Measuring the Perceived Importance of Time- and Frequency-divided Speech Blocks for Transmitting over Packet Networks

Yusuke Hiwasaki, Toru Morinaga*, Jotaro Ikedo†, and Akitoshi Kataoka

NTT Cyber Space Laboratories, NTT Corporation, Tokyo Japan
hiwasaki.yusuke@lab.ntt.co.jp

Abstract

This paper presents a way to calculate the perceived importance of speech segments as a single value criterion, using a linear regression model. Unlike the commonly used voice activity detection (VAD) algorithms, this method allows us to obtain a finer priority granularity of speech segments. This can be used in conjunction with frequency scalable speech coding techniques and IP QoS techniques to achieve efficient and quality-controlled voice transmission. A simple linear regression model is used to calculate the estimated mean opinion score (MOS) of the various cases of missing speech segments.

1. Introduction

Voice over IP (VoIP) communication has for some years been under consideration as an alternative to traditional telephone networks. However, IP networks are based on a best-effort policy which was initially designed to meet the demand for ordinary file transmissions. The best-effort network means that if the end-to-end bandwidth of the network is high enough, all packets are transmitted without problems. However, if any one of the intermediate relaying nodes (a router) is congested, this can cause a problem since the packet will either be queued or dropped at the node. This property is not really suitable for real-time applications such as voice communication, where data packets must be transmitted with as small delay as possible to obtain comfortably responsive conversation. This requires the receiving buffer length to be short. However, packets that are delayed beyond the buffer boundary are discarded.

To solve the above problem, we have seen the emergence of IP QoS techniques, such as DiffServ [1], in which packets are assigned priority levels and handled accordingly. That is, when DiffServ is applied to a network, each packet entering the network is graded and marked according to the type of the packet, port number, and IP address. All routers transfer higher priority packets before the lower ones, thus implementing QoS control for the packets.

It should be noted that these QoS techniques are intended to control priority per flow. However, since speech segments are encoded and packed in unit of a certain length of time, QoS control can be used to make packet transmission more efficient if we can grade the packets according to the speech state. This has typically been done by using voice activity detection (VAD) algorithms [2] and discontinuous transmission (DTX). However, these techniques only produce a binary decision for each segment, that is, whether or not each segment is an active speech segment, and this requires careful threshold tuning.

On the other hand, scalable and multiple-description coding (MDC) are considered useful as tools for the coding of speech signals on IP networks [3]. 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 entire bit-stream. Such flexibility will be useful because it allows a single encoder to meet a variety of bit-rate and fidelity requirements. However, attempts to demonstrate the utility of such techniques have been few [4].

This paper describes a novel way to calculate the perceived importance of speech segments as a single value. This tool is useful in that it provides a finer granularity of speech priority, and gives us greater freedom of control over speech quality, particularly when transmission is through a QoS-aware IP network. The method can be applied to any coding scheme, including multiple-description coding.

2. Coding aspect

In this paper, we assume that the codec realizes a frequency-scalable coding. This involves sub-band encoding, that is, separation of a wide-band speech signal into two or more band-passed signals by using an analysis filter bank and separately encoding of the signal thus produced. In decoding, the wide-band signal is reconstructed from the sub-band signals by using a synthesis filter bank.

The ITU-T coding standard G.722 [5] provides typical implementation of such a coding scheme, and is illustrated in Fig. 1. The encoder is fed a 16-bit-encoded and 16-kHz-sampled PCM signal, which it separates into lower $x_L$ and higher $x_H$ sub-band signals 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 indices $i_L$ and $i_H$, which are then multiplexed and transmitted.

Figure 1: G.722 encoder and decoder.
us an overall bit-rate of 64 kbit/s.

3. Estimated mean opinion score (MOS) value

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. How this applies in the case of G.722 is illustrated in Fig. 2. The aim of our study is to formulate a way of calculating the importance of each individual block. We can then re-phrase the objective as estimation of the degree of perceived degradation when a block is lost. Formally, each block is a signal vector $s_{k,f}[n]$ where $1 \leq n \leq N$ and $N$ is the total number of samples within the frame, with frame index $k$ for frequency band $f$. In this study, we use a linear-regression model and define a single-value measure for each block as the following equation:

$$y[k,f] = \alpha_0 + \sum_{r=1}^{R} \alpha_r \tilde{x}_r[k,f].$$  \hspace{1cm} (1)

Here, $y[k,f]$ is the objective value, that is, the estimated MOS value when a block $s_{k,f}$ is missing. $\alpha_r$ are the regression coefficients, $\tilde{x}_r[k,f]$ is the normalized version of the explanatory variables that describe features of the signal block $s_{k,f}$, and $R$ is the number of explanatory variables.

