A Bayesian-Maximum Entropy Approach to Subjective Voice Quality Testing

Ali E. Abbas, Athina P. Markopoulou

{aliabbas, amarko}@stanford.edu

Department of Management Science and Engineering-Stanford University
Department of Electrical Engineering-Stanford University

Abstract. In order to assess the performance of Internet telephony, it is often necessary to translate network impairments (such as packet loss, delay and jitter) into human perceived quality (which is quantified in terms of subjective voice quality ratings). Subjective quality testing is expensive and typically involves a large number of questions and humans. It is therefore important to design simple and reliable subjective testing experiments. This paper presents a method to assess the subjective quality of a number of speech samples that have incurred various degrees of the same network impairment. Questions are asked according to an adaptive algorithm until all voice ratings are elicited within a desired accuracy. Our algorithm (i) uses information theory to minimize the expected number of questions needed and (ii) uses binary questions, which are simpler than the types of questions used by standard subjective testing procedures.

1. INTRODUCTION

In the last decades, voice communication is taking place not only over the traditional telephone network, but also over wireless cellular networks and over the Internet. Transmission of voice over a communication network is subject to various impairments. To assess the effects of these impairments, as perceived by the users, formal testing procedures have been developed and standardized, known as "subjective voice quality testing". Subjective quality testing consists of asking the opinion of people on the quality of speech samples. In general, subjective evaluation is considered expensive, because it requires special lab conditions, a large number of questions and listeners. Therefore, designing a simple reliable experiment is of major importance. More specifically, it is desirable to elicit the opinion of the humans involved, as accurately as possible, using the least number of questions.

We observe that the problem of subjective voice quality evaluation fits well within the realm of utility elicitation techniques in decision analysis. The analogy arises from the properties of the voice quality rating curves that increase monotonically as the impairments decrease (or decrease monotonically as the impairments increase). If, in addition, the rating curves are normalized from zero to one over a continuous scale, then many of the algorithms proposed for utility elicitation can be applied to the elicitation of subjective voice quality ratings. In this paper, we apply a Bayesian-maximum entropy technique, developed for utility elicitation in [1], to elicit the opinion of a single person on the quality of a number of samples, impaired with increased amount of (the same type of) impairment. Our approach can assist people working in Speech Subjective Testing Labs, to design efficient and simpler questionnaires. In order to demonstrate our algorithm, we simulate a widely known (and standardized) experiment, in which 10-packet loss conditions were translated to speech quality. Our results show that we obtain the correct values of subjective ratings after only a few binary questions.

2. APPLICATION CONTEXT

2.1 Voice transmission over a network

We are all familiar with the use of telephony for communication. Traditionally, telephone calls take place over the public switched telephone network. In the last decades, wireless and cellular networks have been used to provide telephony services to mobile users. Finally, the Internet infrastructure has enabled cheap –and often free-voice communications, typically at low quality.

A "good" telephone call is one in which the participants can communicate without difficulties, or annoying and distracting effects. However, when taking place over the above-mentioned communication networks, voice conversations are subject to various impairments. For example, we are all familiar with delay impairments that may occur during cross-Atlantic phone calls: the response from the other person is long, and the participants end up talking at the same time or talking in turns. Echo impairments may also become noticeable when delay is high. Another possible impairment is speech distortion: modern encoders achieve high compression at the cost of distortion of the speech signal. Parts of the speech signal may be further lost during transmission over lossy networks, such as wireless network or the Internet during congestion.

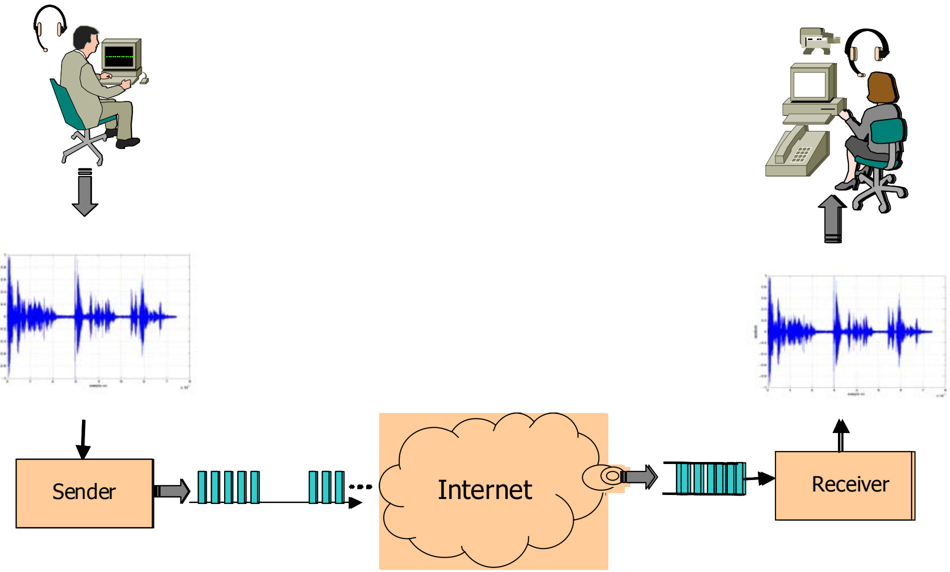

FIGURE 1. Example of a call taking place over the Internet.

