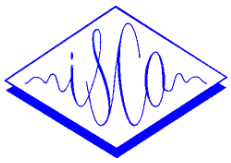

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

AN OBJECTIVE MEASURE FOR QUALITATIVELY ASSESSING LOW-BIT-RATE CODED SPEECH

Toshiro Watanabe and Shinji Hayashi

NTT Human Interface Laboratories
3-9-11, Midori-cho, Musashino-shi, Tokyo 180, Japan

ABSTRACT

This paper reports an objective measure for assessing low-bit-rate coded speech. A model for this objective measure, in which several known features of the perceptual processing of speech sounds by the human ear are emulated, is based on the Hertz-to-Bark transformation, on critical-band filtering with a preemphasis for boosting higher frequencies, on nonlinear conversion for subjective loudness, and on temporal (forward) masking process. The Bark spectral distortion rating (BSDR) is computed for each 10-20 ms segment of the original and coded speech. The effectiveness of this measure was validated by regression analysis between the computed BSDR values and subjective MOS ratings obtained for a large number of utterances coded by several versions of a CELP coder and a VSELP coder under three degraded conditions: input speech levels, transmission error rates, and background noise levels. The BSDR values correspond better to MOS ratings than several commonly used measures. As a result, BSDR can be used to accurately predict subjective scores.

1. INTRODUCTION

At the final stage of the standardization of a speech coder, the speech quality should be represented in terms of mean opinion score (MOS) and opinion equivalent Q which are obtained only from expensive subjective tests. However, during the development stage of coding, these methods are too elaborate and time consuming for frequent and rapid evaluation of coder performance. Objective measures which correspond well to subjective measures, especially to MOS, are greatly needed because of the inconvenience of subjective measures. In 64-16 kb/s speech coders, the LPC cepstral distance (CD) can be used to estimate subjective quality of several kinds of distorted speech with good accuracy [1]. However, in practice this measure has not given sufficiently accurate estimation for less than 8-kb/s CELP-type coders.

An objective measure based on an auditory model has been proposed, in which the correlation coefficient between the objective estimation and MOS is higher than for other traditional measures [2]. Moreover, a variety of weighted distance measures on Bark spectra have been investigated using the same model [3]. Other objective measures have also been investigated [4, 5].

This paper proposes a new objective quality measure called Bark spectral distortion rating or BSDR, in which two types of masking effect are taken into account. A detailed model for this measure is described in Section 2. Experimental results given in Section 3 show that BSDR values correspond better to subjective quality in terms of MOS than other traditionally used objective measures. A quasi-optimal parameter set for controlling masked volume is then determined in order to correspond as closely as possible to the subjective quality for several kinds of distorted speech. It is shown that the estimated MOS from the optimized BSDR corresponds well with the experimental MOS in Section 4.

2. OBJECTIVE QUALITY ASSESSING MODEL

Fig. 1(a) shows an outline of the objective quality evaluation process based on an auditory model [2, 6] and more detailed contents of the loudness computation part are illustrated in Fig. 1(b). Each part in this figure will be explained in turn.

2.1 FFT

An input speech vector x(t) multiplied by the Hamming window with 50% overlap is transformed into the frequency domain, then the power spectra P(ω) is computed.

2.2 Critical-band filtering and Preemphasis

The power spectra vector P(ω) is multiplied by the coefficients of a bank of critical-band filters with a preemphasis to be transformed into the Bark spectra vector Ξ(n) [6]. This bank consists of 15 filters, which are spaced at intervals of about 1 Bark, and it covers the telephony frequency band (0.3-3.4 kHz). Figure 2 shows the characteristics of the bank of critical-band filters.

Since these filters overlap one another and subjective loudness is assumed to be related to be summation of their outputs, it may be somewhat higher than the actual loudness. In order to investigate this masking effect, a threshold, Tf, is set as a parameter which is in proportion to each filter's amplitude at a central frequency, as shown in Fig. 2. Only the coefficients above the threshold are useful for computing Ξ(n).

