IEICE TRANS. COMMUN., VOL.E89–B, NO.2 FEBRUARY 2006

PAPER Special Section on Multimedia QoS Evaluation and Management Technologies

Measuring the Perceived Importance of Speech Segments for Transmission over IP Networks

Yusuke HIWASAKI[†], *Member*, **Toru MORINAGA**[†*], *Nonmember*, **Jotaro IKEDO**[†**], *and* **Akitoshi KATAOKA**[†], *Members*

SUMMARY This paper presents a way of using a linear regression model to produce a single-valued criterion that indicates the perceived importance of each block in a stream of speech blocks. This method is superior to the conventional approach, voice activity detection (VAD), in that it provides a dynamically changing priority value for speech segments with finer granularity. The approach can be used in conjunction with scalable speech coding techniques in the context of IP QoS services to achieve a flexible form of quality control for speech transmission. A simple linear regression model is used to estimate a mean opinion score (MOS) of the various cases of missing speech segments. The estimated MOS is a continuous value that can be mapped to priority levels with arbitrary granularity. Through subjective evaluation, we show the validity of the calculated priority values.

key words: *scalable speech coding, multiple description coding, packet networks, VAD, QoS, VoIP, estimated MOS, linear regression model*

1. Introduction

Voice over IP (VoIP) communication has for some years been under consideration as an alternative to traditional PSTN. However, IP networks are based on a best-effort policy which was initially designed to meet the requirements of simple file transmission and is thus not really suitable for transmitting media data in real-time applications. To solve this problem, we have seen the emergence of IP QoS techniques, such as RSVP (resource ReSerVation Protocol) [1] and Diffserv (Differentiated services) [2], [3], in which packets are assigned priority levels and handled accordingly. In the latter case, all routers are provided with QoS policies that give priority to the transmission of real-time packets, thus implementing QoS control for these packets.

Discontinuous transmission (DTX) techniques, which are mainly used in conjunction with voice activity detection (VAD) algorithms, are also important [4]–[6]. In these techniques, the focus is on voice activity, and the intention is to reduce the average transmission bandwidth by transmitting only the active segments that include speech activity. However, a VAD only produces a simple binary decision; this is inflexible, particularly in that it does not take advantage of the multiple priority levels that IP QoS techniques can provide. Also, misjudgments in VAD can lead to speech quality degradation, and this is particularly significant when background noise is present. A refinement of the VAD algorithm is a speech classifier, which can classify the state of active speech in finer resolutions, such as voiced, unvoiced, and onset [7]. However, such algorithms are only bound to detect the state of the speech source, and the classification results do not necessarily reflect the degree of perceived importance of each per speech segment.

On the other hand, scalable [8] and multiple-description coding (MDC) are considered useful as tools for the coding of speech signals on IP networks [9]–[11]. In these algorithms, the encoder generates the bit-stream in a layered manner so that the decoder can reconstruct the speech from a subset of the bits in the stream. Such flexibility is useful because it allows a single encoder to meet various bit-rate and fidelity requirements. However, there have been few attempts to demonstrate the utility of such techniques [12].

In this paper, we propose a way to measure the instantaneous priority of speech segments as a means of taking advantage of scalable speech coding in the context of IP QoS techniques. The technique is a novel way to calculate a single value that reflects the perceived importance of each speech segment, which will vary from segment to segment [13], [14]. This tool is useful in that it provides a fine granularity of speech priority, and we can control speech quality more precisely; this is particularly useful when transmission is through a QoS-aware IP network. The priority is calculated using a linear regression model, and the parameters are optimized automatically by means of a least-squares fit, not by tedious and fallible manual tuning.

The rest of the paper is organized as follows. Section 2 gives an example of a transmission scheme in which the instantaneous priority calculation method is used. Section 3 describes the calculation of speech-segments priority. Section 4 covers the evaluation of the priority calculation and shows that the proposed form of priority grading is valid. We conclude the paper with Sect. 5.

2. Transmitting Scalable Bitstreams over QoS-Aware Networks

A trivial way to apply a scalable codec in an IP environment is to use it as a variable bit-rate codec; that is, to adaptively change the bit-rate according to the level of end-to-end network congestion. Although RTCP [15] and other methods are available as means for determining the congestion level,

Manuscript received May 2, 2005.
Manuscript revised August 11, 2005.
[†]The authors are with NTT Cyber Space Laboratories, NTT Corporation, Musashino-shi, 180-8585 Japan.
[*]Presently, with Plala Networks Inc.
[**]Presently, with NTT Resonant Inc.
DOI: 10.1093/ietcom/e89–b.2.326

