# Listening-Time Relationships in a Subjective Speech Quality Experiment

*Stephen D. Voran*

*Institute for Telecommunication Sciences*

*1-303-497-3839*

*svoran@its.bldrdoc.gov*

**Abstract**

We have designed, conducted, and analyzed a subjective speech quality experiment with unrestricted timing where subjects can vote whenever their opinions are fully formed, rather than at fixed time intervals. Analysis of the resulting listening times reveals that subjects tend to listen for a longer time before approving a recording and for a shorter time before rejecting a recording. This listening-time difference tends to increase for poorer quality systems and for more critical subjects. We present a mathematical model that reproduces these results. In addition, subjects operate more quickly as they move through the experiment.

**Keywords**

Listening time, speech quality, subjective test

## 1 Introduction

ITU-T Recommendations P.800 and P.830 [1,2] provide formal specifications for the conduct of subjective listening and conversation experiments. By far the most commonly implemented listening experiments require experiment subjects to give an opinion after a stimulus has been completely played. These experiments have restricted or forced timing in the sense that the playing time (and we assume the listening time) is constant for all subjects. This adds a degree of control to the experiment which is highly desirable. Restricted timing can also simplify experiment implementation. But control and simplicity are often gained at the expense of realism.

In a non-experimental environment, users form opinions and, more importantly, may even act upon those opinions, without artificial time restrictions. Important actions could include terminating a call or other application, and changing operating modes, channels, services, or providers. One could argue that it is most relevant to know what opinion a user has formed *whenever that opinion is fully formed*, rather than at arbitrary time intervals.

Objective quality assessment tools commonly seek to emulate the human responses gathered in formal subjective experiments. Thus the discussion of restricted and unrestricted timing in quality assessment ultimately relates to both subjective and objective quality assessment.

We have designed, conducted, and analyzed a subjective speech quality experiment with unrestricted timing. In this case, unrestricted timing means that subjects can vote on a relatively long (32 sec) recording at any time after it has started, and they can restart a recording at any time if they wish to hear it again. Once a subject votes, the experiment moves on to the next recording. The expectation behind this approach is that subjects can vote once their opinions are fully formed to their individual satisfaction, rather than at fixed, forced time intervals. This could also be called a self-paced subjective experiment.

Analysis of playing times and the votes given reveals several listening time relationships. These relationships include "experiment acceleration," "cautious approval," and others. This paper includes a simplified mathematical model for a subject's voting process that reproduces some of the observed relationships.

## 2 Experiment Overview

Subjects in the experiment were 35 professionals from the fields of law enforcement, fire, and emergency medical services. Only one female subject was available for this experiment. Subjects were presented with recordings and were asked "Is the speech quality suitable for mission-critical communications?" The binary responses allowed were "yes" and "no." Subjects could vote at any time after a recording had started. A recording could be restarted at any

time, including after it had finished playing. Recordings ranged in length from 30.0 to 39.2 seconds, with an average length of 31.7 seconds.

For each subject and each recording played, the software controlling the experiment collected both the subject's vote ("yes" or "no") and the total amount of time that the subject had allowed that recording to play. The playing times associated with multiple starts, if any, were added together to give the total playing time. We necessarily assume that whenever a recording is playing, the subject is listening, and thus we refer to this measured playing time as "listening time" in the remainder of this paper.

Each session of the experiment included 88 recordings, one each from 88 different systems-under-test. The same 88 systems were included in each session, but were played in a different random order for every session and for every subject. Each subject participated in four sessions, and thus provided 4 votes for each system, resulting in a total of 88×4=352 votes. The total number of votes collected for each session was 88×35=3080 votes. The grand total number of votes collected in the experiment was 352×35=3080×4=12,320 votes. The key variables in this experiment were system, session, and subject.

Each subject was seated at a table (1.5 m × 0.8 m) located in a sound-isolated room. Each subject heard the speech recordings through a speaker on the same table and located about 0.6 m in front of the subject. Subjects were encouraged to adjust the level of the speech recording to the preferred listening level at any time. Background noise was played through an additional pair of speakers, located at the edge of the room, about 1.8 m in front of the subject. Subjects could not adjust the level of the background noise. Rather it was fixed at a nominal level of 60 dBA SPL for Sessions 1 and 2, and 45 dBA SPL for Sessions 3 and 4. Subjects were instructed to vote based on the speech quality and to ignore the background noise to the extent possible. All subjects heard the four sessions in the natural sequence (1, 2, 3, 4).

