\_LTP) used to code Mandarin speech, shown in Fig 1.

In the CRPE\_LTP coder, the approach to represent the speech signals  $s(n)$  is to use the speech production model in which speech is viewed as the results of passing an excitation,  $e(n)$  through a linear time-varying filter (LPC),  $h(n)$ , that models the resonant characteristics of the speech spectral envelope. The  $h(n)$  is represented by 8 LPC coefficients which are quantized in the form of Log.-Area Ratios. According to properties of speech signals, the speech signals are divided into three categories: *Silence*, *Unvoiced* and *Voiced*. In voiced speech parts, quasi-periodicity exists, which will be extracted by a pitch predictor filter.

At the receiving end, the information bits are decoded and hence, the model parameters are recovered. At the decoder, the voiced speech parts of excitation are recovered using a pitch synthesis filter, then sent to LPC synthesis

filter. The unvoiced parts of excitation are transferred directly to LPC synthesis filter. The silence parts of speech signals are replaced by zero frames, for comfort noise treatment later.

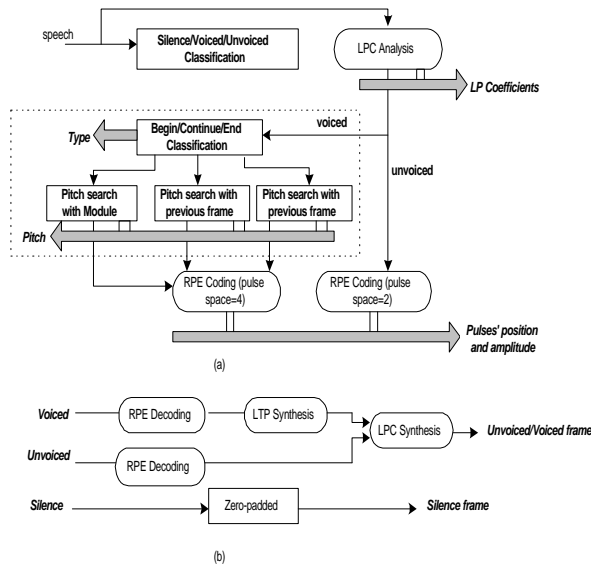


Figure 1: Simplified block diagram of CRPE\_LTP speech coder (a) encoder (b) decoder

### A. Silence/Voiced/Unvoiced Classification

The classification is important for the CRPE\_LTP coder. Its failure will lead to poor coding quality. We therefore accept the algorithm proposed by Sassan Ahmadi and Andreas S.Spanias [9]. In this algorithm, classification is made by short time zero-cross rate, short time energy, and cepstral peaks. Experiment evidence shows the performance of the algorithm is good, although complexity is relatively high.

### B. Begin/Continue/End Frames of voiced speech Classification (BF/CF/EF)

We propose a forward-backward search algorithm to classify the voiced speech frame into three classes, which are shown as follows:

The motivation that we classify voiced frames is that

```

IF (i-1)-th frame is unvoiced AND (i+K)-th frame is
voiced
    THEN i-th, (i+1),..., (i+K-1) frame belongs to
    BF
ELSE
    IF (i-1)-th frame belongs to CF AND (i+M)-th
    frame is unvoiced
        THEN i-th, (i+1),..., (i+M-1) frame
        belongs to EF
    ELSE
        i-th frame belong to CF
    
```

quasi-periodicity obviously exists in some frames, not in other frames. We consider the effect of LTP may be better if we process them separately. Here, we introduce two parameters K and M whose function is to control the scope of each class of frames. Here, we select  $K=3$  and  $M=1$ .

### C. Pitch Determination

Currently, autocorrelation and cross-correlation computation are main pitch determination methods. The key to them is the similarity between frames. If the similarity is very little, performance of these methods will be poor. To improve the performance of these methods, it must be that pitch is searched in two frames with quasi-periodicity. So we propose a *two-state pitch search method*. For Begin frames of voiced speech, we compute the cross-correlation value with a frame stored in a history module. For Continue frames and End frames, we compute the cross-correlation value with previous frames. The method can be shown in Fig 2. The module pulses are updated from the CF.

