Vector Quantization of Pitch Information in Mandarin Speech

SIN-HORNG CHEN AND YIH-RU WANG

Abstract—**By taking advantage of the simple tone structure of pitch contours in Mandarin speech, pitch information is orthogonally transformed and vector quantized. An average bit rate of 0.78 bits/frame (34.67 bits/s) for voiced sounds was achieved.**

I. Introduction

In speech vocoding, parameters of a speech production model are quantized for efficient transmission or storage. They include coefficients of an all-pole filter which models the vocal tract, and gain and pitch period of an impulse-train excitation source for voiced sounds or gain of noise-like excitation for unvoiced sounds. For English speech, the pitch contour is, in general, a smooth curve which can be efficiently quantized by using coding techniques such as ADM [1]. For Mandarin speech, we can take advantage of the simple tone structure and quantize the pitch contour more efficiently by using vector quantization techniques.

Mandarin is a tone language. Each word is pronounced as a monosyllable according not only to its phonetic sign but also to its tonality. There are only five basic tones in Mandarin speech, namely, Tone 1 (high-level tone, with symbol "−"), Tone 2 (midrising tone—"/"), Tone 3 (midfalling-rising tone—"√"), Tone 4 (high-falling tone—"\"), and Tone 5 (Neutral tone—"·"). The information of the tonality of a word mainly appears on its pitch contour [2] so that we have only five basic types (shapes) of pitch contour for Mandarin words. The basic shapes of pitch contour for Tone 1 to Tone 4 are shown in Fig. 1. They are more regularly pronounced comparing with Tone 5. The pitch contour of Tone 5 varies and is influenced by its adjacent tones. However, it is much less important in Mandarin speech pronounciation because it usually sounds brief, light and is relatively infrequently used. Thus, Tone 5 is not needed to specially consider in the coding of pitch contours. Besides the phonetic factor of pitch contour shape mentioned above, in practice, variations also come from speaking habit, pitch level and even the mood of the speaker. It also depends on text content. Despite of those variations, only a very limited number of representative pitch contour patterns of words can be found in Mandarin conversation. Therefore, pitch information can be represented by the shape and the length of pitch contour segment word-by-word instead of specifying it frame-by-frame.

In segmenting the pitch contour of an utterance, a voiced/unvoiced (V/U) decision can divided the pitch contour of an utterance into segments. However, coarticulation can make the pitch contour of several contiguous words be connected together and form a single segment. Segmenting such a connected pitch segment into word-periods is usually a difficult task. Fortunately, some distinct characteristics of the structure of Mandarin speech are helpful to circumvent the obstacle of pitch contour segmentation. First, every Mandarin word is a monosyllable. It primarily consists of a vowel or diphthong nucleus which may be followed by a nasal or preceded with a consonant. The location of an unvoiced consonant which is relatively easy to detect by a V/U decision becomes a natural boundary of words. Second, words can be concatenated together to form multisyllabic phrases. Mandarin speech usually uses phrases as the basic pronunciation units. According to the statistics of a typical Mandarin dictionary, about 90% of phrases are monosyllabic or disyllabic phrases. The interphrase connection of pitch contour is usually not serious. The most common exception occurs when a phrase is followed by a word of Tone 5. But because Tone 5 is of little perceptual importance, this effect can be neglected in our coding consideration. Consequently, most multiword pitch segments to be considered are formed by connecting pitch contours of two words. Last, when the pitch contour of two words are connected together, their levels will be adjusted to form a continuous pitch segment. Besides, their shapes will be distorted slightly to make the pitch contour more smooth. From above discussions, a V/U decision is almost good enough for pitch contour segmentation and is referred to as coarse segmentation. The resulting pitch segments can hence be represented by using finite patterns.

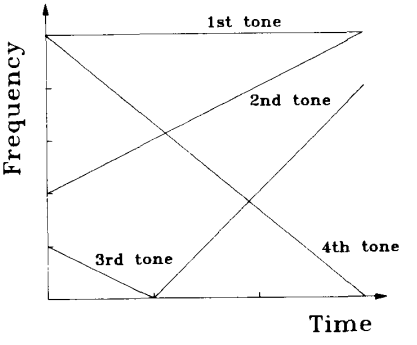

Fig. 1. Standard pitch frequency contours of four tones.

In this paper, we propose a method to quantize the shape of pitch contour segment of Mandarin speech by using orthogonal polynomial representation and vector quantization techniques.