Other than these particular features, required by the specific project at hand, the experiment conformed to the conventions set out in [1].

# 3 Experiment Results

## 3.1 System Scores

The experiment generated a total of 12,320 votes, and 60.5% of these were "yes" votes. We treat these votes as Bernoulli trials and thus consider that for each system there is an underlying value $p$ that represents the probability of a "yes" vote for that system. The maximum likelihood estimate for $p$ is simply the fraction of "yes" votes and we can also view this value as a subjective score for the system. Larger values of $p$ indicate that a greater fraction of subjects find the system suitable and this can be equated with a higher quality system.

Confidence intervals for this estimate are more complex and described in [3]. Figure 1 shows the resulting overall estimates of $p$ and the associated 95% confidence intervals for the 88 systems in this experiment, after sorting.

## 3.2 Subject Listening Times and Consistency

Because the experiment timing was unrestricted, each subject could move through the experiment at his or her preferred speed. The fastest subject had a total (all four sessions) listening time of 18.4 minutes, and an average listening time of 3.1 seconds/recording. The slowest subject had a total listening time of 107 minutes, and an average listening time of 18.2 seconds/recording.

Subjects were offered a break time between each of the four sessions; some accepted and some did not. The total time taken between sessions by subjects ranged from 4 minutes (effectively no breaks at all) to 24 minutes. The experiment did not reveal any relationship between listening times and break times.

Given these wide ranges, it is natural to ask if any of the subjects did a better or worse job than others. Each subject's opinions are correct by definition, but a per-subject figure of merit can be based on a subject's internal consistency. Since each subject rated each system four times, these ratings can be used to form a simple measure of internal consistency for each subject. The subjects with the highest internal consistency had mean listening times across a very wide range (5 to 18 seconds/recording). The three subjects who had mean listening times below this range (3-5 seconds/recording) had average internal consistency. The subjects with the worst internal consistency had mean listening times ranging from 8 to 14 seconds/recording. Note that these listening times are restricted to the middle of the observed range.

Analyses of internal consistency versus break times, and internal consistency versus total (listening plus break) times yield results very similar to those noted above. Specifically, the most consistent subjects can be very fast, very slow, or moderate. The least consistent subjects tend to be moderate.

## 3.3 Listening Times by System

Figure 2 shows the relationship between mean (across all sessions and subjects) listening time and the quality of the system-under-test. This figure has one asterisk for each system. Horizontal position indicates the estimated value of $p$ for the system and vertical position gives the mean listening time for that system and the 95% confidence interval for that mean. Throughout this paper, 95% confidence intervals on mean listening times are calculated as 1.96 times the standard error, as is customary. This is an

approximation — since all listening times are positive, their distribution can only approximate the normal distribution.

This figure indicates that on average, subjects listen longest (11 sec) to the medium-high quality systems ($0.6 < p < 0.8$). Listening time drops dramatically for the systems below this range, and it drops slightly for the systems above this range.

### 3.4 Experiment Acceleration

Figure 3 shows histograms for the 3080 listening times associated with the 3080 votes in each of Session 1 and Session 4. The tails of these histograms extend out to 64 seconds, but they contain no interesting features.

The listening times for these two sessions are different. In Session 1, there are numbers of cases where recordings are played in their entirety, leading to the secondary peak located just beyond 30 seconds. By the time subjects are performing Session 4, very few recordings are played in their entirety. In general, Session 1 shows a relative frequency that is greater than Session 4 for listening times beyond 15 seconds. On the other hand Session 4 shows a relative frequency that is greater than Session 1 for listening times in the 0 to 12 second range.

Figure 4 summarizes this effect and includes results from Sessions 2 and 3 as well. This figure shows the mean listening time and a 95% confidence interval on that mean for each of the four sessions. On average, Session 2 progresses faster than Session 1, Session 3 progresses faster than Session 2, and Session 4 progresses faster than Session 3. Each of these differences is significant at the 95% level. Since the experiment speed is increasing, we call this "experiment acceleration."

Several possible sources of experiment acceleration come to mind immediately. The first is fatigue or impatience: perhaps as an experiment continues on, subjects who accelerate choose to operate more quickly in order to hasten the end of the experiment. A second is learning: perhaps as the experiment continues on, subjects who accelerate have become more efficient at performing the required task. Another potential source is offered in Subsection 3.9. In general, perhaps a mix of these and other sources produces experiment acceleration.