Fig.1 shows an example of a conversation that is taking place over the Internet. The person at the left is sitting on his desktop and is talking to the person at the right. The “sender” software on his desktop performs the following functions: records the voice, encodes the speech sample, puts it in packets of equal size and sends them over the Internet at fixed time intervals. As they go through the Internet, some packets may be lost and some packets may be delayed more than others. At the other end, the “receiver” software performs the following functions: receives the packets, puts them in a buffer to be played out at fixed intervals, reconstructs the information of any lost packets and plays out the reconstructed voice signal at the speaker or headphone. Similar steps happen from right to left when the second person speaks.

2.2 Subjective Quality Testing

The ultimate judges for the quality of a conversation are the users themselves and the most appropriate metric is their opinion; thus the term “subjective quality” has been associated with the ratings. Standards have been published by the International Telecommunications Union (ITU-T), and in particular by Study Group 12 (SG-12), in order to define subjective quality and to specify the ways to measure it. Most tests are carried out by interviewing people in a standardized test environment.

The ITU-T Recommendation is ITU-T P.800 [2] defines the most commonly used subjective metric: MOS, which stands for “Mean Opinion Score”. In these tests, a set of K listeners, listen to speech samples and give them a rating on a discrete scale from 1 to 5 (1 is the worst and 5 is the best possible rating). These ratings are then averaged over the K listeners, and this average is the rating for the sample, the Mean Opinion Score. MOS greater than 3.6 is considered acceptable for today’s Internet, while MOS above 4.0 is desirable for the quality to be comparable to the traditional telephone network. The above rating system is summarized in Fig.2, which is taken from [3].

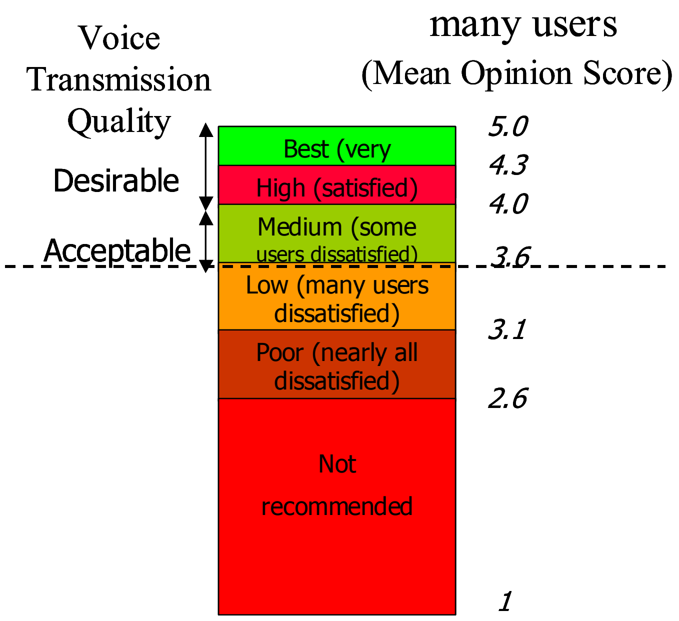
Voice Transmission Quality
many users
(Mean Opinion Score)
Desirable
Acceptable
Best (very
High (satisfied)
Medium (some users dissatisfied)
Low (many users dissatisfied)
Poor (nearly all dissatisfied)
Not recommended
5.0
4.3
4.0
3.6
3.1
2.6
1

FIGURE 2. Subjective Voice Quality and Mean Opinion Score.

Special labs exist for the purpose of conducting subjective quality tests, according to the ITU-T standards, such as COMSAT. Also, large telecommunication companies, such as Nortel, AT&T and Lucent, have their own labs and a group of specialists dedicated to subjective testing.

2.3 Mapping Network Impairments to subjective quality

The reader might have already noticed that there is a gap between the network impairments (section 2.1) and subjective quality (section 2.2). On one hand, network impairments can be described and measured using objective parameters. On the other hand, subjective quality is really what matters but requires asking the opinion of a number of humans, which may not always be feasible or desirable. As an example of objective parameters vs. subjective quality, Fig.3a shows an original speech sample. Fig. 3b shows the same sample, with 50% of it (shown in red) considered lost in the network. When the sample of Fig. 3b is played out, the missing parts will be perceived as interruptions or noise. The more packets are lost, the more annoying the perceived effect. However, how much more annoying? Does the annoyance increases linearly or with the square of packet loss? Fig. 3c shows another kind of impairment: the speech sample is received correctly but it is played out twice as fast. This will be perceived as a change in the pitch of voice. However, how annoying is this effect? How does the annoyance increase with playing speed? E.g. does double speed cause double or more annoyance?