In our study, we adopt the following three $(R=3)$ variables to describe the perceptually important features.

$$x_1[k,f] = \log_{10}(E[(s_{k,f}[n])^2])$$  \hspace{1cm} (2)

$$x_2[k,f] = x_1[k,f] - \log_{10} \sum_{f=1}^{F} E[(s_{k,f}[n])^2]$$  \hspace{1cm} (3)

$$x_3[k,f] = \max_p (\rho_{k,f}[\tau]),$$  \hspace{1cm} (4)

where $\rho_{k,f}[\tau] (20 \leq \tau \leq 150)$ is the windowed autocorrelation function of signal block $s_{k,f}[n]$ and $F$ is the number of frequency bands. 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 whole signal, and $x_3$ is the periodicity of the signal. In order to normalize the effects of the explanatory variables on the criterion variable $y$, each variable was normalized to have a mean of zero and variance of one. This normalization is done by using

$$\tilde{x}_r = \frac{x_r - \mu_r}{\sigma_r},$$  \hspace{1cm} (5)

where $\mu_r$ and $\sigma_r$ 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).

4. Calculation of regression coefficients

To use the above model, we have to find the right regression coefficients $\alpha_r$. To do this, we performed MOS tests to obtain the observed MOS values $\hat{y}[k,f]$ when a signal block with known features $x_r[k,f]$ had been erased, and then applied least squares fit to calculate $\alpha_r$.

The first step is to artificially erase blocks from the 16-kHz sampled PCM speech signals and perform MOS tests to measure the subjective quality after the erasure. To separate higher band lower bands, we used the same QMF as in G.722. Erasures must be performed carefully, because simply applying a rectangular window to erase a block of signal can cause a very annoying artifact that strongly affects the MOS results. To avoid this, based on the fact that most speech coding methods use inter-frame prediction thus does not result in abrupt zero filling of the output signal, we decided to apply the trapezoidal window shown in Fig. 3.

As speech materials, we used 10 (5 female and 5 male) Japanese speech sets under clean and speech under the above-described background noise conditions. Each signal was an 8-second two-sentence speech material. Block erasure was applied to numerous speech segments, and speech samples subjected to this artificial erasure were 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 observed MOS values $\hat{y}[k,f]$ are known, a least-squares fit is used to find $\alpha_r$ that minimizes the sum of the estimation error. This is shown in the following equation:

$$E[(y[k,f] - \hat{y}[k,f])^2] \rightarrow 0,$$  \hspace{1cm} (6)

where $\hat{y}$ is the observed MOS value. To see the effect of each explanatory variable, we have optimized the regression coefficients for the seven possible combinations. The combinations and the results of optimization are given in Table 1.

If we recall Equation (1), the meaning of the results is that $\alpha_0$ is the intercept of the MOS value, in other words, the average. All coefficients except $\alpha_0$ should be negative, because all positive explanatory variables shift $y[k,f]$ in the direction of lower score. A lower estimated MOS score for a block thus means that the block is more important and has a correspondingly high priority.

Table 1 shows that the average squared error is minimized (and the contribution rate is maximized) by using all three explanatory variables, $x_1, x_2$ and $x_3$ (#4). The second best result is obtained by using $x_1$ and $x_2$ (#2). Using all explanatory variables thus gives the best result.
Table 1: Optimized regression coefficients $\alpha_r$.

<table>
<thead>
<tr>
<th>#</th>
<th>features used</th>
<th>$\alpha_0$</th>
<th>$\alpha_1$</th>
<th>$\alpha_2$</th>
<th>$\alpha_3$</th>
<th>average squared error</th>
<th>contribution rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$x_1$ only</td>
<td>3.15</td>
<td>-0.75</td>
<td>-</td>
<td>-</td>
<td>0.45</td>
<td>0.62</td>
</tr>
<tr>
<td>2</td>
<td>$x_2$ only</td>
<td>3.24</td>
<td>-0.86</td>
<td>-</td>
<td>-</td>
<td>0.47</td>
<td>0.61</td>
</tr>
<tr>
<td>3</td>
<td>$x_3$ only</td>
<td>3.12</td>
<td>-</td>
<td>-0.74</td>
<td>-</td>
<td>0.60</td>
<td>0.50</td>
</tr>
<tr>
<td>4</td>
<td>$x_1$ and $x_2$</td>
<td>3.19</td>
<td>-0.45</td>
<td>-0.49</td>
<td>-</td>
<td>0.34</td>
<td>0.72</td>
</tr>
<tr>
<td>5</td>
<td>$x_1$ and $x_3$</td>
<td>3.13</td>
<td>-0.55</td>
<td>-0.31</td>
<td>-</td>
<td>0.40</td>
<td>0.67</td>
</tr>
<tr>
<td>6</td>
<td>$x_2$ and $x_3$</td>
<td>3.19</td>
<td>-0.61</td>
<td>-0.36</td>
<td>-</td>
<td>0.39</td>
<td>0.68</td>
</tr>
<tr>
<td>7</td>
<td>$x_1$, $x_2$ and $x_3$</td>
<td>3.17</td>
<td>-0.37</td>
<td>-0.43</td>
<td>-0.19</td>
<td>0.32</td>
<td>0.73</td>
</tr>
</tbody>
</table>

Table 2: The average bit-erasure rate of tested conditions.

