Speech Recognition Performance as an Effective Perceived Quality Predictor

Wenyu Jiang, Henning Schulzrinne
Columbia University
Department of Computer Science
1214 Amsterdam Ave, Mail Code 0401
New York, NY 10027, USA
Email: {wenyu,hgs}@cs.columbia.edu

***Abstract*—**

Determining the perceived quality of packet audio under packet loss usually requires human-based Mean Opinion Score (MOS) listening tests. We propose a new MOS estimation method based on machine speech recognition. Its automated, machine-based nature facilitates real-time monitoring of transmission quality without the need to conduct time-consuming listening tests. Our evaluation of this new method shows that it can use the word recognition ratio metric to reliably predict perceived quality.

In particular, we find that although the absolute word recognition ratio of a speech recognizer may vary depending on the speaker, the relative word recognition ratio, obtained by dividing the absolute word recognition ratio with its own value at 0% loss, is speaker-independent. Therefore the relative word recognition ratio is well suited as a universal, speaker-independent MOS predictor.

Finally we have also conducted human-based word recognition tests and examined its relationship with machine-based recognition results. Our analysis shows that they are correlated although not very linearly. Also we find that human-based word recognition ratio does not degrade significantly once packet loss is large ($\geqslant$ 10%).

Keywords: perceived quality; speech recognition; packet audio; Internet telephony; subjective listening test; speech intelligibility; quality of service

I. Introduction

Voice over IP (VoIP) based on packet audio is becoming a popular service due to the cost savings and new services it can provide. However, due to the best-effort nature of the public Internet, packet audio sent over the Internet is subject to loss and delay jitter. This affects the audio quality as perceived by the end user. Determining the resulting quality generally requires human-based, subjective listening tests.

The most common type of listening tests are the Mean Opinion Score (MOS) tests [10]. In a MOS test, each listener rates an audio clip with a value called an *opinion score*. It can take on one of the following values: 5 (Excellent), 4 (Good), 3 (Fair), 2 (Poor) and 1 (Bad). The resulting average across listeners is called the *mean opinion score* (MOS).

MOS is the most widely used metric for perceived quality. However, MOS tests require human subjects and are time consuming. For network service providers such as voice over IP carriers, it is important to monitor in real-time the service quality as perceived by end users. So it is desirable to be able to predict perceived quality in an automatic and timely fashion.

Therefore we propose a new method of estimating perceived quality based on speech recognition performance. We have evaluated this scheme on the IBM ViaVoice [6] speech engine and found that speech recognition performance based on *word recognition ratio* (denoted as R) can be reliably mapped to speech quality (MOS). R is defined as the percentage of words that are recognized correctly, as shown in Equation (1). It is also termed the *absolute* word recognition ratio (R_{abs}) in this paper.

$$R_{abs} = \frac{\text{number of correctly recognized words}}{\text{total number of spoken words}} \tag{1}$$

When the network has a packet loss probability p, the recognition ratio is denoted as $R_{abs}(p)$. By measuring a curve for $R_{abs}(p)$ and another curve for $MOS(p)$, we can combine them to produce a new mapping from R_{abs} to MOS, i.e., the mapping function $MOS(R_{abs})$. We have conducted MOS listening tests and machine-based recognition tests to confirm the feasibility of $MOS(R_{abs})$ mapping.

Usually a speech recognizer's performance depends significantly on the speaker, due to accent, talk speed and other factors. Therefore a mapping function from R_{abs} to MOS calibrated for one speaker may not work well for another speaker. To tackle this problem, we use the *relative* word recognition ratio, denoted as R_{rel}. It is obtained by dividing the absolute word recognition ratio with its own value under zero loss condition, as shown in Equation (2), where p is the packet loss probability:

$$R_{rel}(p) = \frac{R_{abs}(p)}{R_{abs}(0\%)} \tag{2}$$

Thus, the relative word recognition ratio always has a value of 100% under 0% loss. Our evaluation shows that the curve of $R_{rel}(p)$ is speaker-independent. Therefore, it is well suited as a universal and unbiased MOS predictor.

To apply speech recognition to real-time quality monitoring, the sender simply transmits a pre-recorded speech clip over the network. The receiver looks up the speaker identity, finds the speaker's pre-calibrated $R_{rel}(0\%)$ value, performs speech recognition and compares the result to the stored original text. Then the receiver maps the measured recognition ratio to MOS, in real-time.