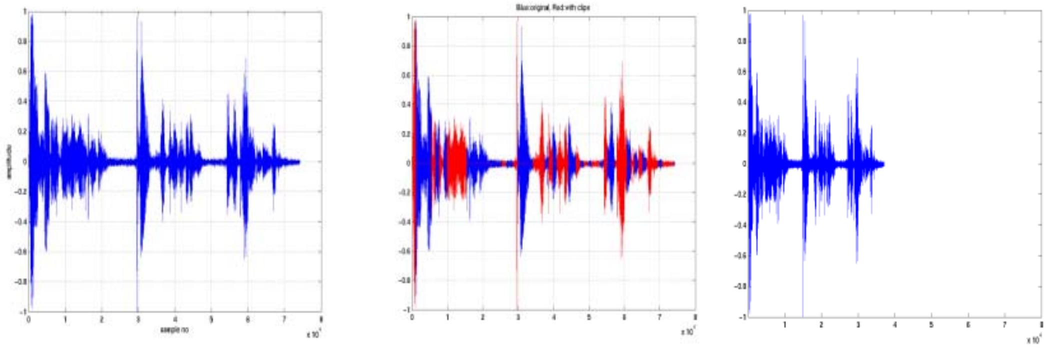

FIGURE 3. Example impairments (a) original sample (b) packet loss (e.g. 50% packets uniformly lost) (c) change in pitch (sample played twice as fast).

Network impairments can be objectively measured by a telephone company using a monitoring device attached to the network and collecting metrics such as packet loss, average delay, delay jitter etc. A customer may also be interested in such measurements to make sure that he indeed received the promised service. A translation from the measured network metrics to subjective quality is useful to understand the effect perceived by the user. As it is practically impossible to have people sitting on the phone and giving statistics about the quality of phone calls on a continuous basis and for all network paths, experiments are conducted in special labs that quantify the relation between the objective and subjective measures.

Fig. 4 shows the results of three such experiments, as provided by the ITU-T recommendations G.107 [4] and G.113 [5]. The purpose was to quantify the effect of packet loss to subjective quality. The designers of the experiment took some standardized samples encoded using ITU-T G.711. They asked a number of people for their opinion on these samples and obtained the MOS for packet loss 0%. Then, they artificially erased 1%, 2% …20% of the speech, considering 10ms continuous segments of speech lost in a uniform or burst way. They asked the opinion of each person for each impaired sample, obtained and plotted the MOS for every packet loss percentage. The scenario for the top curve uses packet loss concealment (PLC) and uniform packet loss. The scenario for the middle curve used packet loss concealment and bursty packet loss. The bottom curve corresponds to uniform packet loss but does not use PLC used at the receiver, thus the greater impairment. For all three scenarios, the curves are monotonically decreasing. This is expected as speech quality can only decrease when loss percentage increases.

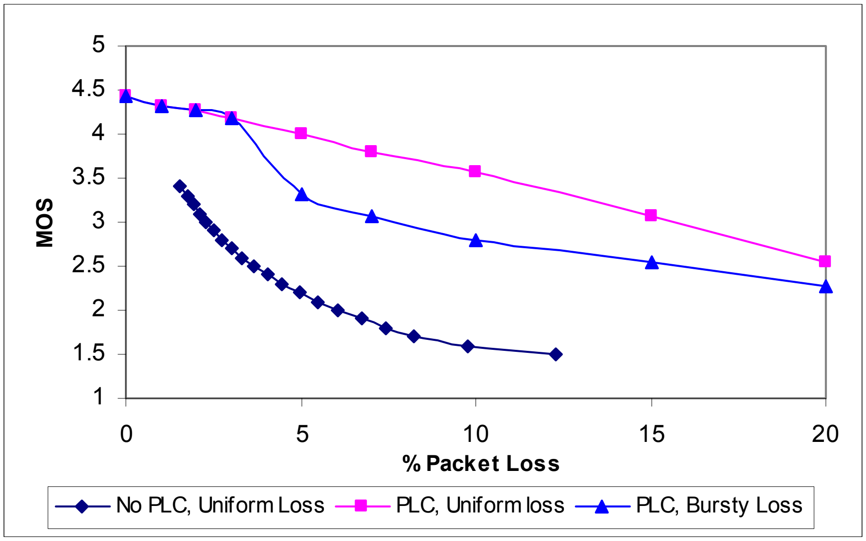

FIGURE 4. Mapping packet loss to subjective quality for G.711 encoded speech. These curves are taken from the ITU-T standards G.107 and G.113.

Although ITU-T P.800 specifies the conditions of the experiment, it does not specify the questionnaire to be asked. The current practice for obtaining MOS-impairment curves is to ask *nK* questions. Indeed, if the number of humans is *n* and the number of impairment conditions is *K*, then *nK* questions are required in total to obtain one of the MOS-Impairment curves.