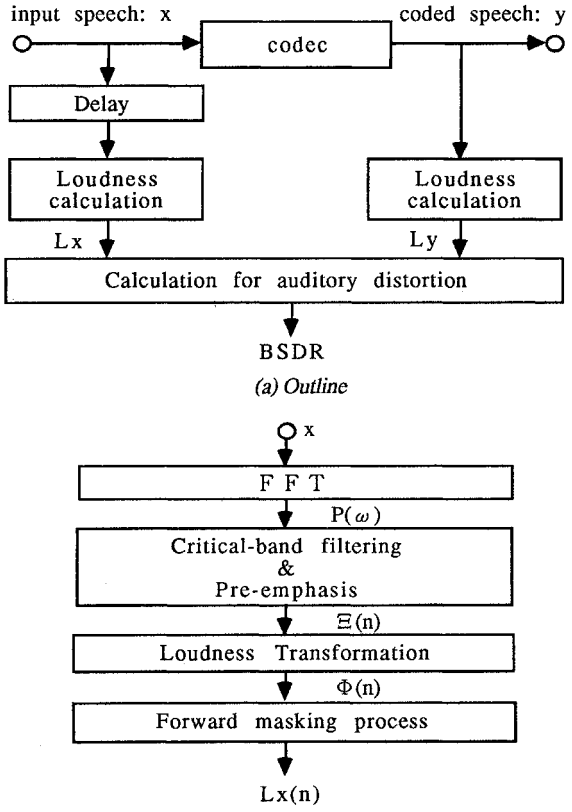

Fig. 1 Block diagram of objective assessment model

2.3 Transformation into loudness

In the next step, $\Xi(n)$ is transformed into the loudness vector $\Phi(n)$ based on the cube-root amplitude compression [6], which is an approximation to the power law of hearing and simulates the nonlinear relationship between the intensity of sound and its received loudness.

2.4 Forward masking process

A forward masking effect occurs when a masker precedes a signal in spite of the fact that the signal and masker are not presented together. This phenomenon is applied to neighboring loudness vectors [7]. The resultant loudness vector $L_i(n)$ for a current frame after considering the forward masking effect is determined by subtracting the masking vector $M_i(n)$ from the loudness vector $\Phi_i(n)$ or by setting zeros as follows:

$$L_i(n) = \begin{cases} \Phi_i(n) - M_i(n) & \Phi_i(n) \geqq M_i(n), \\ 0 & \Phi_i(n) < M_i(n). \end{cases} \quad (1)$$

Here, the masking vector $M_i(n)$ has already been computed through a subsequent process in the previous frame. After the computation of the Bark spectra $\Phi_{i-1}(n)$, it is compared with the masking vector $M_{i-1}(n)$. The temporary masking vector for the next frame is then reset as follows:

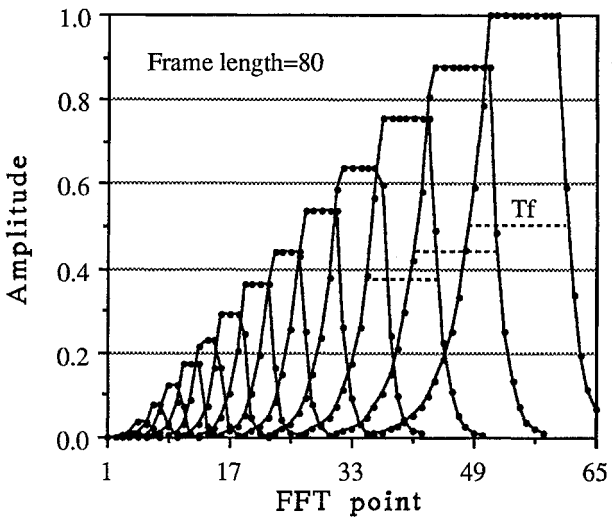

Fig. 2 Characteristics of critical-band filters

$$M'_{i-1}(n) = \max\{M_{i-1}(n), \Phi_{i-1}(n)\}. \quad (2)$$

The actual masking vector of the next frame decreases exponen-tially as follows:

$$M_i(n) = M'_{i-1}(n) \cdot \exp(-\text{nwind} / \tau), \quad (3)$$

where nwind is the frame length and τ is the time constant.

