Front. Electr. Electron. Eng. China 2008, 3(2): 204–207
DOI 10.1007/s11460-008-0043-1

RESEARCH ARTICLE

Xiaoyan DANG, Kun TANG

DCT_M model for excitation parameter in low bit rate vocoder

Abstract The description precision of an excitation signal greatly influences the quality of reconstructed speech in low bit rate vocoders. To improve the reconstruction quality, the DCT_M model is proposed to express the excitation spectral parameter, which transforms the variable length vector to fixed dimension vector through DCT transformation. It then quantizes the fixed length vector using multi-stage vector quantization. Tests show that the proposed method can keep the shape of the entire spectral envelope and reduce model error thus greatly improve the description precision. Test results in the sine excitation linear prediction (SELP) vocoder show that the DCT_M model can improve the naturalness of reconstructed speech, with subjective test score of 65%.

Keywords vocoder, DCT_M model, model error

1 Introduction

Most low bit rate speech codecs use parameter coding, including vocal track and excitation parameter. The description precision of the excitation parameter influences the naturalness of reconstructed speech greatly, wherein the bottleneck of the excitation expression is the description of excitation spectral amplitude because its length is variable and changes with pitch. The excitation spectral description is a variable length vector quantization problem. This paper focuses on the problem of the variable length excitation spectral amplitude parameter model and quantization.

Variable length vector is generally changed into a fixed length and quantized thereafter. Although there are many dimension reducing methods, most low bit rate speech codecs like the sine excitation linear prediction (SELP) [1] use the prior-10 model; some others bring forward the 2-stage [2] model to improve the prior-10 model, considering the whole spectral shape that makes sense. Current methods like Li's WNSTVQ [3] and others proposed in Refs. [4, 5] also handle this problem well.

The quantization of the excitation spectral parameter is very important in low bit rate vocoders. The commonly used prior-10 model is used by vocoders like SELP for its convenience and its accords to human hearing properties. However, the prior-10 model only conducts simple truncation to the variable vector, losing all the intermediate and high frequency information. The 2-stage model offsets the error of prior-10 model to some extent, but it still stays in the area of linear truncation. Its simple but coarse dimension transformation method cannot totally resolve the problem of the spectral shape difference compared to the original spectral shape. On the other hand, the WSNVQ method is too complicated to put into practice.

We propose the DCT_M model here to describe the excitation spectral amplitude. The DCT_M model uses the DCT model to change the variable length to M, and then quantize it. Testing shows that the DCT_M model is simple and can preserve the whole spectral shape very well. Compared to prior-10 model and the 2-stage model, the description precision is greatly improved while the excitation error is reduced, thus finally the subjective quality is improved.

2 DCT_M model

The process of the DCT_M model is as follows (Fig. 1). The encoder uses DCT transformation to change the variable length vector into a fixed length M, and then quantizes the fixed length M dimension vector. The decoder unquantizes and reconstructs the variable length vector using IDCT transformation.

2.1 Encoding model

Suppose $\boldsymbol{x}$ is the excitation spectral amplitude vector. Its dimension N changes with the pitch signal P. Under the sampling rate of 8 KHz, N is equal to $P/2$. At the encoding end, we use the DCT_M model to change $\boldsymbol{x}$ into M

Translated from *Journal of Tsinghua University* (*Science and Technology*), 2007, 47(4): 578–580 [译自: 清华大学学报 (自然科学版)]

Xiaoyan DANG (✉), Kun TANG
State Key Laboratory on Microwave and Digital Communications, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
E-mail: dxy03@mails.tsinghua.edu.cn