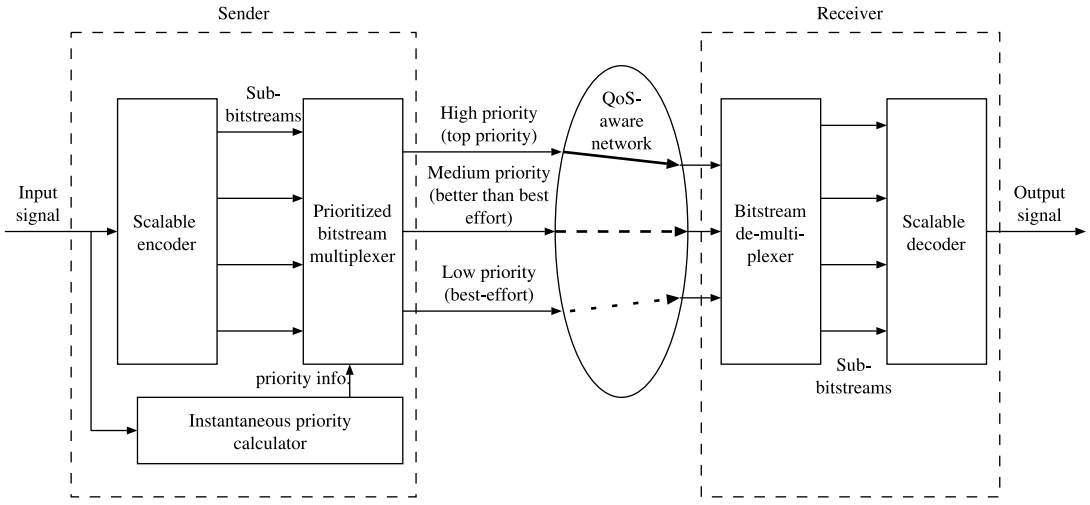


Fig. 1 Applying instantaneous priority calculation in a packet-transmission system.

such methods only provide statistics on the past, so delays in adapting to changes are unavoidable. This introduces degradation, which is particularly disruptive in the case of real-time applications.

A possible solution for a QoS-aware network is to divide each bitstream into smaller units, and assign a priority level to each. For example, the essential core bitstream is sent with a higher priority level, while bitstreams that provide enhanced quality are sent with lower priority levels. When the priority of sent bitstreams is explicitly indicated, any congested node (router) can itself judge whether a packet can be queued or should be sent immediately. In this way, quality can be maintained while adaptation to congestion is performed right at the node.

Real-time speech transmission is typically handled by DTX in conjunction with VAD, since this approach takes advantage of the dynamic nature of speech. Assigning each bitstream a dynamically variable priority is a logical refinement. For lower priority time segments, the priority of the core bitstream may be lowered to reduce the average bitrate at the higher priority level. Figure 1 is a block diagram that shows the application of instantaneous priority calculation for the individual sub-bitstreams in a hypothetical transmission system. In this system, a per-time-frame (instantaneous) priority level is assigned to each sub-bitstream. The sub-bitstreams are then tagged to reflect the priority levels, after which they are multiplexed. Packet sending is then governed by the priority levels. When the network is not congested, all packets are sent. When the network becomes congested, lower-priority packets, that is, those that will have the least effect on the subjective quality, begin to be dropped. Note that this idea is not restricted to speech coding, but is applicable to the transmission of data for other media, such as moving pictures.

The basic idea of such a transmission system has been presented previously as *priority discarding* [16], in which speech packets are assigned with delivery priorities by means of speech classifiers and are discarded according to their priorities when the network is congested. However, as stated before, the speech classifiers only indicate the source state, and their results do not directly reflect the subjective quality resulting from the packet losses.

3. Instantaneous Priority Calculation

To formulate a way of calculating the importance of each bitstream from instant to instant, we can re-phrase the objective as the estimation of the degree of perceived degradation if a packet is lost. Since the focus here is on perceived quality, the approach is in contrast to that seen in VAD, where the focus is on the speech state of the speaker.

3.1 Coding Scheme

Before going into the details about the method of priority calculation, we will define the framework used in the body of this paper. We assume that the codec provides a frequency-scalable form of coding. That is, the codec must incorporate sub-band encoding, where the wide-band speech signal is separated into two or more band-passed signals by an analysis filter bank, and must separately encode the signals thus produced. In decoding, the wide-band signal is reconstructed from the sub-band signals by a synthesis filter bank. Since the packing of speech signals at 20-ms intervals is a common practice, we can assume that the speech-signal data is both frequency- and time-divided into *blocks*. Figure 2 is a schematic view of how speech can be segmented into blocks. Here, f represents the frequency-band index and k represents the time-segment index.

The ITU-T coding standard G.722 [17] which is a typical implementation of such a coding scheme, is illustrated in Fig. 3. The encoder is fed a 16-bit-encoded and 16-kHz-sampled PCM signal, which it separates into lower s_L and higher s_H sub-band signals by using a quadrature mirror filter bank (QMF). The respective signals are encoded by a 6-bit and a 2-bit AD-PCM quantizer, producing output code

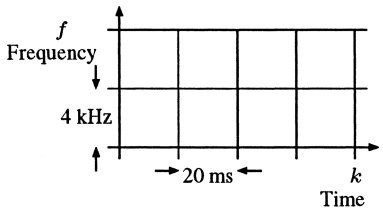


Fig. 2 Division of speech into blocks.

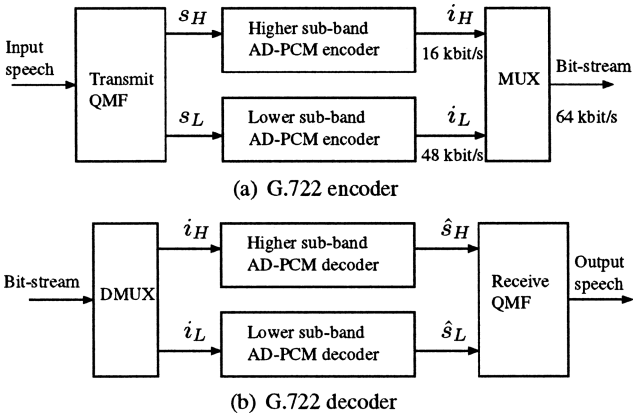


Fig. 3 ITU-T G.722 encoder and decoder.

streams i_L and i_H, which are then multiplexed and transmitted. This gives us an overall bit-rate of 64 kbit/s[†].