Clearly, there are several issues with this practice.

First, the number of questions is large. In this paper, we are able to minimize the number of questions per listener, using *entropy-coding principles*. We are able to further reduce the number of questions per listener, capitalizing on the observation that impairment conditions are ordered and *quality decreases with increasing degree of impairment* (e.g. 5% packet loss will lead to MOS no larger then that of 2% loss). Therefore, the rating of each speech sample provides bounds for the other ordered samples. E.g. if a human subject believes that the speech sample with 2% loss sounds worse than quality 3 (i.e. (MOS(2%)<=3), this implies that all samples with more loss than 2% should also have a quality less than 3 (i.e. (MOS(l%)<=3 for all l>=2), assuming of course that the human subject gives consistent answers.

Second, the questions asked according to ITU-T P.800 are not easy to answer. E.g. it is often difficult to decide whether a sample should be rated with 2 or 3 in the 1-5 discrete scale. Furthermore, there is an inherent tradeoff between granularity in the scale and repeatability/consistency of the experiments. In our approach, we replace the 1-5 discrete scale with a 0-1 continuous scale (using the Von Neuman and Morgenstern Standard Gamble approach) and at the same time we replace the ITU-T P.800 test questions, with easier binary questions. E.g. we are asking whether a sample sounds better or worse then 3.5 instead of asking an exact number among 1,2,3,4 or 5. The binary question has two advantages: (i) it is easier for the human subject to answer and at the same time (ii) it allows for increased granularity, using the entire spectrum [1,5] rather than 5 numbers.

Finally, there has been extended criticism on many aspects of the ITU-T P.800 procedures. For example in [6], the discrete scale is criticized for both the difficulty of answering and its coarse granularity; a continuous scale is proposed instead. Furthermore, the paper argues that the terms corresponding to the 5 numbers have different interpretations in different languages.

In the next sections we formulate and solve the problem of minimizing the expected number of questions per listener, to elicit a given Quality-Impairment curve within a desired accuracy.

3. PROBLEM FORMULATION

Let us consider a single listener and let us assume that s/he gives consistent answers according to a quality (Q)-impairment (I) curve, unknown to us. This means that the same person will not rate a sample with 2% loss as worse than quality 3 and another sample with 5%>2% loss as better than 4. (In future work we extend this work to incorporate mistakes and inconsistencies).

We know that the impairment conditions are ordered such that ($I_0<I_1<I_2<\ldots<I_K$) and we also know that quality is decreasing with increasing impairment, i.e. $Q_0>=Q_1>=\ldots Q_K$. We would like to elicit the quality rating of each sample using the minimum expected number of questions, as accurately as possible and using questions that are easy to answer.

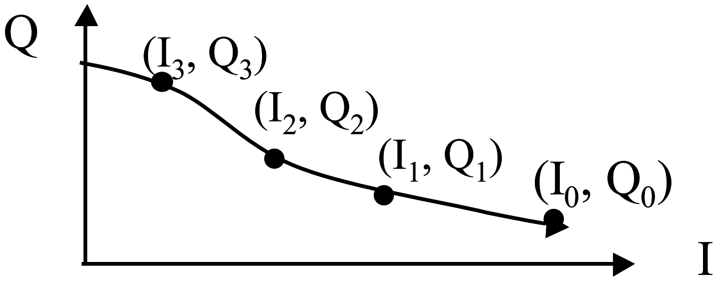

FIGURE 5. Voice quality rating curves.

A characteristic that makes the ratings of a single listener particularly similar to utility values is that they are inherently ordered: the listener perceives worse quality with increasing degrees of impairment.

In this paper, we exploit this ordering property to calculate the maximum-entropy joint distribution for each listener's "*j*" ratings: Q_{ij}, *i=1,…,k, j=1,…n*. We use entropy-coding principles to minimize the expected number of questions needed to obtain the Q_{ij}'s within a desired resolution. The number of questions can be further reduced by testing only a subset of the *k* samples and by using the maximum entropy principle to infer the remaining ratings. In the next section, we present a solution to this elicitation problem and a geometric representation for all points Q_i, , *i=1,…,k* provided by the same listener *(j)*. For the rest of the paper, we refer to those points as the single listener's Q-I curve: $(I_i,\ Q_i)$, *i=1,…,k*. The same procedure could be applied to all listeners.

4. SOLUTION

4.1 Geometric Interpretation and Observations

Note that the properties of quality rating curves provide for a geometric interpretation of the subjective voice quality-rating problem. For example, consider the two points Q$_2$ and Q$_1$ on the quality-rating curve of Fig. 6(a). These two points can be represented as a point $Q=(Q_1,Q_2)$ in the shaded region shown in Fig. 6(b), which we call the quality rating volume. Furthermore, if the quality rating curve is normalized to range from zero to one, then the sample with the highest impairment and that with the lowest impairment have quality ratings of 0 and 1 respectively.

