Journal of Southeast University (English Edition) Vol. 25, No. 2, pp. 152 – 155 June 2009 ISSN 1003—7985

Frame erasure concealment in wideband speech coding based on large hidden Markov model

Wang Shikui[1, 2] Tang Yibin[1] You Hongyan[1] Wu Zhenyang[1]

([1] School of Information Science and Engineering, Southeast University, Nanjing 210096, China)
([2] School of Physics and Electronic Information, Anhui Normal University, Wuhu 241000, China)

Abstract: Frame erasure concealment is studied to solve the problem of rapid speech quality reduction due to the loss of speech parameters during speech transmission. A large hidden Markov model is applied to model the immittance spectral frequency (ISF) parameters in AMR-WB codec to optimally estimate the lost ISFs based on the minimum mean square error (MMSE) rule. The estimated ISFs are weighted with the ones of their previous neighbors to smooth the speech, resulting in the actual concealed ISF vectors. They are used instead of the lost ISFs in the speech synthesis on the receiver. Comparison is made between the speech concealed by this algorithm and by Annex I of G. 722. 2 specification, and simulation shows that the proposed concealment algorithm can lead to better performance in terms of frequency-weighted spectral distortion and signal-to-noise ratio compared to the baseline method, with an increase of 2. 41 dB in signal-to-noise ratio (SNR) and a reduction of 0. 885 dB in frequency-weighted spectral distortion.

Key words: frame erasure concealment; wideband speech; large hidden Markov model; immittance spectral frequency (ISF) parameter

In the transmission of speech signals on the Internet, some packets may be lost. We may apply the commonly used methods to deal with the packet loss, namely, automatic retransmission request (ARQ) and forward error correction (FEC). Due to the instant property of voice communication, ARQ is not applicable to this problem because it takes extra time. FEC needs to increase the transmission bandwidth, and it is also not practicable because of the limited bandwidth on the Internet currently. So frame erasure concealment is usually made by reconstructing the lost frame parameters of coding speech on the receiver.

The simplest receiver-based frame erasure concealment is to repeat the parameters of the previous frame, and it makes use of the short-term stability of speech . Ehsan and Kubin[1] put forward a measure of the short-term stability of speech. The repetition-based approach is straightforward, but it does not utilize the statistical evolution of speech and thus has poor effect. Vaillancourt et al. [2] presented a method for re-synchronizing the glottal pulse after an erased frame. The method can be applied with or without additional side information. Thyssen et al. [3] brought forward methods to update the G. 722 subband decoder state memory during frame erasure.

In this paper, a large hidden Markov model (LHMM) is used to model the immittance spectral frequency (ISF) qarameters in the AMR-WB speech coder. The lost ISFs are optimally estimated based on the MMSE rule. The speech segments concealed with different methods are compared, and simulation shows that this algorithm can lead to better performance in terms of frequency-weighted spectral distortion and signal-to-noise ratio compared with the baseline method.

1 Baseline Frame Erasure Concealment

In AMR-WB G. 722. 2[4], every speech segment of 20 ms is divided into four subframes to be analyzed, and parameters are obtained for each subframe, including linear prediction (LP) filter coefficients, adaptive codebooks and their gains, fixed codebooks and their gains, etc. The LP filter coefficients are converted to the immittance spectral pair (ISP) representation for quantization and interpolation purposes. The LP filter coefficients are quantized using the ISP representation in the frequency domain, that is, ISF. The ISF vector is given by $\boldsymbol{f}^{\mathrm{T}} = \{f_0, f_1, \dots, f_{15}\}$ with T denoting transpose.

The frame erasure concealment in AMR-WB G. 722. 2 Annex I[5] is used as the baseline. It is based on the state machine as shown in Fig. 1.