3.2 Estimated Mean Opinion Score (MOS) Values

In this section, we will introduce the way of estimating the degree of perceived degradation if a frequency- and time-divided block is lost, using a linear regression model based on the signal features. Formally, each block is a signal vector $\boldsymbol{s}_{k,f}[n]$, where $1 \leq n \leq N$ and N is the total number of samples within the frame with frame index k for frequency band f.

We used the following three perceptually important features that can be calculated from the signal:

$$x_1[k,f] = \log_{10}\left(\mathrm{E}\left[(\boldsymbol{s}_{k,f}[n])^2\right]\right) \tag{1}$$

$$x_2[k,f] = x_1[k,f] - \log_{10}\sum_{f=1}^{F}\mathrm{E}\left[(\boldsymbol{s}_{k,f}[n])^2\right] \tag{2}$$

$$x_3[k,f] = \max(\rho_{k,f}[\tau]), \tag{3}$$

where $\rho_{k,f}[\tau](20 \leq \tau \leq 150)$ is the windowed autocorrelation function of signal block $\boldsymbol{s}_{k,f}[n]$, F is the number of frequency bands, and $\mathrm{E}[\cdot]$ denotes an expectation. It is easily seen that x_1 is the average logarithmic power of the signal, x_2 is the power relative to that of the signal as a whole, and x_3 is the periodicity of the signal. To see how each feature affects the criterion, we normalized each x_r to have a mean of zero and variance of one. This normalization was done using

$$\tilde{x}_r = \frac{x_r - \mu_r}{\sigma_r^2}, \tag{4}$$

where μ_r and σ_r are the 1st- and 2nd-order moments of each x_r, respectively. To calculate the 1st- and 2nd-order moments of each explanatory variable, we used 220,000 frames of speech material, including clean speech with average power scaled to 26 dB below the 16-bit saturation amplitude, and speech with added background car noise (−15-dB relative), babble (−20-dB relative), and interfering speech (−20-dB relative).

To obtain a single-valued measure for each block, we define the linear regression model as follows:

$$y[k,f] = \alpha_0 + \sum_{r=1}^{R}\alpha_r\tilde{x}_r[k,f]. \tag{5}$$

Here, $y[k,f]$ is the objective value, that is the estimated MOS value when a block $\boldsymbol{s}_{k,f}$ is missing, α_r are the regression coefficients, $\tilde{x}_r[k,f]$ are normalized versions of the explanatory variables which are the above signal features $x_r[k,f]$, and $R(=3)$ is the number of coefficients and explanatory variables. Since $y[k,f]$ denotes the estimated MOS value, the absence of an important signal block $\boldsymbol{s}_{k,f}$ would lead to a lower subjective MOS score. By inspecting Eqs. (1), (2), and (3) closer, it is obvious that any signal block that has high values for these three explanatory variables may be considered to be important. This means that all regression coefficients α_r are expected to be negative, since a higher value would contribute to a lower $y[k,f]$. The actual values of α_r and how they were calculated will be given in the following section.

3.3 Calculation of Regression Coefficients

To use the above model, we need to find the precise regression coefficients α_r that can correctly reflect the subjective scores to give an objective measure. To do this, we performed MOS tests to obtain the empirical MOS score $\hat{y}[k,f]$ when a signal block with known features $x_r[k,f]$ had been erased, and then used a least-squares fit to calculate α_r. In this way, we avoided the manual parameter tunings often required for VAD and speech classifiers.

The first step is to artificially erase blocks from the 16-kHz sampled PCM speech signals and perform MOS tests to measure the post-erasure subjective quality. To separate higher and lower bands, we used the same QMF as in G.722. Erasure must be performed with care, because simply applying a rectangular window to erase a block from the signal can lead to a very annoying artifact which strongly affects the MOS results. To avoid this, we took advantage of the fact that most speech coding methods use inter-frame prediction and thus do not produce abrupt transitions to and from zero for the output signal, we decided to apply the trapezoidal window shown in Fig. 4.

As speech materials, we used 10 (5 female and 5 male) Japanese speech sets, with both clean versions and versions under the above-described background-noise condi-

[†]Although the decoder can also be operated in 48- or 56-kbit/s modes, we used the 64-kbit/s mode throughout this study.

tions. Each signal was an 8-second two-sentence portion of speech material. Block erasure was applied to numerous speech segments in each speech sample. The resulting samples were then evaluated by 24 non-experts. They were asked to assign each item of speech material a grade ranging from 1 (very poor) to 5 (very good), that is, to follow the standard MOS evaluation procedure.