Note that the background noise level is constant for Sessions 1 and 2, it drops between Sessions 2 and 3, and is constant for Sessions 3 and 4. Thus the background noise level has the potential to confound the results shown in Figure 4. Note however that Figure 4 shows acceleration between Sessions 1 and 2 (no change in background noise level) and acceleration between Sessions 3 and 4 (no change in background noise level). In light of these two accelerations, the acceleration between Sessions 2 and 3 (background noise does change) does not seem exceptionally large or small. It appears that timing changes due to the background noise level change are small compared to the natural acceleration associated with progressing through the four sessions of the experiment.

Analyses did not reveal any trends linking acceleration to particular systems or classes of systems in the experiment.

### 3.5 Subject Acceleration

A per-subject analysis reveals that only about half of the subjects contribute to experiment acceleration. This result is based on the comparison of the per-subject mean listening times for Sessions 1 and 4 in light of the corresponding 95% confidence intervals. Seventeen of the 35 subjects (49%) have Session 4 mean listening times that are statistically significantly shorter than their Session 1 mean listening times. The differences range from 2 to 22 seconds/recording. These 17 subjects contribute to experiment acceleration.

Thirteen of the 35 subjects have Session 4 mean listening times that are statistically equivalent to their Session 1 mean listening times. These 13 subjects do not contribute to experiment acceleration. Finally, 5 of the 35 subjects have Session 4 mean listening times that are statistically significantly longer than their Session 1 mean listening times. These five subjects actually slow down as the experiment proceeds. However, the increases in listening times are comparatively small and range from 2 to 5 seconds/recording.

The experiment also revealed a relationship between subject speed and subject acceleration. Subjects who performed the experiment more slowly overall tended to display more acceleration (e.g., a subject with a grand mean listening time of 18 sec/recording, a Session 1 mean listening time of 30 sec/recording, and a Session 4 mean listening time of 14 sec/recording). On the other hand, subjects who performed the experiment more quickly overall tended to show little or no acceleration, or even some deceleration (e.g., a subject with a grand mean listening time of 3 sec/recording, matched by Session 1 and Session 4 mean listening times of 3 sec/recording each).

### 3.6 Cautious Approval

Figure 5 expands on Figure 4. It shows the mean listening times (and 95% confidence intervals) before "no" votes and before "yes" votes, for all four sessions. In all four cases the "yes" listening time is significantly greater than the "no" listening time, and that difference is nearly constant at about 3 seconds. This 3 second difference is about 31% of the grand average listening time of 9.8 seconds. On average, subjects wish to listen longer to a recording before approving it (voting "yes") than before rejecting it (voting "no"). We call this relationship "cautious approval/rapid rejection" or just "cautious approval" for short. This might be viewed as a natural corollary to the fact that subjects respond more quickly to speech quality decreases and more slowly to speech-quality increases [5].

Combining data from all four sessions yields the two listening time histograms shown in Figure 6. The mean listening time before "no" votes is 7.9 seconds, while the mean listening time before "yes" votes is 11.1 seconds. The 95% confidence intervals on these mean values are [7.7,8.1] and [10.9,11.3] respectively. These intervals are disjoint and the mean listening times before "no" and "yes" votes are different at the 95% level.

The most common listening times before "no" votes are in the 2-3 second interval but the most common listening times before "yes" votes are in the 7-8 second interval. From 0 to 7 seconds "no" votes are more likely, and from 7 to 21 seconds "yes" votes are more likely. Some "yes" votes are given only after the entire recording is heard, and this creates the secondary peak located just beyond 30 seconds.

Integration of these histograms reveals that the first 5 seconds include about 50% of the "no" votes but only 20% of the "yes" votes. In the first 10 seconds, about 75% of the "no" votes but only 50% of the "yes" votes have been given.

It is important to note that the subjective experiment included a wide range of systems and speech qualities. Some systems had temporal variation in speech quality (e.g., high quality speech coding with infrequent but serious channel impairments) while others had more constant speech quality (e.g., speech coders operating over clear channels.) With each recording, subjects had no way of knowing if the first few seconds would be representative, or if the speech quality might change later in the recording. Given this variety and unpredictability, subjects had to rely completely on their individual processes for forming opinions over time. The cautious approval described in this subsection reflects average behavior across all 35 of the individual subjects' opinion-forming processes and across all 88 of the systems-under-test.