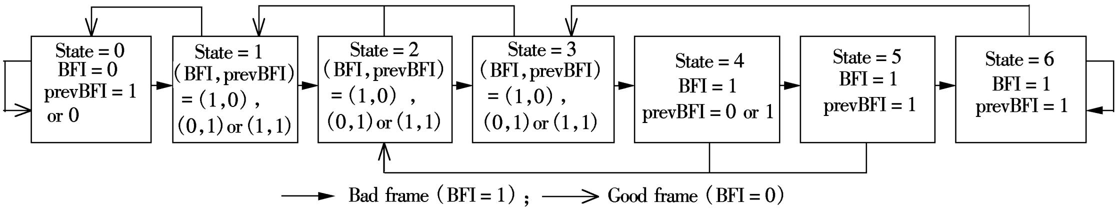

State = 0
BFI = 0
prevBFI = 1 or 0
State = 1
(BFI, prevBFI) = (1,0), (0,1) or (1,1)
State = 2
(BFI, prevBFI) = (1,0), (0,1) or (1,1)
State = 3
(BFI, prevBFI) = (1,0), (0,1) or (1,1)
State = 4
BFI = 1
prevBFI = 0 or 1
State = 5
BFI = 1
prevBFI = 1
State = 6
BFI = 1
prevBFI = 1
Bad frame (BFI = 1); Good frame (BFI = 0)

Fig. 1 State machine

The system starts in state 0. Once a bad frame is detected, the state counter is incremented by one and is saturated when it reaches state 6. And once a good speech frame is detected, the state counter is decreased by one. When BFI = 1 (BFI: bad frame identifier), prevBFI = 0 or 1, state = 1, 2, …, 6, an error is detected in the received speech frame and the frame erasure concealment is started by

Received 2008-11-04.

Biographies: Wang Shikui (1970—), male, graduate; Wu Zhenyang (corresponding author), male, professor, zhenyang@ seu. edu. cn.

Foundation items: The Science Foundation of Southeast University (No. XJ0704268), the Natural Science Foundation of the Education Department of Anhui Province (No. KJ2007B088).

Citation: Wang Shikui, Tang Yibin, You Hongyan, et al. Frame erasure concealment in wideband speech coding based on large hidden Markov model [J]. Journal of Southeast University (English Edition), 2009, 25 (2): 152 – 155.

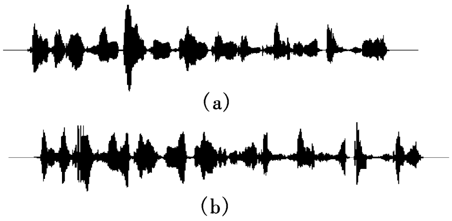
$$\mathbf{ISF}_{\mathrm{q}}(i) = \alpha\mathbf{past\_ISF}_{\mathrm{q}}(i) + (1-\alpha)\mathbf{ISF}_{\mathrm{mean}}(i) \quad i = 0, 1, \ldots, 16 \tag{1}$$

where $\alpha = 0.9$; $\mathbf{ISF}_{\mathrm{q}}(i)$ is an ISF vector for a current frame; $\mathbf{past\_ISF}_{\mathrm{q}}(i)$ is an ISF vector from the previous frame. $\mathbf{ISF}_{\mathrm{mean}}(i)$ is a combination of adaptive mean and constant mean ISF vectors in the following manner:

$$\mathbf{ISF}_{\mathrm{mean}}(i) = \beta\,\mathbf{ISF}_{\mathrm{const\_mean}}(i) + (1-\beta)\mathbf{ISF}_{\mathrm{adaptive\_mean}}(i) \quad i = 0, 1, \ldots, 16 \tag{2}$$

where $\beta = 0.75$; $\mathbf{ISF}_{\mathrm{adaptive\_mean}}(i) = \frac{1}{3}\sum_{i=0}^{2}\mathbf{past\_ISF}_{\mathrm{q}}(i)$ and is updated whenever BFI = 0; $\mathbf{ISF}_{\mathrm{const\_mean}}(i)$ is a vector containing long time average of ISF vectors.

AMR-WB speech coding is based on code-excited linear prediction (CELP). Due to the prediction property, error propagation will occur when some frames are lost. Fig. 2 shows that, when the frame erasure rate is 20%, the waveforms of the concealed speech segment and the original one are clearly different. There are annoying artifacts in the concealed speech.