Since all explanatory variables $\tilde{x}_r[k, f]$ of the block in conjunction with the empirical MOS score $\hat{y}[k, f]$ were known, a least-squares fit was used to find the α_r values that minimized the total error in the estimation. This is shown in the following equation:

$$\mathrm{E}\left[(y[k, f] - \hat{y}[k, f])^2\right] \to 0, \tag{6}$$

where $\hat{y}$ is the empirical MOS value. To evaluate the contribution of each explanatory variable, we optimized the regression coefficients for each of the seven possible combinations. The combinations and results of optimization are given in Table 1. To see the effectiveness of the optimization, we also calculated the mean squared error and the contribution rate. The mean squared error is calculated as

$$e = \mathrm{E}\left[(y[k, f] - \hat{y}[k, f])^2\right], \tag{7}$$

and the contribution rate is calculated as

$$c_r = \frac{\sigma_y^2}{\sigma_{\hat{y}}^2} \tag{8}$$

where the numerator σ_y^2 and the denominator $\sigma_{\hat{y}}^2$ are the 2nd-order moments of the empirical MOS value and the estimated MOS value, respectively. The contribution rate ranges between 0.0 and 1.0, with the model having a better fit if the contribution rate is closer to 1.0.

Recalling Eq. (5), the meaning of the results is that α_0 is the intercept of the MOS value, in other words, the average. As described in the previous section, all coefficients except α_0 should be negative, because all positive explanatory variables shift $y[k, f]$ in the direction of lower score. A lower estimated MOS value for a block thus means that the block is more important and has a correspondingly higher priority.

Table 1 shows that the mean squared error is minimized (and the contribution rate is maximized) by using all three explanatory variables, x_1, x_2, and x_3 (#7). The second-best result is obtained by using x_1 and x_2 (#4). Thus, using all explanatory variables gives the best result.

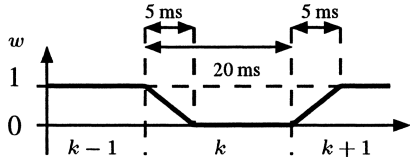


Fig. 4 Window function w when the k-th frame is to be erased.

3.4 Calculated Results

Figure 5 shows an example of the explanatory variables and estimated MOS values for a clean female-speech signal: "Oak is strong and also gives shade." In Fig. 5(a), the amplitude of the input signal in 16-bit linear PCM is shown. The solid and dotted lines in Figs. 5(b), 5(c), and 5(d) represent the explanatory variables in the lower and higher band, respectively. In Fig. 5(e), the corresponding priorities are shown. Note that all explanatory variables have been normalized in the way described in the previous section. In general, the blocks in the lower band have lower estimated MOS values than those in the higher band, showing the greater importance of the lower-band blocks. However, this is reversed in the case of consonants, such as fricatives. Here, the higher band becomes dominant since it contains more power than the lower band, indicating the importance of both bands in maintaining the quality of wide-band speech.

Although the figure shows that the estimated MOS value has a strong correlation with the absolute power, we cannot rely solely on this parameter, since we can expect a range of input levels. Using only the absolute power would make all lower input-level speech fall into the lower priority level, which is obviously undesirable.

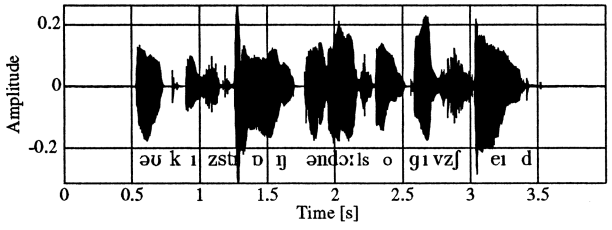
Numerous features of a speech signal can be taken into account in calculating the degree of degradation, including the number of zero crossings, first reflection-coefficient, power continuity, and the degree of inter-frame prediction used in the coding scheme. The method presented here is a linear regression model to which other features can easily be added as new explanatory variables. Instead of the tedious tuning of parameters and thresholds, the basis is computational optimization according to Eq. (6). The method has a further advantage in that we can objectively judge the adequacy of any added variables by evaluating the errors in estimation.

The most common way to compensate for the speech degradation caused by network congestion is using coding schemes utilized with packet-loss concealment (PLC) algo-

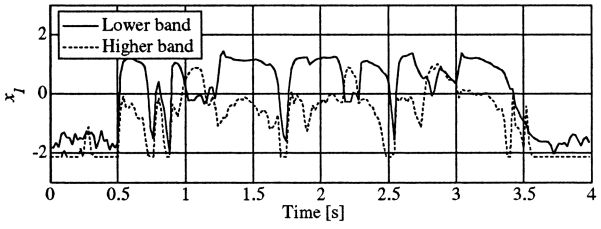
Table 1 Optimized regression coefficients α_r.

#	Features used	α_0	α_1	α_2	α_3	Mean squared error (e)	Contribution rate (c_r)
1	x_1 only	3.15	− 0.75	–	–	0.45	0.62
2	x_2 only	3.24	–	− 0.86	–	0.47	0.61
3	x_3 only	3.12	–	–	− 0.74	0.60	0.50
4	x_1 and x_2	3.19	− 0.45	− 0.49	–	0.34	0.72
5	x_1 and x_3	3.13	− 0.55	–	− 0.31	0.40	0.67
6	x_2 and x_3	3.19	–	− 0.61	− 0.36	0.39	0.68
7	x_1, x_2 and x_3	3.17	− 0.37	− 0.43	− 0.19	0.32	0.73

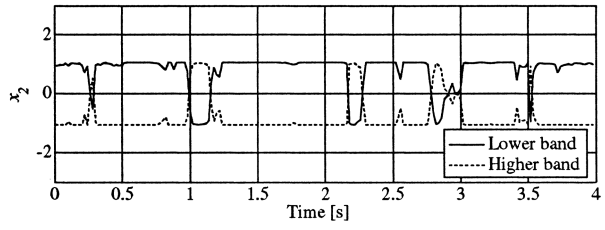
rithms such as ITU-T G.711 Appendix I [18]. The priority calculation model proposed in this paper is also applicable to such cases, because the empirical MOS scores $\hat{y}$ can still be obtained using the scheme with PLC, and the optimization of the regression coefficients can be carried out in the way described in Sect. 3.3.