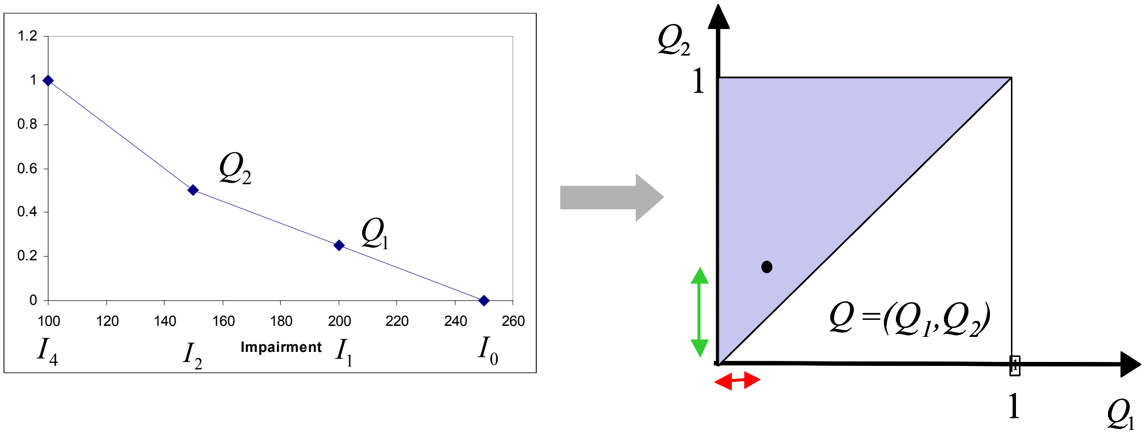

FIGURE 6. (a) Voice quality rating curves. (b) Vector space representation.

Note that any point in the quality volume represents the ordering of the given impairments but assigns different quality rating values to them. For example, the point (Q_1,Q_2) suggests that the rating of $Q_2 \geq Q_1$ and that the impairment of $I_2 \leq I_1$ but any values of quality ratings in the shaded region satisfy this condition. In other words, if we only know the order of the impaired samples, and would like to guess the location of the quality-rating vector, it is reasonable to assume that its location is uniformly distributed over the quality volume. By thinking of the quality rating points as ordered statistics, we can deduce the following properties.

Quality rating assignment given the order of the samples:

The maximum entropy marginal distributions for quality values of a set of K ordered impairments, is the family of Beta distributions, *Beta (j,K-j*-1), *j=1,...K-2*. The mean of these distributions, *j/(K-1),* is the quality value of each impairment, *j*. Impairments *j=0* and *j=K-1* are given deterministic with values 0 and 1 respectively.**Example**

Consider a quality-rating curve with 4 impairments. The sample with lowest rating and that with the highest rating have values of 0 and 1 respectively. As discussed above, the remaining two samples have marginal distributions of Beta(1,2) and Beta(2,1). These distributions are shown in Fig. 7. The quality value assignment for the 4 samples starting from worst to best is thus equal to: Q=*(0, 1/3, 2/3,1)*.

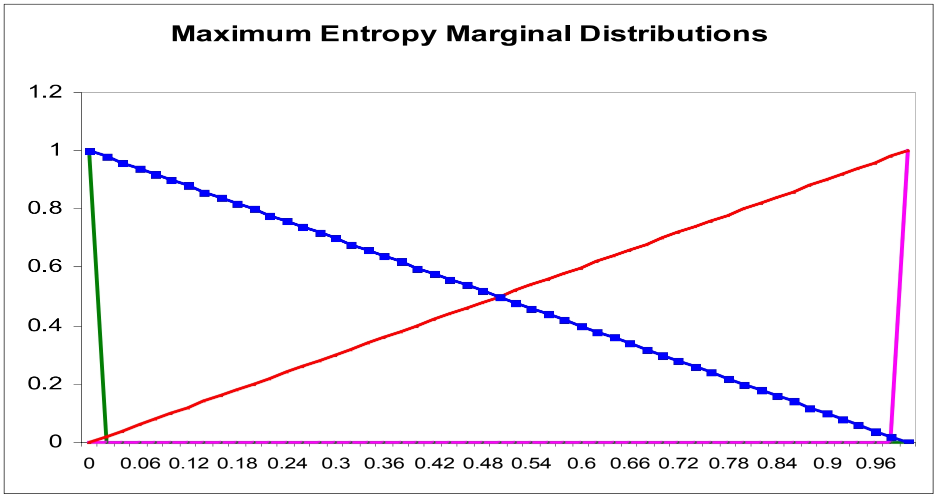

FIGURE 7. Marginal distributions for ordered samples.