II. Orthogonal Polynomial Representation and Vector Quantization of Pitch Contour

The speech signal is first low-pass filtered by using a 6th-order elliptic filter, sampled at an 8 KHz rate, and A/D-converted into 12 bit data. Then a modified AMDF [3] algorithm is employed to detect the pitch period for each 22.5 ms frame (180 samples). The voiced/unvoiced decision is done by using a statistical pattern recognition approach [4] with the features—energy, zero-crossing rate, normalized prediction error, and minimum value in the normalized modified AMDF curve. This V/U decision is also served to segment the pitch contour of an utterance. Pitch contour is then divided into segments. Most of them consist of one or two words. Only a few contains more than two words. Then a smoothing procedure is performed to correct all the double, triple, and half pitch errors which lie far away from the smoothed pitch contour formed by all other pitch values. The smoothing procedure is done on a segment-by-segment basis. First, the pitch mean of a segment is found. Then, the difference of pitch values in two contiguous frames is examined. If it is greater than a predetermined threshold, the one lies farther away from the mean value is treated as a double, triple, or half pitch error and corrected. The above process is done

Paper approved by the Editor for Quantization, Speech/Image Coding of the IEEE Communications Society. Manuscript received December 22, 1987; revised October 16, 1989. This work was supported by the National Science Council (NSC), Taiwan, Republic of China.

The authors are with the Department of Communication Engineering and Center for Telecommunication Research, National Chiao University, Hsinchu, Taiwan, Republic of China.

IEEE Log Number 9037884.

twice, forward and backward, for each segment in order to ensure the smoothness of pitch contour.

Due to the smoothness of a pitch contour segment, the first four discrete orthogonal polynomials are chosen as the basis functions to represent it. These polynomials are normalized, in length, to [0, 1] and can be expressed as

$$\phi_0\left(\frac{i}{N}\right) = 1$$

$$\phi_1\left(\frac{i}{N}\right) = \left[\frac{12 \cdot N}{(N+2)}\right]^{1/2}\left[\left(\frac{i}{N}\right) - \frac{1}{2}\right]$$

$$\phi_2\left(\frac{i}{N}\right) = \left[\frac{180 \cdot N^3}{(N-1)(N+2)(N+3)}\right]^{1/2} \cdot \left[\left(\frac{i}{N}\right)^2 - \left(\frac{i}{N}\right) + \frac{N-1}{6 \cdot N}\right]$$

$$\phi_3\left(\frac{i}{N}\right) = \left[\frac{2800 \cdot N^5}{(N-1)(N-2)(N+2)(N+3)(N+4)}\right]^{1/2} \cdot \left[\left(\frac{i}{N}\right)^3 - \frac{3}{2}\left(\frac{i}{N}\right)^2 + \frac{6N^2 - 3N + 2}{10 \cdot N^2}\left(\frac{i}{N}\right) - \frac{(N-1)(N-2)}{20 \cdot N^2}\right]$$

for $0 \le i \le N$ where $N+1$ is the length of the pitch contour and $N \ge 3$.

These basis functions are, in fact, discrete Legendre polynomials. They are chosen to represent the pitch contour because they resemble to the basic pitch contour patterns. A pitch contour segment, $f(i/N)$, can then be approximated as

$$\hat{f}\left(\frac{i}{N}\right) = \sum_{j=0}^{3} a_j \cdot \phi_j\left(\frac{i}{N}\right), \quad 0 \le i \le N$$

where

$$a_j = \frac{1}{N+1}\sum_{i=0}^{N} f\left(\frac{i}{N}\right) \cdot \phi_j\left(\frac{i}{N}\right).$$

For the case of a single-word segment, the reconstructed pitch contour $\hat{f}(i/N)$ for each of the four basic tones will not lose much information since orthogonal polynomials up to degree of three are used to fit it. The orthogonal representation simply smooths the original pitch contour. But for some multiword segments, the results of orthogonal expansion may be awful. A finer segmentation can be used for these segments to divide them into subsegments. These subsegments can then be better orthogonally expanded individually.

The four coefficients obtained from the above orthogonal transformation for each segment form a vector which will be quantized by using vector quantization technique. The distortion measure used in vector quantization is defined as

$$D(A, A') = (A - A')^T (A - A')$$

where A and A' are two coefficient vectors. This distortion measure is equivalent to the mean square distance defined on $\hat{f}(i/N)$ and $\hat{f}'(i/N)$. The codebook is designed using the well-known LBG algorithm [5] with binary codeword-splitting for initial codebook generation. In the encoding phase, pitch contour segments are encoded by using a full codebook search.

III. Experiment Results

Speech of 9 male and 8 female speakers was used to train the system. Each speaker spoke 10 sentences with a total length of about 40 s. There are total 1075 training pitch contour segments and the average length of segment is 14.1 frames. Among them, about 76.4% are single-word segments, 18.7% are two-word segments, and 4.9% are multiword segments. By using the orthogonal expansion, the root mean square error of the reconstructed pitch period for the training utterances is 1.95 sampling periods (0.24 ms) per frame.