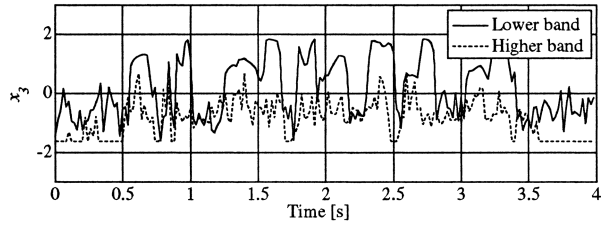
(a) Amplitude



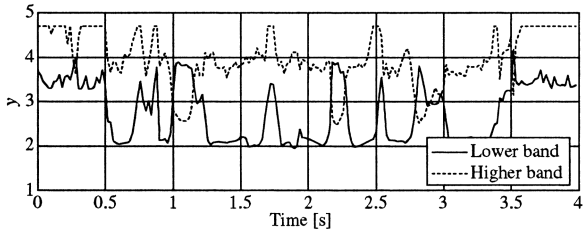
(b) Absolute power (normalized)



(c) Relative power (normalized)



(d) Periodicity (normalized)



(e) Estimated MOS

Fig. 5 Example of power, periodicity, and estimated MOS values for a female speech signal: "Oak is strong and also gives shade."

4. Evaluation of the Instantaneous Priority Calculation

4.1 System Setup

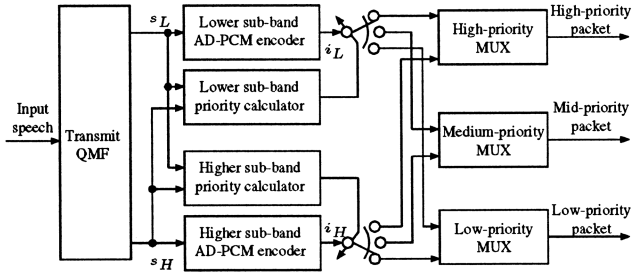
To evaluate the accuracy of the above model, we integrated the priority calculation algorithm with a G.722 codec and performed subjective evaluation tests on signals that had been subjected to random erasure, controlled for particular estimated MOS values. Block diagrams of the encoder and decoder are given in Fig. 6.

Here, the frame length was again set to 20 ms, and the output codes from the lower and higher bands were separately packed into three different priority groups by using the following condition mapping:

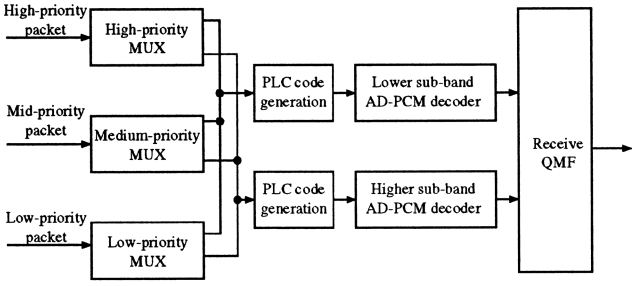
- High priority: $y[k, f] \leq 2.5$
- Medium priority: $2.5 < y[k, f] \leq 3.5$
- Low priority: $3.5 < y[k, f]$.

We used all of the explanatory variables in the priority calculation, and thus set the regression coefficients to the optimized values given in the seventh row (#7) of Table 1.

Since the G.722 algorithm was defined before the use of IP networks for voice transmission was under serious consideration, the recommendation does not specify how packet losses should be handled. For this purpose, we modified the G.722 decoder such that if a frame erasure occurs, all state variables of the adaptive-differential de-quantizer are reset to '0' and the lower- and higher-band bitstreams are replaced by '111101' (6 bits) and '11' (2 bits), respectively. This makes the sub-band decoder output amplitude converge to 0 at the erased frame.



(a) G.722 encoder modified to handle the three priority bitstreams



(b) G.722 decoder modified through integration of priority calculation

Fig. 6 G.722 codec modified to generate and handle priority-assigned bitstreams.

Table 2 Proportions of blocks in the three priority groups.

Condition / priority	High	Medium	Low
Nominal level	35.13%	22.03%	42.83%
Lower level	29.35%	24.63%	46.02%
Car noise	34.68%	40.75%	24.57%
Office noise	35.10%	37.57%	27.33%
Average	33.57%	31.25%	35.19%

Table 3 Statistics on voice activity.