Finally, in addition to MOS listening tests, we have conducted human-based speech intelligibility tests. We ask the listeners to both rate a test audio clip (which is averaged to produce MOS) and transcribe it (which is used to measure human word recognition performance). Although there have been studies on specialized speech intelligibility tests, such as DRT (diagnostic rhyme test) [13], we are not aware of any study involving transcription of normal English sentences.

Our test results confirm that there is a mapping between machine and human word recognition ratio as well, although the mapping is not very linear. The detailed results are in section IV-E.

The remainder of this paper is organized as follows. Section II discusses related work. Section III describes the test setup we use to evaluate speech recognition performance. Section IV presents the evaluation results and discusses the results. Section V shows how to use speech recognition for quality monitoring. Section VI concludes the paper.

II. Related Work

The E-model [12] provides an analytical model for estimating perceived quality based on network performance. A transmission impairment such as packet loss is mapped to an impairment score and then to MOS. The ITU recommendation G.108 [12] provides a few standard mappings between packet loss and MOS for several commonly used voice codecs, as shown in Figure 1.

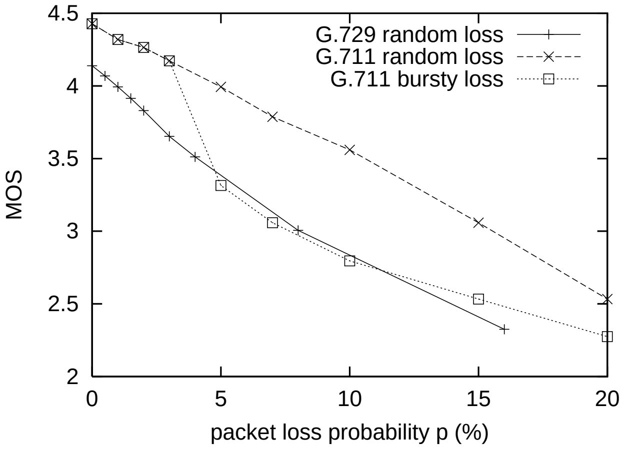


Fig. 1. E-model mapping from loss to MOS based on G.108 data

One problem with the E-model is that the loss-to-MOS mapping depends on both the voice codec and loss pattern. For example, the G.711 [11], [5] codec is more resilient to loss than other codecs, therefore its random loss MOS curve is higher than the random loss MOS curve for G.729 [9], as illustrated in Figure 1. Loss pattern affects quality as well. A burstier loss pattern [19], [2] usually implies lower quality, as seen in Figure 1 on G.711 between random and bursty loss [1]. To accurately estimate perceived quality, we need to calibrate a loss-to-MOS mapping for every combination of codecs and loss patterns that we may want to measure. The calibration requires human-based MOS listening tests, a time-consuming effort.

Chernick et al [4], [3] have evaluated the performance of a speech recognizer with the DoD-CELP [15] codec. Our study differs from theirs in four aspects. First, they focus on how bit error rate affects speech quality. We evaluate the effect of packet loss rate instead, because in the Internet, bit error is extremely rare except in wireless access networks. Second, their MOS tests have only two listeners per audio clip, whereas our tests involve 22 listeners, as detailed in section III-B. Therefore our MOS results are much more accurate. Thirdly, we use the word recognition ratio as our performance metric, whereas they use phoneme recognition ratio. This is detailed in section III-A. Finally, our evaluation also includes human recognition performance, whereas Chernick et al [4], [3] studies machine recognition performance only.

III. Speech Recognition Performance Evaluation Setup

A. *Speech Recognition Engine Setup*

We choose to evaluate the IBM ViaVoice [6] engine because it is a well known speech recognition product and it has a well documented SDK. We use the ViaVoice runtime engine and SDK on Linux to program the software required for speech recognition test.

First, we train the speech engine to adapt to a particular user's voice. To do so, we ask two native English speakers (A and B) to record their voice by reading two pre-defined training scripts (1 and 2) in ViaVoice. Script 1 is used for voice training, whereas script 2 is used for speech recognition testing.

We choose to evaluate the quality of G.729 [9] codec under packet loss because G.729 is a commonly-used codec and is representative of the CELP [1] family codec. The low bitrate nature of G.729 (8 kb/s compared to G.711 at 64 kb/s) makes it well suited for voice over IP due to reduced network load and increased bandwidth efficiency. Notice that we evaluate only G.729 performance without forward error correction (FEC) [17]. We have performed MOS listening tests with FEC in [14].

When training the ViaVoice engine, the training audio is first processed with the G.729 voice codec. Then during testing, the test audio is processed with G.729 along with a simulation of packet losses. The engine then performs speech recognition on the test audio and outputs the result to a log file.