4.2 Algorithm

Given K ordered samples, the problem of eliciting the subjective voice quality ratings is equivalent to searching for the quality vector over the quality volume. If we require binary questions for the search, then the optimal question-selection process partitions the quality volume into two halves each time with equal probability. If the distribution over the quality volume is uniform, this implies that the optimal question-selection process partitions the quality volume into two geometric halves. This is shown in Fig.8. The partitioning of the quality volume into two geometric halves is simple for two dimensions but gets more complicated as the dimensionality of the quality vector increases. Fortunately, the break point needed to partition the quality volume for each sample is, by definition, equal to the median of its marginal distribution at each stage. Recall that the marginal distributions at the start of the questionnaire are the Beta distributions described above. At any given stage the marginal distributions can be obtained by marginalizing the uniform joint distribution over the new bounds for the rating values. Alternatively, the marginal distributions can be updated by conditioning on the new bounds using Bayes rule, or by using Monte Carlo simulation, where we generate uniform samples on the quality volume and select the samples that lie in the current search space.

Note also from Fig.8c and Fig.8d that the optimal partition of the quality volume is not unique (any sample can partition the quality volume into two equal halves). The algorithm will give preference to the samples whose quality ratings we are most uncertain about (whose marginal distribution has the highest entropy).

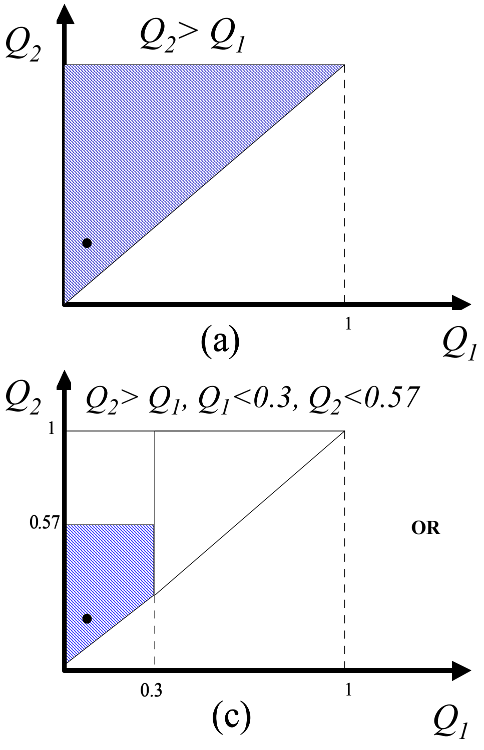

FIGURE 8. Partitioning the quality rating volume.

We now describe the steps of the algorithm using the steps shown in Alg.1 below and the flow chart of Fig. 9.

Alg.1 Algorithm for voice quality ratings elicitation

```
Initially
    • Use the maximum entropy principle to compute the marginal distributions for the quality of each sample.
While “stopping criterion” is not met:
    • Select the sample, $S_{max}$ , whose marginal distribution has the highest entropy
    • Partition this distribution at the median, $M$.
    • Ask the binary question: “Is $Q(S_{max})>M$ ?
    • Update the marginal distributions of each sample
```

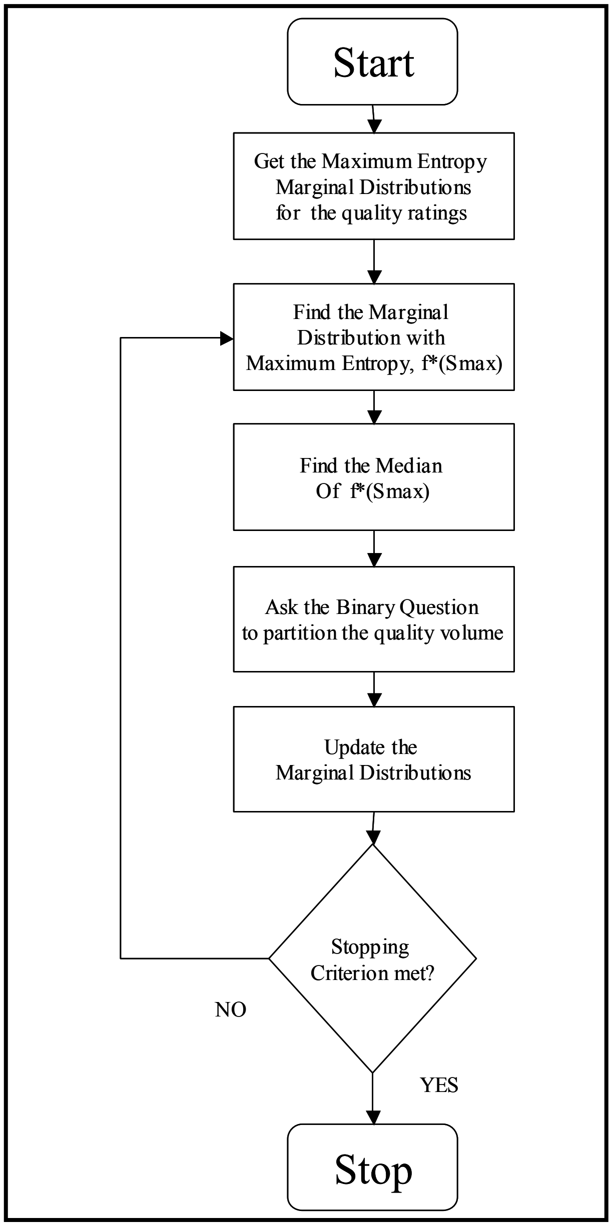
Start
Get the Maximum Entropy Marginal Distributions for the quality ratings
Find the Marginal Distribution with Maximum Entropy, f*(Smax)
Find the Median Of f*(Smax)
Ask the Binary Question to partition the quality volume
Update the Marginal Distributions
Stopping Criterion met?
NO
YES
Stop