VAD state	Ratio
Active	27.76%
Inactive	72.24%

4.2 Speech Samples

For the subjective evaluation test, we used a set of eight (four female and four male) speech samples as clean speech signals. Each sample was 8 seconds long and contained two sentences spoken in Japanese. The average power in each case was set to −26 dB from the 16-bit saturation amplitude (nominal level). Note that the samples were different from those used in Sect. 3.3, and thus they were open data. To see the effect of instantaneous priority calculation under various speech conditions, we also tested low-amplitude versions of the same materials (−36 dB from the 16-bit saturation amplitude), and versions to which office noise or car noise had been added (−20 dB relative to the speech signal). Table 2 shows the percentage of the blocks in each priority level. Although the thresholds described above were selected arbitrarily, the average result was roughly one-third of the blocks being assigned to each priority level.

Across all three groups, 5.0% of all speech blocks were randomly erased. Since we were using 8-second speech materials, 5.0% block-erasure is equivalent to erasing 40 out of 800 blocks (400 frames for each of the two bands). Although a finer granularity is possible, we only set up three priority groups for this evaluation. This is because the number of blocks that belong to the respective priority levels would become too few if we used the 8-second-long speech materials.

For comparison, we included samples that had been subjected to random block erasure at rates of 1.25, 2.5, 5.0, and 10.0%. We also manually labeled the same samples with voice-activity indications. The proportions of active and inactive segments are shown in Table 3. In these cases, 5.0% random block erasure was applied to either active or inactive segments. For speech signals with background noise, we used the same VAD label as in the original clean speech conditions, assuming that ideal VAD was performed, i.e., the VAD simulation could completely distinguish between the speech section and the background noise. Here, we only used the simple VAD binary priority grading and did not use elaborate speech classifiers, because the mapping of speech state classification to priority is not straightforward and has not been established.

Since G.722 coding allocates different bit-rates to the lower (48 kbit/s) and higher (16 kbit/s) frequency bands, it is useful to see how each of the block-erasure conditions affects the bit-rate: a result in this form is given in Table 4. This table shows that 5.0% block-erasure conditions in all priority groups (high, medium, and low) fell within the range of bit erasure-rates from 10.0% to 2.5%. For VAD conditions, the erasure rate in terms of blocks was the same as the erasure rate in terms of bit-rate, because both higher and lower bands were erased simultaneously.

Table 4 Average bit erasure-rates used as test conditions with G.722.

Erasure rate of blocks	Bit erasure-rate
High priority 5.0%	7.5%
Medium priority 5.0%	6.2%
Low priority 5.0%	3.3%
10.0% random	9.8%
5.0% random	5.1%
2.5% random	2.5%
1.25% random	1.2%
VAD active 5.0%	5.0%
VAD inactive 5.0%	5.0%

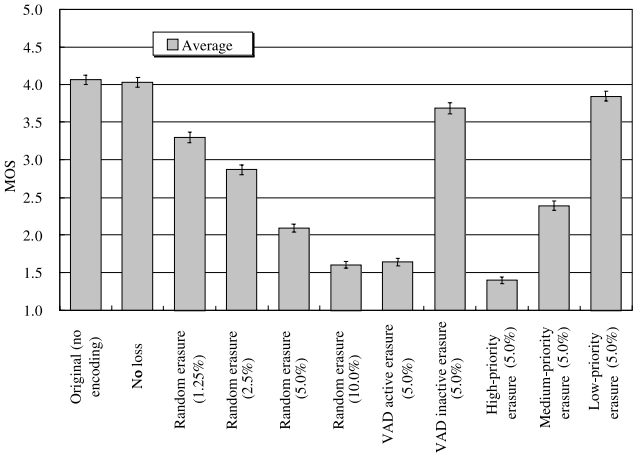


Fig. 7 Overall results of evaluation for all conditions.

4.3 Results

Using the standard MOS procedure, the speech samples described above were evaluated by 24 non-experts.

The results of overall evaluation testing, averaged over all speech conditions, are plotted in Fig. 7, together with the 95% confidence intervals. The results indicate distinct differences between the three priority groups. Erasing 5% of low-priority blocks produced better scores than the random erasure of 1.25% of blocks, while erasing 5% of high-priority blocks produced worse scores than the random erasure of 10% of all blocks. The results for erasing 5% of medium-priority blocks were slightly better than those for random erasure of 5% of all blocks. These results show that the calculation of priority was generally sound. In terms of bit erasure-rate (erasure rate as a percentage of bit-rate), referring back to Table 4, erasing all three priority blocks fell within the range of randomly erased 2.5% and 10.0% con-

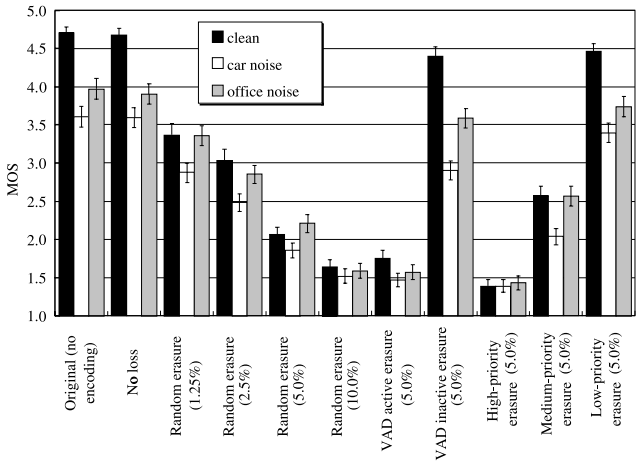


Fig. 8 Results of evaluation for noisy conditions.