By comparing the dictated text with the original script, we obtain R_{abs}, the *absolute word recognition ratio*, the percentage

[1] The G.108 recommendation, however, does not define what burstiness means or specify the degree of burstiness

of words that are correctly dictated by the speech engine. The text comparison is automated using the "wordscore" tool from U.C. Berkeley [18]. "wordscore" reads a reference text string and a modified string, then outputs the number of word insertions, deletions, and substitutions. These numbers all count toward calculation of the word recognition ratio. R_{abs}, along with the *relative* word recognition ratio R_{rel}, is the performance metric we will use to predict perceived quality (in MOS).

We first ask speaker A to read script 2. His reading is recorded and then split into 25 audio clips. These clips are processed using the G.729 codec under five different simulated loss conditions with 0%, 2%, 5%, 10% and 15% loss. This produces five audio clips per loss condition and reduces measurement noise due to any peculiarity of a particular audio clip. Later we also replicate the same test for speaker B based on the same text script and create another 25 audio clips. When testing audio clips of speaker B, we instruct the ViaVoice engine to use speaker B's voice model instead of speaker A's.

An alternative performance metric for speech recognition is the phoneme recognition ratio [4]. However, it has two drawbacks. First, it is less obvious to human users. Second, a speech recognition engine may use grammars and language rules to make a best guess of the input speech. After the engine applies grammar correction and language heuristics, the resulting sentence may no longer have the same phonemes as the original ones. Therefore we do not use the phoneme recognition ratio in our test.

B. MOS Listening Test Setup

Since our goal is to examine whether speech recognition performance can reliably predict perceived quality, we have performed MOS listening test as well. We use the same 25 audio clips for both the ViaVoice recognition test and the MOS listening test. A total of 22 listeners have participated in the MOS test. The MOS for each audio clip is then also averaged, with five clips per loss condition, in the same way the recognition ratio is averaged. The standard deviation of the MOS values is about 0.7 MOS on average. The corresponding 90% confidence interval is 0.11 MOS on average, which is reasonably accurate.

In addition to MOS testing, we also ask the listeners to transcribe the text for all test audio clips. Then we analyze the corresponding human absolute word recognition ratio with respect to packet loss.

IV. Speech Recognition Evaluation Results

A. Absolute Recognition Ratio vs. MOS

The first curve we obtain is from the MOS listening test, as in Figure 2. Not surprisingly, G.729's MOS decreases monotonically with respect to packet loss probability.

Figure 3 describes the result from the machine recognition test on ViaVoice. The absolute word recognition ratio R_{abs} is calculated according to Equation (1). Apparently, the recognition ratio also decreases monotonically with respect to loss.