FIGURE 9. Flow chart for quality-rating algorithm.

Stopping Criteria

Let us now discuss the stopping criterion for the algorithm. From the properties of entropy-coding theory, an expression for the stopping criterion can be obtained when the expected number of questions needed to locate the quality vector within a hypercube of side length (Δ) has been asked. This is achieved, by the discretizing the quality volume into hypercubes of side length equal to Δ and calculating the discrete form entropy expression, where

$$\text{Expected number of questions} \cong \log\text{ (number of hypercubes)}$$

$$= \log\left(\frac{\text{Quality volume}}{\text{volume of hyperubes}}\right) = \log\left(\frac{\frac{1}{(K-2)!}}{\Delta^{K}}\right).$$

5. APPLYING THE ALGORITHM

To demonstrate the application of the algorithm, we "simulate" a subjective test in the following sense. We consider the experiments whose results are shown in one of the curves of Fig.4. The algorithm poses the questions that would be given to a human subject, participating in the test in the lab. We assume that the human answers each question according to one of the curves. In other words, we assume that each curve captures perfectly the preferences of the subject. The "simulation" proceeds iteratively as follows. The algorithm asks the first question; the (simulated) human consults the curve and answers with "greater" or "less"; based on this answer the algorithm asks the next question; and so on. We see that our algorithm manages to reproduce the original Q-I curve after only a few binary questions

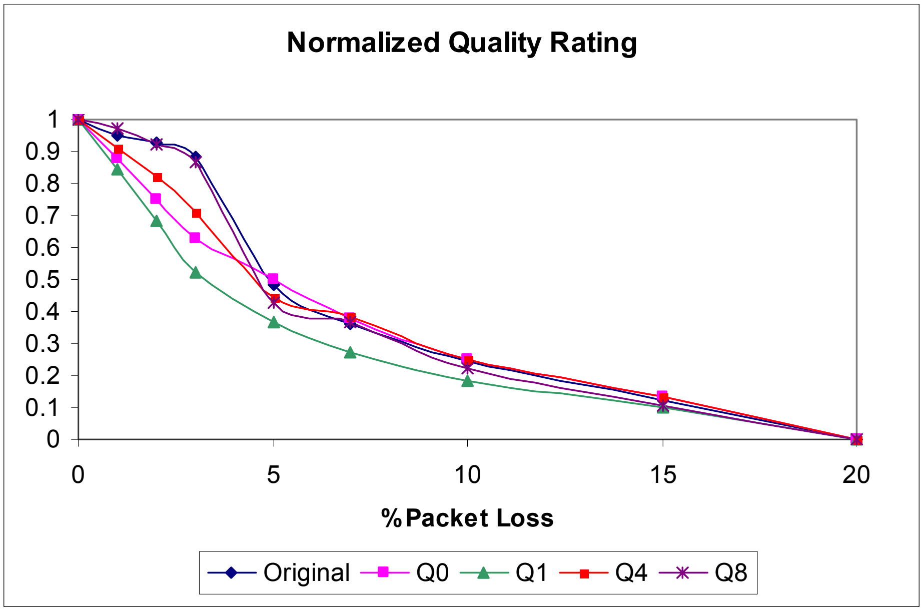

FIGURE 10. Original and elicited Q-I curves for the "bursty loss" experiment.

Example 1

First, consider the middle curve of Fig.4, which corresponds to the bursty loss scenario. It is re-drawn in Fig.10 below, and labeled as "original" curve. In the same figure, we now plot our estimate for the Q-I curve after 0 questions (labeled as "Q0"), 1 question (Q1), 4 questions (Q4), and 8 questions (Q8). We can visually see that our algorithm successfully reproduces the original Q-I curve, after only a few binary questions, much earlier than the exhaustive traditional approach.

To quantify the "closeness" between the elicited and the original curve after each question, we define the distance/error between an elicited and the real curve, as the mean square (or Euclidean) distance between the two Q vectors. Fig. 11 shows that the error decreases asymptotically with the number of questions.

It is interesting to note that the error decreases asymptotically, because each answer divides the volume in two. However, the error does not necessarily decrease monotonically with the number of questions. For example, in Fig.11, the error increases after the first question. This means that, we cut the volume in half, but the distance between the new estimate and the true value is larger than before. However, the algorithm is guaranteed to converge asymptotically.