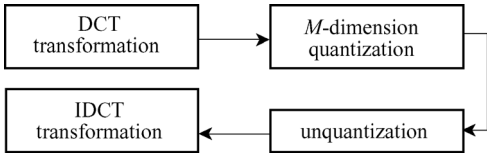

Fig. 1 DCT_M model

dimension vector $\boldsymbol{x}_M$, and then quantize it. Suppose $\boldsymbol{A}_{N\times M}$ is the dimension transformation matrix used in the DCT_M model, which is automatically created according to the pitch signal. We can get the fixed length vector $\boldsymbol{x}_M$ according to $\boldsymbol{x}_M = \boldsymbol{A}^{\mathrm{T}} \times \boldsymbol{x}$, and then $\boldsymbol{A}$ can be expressed as:

$$\boldsymbol{A} = \begin{cases} (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_M), & N \geqslant M \\ (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N | \boldsymbol{0}), & N < M \end{cases}, \tag{1}$$

where $\boldsymbol{0}$ is a zero matrix, $\boldsymbol{a}_i$ is the radix of the diagonal transformation:

$$a_{i,n} = \begin{cases} \sqrt{\dfrac{2}{N}} \cos \dfrac{(2(n-1)+1)\pi(i-1)}{2N}, & i = 1, \\ \sqrt{\dfrac{1}{N}} \cos \dfrac{(2(n-1)+1)\pi(i-1)}{2N}, & i = 2,3,\ldots,M, \end{cases}$$
$$n = 1, 2, \ldots, N. \tag{2}$$

$\boldsymbol{A}$ is an M-order dimension-cut matrix when $N \geqslant M$, and an N-order dimension-expand matrix when $N < M$.

2.2 Fixed length vector quantization

The variable length vector $\boldsymbol{x}$ is transformed to the fixed length M using DCT_M, where subscript M represents the fixed length after DCT_M transformation. The multi-stage vector quantization (MSVQ) is used to quantize the fixed length vector $\boldsymbol{x}_M$, and the quantized vector is $\boldsymbol{y} = Q(\boldsymbol{x}_M)$. We use simulating annealing method on a large database to get the MSVQ codebook.

2.3 Decoding model

Suppose that the vector derived from the unquantization at the decoder is $\boldsymbol{y}$. Create transformation matrix $\boldsymbol{B}$ with the transmitted pitch parameter P. If the pitch parameter is transmitted without error, we can find that $\boldsymbol{B} = \boldsymbol{A}$, then we can get the reconstructed excitation spectral amplitude variable vector using $\boldsymbol{x} = \boldsymbol{B} \cdot \boldsymbol{y}$.

2.4 Error analysis

Suppose that the process of dimension fixing and the fixed length VQ is independent of each other. We can find that total error D_{m} can be divided into 2 parts of model error D_{m} and quantization error D_{q} expressed as $D = D_{\mathrm{m}} + D_{\mathrm{q}}$.

The model error D_{m} is imported in dimension truncation when fixing the vector length and the number of D_{m} is decided by transformation matrix $\boldsymbol{A}$. However, quantization error D_{q} is imported in quantizing the fixed length vector. The description precision of the aimed vector can be analyzed in terms of model effectiveness and quantization effectiveness, and we will only focus on the model error in this paper.

3 Comparison of several models

3.1 Description precise analysis of single frame

We select the fixing length of M of 30 in the testing. Under this setting, when the original vector length of the excitation spectral amplitude parameter is less than 30, we can recover the variable length vector losslessly because its model error is 0; when the vector length is larger than 30, we can recover the variable length vector with high precision but rather lossless.

Figure 2 shows three different description models on the same frame excitation signal. Here we use the modeled value without quantization process to get rid of the difference brought by the quantization error. To let the figure make sense, here we choose a frame with the excitation spectral amplitude length of $N = 44$, which means that the model error is bigger than 0 and the DCT_M model is approaching the original spectra with some errors. We can imagine that if we choose a frame whose length $N \leqslant 30$, the model error is 0, which will make the DCT_M model totally overlap the original spectra, which is unfavorable to observe.