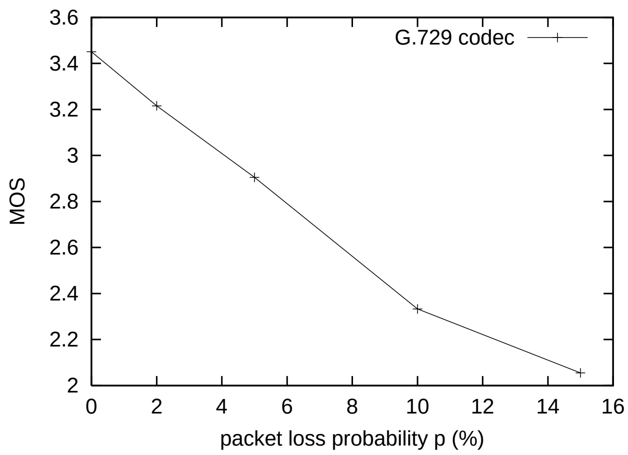


Fig. 2. Impact of packet loss on audio quality

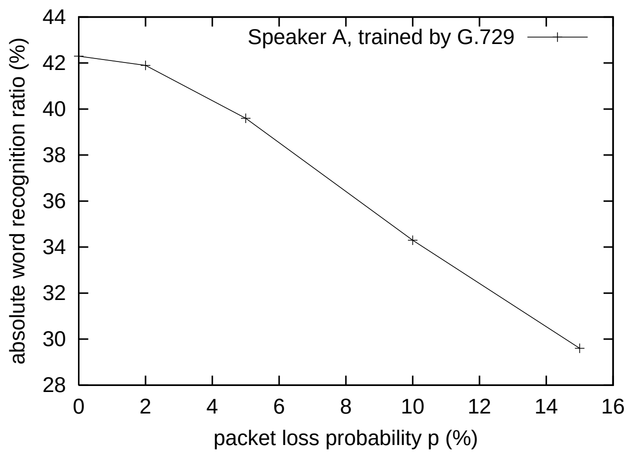


Fig. 3. Impact of packet loss on machine speech recognition, speaker A

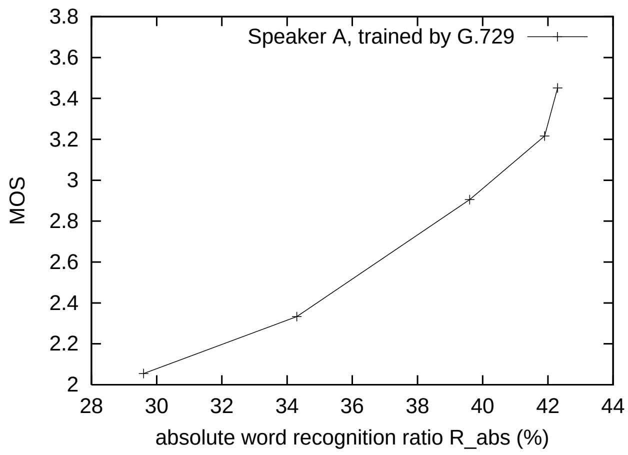
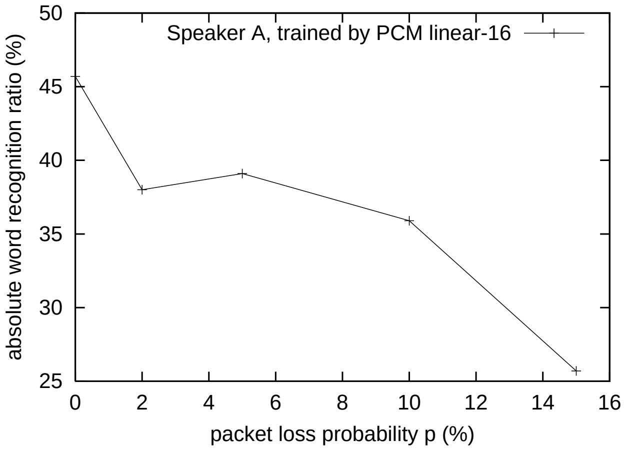


Fig. 4. Mapping from speech recognition performance to MOS

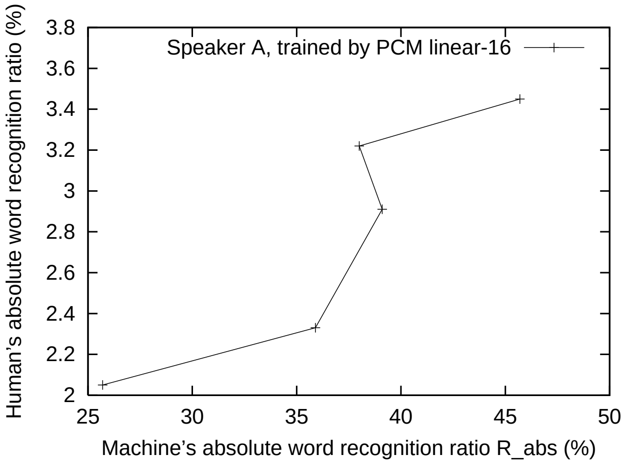
Since both MOS and recognition (R_{abs}) curves are monotonic with respect to packet loss probability p, it indicates a 1-to-1 mapping between MOS and R_{abs}. This is indeed the case after we eliminate the middle variable p and combine the two curves into one, as in Figure 4.

The resulting $MOS(R_{abs})$ curve is still monotonic. Therefore, speech recognition performance can be used to reliably predict perceived quality in terms of MOS.

B. Importance of Input Audio Coding during Voice Training



(a) R_{abs} vs. loss p



(b) Mapping from R_{abs} to MOS

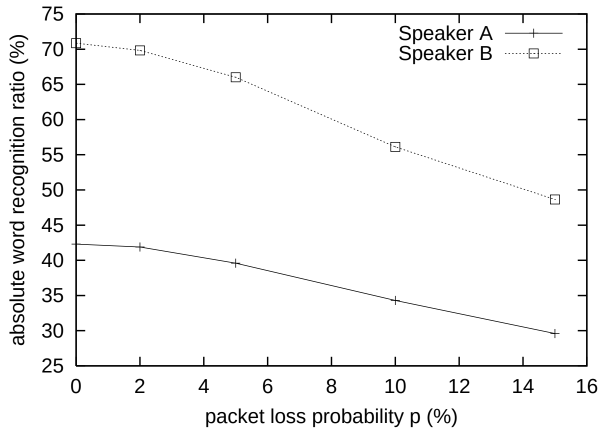
Fig. 5. Importance of voice training using the same codec as test audio clips: irregular results come up if training is by PCM linear-16 instead of G.729