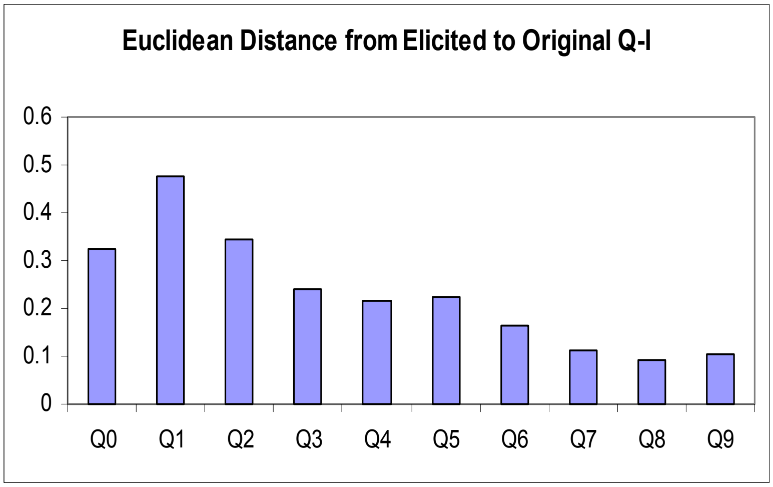

FIGURE 11. Error after every question for the "bursty loss" experiment.

Example 2

As a second example, we considered the top curve of Fig.4, which corresponds to the uniform loss scenario. We repeat a similar simulation. Similar to Fig.10, we plotted the estimated Q-I curve after each question and we observed that it converged fast to the original one. We omit this figure here and we show only the error that quantifies the distance of our estimate from the true Q-I curve. Fig. 12 shows the error with the number of questions decreasing more smoothly in this case.

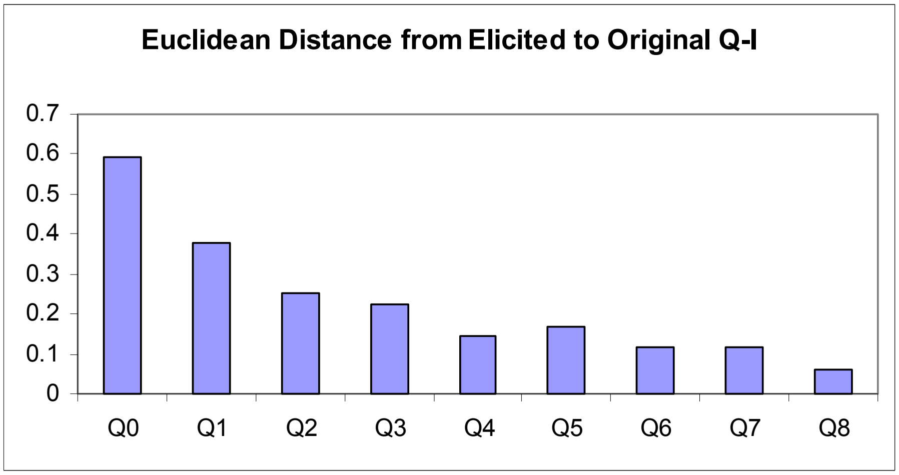

FIGURE 12. Error after every question, for the "uniform loss" experiment.

6. CONCLUSIONS

In this paper, we designed a questionnaire to elicit the Quality-Impairment curve of a single human subject, using the minimum expected number of questions. We demonstrated the efficiency of our algorithm using simulation on some standard MOS-loss curves. The same approach could be followed for other impairments such as delay or jitter.

In this work, we assumed that the human subject gives consistent answers according to a unique Q-I curve. Future work will continue in two directions. First, we are considering experimental validation with actual subjective testing. Second, we are extending our algorithm to deal with imperfect responses from the human subjects.

ACKNOWLEDGMENTS

The authors would like to thank Alan Clark from Telchemy for useful discussions on subjective voice quality. The authors would also like to thank Gary Erickson and the organizers of MaxEnt for providing the scholarship for participating in the workshop.

REFERENCES

1. A.Abbas, "Entropy Methods in Decision Analysis", *Ph.D. Thesis, Stanford University*, 2003.
2. *ITU-T Recommendation P.800*, "Methods for subjective determination of transmissions quality", aug.1196.
3. A. Markopoulou, "Assessing the Quality of Multimedia communications over the Internet", *Ph.D. Thesis, Stanford University*, 2003.
4. *ITU-T Recommendation G.107*, "The Emodel – a computational model for use in transmission planning", Dec. 1998.
5. *ITU-T Recommendation G.113*, "Transmission impairments due to speech processing", Feb. 2001
6. A. Watson & M. A. Sasse, "Measuring perceived quality of speech video and multimedia conferencing applications", *in Proc. of ACM Multimedia'98*, Bristol, UK, Sept. 1998, pp 55-60.