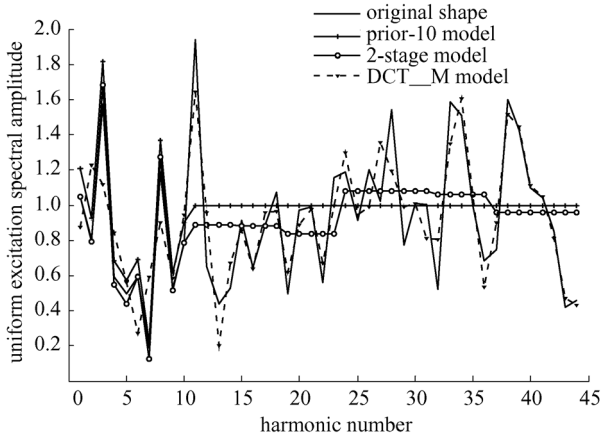

Fig. 2 Comparison of prior-10, 2-stage and DCT_M model

We can see from Fig. 2 that these three models almost overlap at those points whose harmonic number order is less than 10, and all these three models can approach the original waveform excellently. However, at those points whose order is larger than 10, the prior-10 model exhibits a straight line; the 2-stage model shows a laddering shape because of its compensation method of segment mean value. The proposed DCT_M model can trace the original waveform very closely, whose description precision is higher than the former prior-10 model and the 2-stage model.

3.2 Statistical description precise analysis

For statistical analysis, we calculate the statistical mean error using Eq. (3).

$$E_{\mathrm{m}}(k)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}w_i(\boldsymbol{a}_i-\tilde{\boldsymbol{a}}_i)^2},$$
$$\bar{E}_{\mathrm{m}}=\frac{1}{N_f}\sum_{k=1}^{N}E_{\mathrm{m}}(k), \tag{3}$$

where $\boldsymbol{a}_i$ is the original variable length vector; $\tilde{\boldsymbol{a}}_i$ is the modeled but not quantized vector; N is the single frame harmonic number;N_f is the total frame number; E_{m} is single frame harmonic error; $\bar{E}_{\mathrm{m}}$ is mean error; w_i denotes perception weighting coefficients, which cater to the properties of the human ear, sensitive to low frequency and insensitive to high frequency. w_i can be obtained from Eq. (4):

$$w_i=\left\{\frac{117}{25+75\left[1+1.4(8i/P)^2\right]}\right\}^2. \tag{4}$$

Statistical error distributions of these three models are shown in Fig. 3.

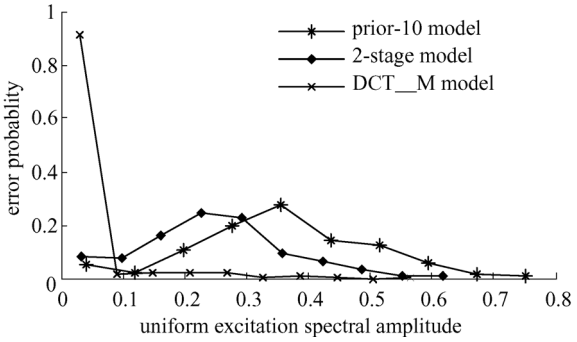

Fig. 3 Statistical error distribution of three models

In Fig. 3 above, the model errors of the 2-stage model gather in the much lower error field compared to the prior-10 model, which behaves as the shifting left of the statistical line's peak area. Furthermore, the statistical errors of the DCT_M model gathers in the very low model error field, with peak area more in the left than the other 2 models. We find that the mean and variance value of the DCT_M model error is reduced greatly, which indicates that the description precision is improved greatly.

4 Testing result

We use a speech database of 107 MB to extract the excitation spectral amplitude parameter, carry out DCT_M transformation to fix their length to be $M = 30$, and then use simulation annealing method to train the 2-stage codebook. We use this 2-stage codebook in the quantization of SELP2.4 kb/s vocoder excitation spectral amplitude parameter, and calculate the spectral distortion (SD) and signal-to-noise ratio (SNR) value according to Eq. (5).