The curves in Figure 3 and 4 are obtained by training the ViaVoice speech engine with G.729 processed audio. We have also experimented with training the ViaVoice engine with PCM linear-16 (direct audio) instead. We find the corresponding loss-recognition ratio curve to be irregular and non-monotonic, as shown in Figure 5(a). If we attempt to eliminate the middle variable p, the resulting $MOS(R_{abs})$ curve will not be a 1-to-1 function. This makes mapping recognition performance to MOS infeasible, as illustrated in Figure 5(b). Therefore, to reliably use speech recognition ratio to predict MOS, the training audio should be processed in the same audio codec as the test audio clips (in our case G.729).

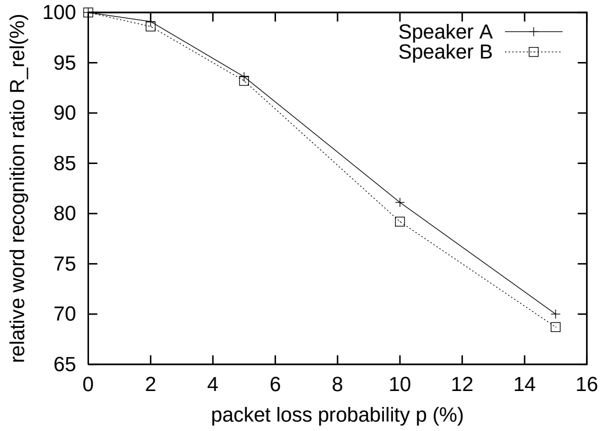
C. Accuracy of Speech Recognition based MOS Predictor

In Figure 4, the rightmost linear segment has a high slope. This segment corresponds to the loss range of 0-2$\%$. The high slope means that even if there is a small change in recognition performance, MOS will change significantly. This is probably because the ViaVoice engine is relatively robust under low loss rates. It also suggests that we should avoid using the absolute word recognition ratio R_{abs} as a MOS predictor if loss is very low, since a small measurement noise in R_{abs} would result in a significantly different predicted MOS. The remaining segments in Figure 4 all have small slopes, therefore MOS prediction within those ranges is much more accurate.

D. Universality of Relative Word Recognition Ratio as MOS Predictor



(a) Absolute word recognition ratio R_{abs} vs. loss p



(b) Relative word recognition ratio R_{rel} vs. loss p

Fig. 6. Suitability of relative word recognition ratio (R_{rel}) as a performance metric that is speaker-independent

The ViaVoice SDK user guide [7, p.49] cites a 90% accuracy for the average speaker without a heavy accent, when the speech is sampled at 22KHz. In comparison, the absolute word recognition ratio in Figure 3 (which is based on speaker A) even at zero loss is quite low, only about 42%. It can be caused by many factors. First, in our test the speech sampling rate is 8KHz instead of 22KHz, because most codecs in VoIP are 8KHz telephone-band codecs. Second, speaker accent can affect recognition performance greatly, and it turns out speaker A talks very fast. The low recognition ratio, however, does not necessarily interfere with MOS prediction at all, as long as the mapping curve between recognition ratio and MOS is monotonic and smooth.

Because the performance of a speech recognizer may differ significantly between speakers, the curve in Figure 3, which is based on speaker A, would be of only limited value if it can only predict quality reliably for a fixed speaker. To examine the dependency of MOS-recognition curve on the speaker, we have replicated the same set of test on speaker B, the second speaker in our evaluation test. The resulting curve is shown in Figure 6(a). The absolute word recognition ratio R_{abs} for speaker B at zero loss is much higher, about 70%. Therefore it is not possible to construct a universal MOS-recognition curve that applies to all speakers based on R_{abs}.

However, both curves in Figure 6(a) have similar trends. Therefore, if we divide each curve by its own recognition ratio at zero loss, that is, $R_{abs}(0\%)$, we obtain two *relative* recognition ratio (R_{rel}) curves both starting at 100% for zero loss, and we would expect the two relative ratio curves to look similar.