(a)

(b)

Fig. 2 Comparison between the original speech and the concealed speech. (a) Original speech; (b) Speech concealed by Annex I

2 LHMM-Based Frame Erasure Concealment

The application of the HMM was first introduced on speech recognition in detail by Rabiner and Juang[6]. Rodbro et al. [7] applied the HMM in the frame erasure concealment in VoIP. A sinusoidal analysis-synthesis model is employed, and there is no error propagation in this model.

In this paper, the states of the HMM may be as many as 64 to 128, so the HMM is called a large hidden Markov model. It has the continuous distribution functions of observation ISFs. The LHMM was first successfully applied in phoneme recognition and very low-rate speech coding[8-9].

2.1 Concealment of ISFs

Due to the prediction property of ISFs, we can only make use of the ISPs, which have been received correctly, to estimate the lost ones. Assume that $t-1$ ISFs, $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}$, have been received correctly. The $t-1$ ISFs are used to estimate the t-th lost ISF. First, we compute the conditional probability $p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1})$ as

$$p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) = \sum_{n=1}^{N} p(\boldsymbol{\varphi}_t, S_t = n \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) = \sum_{n=1}^{N} p(\boldsymbol{\varphi}_t \mid S_t = n) p(S_t = n \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) \tag{3}$$

The second equality is due to the first-order Markov assumption, with N denoting the state number. A forward variable is defined as

$$\alpha_t(n) = p(S_t = n \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) \tag{4}$$

and a backward variable as

$$\beta_t(n) = p(S_t = n, \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}, \boldsymbol{\varphi}_t) \tag{5}$$

They are the different states of the HMM. When $\beta_{t-1}(n)$ $(n = 1, 2, \ldots, N)$ are available,

$$\alpha_t(n) = \sum_{m=1}^{N} a_{mn}\beta_{t-1}(m) \tag{6}$$

where a_{mn} is the transition probability, and $a_{mn} = p(S_t = n \mid S_{t-1} = m)$.

With similar manipulation, when $\alpha_t(n)$ and $\boldsymbol{\varphi}_t$ are available, $\beta_t(n)$ can be calculated by

$$\beta_t(n) = p(S_t = n \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_t) = \frac{p(S_t = n \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) p(\boldsymbol{\varphi}_t \mid S_t = n)}{\sum_{n=1}^{N} p(S_t = n \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) p(\boldsymbol{\varphi}_t \mid S_t = n)}$$

namely

$$\beta_t(n) = \frac{\alpha_{t-1}(n) p(\boldsymbol{\varphi}_t \mid S_t = n)}{\sum_{n=1}^{N} \alpha_{t-1}(n) p(\boldsymbol{\varphi}_t \mid S_t = n)} \tag{7}$$

Assuming that $\alpha_t(n)$ is the initial state distribution, with Eq. (6) and Eq. (7), we may obtain $\alpha_t(n)$ recursively.

In Eq. (3), the conditional probability, $p(\boldsymbol{\varphi}_t \mid S_t = n)$, can be acquired through a training procedure. It is modeled with the Gaussian mixture model as

$$p(\boldsymbol{\varphi}_t \mid S_t = n) = \sum_{k=1}^{M} c_{nk} N(\boldsymbol{\varphi}_t, \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}) \tag{8}$$

where M denotes the number of mixture Gaussian functions, which is usually confined in the range of 2 to 10. c_{nk} is the weighted coefficient of the k-th Gaussian function, with $\sum_{k=1}^{M} c_{nk} = 1$. $\boldsymbol{\mu}_{nk}$ and $\boldsymbol{\Sigma}_{nk}$ are the mean and covariance matrices of the k-th Gaussian function, and $\boldsymbol{\Sigma}_{nk}$ is often chosen to be diagonal. Parameter set $\{c_{nk}, \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}\}$ is trained with the EM algorithm as

$$\{\hat{c}_{nk}, \hat{\boldsymbol{\mu}}_{nk}, \hat{\boldsymbol{\Sigma}}_{nk}\} = \arg\max_{c_{nk}, \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk}} E_{\mathrm{S}}[\log p(\boldsymbol{\varphi}_t / c_{nk}, \boldsymbol{\mu}_{nk}, \boldsymbol{\Sigma}_{nk})] \tag{9}$$