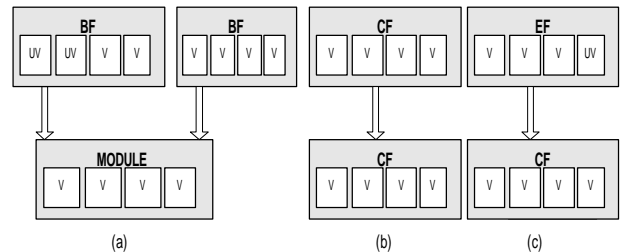


Figure 2 Two-state pitch search model (a) frames in BF search pitch pulses from Modules (b) frames in CF search pitch pulse from previous frames in CF (c) frames in EF search pitch pulse from frames in CF

### D. RPE coding

In RPE coding technique, a candidature sequence of pulses, where the distance between contiguous pulses is equal, is selected from input pulses. For example, 13 pulses are selected from 40 pulses according to the amount of average energy in GSM RPE\_LTP coder. The advantage of RPE coding technique is low computational complexity. However, the selected pulses may not be optimal for their constant distance. Assuming the set of arriving pulses is  $\{v(1), v(2), \dots, v(N)\}$ , where N is the number of pulses, if we divide the pulses into  $k$  parts, where the number of pulses in every part is the same, the following can be deduced:

$$L = \left\lfloor \frac{N}{k * p} \right\rfloor \quad (1)$$

where  $L$  is the number of pulses selected in each partition,  $p$  is the pulse space.

Fig.3 represents the pulses selected by  $k=1 \& p=3$ ,  $k=2 \& p=3$ ,  $k=1 \& p=4$ ,  $k=2 \& p=4$  when  $N=40$ . Although the number of pulses by  $k=2 \& p=4$  is 10, which is less than by

$k=1&p=3$  (GSM configuration), the performance of RPE is better.

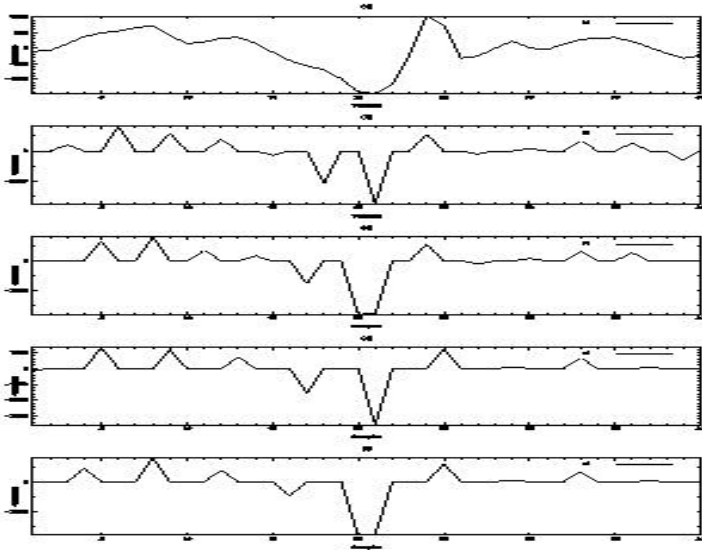


Figure 3 The pulse selection in RPE (a) original LTP residual (b) pulses selected by  $k=1&p=3$  (c) pulses selected by  $k=2&p=3$  (d) pulses selected by  $k=1&p=4$  (e) pulses selected by  $k=2&p=4$

The pulse spaces in RPE coding for voiced and unvoiced speech are different. The first reason is that RPE coding is more effective for unvoiced speech than for voiced speech. There is little influence for voiced speech if we take larger pulse spaces and suitable partition. However, if we take a smaller pulse space for unvoiced speech, the intelligibility of unvoiced speech will be greatly improved. The second reason is that the unvoiced consonants exist in most Mandarin syllables. Their intelligibility is very important for coder performance. So, for unvoiced speech, we selected  $k=1&p=2$ , rather than  $k=1&p=3$  in GSM. For voiced speech, we selected  $k=2&p=4$ .

#### E. Channel Error Control

Generally, the types of channel error in packet networks include packets missing and packets disordered. The packet disorder may be caused by protocols. For example, in connectionless-oriented packet networks, the arriving sequence of packets may be different from the leaving sequence of packets. This kind of error will cause unintelligibility of the reconstructed speech signal. The packet missing may cause more serious problems in the intelligibility of speech signal. Because of the popular utilization of segmentation methods, one speech frame may be segmented into several cells or several speech frames are combined into one packet frame. If one packet is missing in transmission, several speech signal frames will be affected.