<table>
<thead>
<tr>
<th>Block-erasure rate</th>
<th>Bit-erasure rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>High-priority 5.0%</td>
<td>7.5%</td>
</tr>
<tr>
<td>Mid-priority 5.0%</td>
<td>6.2%</td>
</tr>
<tr>
<td>Low-priority 5.0%</td>
<td>3.3%</td>
</tr>
<tr>
<td>10.0% random</td>
<td>9.8%</td>
</tr>
<tr>
<td>5.0% random</td>
<td>5.1%</td>
</tr>
<tr>
<td>2.5% random</td>
<td>2.5%</td>
</tr>
<tr>
<td>1.25% random</td>
<td>1.2%</td>
</tr>
</tbody>
</table>

5. Evaluation

To evaluate the accuracy of the above model, we integrated the algorithm with a G.722 coder and performed subjective evaluation tests on signals that had been subjected to random erasures, controlled for particular estimated MOS values. The frame length was again set to 20 ms, and we assumed that the lower and higher bands are separately packed into different packets. We used all of the explanatory variables in this evaluation, and higher bands are separately packed into different packets. The frame length was again set to 20 ms, and we assumed that the lower and higher bands are separately packed into different packets. We used all of the explanatory variables in this evaluation, and set the regression coefficients as shown in #7 of Table 1. We then grouped the blocks into three sets, according to the estimated MOS values, as follows:

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

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 band). Although finer granularity of priority is possible, we only set up three groups. Otherwise, with the 8-second speech materials, we were sometimes unable to obtain enough blocks in each priority level. For comparison, we included samples subjected to random block erasure at the rates of 1.25%, 2.5%, 5.0% and 10.0%. To see the effect of various speech conditions, the speech material used were clean condition (-26 dB from the 16-bit saturation amplitude), added office noise or car noise (-20 dB relative), and clean low-level (-36 dB the 16-bit saturation amplitude). The decoder output was evaluated by 24 non-experts, using the standard MOS procedure.

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, and a result in this form is given in Table 2. This table shows that all prioritized (high, medium and low) 5.0% block-erasure conditions fall within the range of bit-erasure rates from 10.0% to 2.5%.

The results of overall evaluation tests, averaged over all speech conditions, are plotted in Fig. 4, together with the 95% confidence intervals. These results show that there are distinct differences between the three priority groups. When compared with the pure random-erasure conditions, erasing low-priority blocks scores better than that of 1.25%-random block-erasures, and the erasing high-priority blocks scores worse than that of 10.0%-random erasures. The results for the medium-priority block erasures are almost equivalent to the 5.0%-random erases. These results show that the calculation of priority was in general adequate.

The results under noisy conditions are plotted in Fig. 5. Since the car noise signal used here has relatively stationary power, the erasures are easily noticed, and so this condition leads to lower scores than the non-stationary office noise and clean conditions. Fig. 6 compares the results of the clean speech under two levels (-26 dB and -36 dB from the 16-bit saturation amplitude). At the lower erasure rates, the performance is worse for the low-level (-36 dB) input signals than the nominal-
subjective scores are bound to fluctuate. That the framework of the MOS test differs in each case, so the differences from those used in the evaluation procedure. This means used in the optimization of regression coefficients in Section 4 because the subjects, block-erasure rates and speech materials assess the adequacy of the estimated MOS values. This is partly In the evaluation presented in the previous section, we did not nominal-level inputs. This is probably because erasures are more easily noticed in the high-priority erasure and 10.0% random-erasure conditions. However, this is reversed in the cases of level (-26 dB) ones. However, this is reversed in the cases of the high-priority erasure and 10.0% random-erasure conditions. This is probably because erasures are more easily noticed in the nominal-level inputs.

To sum up, the results under the various speech conditions show consistent general trends in overall performance, and thus we can conclude that the method does not have much background noise or level dependencies.

6. Discussions

In the evaluation presented in the previous section, we did not assess the adequacy of the estimated MOS values. This is partly because the subjects, block-erasure rates and speech materials used in the optimization of regression coefficients in Section 4 differ from those used in the evaluation procedure. This means that the framework of the MOS test differs in each case, so the subjective scores are bound to fluctuate.

Fig. 7 shows an example of estimated MOS values for a clean female speech signal: “Oak is strong and gives shade.” 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, degree of inter-frame prediction used in the coding scheme, and more. The method presented here is a linear regression model to which other features can easily be added as new explanatory variables. This does not involve the tedious tuning of parameters and thresholds, but rather computational optimization based on Equation (6). This method has a further advantage in that we can objectively judge the adequacy of the added variables by evaluating the estimation errors and contribution rates.

7. Conclusions

We have presented a way to quantify the perceptual importance of speech segments. The method is based on a simple linear regression model with the logarithmic power, power proportion, and periodicity as the features. The adequacy of these features was verified by 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 integrated this method with the ITU-T G.722 coding scheme, and found that it provides adequate priority grading of blocks.

8. Acknowledgments

The authors would like to thank Hisashi Ohara for guiding our research and members of the Acoustic Information Group for their helpful advice and valuable discussions.

9. References