### 3.7 Subject Cautious Approval

Figure 7 shows mean (across all systems and sessions) "yes" and "no" listening times and the associated 95% confidence intervals for each of the 35 subjects that participated in the experiment. These results are sorted from the fastest subject (left) to the slowest subject (right). This figure shows that cautious approval is a property that can be found in faster and slower subjects. Specifically 26 of the 35 subjects (74%) have average "yes" times that are statistically greater than their average "no" times, at the 95% confidence level. Only one subject has an average "no" time that is statistically greater than his average "yes" time. For the remaining 8 subjects, "yes" and "no" times are statistically equivalent.

The majority of the subjects in this experiment display cautious approval and this relationship seems consistent with some intuitive notions of speech quality judgement: Typically a "no" vote requires only the detection of one or more impairments that singly or cumulatively reach a subject's rejection threshold. On the other hand a "yes" vote typically requires verification that no such impairments exist, or that they are sufficiently small. Indeed it seems that the time required for the first task (detecting impairments) would generally be less than the time required for the second task (verifying absence of impairments).

A highly simplified mathematical model for an individual subject's voting rules can be used to further illustrate this argument. Suppose a given subject is willing to listen to a recording for $T$ seconds. Further suppose that the subject has an impairment accumulation function $A(t)$ that accounts for the perceived net effect of the impairments from the start of the recording up to time $t$. Finally, assume that the subject's vote is based on comparison of this net effect and a rejection threshold $R$:

$$\begin{aligned} A(T) \geq R &\rightarrow \text{Reject} \\ A(T) < R &\rightarrow \text{Approve.} \end{aligned} \tag{1}$$

If $A(t)$ is a non-decreasing function of $t$, then this subject can take a shortcut in some cases and remain consistent with the rules in (1):

$$t \leq T, \quad A(t) \geq R \rightarrow \text{Reject.} \tag{2}$$

That is, as soon as the cumulative effect of the impairments reaches the rejection threshold, the subject can reject the recording; there is no need to listen for the full $T$ seconds. Examples of potential accumulation functions that are non-decreasing are the maximum function,

$$A(t_0) = \max_{0 \leq t \leq t_0} \left( I\left(s(t)\right) \right), \tag{3}$$

and the integral,

$$A(t_0) = \int_0^{t_0} I\left(s(t)\right) dt. \tag{4}$$

Here $I(\bullet)$ is a non-negative impairment function that quantifies the subject's perception of impairment, and $s(t)$ is the recorded speech signal.

Even if $A(t)$ is not non-decreasing, this subject may still be able to take a shortcut and remain consistent with the rules in (1): As $t$ approaches $T$, if $A(t)$ is large enough, the recording may be "unredeemable." That is, the subject may realize that even perfect speech quality for the remaining $T$-$t$ seconds will not reduce $A(t)$ below the rejection threshold $R$. Thus the subject can reject the recording at time $t < T$. An example of an accumulation function that is not non-decreasing is the mean

$$A(t_0) = \frac{1}{t_0} \int_0^{t_0} I\left(s(t)\right) dt. \tag{5}$$

The experiment described in [4] indicates that in some cases at least, an experimentally determined accumulation function is bounded by the mean function and the maximum function:

$$\frac{1}{t_0}\int_0^{t_0} I\left(s(t)\right)dt < A(t_0) < \max_{0\le t\le t_0}\left(I\left(s(t)\right)\right). \qquad (6)$$

The simplified mathematical model illustrates how subjects can take shortcuts when rejecting a recording so it is most literally described by the "rapid rejection" side of the "cautious approval/rapid rejection" relationship.

Subjects clearly have opportunities to take shortcuts before rejecting a recording, but it is much harder to imagine scenarios where subjects can approve a recording at $t < T$ seconds without risking violation of the rules in (1). No matter how small $A(t)$ is and how close $t$ gets to $T$, there remains a chance that the final $T$-$t$ seconds will include an impairment so severe that $R \le A(T)$. When $t < T$ seconds remain in a recording, it may seem to subjects that the best case scenario for that remaining time is bounded, but the worst case scenario is unbounded.

Figure 8 provides a graphical indication of a relationship between how critical a subject is, and how much cautious approval a subject displays. This figure has one asterisk for each subject. Horizontal position indicates the mean time difference between "yes" listening times and "no" listening times; subjects with greater cautious approval are towards the right. The vertical position shows the fraction of the subject's 352 votes that were "yes" votes; more critical subjects are towards the bottom. This figure indicates that subjects who are less critical also tend to exhibit less cautious approval.

The simplified mathematical model introduced above can reproduce this relationship. If a given subject has an unusually high value of $R$, but a typical value of $T$, then that subject will be less critical than normal since larger values of $A(T)$ are needed in order to exceed $R$. It will also take longer for $A(t)$ to exceed $R$, so less time can be saved by giving "no" votes early ($t < T$). This reduces the spread between "no" and "yes" listening times, and thus reduces the magnitude of cautious approval displayed by that subject.