The structure of Mandarin speech is simple, shown in Fig.4. It can be utilized to predict the packet errors in channels, or to reconstruct the order of disordered packet.

By state transfer map, some packet errors can be detected. For example, if the current frame belongs to UV and the next frame belongs to CF, then the frame belonging to BF must be missing. The method also has obvious limitations. If one or several frames belonging to certain kinds of frames missing, the method can not detect errors. However, other methods can be incorporated to enhance the method, such as using the specific tone contours of mandarin speech to predict these kinds of errors. Fig.5 presents the pitch contour of each of the first four tones.

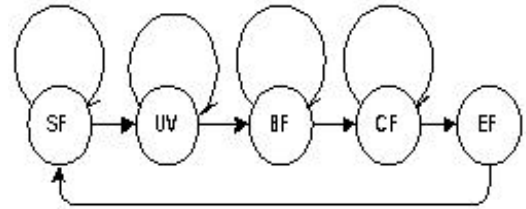


Figure 4 The state transfer map of Mandarin speech

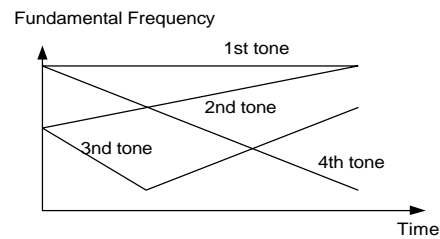


Figure 5 Standard F0 contour patterns of the first four tones

#### IV. CLTP\_RPE CONFIGURATION

In CRPE\_LTP coder, the update rate for frames is 20ms (160 samples at a sampling rate of 8K Hz). The update rate for sub-frames is 5ms. The detailed parameters for CRPE\_LTP are shown in table I.

	Method	Bit Allocation/Frame
LPC Coefficients	8 <sup>th</sup> Order LP predictor	36 bits(6,6,5,5,4,4,3,3)
Type	S/UV/V and BF/CF/EF classification	2 bits (00-UV; 01-BF; 10-CF; 11-EF) *4
LTP for voiced	Two-state Pitch Search	Lag: 7 ; Gain: 2
RPE for Voiced	Pulse Space = 4	Grid Position: 2 ; Gain: 6; Pulse Amplitude: 3*10
RPE for Unvoiced	Pulse Space = 2	Grid Position: 1 ; Gain: 6; Pulse Amplitude: 3*20
Silence Frame	The content of the frame is "1111111111111111"	

Table I The configuration of CRPE\_LTP

## V. PERFORMANCE

Comparing with the original RPE\_LTP coder, the characteristics of CRPE\_LTP are:

1. It differentiates frames according to their voicing characteristics, and processes them separately.
2. For unvoiced speech frames, RPE coding method is different from that in voiced speech frames. There is no LTP analysis. Because there is no obvious quasi-periodicity in unvoiced frames, the effect of LTP is small. Without LTP, we can save more bits for RPE coding.
3. For voiced speech frames, the frames are divided into three types of frame: Beginning Frames (BF), Continuing Frames (CF) and Ending Frames (EF). BF refer to frames located at the beginning of voiced speech. The characteristic of these frames is unobvious quasi-periodicity. CF refers to the frames located in the middle of voiced speech, characterized by obvious quasi-periodicity. EF refers to frames located at the end of voiced speech. The characteristics of EF are sometime quasi-periodicity, sometime not. BF mostly relative to the intelligibility of consonants because of coarticulation. Whilst CF and EF are helpful for the intelligibility of vowels and tones. It uses different pitch determination algorithms for frames whose characteristics are different.
4. To predict packet errors in speech transmission, the phonetic characteristics of Mandarin speech are firstly incorporated into the decoder.

We will compare the performance of CRPE\_LTP with that of GSM RPE\_LTP. CRPE\_LTP coding handles the voiced speech and unvoiced speech separately. The possible number of bits coding one frame is 16 (SF); 228 (BF/CF/EF); 308 (UV). In the GSM RPE\_LTP coding, the number of bits per frame is 260. Although the bits paid for the unvoiced speech in CRPE\_LTP are more than these in GSM RPE\_LTP, we can say that the total bit rate of CRPE\_LTP is smaller than that of GSM RPE\_LTP for almost all speech transmission. The first reason is that the length of unvoiced speech is less than that of voiced speech for one Mandarin syllable. The number of bits for voiced speech in CRPE\_LTP is much smaller than that in GSM RPE\_LTP. The increased bits for unvoiced speech could be compensated by the decreased bits for voiced speech. The second reason is that the silence frames in CRPE\_LTP only use 16 bits, far smaller than 260 bits if they are in GSM RPE\_LTP.