TABLE I
The Root Mean Square Distortion of the Reconstructed Pitch Period Obtained by Using VQ's for Encoding $(a_0, a_1, a_2, a_3)^T$

unit : sampling period/frame

number of codewords	distortion/frame			
	coarse segmentation		fine segmentation	
	inside training set	outside training set	inside training set	outside training set
64	3.94	4.34	3.51	4.07
128	3.55	3.97	3.14	3.61

First, 6 and 7 bit VQ's were used to encode $(a_0, a_1, a_2, a_3)^T$. The root mean square errors of the reconstructed pitch period for these VQ's are listed in Table I. The results for the case of fine segmentation which will be discussed later are also listed in this table. Fig. 2 is an example showing the effectiveness of our VQ coding schemes. Fig. 2(a) displays the original pitch contour of a test utterance which is inside the training set. Orthogonal polynomial representation of this pitch contour is shown in Fig. 2(b). All pitch contour segments except the last one are fit well by the orthogonal expansion. The distortion for the last segment is perceivable although it is still acceptable. As discussed later, this can be improved by using a finer segmentation. Reconstructed pitch contours using 6 and 7 bit VQ's are drawn in Fig. 2(c) and (d), respectively. From this example we see that the quantization errors except for the first frames of both the first and the third segments were small for both 6 and 7 bit VQ's. Because these two frames correspond to unvoiced-to-voiced transitions, they do not have obvious periodicity and are perceptually less relevant. Therefore, large quantization distortions in these two frames were not perceivable in the synthesized speech.

Alternatively, the mean of each pitch contour segment (coefficient a_0) can be separated from the feature vector and encoded independently by using a 6 bit uniform scalar quantizer. Then VQ's of 4, 5, and 6 bit were used to encode $(a_1, a_2, a_3)^T$. The distortions of these VQ's are listed in Table II. Comparing the results shown in Tables I and II, significant improvements in performance were achieved for mean-separated VQ's but with higher bit rates. The reconstructed pitch contour of Fig. 2(a) using 6 bit VQ for $(a_1, a_2, a_3)^T$ is drawn in Fig. 3. By comparing to Fig. 2(c), improvements on both the first and the third segments were achieved in Fig. 3.

It should be noted that the second and the last segments in pitch contour of Fig. 2(a) represent the connected pitch contour of a disyllabic phrase. Nevertheless, the quantization result is good for the second segment and acceptable for the last. Although the V/U decision cannot divide some pitch contour segments of a sentence into word-periods, the quantization result is still acceptable. In Fig. 2(b), the last segment was distorted after orthogonal expansion. The first four discrete Legendre polynomials could not fit the pitch contour well. A finer segmentation can be taken in order to make a better reconstructed pitch contour.

In finer segmentation, the original pitch contour was segmented when the distortion of the orthogonal transformed pitch contour was greater than a threshold value. A boundary point was detected by finding the maximum distortion in the central part of the connected segment. Although this procedure cannot divide all multiword segments into word-periods, the pitch contour segments will become smoother such that the distortion due to orthogonal transformation will be reduced. In our experiment, the number of pitch contour segments increases to 1145 for the training utterances using the fine segmentation. The average length was accordingly reduced to 13.24 frames/segment and the root mean square distortion of orthogonal expansion was reduced to 1.38. The resulting distortion of VQ-

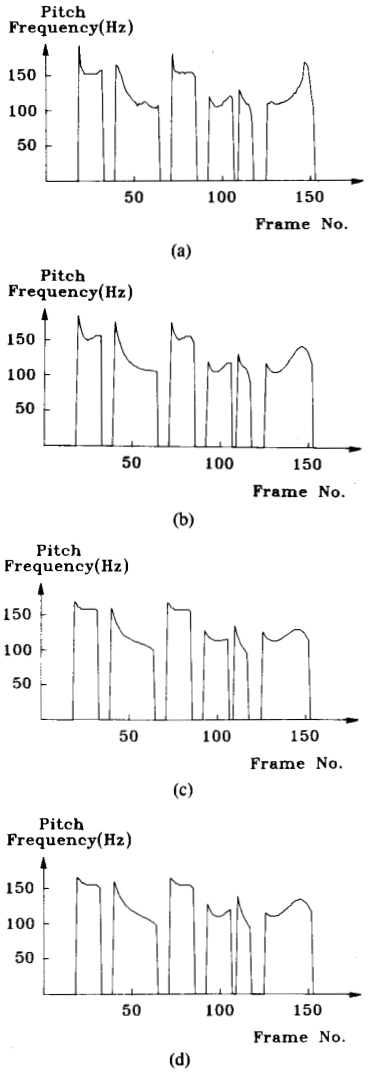

Fig. 2. An example of orthogonal expansion and VQ: (a) the original pitch contour; (b) the reconstructed pitch contour using orthogonal polynomial expansion; and the reconstructed pitch contours using (c) 6 bit and (d) 7 bit VQ's. (The sentence is "[ta –][tzuo\ le ·][jin –][chian/][de ·] [nwu/li \]." Every [ ] represents one pitch contour segment.)