On the condition that the past $t-1$ ISFs, $\boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}$, are received, the density function of the lost ISF, $\boldsymbol{\varphi}_t$, is determined by

$$p(\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}) = \sum_{n=1}^{N} [\alpha_t(n) \sum_{k=1}^{M} \hat{c}_{nk} N(\boldsymbol{\varphi}_t, \hat{\boldsymbol{\mu}}_{nk}, \hat{\boldsymbol{\Sigma}}_{nk})] \tag{10}$$

Based on the MMSE rule, the optimal estimation of $\boldsymbol{\varphi}_t$ is

$$\tilde{\boldsymbol{\varphi}}_t = E[\boldsymbol{\varphi}_t \mid \boldsymbol{\varphi}_1, \ldots, \boldsymbol{\varphi}_{t-1}] = \sum_{n=1}^{N} \alpha_t(n) (\sum_{k=1}^{M} \hat{c}_{nk} \hat{\boldsymbol{\mu}}_{nk}) \tag{11}$$

In order to smooth the decoded speech, the estimated ISF is weighted with that of the previous frame:

$$\hat{\boldsymbol{\varphi}}_t = 0.8\boldsymbol{\varphi}_{t-1} + 0.2\tilde{\boldsymbol{\varphi}}_t \tag{12}$$

2.2 Computational complexity and memory requirements

We assume that two consecutive frames at most are lost. The number of Gaussian mixture functions is $M=3$ and the state number $N=64$. Six correctly received frames are used to estimate the lost parameters. The HMM model can be trained off-line, so it is not included in the computation. The computational complexity and memory requirements are estimated as below:

The majority of the calculation comes from computing $\alpha_t(n)$ and $\beta_t(n)$ recursively. The multiplication amounts to about $N^2D(k_1+k_2)$, and an order of magnitude is 10^6 in this scenario, where D denotes the dimension. The computational complexity in this paper is much higher than that in the baseline algorithm, while the extra algorithmic delay does not hinder ordinary voice communication. Memory consumption results from three training parameters: transition matrix $\boldsymbol{A}$ (N^2 words), mean-value vector $\boldsymbol{\mu}_{nk}$ (MND words), and covariance matrix $\boldsymbol{\Sigma}_{nk}$ (MN^2D words). Add in some temporary computation results, and then the total memory requirement is no more than 3×10^5 words.

3 Evaluation Standards

Two standards are used to compare the proposed algorithm with the baseline:

1) Signal-to-noise ratio(SNR) SNR is used to evaluate the speech difference between the processed speech and the initial speech,

$$\mathrm{SNR} = 10\lg \frac{\sum_{n=1}^{N} x_n^2}{\sum_{n=1}^{N} (x_n - \hat{x}_n)^2} \tag{13}$$

where N is the number of speech samples. As for a segment speech of T s, $N = 16\,000T$; x_n is the sample amplitude of the processed speech, and $\hat{x}_n$ is the amplitude of the concealed speech.

2) Spectral distortion(SD) The other standard is the perceptually-based frequency-weighted spectral distortion[10]. With a small modification, SD between the concealed and the initial ISFs of wideband speech is obtained by

$$\mathrm{SD}_{\mathrm{fw}}(A_c(z), A(z)) = \sqrt{\frac{1}{W_0'}\sum_{f=0}^{7\,000} |W_B(f)|^2 \left| 10\lg\frac{|A_c(f)|^2}{|A(f)|^2} \right|^2} \tag{14}$$

where $A_c(z)$ and $A(z)$ are the concealed and the initial LPC filters; 7 000 denotes the speech bandwidth(Hz); W_0' is the sum of the weighting factors, and the weights use the Bark weighting defined by

$$W_B(f) = \frac{1}{25 + 75\left[1 + 1.4\left(\frac{f}{1\,000}\right)^2\right]^{0.69}} \tag{15}$$

4 Simulation

4.1 Gilbert channel

The Gilbert channel[11] (see Fig. 3) is adopted to model the Internet channel. P, p, Q, q and e are given different values to describe the channels. G indicates good state, and B indicates bad state. P, p, Q and q are the transition probabilities between states, with $Q+P=1$ and $p+q=1$. e is the frame loss ratio.