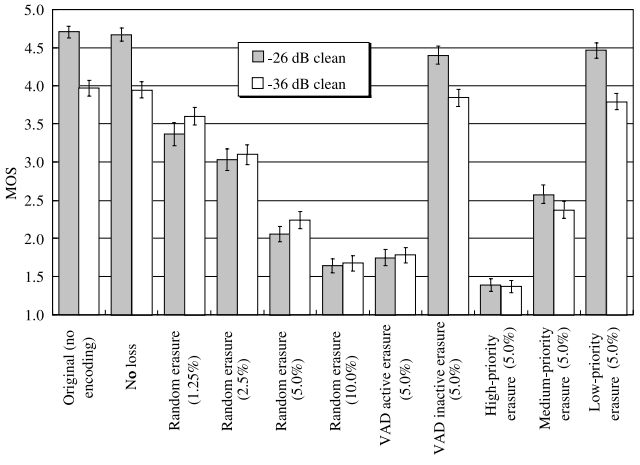


Fig. 9 Results of evaluation for the two input level conditions.

ditions, indicating that the results are valid. However, when we compare with the results for VAD, it is difficult to say that the proposed method is superior to VAD, because the erasure rates in terms of bit-rate were not equal: as seen in Table 4, the bit erasure-rate of erasing high-priority blocks (7.5%) was more than that of erasing the voice active blocks (5.0%), and that of low-priority blocks (3.3%) was less than the voice inactive blocks (5.0%).

The results under noisy conditions are plotted in Fig. 8. Since the car-noise signal used here had relatively stationary power, the erasures were easily noticed, so this condition led to lower scores than the non-stationary office noise and clean conditions. Although the degree of degradation varied across the conditions, the relationships between the various erasure conditions still remained the same as those shown in Fig. 7, which indicates that priority estimation was successful. It should be noted that, in the car noise condition, the low-priority erasures performed significantly better than the VAD inactive erasures. This is because the losses in stationary background noise sections, that is, the inactive sections, could be perceived more easily. This indicates that the VAD or speech classifiers are not capable of directly reflecting the perceived importance of the input signal. A "noise classifier" is a possible solution to this problem, but it is not so desirable because it would further complicate source-state-to-priority mapping problems.

In Fig. 9, we compare the results for clean speech at the two levels (−26 dB and −36 dB from the 16-bit saturation amplitude). Comparing these two input levels, we see that the MOS scores range is narrower for the low-level (−36 dB) input. This is probably due to the fact that the speech signal becomes less comprehensible when the signal level drops, and the scores at the higher end were saturated at around 4.0 even for original and no-loss conditions. On the other hand, the low-level input signal performed better in the random erasures conditions, because erasures in the nominal-level inputs were more perceptible. Inspecting the scores for priority grading, we find that the relationship between the scores of all loss conditions is generally consistent with the average performance shown in Fig. 7. The only inconsistency between the low-level and the nominal-level input is that the low-priority 5.0% erasures in −36-dB signal input performed only slightly, but not significantly, worse than the VAD inactive erasures.

To sum up, the results under the various speech conditions consistently show the effectiveness of the method of priority calculation and indicate that the method is relatively robust against background noise and is independent of the input level. Compared with VAD classification, our method can provide a finer and more adequate granularity in determining the dynamically changing perceptual importance of speech blocks, even under stationary background noise conditions.

In this evaluation, we did not assess the adequacy of the estimated MOS values as an absolute measure. This is because the speech samples were open data; that is, the block erasure-rates and speech materials used in the optimization of regression coefficients in Sect. 3.3 differed from those used in the evaluation procedure. This means that the results of this MOS test may not have matched the first one described in Sect. 3.3. The important point is that we have obtained a good measure that can be used to estimate the relative importance of blocks within a speech signal.

5. Conclusion

We have presented a way to quantify the perceptual importance of speech blocks and applied the method in frequency scalable coding. The method is based on a simple linear regression model with the logarithmic power, power proportion, and periodicity as features. This method has advantages over the conventional approach, VAD, in that it is based on an objective model of perceived quality rather than on the speech state.The adequacy of the features was verified through regression analysis, which objectively demonstrated that using all of the features gives the best estimates. We have conducted a subjective evaluation of this method when it is integrated with the ITU-T G.722 codec,

and found that it provides a practical priority grading for speech blocks.

Acknowledgments

The authors would like to thank Hisashi Ohara for guiding the research, and members of the Acoustic Information Group for their helpful advice and for valuable discussions. We would also like to express our gratitude to the anonymous reviewers for constructive comments.

References

[1] R. Braden, L. Zhang, S. Berson, S. Herzog, and S. Jamin, "Resource reservation protocol (RSVP)—Version 1 functional specification," RFC 2205 (Proposed Standard), Sept. 1997. Updated by RFCs 2750, 3936.

[2] S. Blake, D. Black, M. Carlson, E. Davies, Z. Wang, and W. Weiss, "An architecture for differentiated service," RFC 2475 (Informational), Dec. 1998. Updated by RFC 3260.

[3] D. Grossman, "New terminology and clarifications for DiffServ," RFC 3260 (Informational), April 2002.