TABLE II
THE ROOT MEAN SQUARE DISTORTION OF THE RECONSTRUCTED PITCH PERIOD OBTAINED BY USING VQ'S FOR ENCODING $(a_1, a_2, a_3)^T$ AND A SEPARATE 6 BIT QUANTIZER FOR a_0

unit : sampling period/frame

number of codewords	distortion/frame			
	coarse segmentation		fine segmentation	
	inside training set	outside training set	inside training set	outside training set
16	2.78	3.38	2.41	3.18
32	2.48	3.15	2.13	2.94
64	2.28	3.00	1.93	2.66

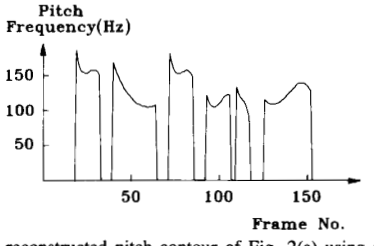

Fig. 3. The reconstructed pitch contour of Fig. 2(a) using a 6 bit VQ to encode the vector $(a_1, a_2, a_3)^T$ and a 6 bit scalar quantizer to encode a_0 for each pitch contour segment.

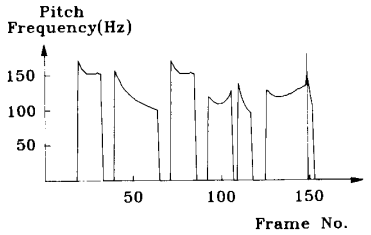

Fig. 4. The reconstructed pitch contour of Fig. 2(a) using a 7 bit VQ to encode $(a_0, a_1, a_2, a_3)^T$ for each pitch contour segment obtained by coarse and fine segmentations. The vertical line in the last segment is used to indicate the boundary of two subsegments obtained by fine segmentation.

TABLE III
THE AVERAGE BIT RATES OF CODERS THAT USE VQ'S FOR ENCODING $(a_0, a_1, a_2, a_3)^T$

number of codewords	average bite rates(bits/frame)	
	coarse segmentation	fine segmentation
64	0.78	0.831
128	0.85	0.906

quantized pitch contours after both coarse and fine segmentations is listed in Tables I and II. Obviously, better performance is obtained by the fine segmentation. Fig. 4 shows the reconstructed pitch contour of Fig. 2(a) using fine segmentation and 7 bit VQ for $(a_0, a_1, a_2, a_3)^T$. It is better than Fig. 2(d). However, discontinuity exists at the boundary point of the last segment in Fig. 4. Because energy around the word boundary points is usually small, the distortion of the reconstructed pitch contour segment is barely perceivable.

A 40 s speech segment of another male speaker outside the training group was also used to test the performance of the procedure. The average distortions are also listed in Tables I and II. By informal listening test, the quality of the reconstructed speech is still good.

In order to completely represent the pitch contour, run-length coding was employed to encode the length information. The length of unvoiced segment is not considered here. For encoding the length information of a voiced segment, 5 bits were needed (4–35 frames linearly). Therefore a total 11–12 bits were used to encode the pitch information for each word in our VQ schemes. The average bit rates are shown in Table III.

Finally, a 10th-order LPC vocoder was implemented to check the performance of this pitch contour coding scheme. For simplicity, the all-pole filter coefficients were not quantized. When the above VQ schemes were used, informal listening tests showed that no distinguishable degradation in speech quality can be perceived for most speakers.

IV. CONCLUSION

In this paper, the tone structure of pitch contour in Mandarin speech has proved to be useful for coding the pitch information. Orthogonal transformation and vector quantization are employed to efficiently encode the pitch contour segment-by-segment instead of frame-by-frame. Bit rates of 0.78 bits/frame (34.67 bits/s) for voiced sounds were achieved in our experiments. It is a variable-rate coding scheme with an average delay of 317 ms.

REFERENCES

[1] S. Roucos, R. Schartz, and J. Makhoul, "Segment quantization for very-low-rate speech coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982, pp. 1565–1568.

[2] L. S. Lee, C. Y. Tseng, J. Huang, and K. J. Chen, "Digital synthesis of Mandarin speech using its special characteristics," *J. Chinese Inst. Eng.*, vol. 6, pp. 107–115, Mar. 1983.

[3] Y. R. Wang, "LPC vocoder using vector quantization in pitch contour," Master thesis, National Chiao Tung Univ., Taiwan, 1987.

[4] B. S. Atal and L. R. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with application to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 201–212, June 1976.

[5] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.