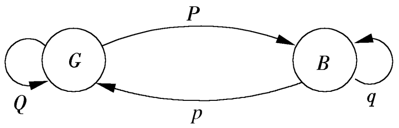


Fig. 3 Gilbert channel model

The parameter configurations of the Gilbert model in this simulation are set in Tab. 1. In the table, the average frame erasure ratio, f_{erasure}, is calculated by

$$f_{\mathrm{erasure}} = \frac{P}{P+p}e \tag{16}$$

Tab. 1 Parameter configurations of the Gilbert channel

Configuration	P	p	e	f_{erasure}/%
1	0.1	0.2	0.6	15.0
2	0.2	0.3	0.6	20.0

4.2 Simulation results

In order to reduce the calculation and accurately describe the ISFs, we choose $N=64$. For simplicity, no more than two consecutive frames are lost. We determine $M=3$. The LHMM is trained in advance. Since there is no ready-made wideband speech database, we use English speech of about 30 min downloaded from the Internet to train. After training, the parameters of the HMM are obtained. It can be seen that the transition matrix dominates diagonally.

In the speech synthesis on the receiver, for the purpose of comparison, all the other lost parameters except the ISFs of the erased frames are concealed according to G. 722. 2 Annex I. Then we choose a segment of speech of about 2 s.

In the simulation of the first parameter configuration, 13 frames are lost; the SNRs of the two algorithms are respectively $\mathrm{SNR}_{\mathrm{I}} = 16.37$ dB, $\mathrm{SNR}_{\mathrm{HMM}} = 18.62$ dB, and the SNR gain between them is 2.25 dB. In the simulation of the second parameter configuration, 18 frames are lost; the SNRs of the two algorithms are respectively $\mathrm{SNR}_{\mathrm{I}} = 16.86$ dB, $\mathrm{SNR}_{\mathrm{HMM}} = 19.43$ dB, and the SNR gain between them is 2.57 dB. The average increase in the SNR is about 2.41 dB in the two parameter configurations.

The spectral distortions of each concealed ISF in the two parameter configurations are displayed in Fig. 4. For the purpose of comparison, the spectral distortions of the ISFs of the baseline are also displayed all together. We can see that the spectral distortions by the proposed concealment is generally less than that by Annex I. In the first simulation, the average difference between them is 0.648 dB. In the second, the average difference is 1.123 dB. The average reduction in spectral distortion is about 0.885 dB in the two cases.

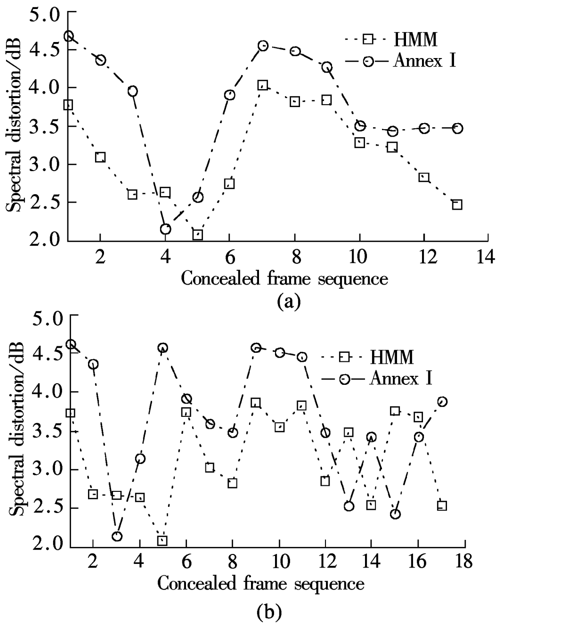


Fig. 4 Spectral distortions. (a) Configuration 1; (b) Configuration 2

5 Conclusion

The traditional frame erasure concealment algorithm makes use of the correlation of parameters to some extent, but it does not apply statistics characteristics and evolution of speech parameters. In this paper, an LHMM is used to model the ISFs of wideband speech, and the correctly-received ISFs are employed to estimate the lost ISFs. Simulation shows that the algorithm has some advantages over the baseline in terms of signal-to-noise ratio and perceptually weighted spectral distortion.