In order to examine the relationship between the forward masking effect and subjective quality, τ is used as a parameter.

2.5 Distortion computation

In the final step of this model, the BSDR value is computed as follows:

$$\text{BSDR} = 10 \cdot \log_{10} (\text{LN} / \text{LD}), \quad (4)$$

where loudness LN and loudness distortion LD are defined as follows:

$$\text{LN} = (1/k) \sum_{i=1}^{k} \sum_{n=1}^{15} [Lx_i(n)]^2, \quad (5)$$

$$\text{LD} = (1/k) \sum_{i=1}^{k} \sum_{n=1}^{15} [Lx_i(n) - Ly_i(n)]^2, \quad (6)$$

where k is the total number of frames in which speech power is greater than the threshold level.

3. PERFORMANCE OF NEW MEASURE

3.1 Subjective tests

In order to compare the subjective quality and the objective evaluation, we used the results of three MOS tests performed during development of CELP-type coders. The first test (Exp. 1) was done to examine the robustness of PSI-CELP [8], which has been adopted as the half-rate

CODEC standard for digital cellular telecommunication in Japan, against several distorted speech conditions:

- Optimal (E0),
- Input speech level of -10 dB and -20 dB lower than optimal condition (L),
- Transmission error rate of 1% and 3% with software and hard-ware decision protection (ES and EH), and
- Background noise level of 15 dB and 30 dB to speech level (N).

The total number of speech samples was 936, including coded speech of 13 slightly different versions of PSI-CELP and VSELP [9]. The second test (Exp. 2) was done to optimize PSI-CELP's parameters using 956 speech samples. The third test (Exp. 3) was done to compare the speech quality of CS-CELP [10], which is a candidate for the standardization of the high-quality 8-kb/s speech coder in ITU-T, with the 32 kb/s standard ADPCM (G. 726), including 44 speech samples.

All tests included several references created by the modulated noise reference unit (MNRU) [11] and the original speech. Every sentence was spoken by at least two male and two female speakers. All utterances were heard by 16 or 24 listeners with one-ear headphones and were assessed based on five categories: excellent (5) to poor (1).

3.2 Comparison with objective measures

In order to examine the relationships between the subjective measure in terms of MOS and six objective measures, second-order polynomial predictions were performed by least squares linear regression.

Table 1 lists the correlation coefficients between MOS and objective evaluation (R), and the standard deviations (s) for MNRU, all speech samples of PSI-CELP and VSELP (*ALL*), and one version of PSI-CELP (*PSI-A*) from Exp. 1. These speech samples were evaluated using BSDR, LPC cepstral distance (CD), signal-to-noise ratio (SNR), segmental SNR (SNRseg), WSNR and segmental WSNR (WSNRseg). The WSNR and WSNRseg measures, which are widely used for CELP-type coders, are calculated based on perceptually weighted mean square errors. In the BSDR computation, the two types of masking effect described in Sections 2.2 and 2.4 were not considered: Tf= 0 and τ=1.

For MNRU, the frequency-domain measures, CD and BSDR, correspond to MOS much better than the time-domain measures, SNR, SNRseg, WSNR, and WSNRseg. If there were no unavoidable phase shifts between the input and output speech caused by up- and down-sampling, and IIR filtering in MNRU module, these time-domain measures evaluated more precisely.

In the case of *ALL* and *PSI-A* data, however, CD data correspond to MOS ratings significantly worse than in the case of MNRU, and the correlation coefficients of the CD data are almost the same or lower than those of the time-domain measures. Therefore, without two types of masking effect, the BSDR measure correlates best with MOS and has the smallest mean square errors for both MNRU and low-bit-rate coders.