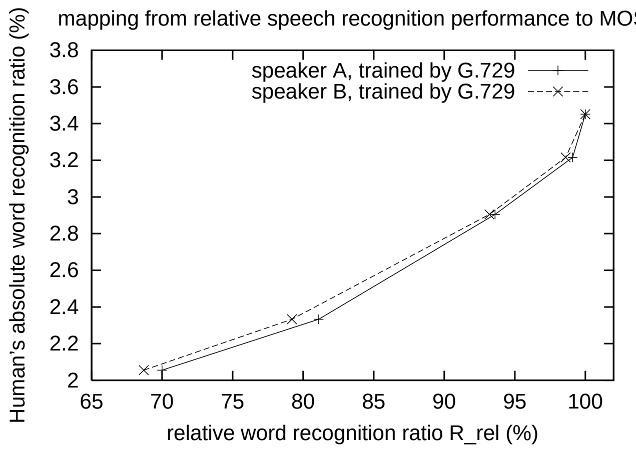


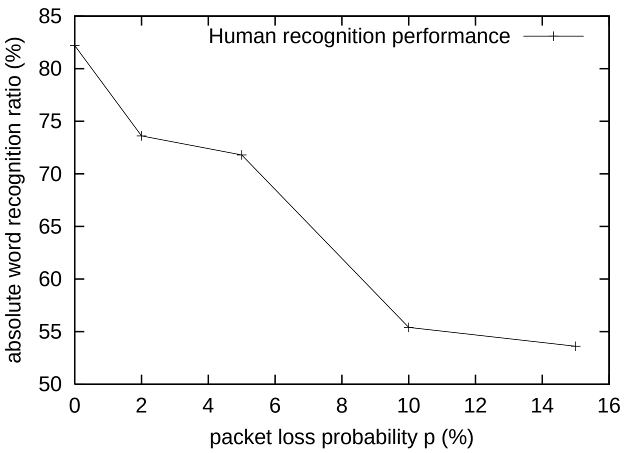
Fig. 7. Universal, speaker-independent MOS prediction based on relative word recognition ratio R_{rel}

Figure 6(b) confirms that this is indeed true. Both R_{rel} curves are very close to each other. Even though speaker dependency affects the absolute word recognition ratio greatly, the relative word recognition ratio R_{rel} is much more universal as a MOS prediction metric. This is illustrated in Figure 7, where the R_{rel} to MOS mapping curves for both speaker A and B are almost identical. Therefore we can use R_{rel} as a *universal*, speaker-independent MOS predictor.

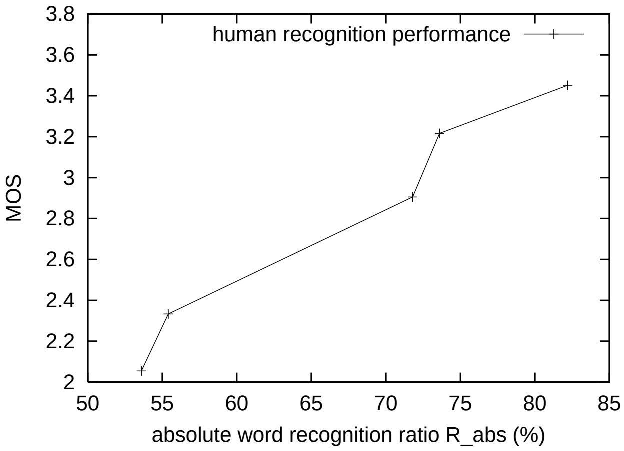
In a real monitoring application, it is not mandatory to use pre-recorded speech from more than one speaker, because the mapping in Figure 7 is speaker-independent. However, the fact that the relative word recognition ratio R_{rel} is speaker-independent could not have been predicted or inferred without performing the experiments in this paper. That is why we have examined two different speakers in our paper.

The two speakers we use in our test have completely different accents. In particular, speaker A talks very fast, whereas speaker B talks slowly and with a heavier accent. One would therefore expect their recognition performance curves to be completely independent. Yet their relative performance curves are very similar, as shown in Figure 7. Consequently, we can be quite confident about the universal nature of the relative word recognition ratio.

E. Human Recognition Performance



(a) Impact of packet loss on human speech recognition



(b) Mapping from human recognition performance to MOS

Fig. 8. Test results of human-based recognition performance

We have also conducted human-based recognition test by asking listeners to transcribe the text for each audio clip. The analysis of the human recognition results leads to the curve in Figure 8.

The curve in Figure 8(a) is not as smooth and linear as the MOS curve in Figure 2. It can be caused by many factors. First, speech intelligibility is not entirely the same as speech quality. Being able to recognize a word does not always imply good quality. Second, a human's ability to recognize words depends highly on his/her familiarity with the context of the speech. Therefore the inherent difference between listeners in cultural/educational background may introduce some variance.