References

[1] Ehsan M S, Kubin G. Frame change ratio: a measure to model short-time stationarity of speech[C]//*Innovations in Information Technology*. Dubai, United Arab Emirates, 2006: 1 − 5.

[2] Vaillancourt T, Jelinek M, Salami R, et al. Efficient frame erasure concealment in predictive speech codecs using glottal pulse resynchronization [C]//*IEEE International Conference on Acoustics, Speech and Signal Processing*. Honolulu, HI, USA, 2007: 1113 − 1116.

[3] Thyssen J, Zopf R, Chen Juin-Hwey, et al. A candidate for the ITU-T G. 722 packet loss concealment standard[C]//*IEEE International Conference on Speech and Signal Processing*. Honolulu, HI, USA, 2007: 549 − 552.

[4] Telecommunication Standardization Sector of ITU. ITU-T recommendation G. 722. 2 wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband(AMR-WB)[S]. 2003.

[5] Telecommunication Standardization Sector of ITU. Wideband coding of speech at around 16 kbit/s using adaptive multi-rate wideband(AMR-WB). Appendix I: Error concealment of erroneous or lost frames[S]. 2002.

[6] Rabiner L, Juang B H. *Fundamentals of speech recognition* [M]. New Jersey: Prentice-Hall, 1993: 321 − 389.

[7] Rodbro C A, Murthi M N, Andersen S V, et al. Hidden Markov model-based packet loss concealment for voice over IP[J]. *IEEE Transactions on Audio, Speech, and Language Processing*, 2006, **14**(5): 1609 − 1623.

[8] Pepper D J, Clements M A. Phonemic recognition using a large hidden Markov model[J]. *IEEE Transactions on Signal Processing*, 1992, **40**(6): 1590 − 1595.

[9] Pepper D J, Clements M A. On the phonetic structure of a large hidden Markov model [C]//*International Conference on Acoustics, Speech, and Signal Processing*. Toronto, Ont, Canada, 1991: 465 − 468.

[10] Collura J S, McCree A, Tremain T E. Perceptually based distortion measurements for spectrum quantization[C]//*IEEE Workshop on Speech Coding for Telecommunications*. 1995: 49 − 50.

[11] Gilbert E N. Capacity of a burst-noise channel[J]. *Ben Syst Tech* Ⅰ, 1960, **39**: 1253 − 1266.

基于大型隐马尔可夫模型的宽带语音丢帧补偿

王仕奎[1,2]　汤一彬[1]　尤红岩[1]　吴镇扬[1]

([1] 东南大学信息科学与工程学院,南京 210096)

([2] 安徽师范大学物理与电子信息学院,芜湖 241000)

摘要:研究了在语音传输过程中由于参数丢失导致语音质量急剧下降的丢帧补偿问题. 利用大规模隐式马尔可夫模型对自适应多速率宽带语音编码(AMR-WB)的 ISF 参数进行建模, 然后对丢失的 ISF 参数进行基于最小均方误差(MMSE)准则的最优估计, 将估计的 ISF 参数和前帧的 ISF 参数进行加权以平滑估计值, 得到补偿的 ISF 参数. 在接收端,利用 ISF 参数的估计值进行语音合成. 将本算法的合成语音和由 G. 722. 2 标准附件 I 的基准补偿的合成语音进行比较, 仿真结果表明, 本补偿算法可以得到更好的性能, 在频率加权谱失真和信噪比这 2 种评价准则上都有所改善, 信噪比提高约 2. 41 dB, 频率加权谱失真下降约 0. 885 dB, 证明了该算法的有效性.

关键词:丢帧补偿;宽带语音;大型隐马尔可夫模型;ISF 参数

中图分类号:TP391