[4] 3rd Generation Partnership Project (3GPP); Technical Specification Group Services and System Aspects, TS 26.094—Adaptive Multi-Rate (AMR) speech Codec; Voice Activity Detector (VAD).

[5] J. Stegmann and G. Schroeder, "Robust voice-activity detection based on the wavelet transform," IEEE Workshop on Speech Coding for Telecommunications, pp.99–100, Pocono Manor, 1997.

[6] S. Gazor and W. Zhang, "A soft voice activity detector based on a Laplacian-Gaussian model," IEEE Trans. Speech Audio Process., vol.11, no.5, pp.498–505, 2003.

[7] J. Stegmann, G. Schroeder, and K.A. Fischer, "Robust classification of speech based on the dyadic wavelet transform with application to CELP coding," Proc. IEEE Int. Conf. Acoust. Speech Sign. Process., pp.546–549, Atlanta, 1996.

[8] W.H.R. Equitz and T.M. Cover, "Successive refinement of information," IEEE Trans. Inf. Theory, vol.37, no.2, pp.269–275, March 1991.

[9] M. Alasti, K. Sayrafian-Pour, A. Ephremides, and N. Farvardin, "Multiple description coding in networks with congestion problem," IEEE Trans. Inf. Theory, vol.47, no.3, pp.891–902, March 2001.

[10] T. Nomura, M. Iwadare, M. Serizawa, and K. Ozawa, "A bitrate and bandwidth scalable CELP coder," Proc. IEEE Int. Conf. Acoust. Speech Sign. Process., pp.341–344, Seattle, 1998.

[11] K. Koishida, V. Cuperman, and A. Gersho, "A 16-kbit/s bandwidth scalable audio coder based on the G.729 standard," Proc. IEEE Int. Conf. Acoust. Speech Signal Process., pp.1149–1152, Istanbul, 2000.

[12] T. Nomura and M. Iwadare, "Voice over IP systems with speech bitrate adaptation based on MPEG-4 wideband CELP," IEEE Workshop on Speech Coding for Telecommunications, pp.132–134, Porvoo, 1999.

[13] T. Morinaga, Y. Hiwasaki, and J. Ikedo, "Perceptual importance of speech blocks divided in frequency and time domain," Proc. Spring Meet. Acoust. Soc. Jpn., 3–3–16, pp.327–328, 2003.

[14] Y. Hiwasaki, T. Morinaga, J. Ikedo, and A. Kataoka, "Measuring the perceived importance of time- and frequency-divided speech blocks for transmitting over packet networks," Proc. INTERSPEECH - ICSLP, pp.637–640, Jeju Island, 2004.

[15] H. Schulzrinne, S. Casner, R. Frederick, and V. Jacobson, "RTP: A transport protocol for real-time applications," RFC 3550 (Standard), July 2003.

[16] D.W. Petr, J.L.A. DaSilva, and V.S. Frost, "Priority discarding of speech in integrated packet networks," IEEE J. Sel. Areas Commun., vol.7, no.5, pp.644–656, 1989.

[17] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.722—7 kHz audio-coding within 64 kbit/s, Nov. 1988.

[18] ITU-T (Telecommunication Standardization Sector, International Telecommunication Union), Geneva, Switzerland, ITU-T G.711 Appendix I—A high quality low-complexity algorithm for packet loss concealment with G.711, Sept. 1999.

Yusuke Hiwasaki received a B.E. degree in instrumentation engineering and an M.E. degree in computer science from Keio University, Yokohama, Japan, in 1993 and 1995, respectively. Since joining NTT Human Interface Laboratories (now Cyber Space Laboratories), Tokyo Japan, in 1995, he has been engaged in the research fields of low bit-rate speech coding and speech coding in voice-over-IP telephony. From 2001 to 2002, he was a guest researcher at Kungliga Tekniska Högskolan (Royal Institute of Technology) in Stockholm, Sweden. He is also a member of the IEEE and the Acoustical Society of Japan (ASJ).

Toru Morinaga received B.E. and M.E. degrees in information engineering from Okayama University, Okayama Japan in 1997 and 1999, respectively. In 1999, he joined NTT Cyber Space Laboratories, Tokyo Japan, and until 2003, he was engaged in the research field of speech coding in voice-over-IP telephony. He is now with Plala Networks Inc.

Jotaro Ikedo received a B.E. degree in electronic engineering and an M.E. degree in electrical engineering from Kogakuin University, Tokyo Japan, in 1989 and 1991, respectively. Since joining NTT in 1991, he has been engaged in R&D concerning low-bit-rate speech coding and wireless transmission. He contributed to the ARIB STD-27 and ITU-T G.729 standards. He is now developing voice-over-IP systems. He is a member of the IEEE and the Acoustical Society of Japan (ASJ).

Akitoshi Kataoka received B.E., M.E., and Ph.D. degrees in electrical engineering from Doshisha University, Kyoto in 1984, 1986, and 1999 respectively. Since joining NTT Laboratories in 1986, he has been engaged in research into noise reduction, acoustic arrays, speech-signal processing, and medium bit-rate speech and wide-band coding algorithms for ITU-T standards. He is a member of the ASJ and the IEEE. He received the Technology Development Award from the ASJ in 1996 and the Prize of the Commissioner of the Japan Patent Office from the Japan Institute of Invention and Innovation in 2003.