Figure 8(a) shows two segments that are relatively flat. The first one occurs between 2% and 5% loss. It suggests that human's recognition ability remains roughly the same when there is some loss but if the loss is not high. The second flat segment is between 10% and 15% loss. This is probably because the speech at 10% loss is already too hard to recognize, therefore a 15% loss may not be much worse. A mapping from human word recognition ratio to MOS is also plotted, in Figure 8(b). The mapping shows that speech intelligibility is not the same as speech quality.

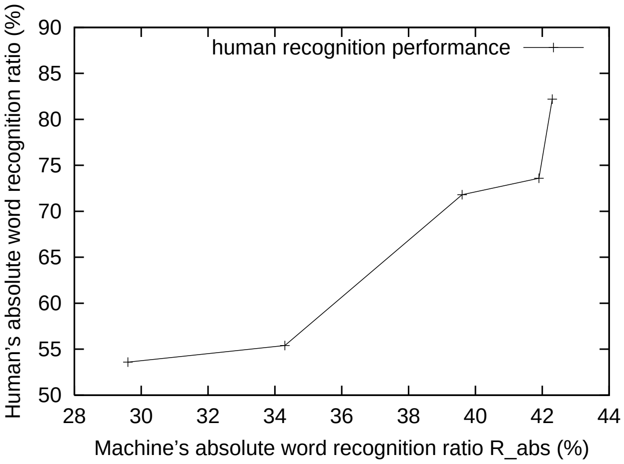


Fig. 9. Predicting human recognition performance using machine word recognition ratio

Finally, by combining both human and machine recognition results, we can establish a new mapping between machine and human word recognition ratio. This is done by eliminating the middle variable p on both Figure 3 and 8(a), and the new mapping is shown in Figure 9. The result indicates that it is possible to predict human speech intelligibility based on machine recognition ratio. The mapping is not close to linear, because human recognition performance is not very linear with respect to loss. In addition, care should be taken for segments with high slopes in Figure 9, because they introduce more prediction errors for the same amount of measurement noise in machine recognition ratio.

V. Application Scenarios

To use speech recognition for real-time quality monitoring, the sender should transmit a pre-recorded speech clip by some speaker. As long as the receiver end knows the original speech text and has calibrated $R_{abs}(0\%)$ for this speaker beforehand, it will be able to compute the current R_{abs} and thereby obtain R_{rel} using Equation (2). In fact, it does not even need to know the current packet loss probability, since the calculation of R_{abs} depends only on the result of speech recognition and the original text. This property makes it suitable for end-to-end black-box measurements. For instance, if using an IP telephony service with a phone-to-phone interface (e.g., an IP calling card), the two ends (analog phones) would not know the packet loss rate, but they can still perform speech recognition and compare to the stored original text.

Then, applying the universal mapping in Figure 7, the receiver will be able to tell how good is the current speech quality.

With this approach, the receiving end need not store the original speech clip itself, but only the original speech text. This is much more scalable when storing many clips, since the speech clips can take up a significant amount of disk space. In fact, even if the receiver stores the original speech clip, the text is still needed because speech recognition is not going to be 100% accurate even with no packet loss.

The approach we take is active measurement, where the sender explicitly generates pre-recorded traffic, controls who the speaker is and what material he/she speaks. Our method is less applicable to passive measurement, firstly because it is difficult to know the identity of the speaker. Even if the speaker is known and his/her $R_{abs}(0\%)$ value already calibrated, the original speech text of an ad-hoc conversation cannot be known in advance. Therefore, the receiver will not be able to estimate the recognition ratio.

VI. Conclusions and Future Work

We present a new method of estimating perceived quality based on speech recognition performance. We have evaluated its effectiveness on the IBM ViaVoice speech engine over a wide range of packet loss rates, and found that the word recognition ratio can serve as a reliable predictor for the Mean Opinion Score (MOS), the most commonly used perceived quality metric.

For this method to be reliable, we find that a speech engine should be trained using audio processed by the same audio codec that we intend to test quality on. Based on our findings, the absolute word recognition ratio (R_{abs}) to MOS predictor is more accurate over higher loss rates (2-15%), but less so for low loss rates (0-2%). This behavior is likely due to the robustness of the speech engine under low packet loss rates.

To examine this MOS predictor's dependency on speakers, we replicated the same set of test on a different speaker. Although the absolute word recognition ratio R_{abs} is much different from the first speaker, the relative word recognition ratio R_{rel}, obtained by dividing the absolute ratio with its ideal maximum value (i.e., the R_{abs} at 0% loss), remains almost the same for different speakers. Therefore, the relative word recognition ratio R_{rel} is well suited as a universal, speaker-independent MOS predictor.