Table 1. Comparison with objective quality measure

coder	measure	R	s
MNRU [72]	BSDR	0.937	0.374
	CD	0.919	0.423
	SNRseg	0.571	0.879
	SNR	0.413	0.976
	WSNRseg	0.797	0.647
	WSNR	0.743	0.717
ALL [936]	BSDR	0.777	0.335
	CD	0.456	0.474
	SNRseg	0.452	0.475
	SNR	0.336	0.502
	WSNRseg	0.375	0.494
	WSNR	0.243	0.517
PSI-A [72]	BSDR	0.869	0.263
	CD	0.431	0.480
	SNRseg	0.660	0.339
	SNR	0.504	0.459
	WSNRseg	0.424	0.481
	WSNR	0.330	0.502

3.3 Optimization of parameters of masking effects

Two types of masking effect were investigated with respect to several frame lengths using results from Exps. 1 and 2. These results are shown in Fig. 3, where three points in each vertical, broken or solid line indicate respectively from the bottom to the top:

(1) The correlation coefficients without any masking effects described in Section 2,

(2) The maximum with the masking effect from neighboring critical-band filters using the parameter of Tf, and

(3) The maximum with the forward masking effect in the case of a fixed Tf in (2).

Figure 3 shows that both types of masking effect are effective in increasing BSDR-MOS correlation for low-bit-rate coded speech in the wide region of frame length. Especially, within 80-160 points (10-20 ms), there is an about 7% increase in the correlation. Figure 3 also shows that there are two large differences in the correlations between both MOS tests: one is when frame length is less than 80 points, and the other is when frame length is grater than 80 points caused by the forward masking effect. We think that these differences depend on the distributions of MOS values. As a result, a quasi-optimal parameter set is given as Tf = 0.6, and τ = 60 for the frame length of 80 points.

4. EVALUATION OF PERFORMANCE

A common second-order polynomial regression curve was first calculated using all MOS and BSDR values from Exps. 1 and 2. For PSI-CELP (CELP-A and -B), VSELP, and CS-CELP, BSDR values were then averaged for every four speech samples uttered by four speakers under the same conditions. Finally, estimated MOS values were obtained

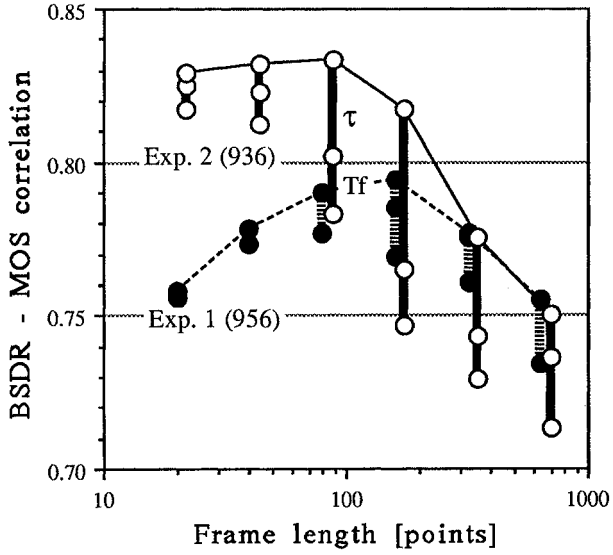

Fig. 3 BSDR-MOS correlations with respect to frame length and two types of masking effect (Tf, τ)

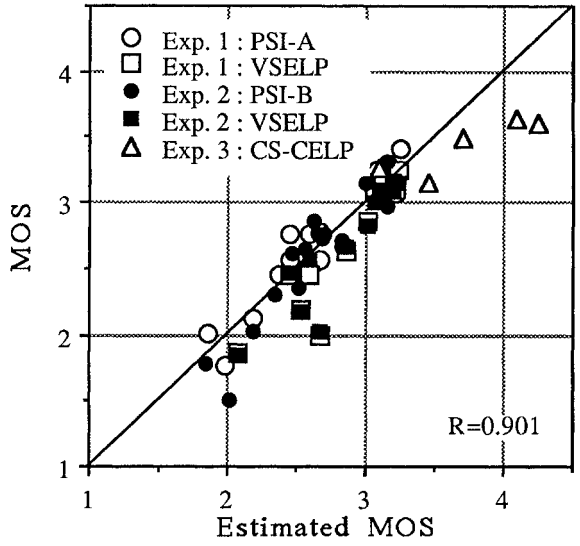

Fig. 4 Relationship between estimated MOS and experimental MOS

from these mean BSDR values. Figure 4 shows the relationship between these estimated and experimental MOS ratings with respect to coders and testing.