$$\mathrm{SD}=\frac{1}{K}\sum_{k}\sqrt{\frac{1}{N_k}\sum_{n=0}^{N_k-1}\left[10\lg\frac{\boldsymbol{x}_k^2}{\hat{\boldsymbol{x}}_k^2}\right]^2},$$
$$\mathrm{SNR}=\frac{1}{K}\sum_{k}\sqrt{\frac{\sum_{n=0}^{N_k-1}\boldsymbol{x}_k^2[n]}{\sum_{n=0}^{N_k-1}(\boldsymbol{x}_k[n]-\hat{\boldsymbol{x}}_k[n])^2}}, \tag{5}$$

where $\boldsymbol{x}_k$ is the original variable length vector; $\hat{\boldsymbol{x}}_k$ is the reconstructed vector; K is the total frame number in the database; N_k is the original vector length; $\boldsymbol{x}_k[n]$ is the nth dimension of $\boldsymbol{x}_k$ and $\hat{\boldsymbol{x}}_k[n]$ is the nth dimension of $\hat{\boldsymbol{x}}_k$.

SD and SNR comparison testing among the prior-10 model, 2-stage model, and the DCT_M model with fixed vector length of 20, 30, and 40 are conducted. To guarantee the fairness of testing, the 2-stage MSVQ with a (8, 8) codebook in the prior-10 model is used. The whole bandwidth error in the 2-stage model is transformed into a 10-order vector, then the prior-10 vector and 10-order error compensation vector is quantized using 2 different 8 bit codebooks, forming a (8, 8) codebook. The DCT_M model quantization uses a (8, 8) 2-stage MSVQ codebook. The 3 models all use (8, 8) size codebook, with the same quantization precision. The testing result is shown in Table 1.

Table 1 SD and SNR comparison of three models

models	SD/dB	SNR/dB	space/word
DCT_ 10 model	3.147	18.735	5120
DCT_ 20 model	2.946	20.603	10240
DCT_ 30 model	2.905	21.181	15360
DCT_ 40 model	2.903	21.019	20480
prior-10 model	4.161	16.725	5120
2-stage model	3.630	18.347	5120

We compromise between SD, SNR and the storage and finally choose DCT_30 to be the optimal algorithm.

4.1 A/B testing for subjective naturality

To get rid of the influence of the other parameters' quantization error, we do not quantize other parameters, but using the extracted value in testing 2.4 kb/s SELP vocoder and we only quantize the DCT_M modeled excitation spectral amplitude parameter. The testing material is a 20-s pure speech, including male and female voices. Ten different listeners are selected to listen to the reconstructed speech, and the testing result is shown in Table 2.

Table 2 A/B testing result of subjective naturalness

	prior-10/person	DCT_30/person	DCT_30 score/%
female	4	6	60
male	3	7	70
total	7	13	65

In Table 2, 65% of the listeners prefer the reconstructed speech using DCT_30 model. We find that DCT_M can improve the subjective naturalness quality of vocoders.

5 Conclusions

This paper focuses on the problem of the description model of the excitation spectral amplitude parameter in low bit-rate vocoders and the DCT_M model is proposed to improve it. Testing results show that this model can greatly reduce model errors, reserve the full-band spectral shape, achieving more precise description than the commonly used prior-10 model and the improved 2-stage model. Subjective testing in 2.4 kb/s vocoder shows that the DCT_M model can improve reconstructed speech quality and gain a subjective A/B score of 65%.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 60272020).

References

1. Li J L. Research on Low Bit Rate Speech Coding Algorithm. Dissertation for the Master Degree. Beijing: Tsinghua University, 2004 (in Chinese)
2. He T H, Zhang J W, Cui H J, et al. A model for describing the excitation parameters of vocoder. Audio Engineering, 2003, (4): 52–55 (in Chinese)
3. Li C, Lupini P, Shlomot E, et al. Coding of variable dimension speech spectral vectors using weighted nonsquare transform vector quantization. IEEE Transactions on Speech and Audio Processing, 2001, 9(6): 622–631
4. Nishiguchi M, Matsumoto J, Wakasuki R, et al. Vector quantized MBE with simplified V/UV division at 3.0 kbit/s. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 1993, 2: 151–154
5. Das A, Rao A V, Gersho A. Variable dimension vector quantization of speech spectra for low rate vocoders. In: Proceedings of the Data Compression Conference, 1994, 420–429