Finally we have also investigated the trend of human word recognition ratio over different network conditions. The results indicate that human recognition performance (speech intelligibility) is related to perceived quality (MOS), although the trend is not close to linear. We find two loss regions where human recognition performance remains relatively flat. The first is when there is some but not very high loss (2-5%). The second is when the loss is too high (10-15%), presumably because recognition is already too difficult with 10% loss. Because humans are good at guessing, speech intelligibility is not necessarily the same as speech understanding, although the latter task should be easier. However, defining speech understanding is a much more complex task than defining say, the word recognition ratio.

Our analysis on the relationship between human and machine word recognition ratio shows that it is possible for a speech recognizer to serve not only as a MOS predictor, but also as a speech intelligibility predictor, although care should be taken for regions where the prediction error may be large.

In this paper we have evaluated speech recognition performance using the G.729 codec. We plan to examine other codecs, such as G.726 ADPCM [8] and GSM [16], and verify whether the universal MOS predictor can be used for these codecs as well.

Finally, another possible extension of this work is to perform quality monitoring in a real world setting, using commercial VoIP products and services.

REFERENCES

[1] John C. Bellamy. *Digital Telephony*. John Wiley & Sons, Inc., third edition, 2000.

[2] Jean-Chrysostome Bolot, Sacha Fosse-Parisis, and Don Towsley. Adaptive FEC-Based error control for interactive audio in the Internet. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, New York, March 1999.

[3] C. Michael Chernick, Stefan Leigh, Keven L. Mills, and Robert Toense. Testing the ability of speech recognizers to measure the effectiveness of encoding algorithms for digital speech transmission. In *IEEE International Military Communications Conference (MILCOM)*, January 1999.

[4] Michael Chernick, Stefan Leigh, Kevin L. Mills, and Robert Toense. Can speech recognizers measure the effectiveness of encoding algorithms for digital speech transmission? Technical report, National Institute of Standards and Technology, 1999.

[5] Working group T1A1.7. Results of a subjective listening test for G.711 with frame erasure concealment. Technical report, Committee T1, May 1999.

[6] IBM. IBM ViaVoice ASR SDK for Linux. Available at http://www.ibm.com/software/speech/dev/sdk_linux.html.

[7] IBM. SMAPI User's Guide: IBM ViaVoice Software Developer's Kit.

[8] International Telecommunication Union. 40, 32, 24, 16 kbit/s adaptive differential pulse code modulation (adpcm). Recommendation G.726, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, December 1990.

[9] International Telecommunication Union. Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear-prediction. Recommendation G.729, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, March 1996.

[10] International Telecommunication Union. Subjective performance assessment of telephone-band and wideband digital codecs. Recommendation P.830, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, February 1996.

[11] International Telecommunication Union. Pulse code modulation (PCM) of voice frequencies. Recommendation G.711, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, November 1998.

[12] International Telecommunication Union. Application of the e-model: A planning guide. Recommendation G.108, Telecommunication Standardization Sector of ITU, Geneva, Switzerland, September 1999.

[13] N. Jayant and P. Noll. Digital coding of waveforms : principles and applications to speech and video, Prentice Hall, 1984.

[14] Wenyu Jiang and Henning Schulzrinne. Perceived quality of packet audio under bursty losses. Technical report CUCS-009-01, Columbia University, Department of Computer Science, 2001.

[15] Office of Technology and Standards. Telecommunications: Analog to digital conversion of radio voice by 4,800 bit/second code excited linear prediction (celp). Federal Standard FS-1016, GSA, Room 6654; 7th & D Street SW; Washington, DC 20407, 1990.

[16] Siegmund M. Redl, Matthias K. Weber, and Malcolm W. Oliphant. *An Introduction to GSM*. Artech House, Boston, 1995.

[17] J. Rosenberg and H. Schulzrinne. An RTP payload format for generic forward error correction. Request for Comments 2733, Internet Engineering Task Force, December 1999.

[18] Chuck Wooters, Andreas Stolcke, and Hiroaki Ogawa. Manual page: wordscore - simple word-error-rate calculation. Technical report, International Computer Science Institute, University of California, Berkeley, 1996. Available at ftp://ftp.icsi.berkeley.edu/pub/real/dpwe/wordscore.tar.gz.

[19] Maya Yajnik, Sue Moon, Jim Kurose, and Don Towsley. Measurement and modelling of the temporal dependence in packet loss. In *Proceedings of the Conference on Computer Communications (IEEE Infocom)*, New York, March 1999.