Since the correlations for CELP-A and VSELP reflect concrete distributions of MOS and BSDR at an optimal point in Fig. 3, the estimated values correspond well with the experimental values. Unknown additional CS-CELP data from Exp. 3, however, correspond somewhat worse because the regression curve does not saturate when MOS is higher than 3.5, but because BSDR values for this coder have some saturation in the same region. With more variety of data in this region, BSDR will be a more useful measure to give indices for coded speech quality.

5. CONCLUSIONS

A new objective measure based on an auditory model has been proposed for estimating speech quality of low-bit-rate coders. The experimental results show the following: that its performance is higher than the other commonly used objective measures, that the estimation errors can be decreased by taking two types of masking effect into the auditory model, and that the estimations correspond well with the subjective quality represented in terms of MOS under several conditions, such as input speech level, transmission error rate, and low background noise level.

ACKNOWLEDGMENTS

We would like to thank to Dr. Nobuhiko Kitawaki and Takao Kaneko of NTT Human Interface Laboratories for their guidance of our research. We are also grateful to Dr. kazunori Mano and Akitoshi Kataoka for providing speech samples and corresponding results of MOS testing.

REFERENCES

[1] N. Kitawaki, H. Nagabuchi and K. Itoh, "Objective quality evaluation for low-bit-rate speech coding systems," *IEEE SAC*, Vol. 6, pp. 242-248, Feb. 1988.

[2] S. Wang, A. Sekey and A. Gersho, "An objective measure for predicting subjective quality of speech coders," *IEEE SAC*, Vol. 10, no. 5, pp. 819-829, June 1992.

[3] K. Nagano and S. Ono, "Weighted distortion for speech quality prediction model," *Proc. Spring Meeting of Acoust. Soc. Jpn*, 1-7-11, May 1993 (in Japanese).

[4] F. Wuppermann, C. Antweiler and M. Kappelan, "Objective Analysis of the GSM Half Rate Speech Codec Candidates," *Proc. EUROSPEECH'93*, pp. 249-252, Sept. 1993.

[5] O. Ghitza, "Auditory Models and Human Performance in Tasks Related to Speech Coding and Speech Recognition," *IEEE Trans. Speech and Audio Processing*, Vol. 2, No. 1, Part II, pp. 115-132, Jan. 1994.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Am.*, Vol. 87 (4), pp. 1738-1752, April 1990.

[7] K. Aikawa, "A phoneme segmentation parameter based on the onset-sensitive auditory neuron model," *IEICE Trans.*, Vol. J71-A, No. 2, pp. 592-600, March 1988 (in Japanese).

[8] S. Miki, K. Mano, H. Ohmuro and T. Moriya, "Pitch Synchronous Innovation CELP (PSI-CELP)," *Proc. EUROSPEECH '93*, Vol. 1, 8.6, pp. 261-264, Sept. 1993.

[9] I. A. Gerson, "Vector sum excited linear prediction (VSELP) speech coding for JAPAN digital cellular," *Technical Report of IEICE*, RCS90-26, pp. 35-40, Nov. 1990.

[10] A. Kataoka, T. Moriya and S. Hayashi, "An 8-kbit/s Speech Coder Based on Conjugate Structure CELP," *Proc. ICASSP-93*, Vol. II of V, pp. II-592-595, April 1993.

[11] ITU-T Recommendation G.191; "Software Tools for Speech & Audio Coding Standardization."