By subjective measurement, we compare the performance of two coding schemes. The 192 words of the Chinese Diagnostic Rhyme Test (CDRT) corpus [2] were used as the test basis. The recordings were made at Nanyang Technological University to evaluate speech compression algorithms. Experimental results show the results of test, in which the performance of CRPE\_LTP is better than that of GSM RPE\_LTP.

Table II represents the performance of CRPR\_LTP and GSM RPE\_LTP by objective measurements. Here, six words {zhan3, zan3, can3, chan3, san3, shan3} are selected,

and their average SD are computed using [11] SD criteria. From table II, the performance of CRPE\_LTP is better than that of GSM RPE\_LTP for these words.

	UV	BF	CF	Aver. SD
GSM RPE_LTP	2.26	1.61	1.64	1.84
CRPE_LTP	1.79	1.49	1.6	1.63

Table II SD performance results for GSM and CRPE\_LTP

## VI. CONCLUSION

As the Internet is developing rapidly, the use of low-bit-rate speech coding for speech transmission services will be more common than that in the traditional telephone networks. The transmission errors are more complex than before. The new conditions require that the performance of speech coders be much better. A new scheme for compression of Mandarin speech is presented in this paper. In the scheme, the allocation of bits is discriminating for different speech frames according to their characteristics.

The structure of Mandarin speech is simple and easy to be utilized to predict some errors caused by speech transmission. In this scheme, the state transfer map of phonemes and pitch contour are firstly utilized. Initial results appear promising.

For continuous Mandarin speech, the tones change very rapidly. The basic pitch contours of four tones just reflect the common change of pitch. Some specific conditions of Mandarin speech production may make the change of pitch complex [10]. More powerful algorithms need designing in the future to cater for this.

## REFERENCE

- [1] Mark E. Perkins, "Speech Transmission Performance Planning in Hybrid IP/SCN Networks", IEEE Communication Magazine, July 1999, pp.126-131.
- [2] Zongge Li, E.C.Tan, I.V.McLoughlin, T.T.Teo (2000), Proposals of Standards for Intelligibility Test of Chinese Speech, IEE Proceedings on Vision, Image and Signal Processing, .
- [3] ETSI 300 961 (1998), Digital Cellular Telecommunication System-Full rate speech-Transcoding (GSM 06.10 version 5.1.1), European Telecommunications Standards Institute.
- [4] Zhang Jialu (1994), Phonetic and Linguistic Features of Spoken Chinese, Proceeding of International Symposium on Speech, Image Processing and Neural Networks, pp. 117-121.
- [5] Sin-hong chen and Yin-ru Wang (1990), Vector Quantization of Pitch Information in Mandarin Speech, IEEE Transactions on Communications, vol. 38, No. 9, pp 1317 - 1320
- [6] E.F.Deprettere and P.Kroon ( 1986), Regular Pulse Excitation – A Novel Approach to Effective and Efficient Multipulse Coding of Speech, IEEE Transaction on ASSP, Vol ASSP-34, No.5, Oct, 1986, pp1054-1063.
- [7] Lin-Shan Lee (1997), "Voice Dictation of Mandarin Chinese", IEEE Signal Processing Magazine, July, 1997, pp63-101.
- [8] Sasan Ahmadi and Andreas S.Spanias(1999), "Cepstrum-Based Pitch Detection Using a New Statistical V/UV Classification Algorithm", IEEE Trans., on SAP, Vol.7, No.3. pp.333-338.
- [9] W.B.Kleijn and K.K.Paliwal, "Multimode and Variable-Rate Coding of Speech", In: Speech Coding and Synthesis, Elseview Science, 1995.
- [10] Lin-shan Lee, Chiu-yu Tseng, Ming Ouh-young, " The Synthesis rules in a Chinese Text-to-Speech System", IEEE Trans. ASSP. Vol.37, No.9, Sep., 1989, pp.1309-1320.
- [11] Paliwal, K.P., Atal, B.S., "Efficient vector quantization of LPC parameters at 24 bits/frame", IEEE Trans., on SAP, Vol.1, No.1, 1993, pp.3-14.