### 3.8 Cautious Approval by System

Cautious approval provides an explanation for the mean listening time decrease for lower quality systems seen in Figure 2. Lower quality systems are those that receive larger proportions of "no" votes. Since "no" votes are given after a shorter listening time than "yes" votes, lower quality systems have lower mean listening times.

Figure 9 is a smoothed and expanded version of Figure 2. It provides a graphical indication of a relationship between systems and cautious approval. The figure shows mean results for groups of systems. The horizontal location indicates the mean value of estimated $p$ for the systems that fall into eight different intervals: $0.1 \cdot k < p \le 0.1 \cdot (k+1)$, $k$=1,2,…,8. The vertical location describes the mean (over that group of systems) listening time per recording before "no" votes and before "yes" votes. The extreme ends of the quality range cannot be used in this analysis because there are insufficient votes of one type or the other at each end of the range.

This figure reemphasizes that cautious approval is associated with the poorer quality systems ($p < 0.5$), where "no" votes dominate. The poorer the system, the more "no" votes it gets and the more quickly these votes are given. In this experiment, the listening time before "no" votes may be approaching an asymptote in the limit of poor quality of 6 or 7 seconds. Also, the poorer the system, the fewer the "yes" votes that it gets, and the more slowly these votes are given. In this experiment, the listening time before "yes" votes may be approaching an asymptote in the limit of poor quality of 14 or 15 seconds.

Better quality systems ($0.5 < p$) have "no" and "yes" listening times that are statistically equivalent, or nearly so. If one chooses to acknowledge the barely significant differences of the right-most data points, then one might speak of cautious approval for poorer systems and "cautious rejection" for better systems. Note however that the cautious approval observed for poorer systems is many times stronger than any cautious rejection that might be present with better systems. If this weak cautious rejection relationship is acknowledged, then the combination of cautious approval and cautious rejection might be described as the "cautious minority vote." That is, longer listening precedes a vote that reflects the minority opinion.

### 3.9 Quality Decline with Time

Figure 10 shows the fraction of "yes" votes and a 95% confidence interval for each session. Each of these results is based on the 3080 total votes given in a session. Recall that each session contained the same 88 systems and used the same 35 subjects, so one could expect that these fractions would be similar for all four sessions.

Figure 10 indicates a weak general trend towards fewer "yes" votes as the experiment progresses (e.g., 64% in Session 1 but only 58% in Session 4) though only some of the differences are statistically significant. That is, the experiment is showing a slight decline in reported quality with time.

Recall that the background noise level does drop between Sessions 2 and 3 and it is possible that this lower level of background noise could unmask some additional impairments in the recordings. However, if this were the main source of the quality decline, one would expect a single quality drop between Sessions 2 and 3 rather than the

continuing quality decline through all four sessions seen in Figure 10.

The trend shown in Figure 10 is weak compared to experiment acceleration but the directions of these two trends are consistent. That is, if some subjects lower their rejection threshold *R* as the experiment progresses, the experiment will accelerate (due to cautious approval) and the reported system quality will decline.

## 4 Summary and Discussion

We have presented listening-time relationships found in a subjective experiment with unrestricted timing. These relationships include experiment acceleration (attributed to about half of the subjects) and cautious approval (attributed to about ¾ of the subjects). We have found that subjects that show more cautious approval also tend to be more critical. We have presented a simplified mathematical model that reproduces cautious approval and its relationship to criticality, and is also consistent with prior work in this area. We have also found that poorer quality systems are heard for shorter times and this is consistent with the cautious approval result. In addition, we have seen that subjects move through the experiment at a wide range of speeds and that speed is not related to internal consistency.

These results lead to a family of interesting questions. An example of a specific question concerns cautious approval: If timing in the present experiment had been restricted so that all opinions were gathered at 9 or 10 seconds (after the mean "no" listening time but before the mean "yes" listening time) would that have shifted opinions in a negative direction?

More generally, how do the results of restricted timing experiments and unrestricted timing experiments compare with each other? Does one of these provide information that is closer to the "truth" or more relevant than the other? In non-experimental environments users generally can form opinions and act on them without any timing restrictions. Does this mean that unrestricted timing experiments provide a better simulation of "real life" than restricted timing tests? Or does this extra degree of freedom just confound the "correct by definition" results that one would get from a restricted timing experiment?

## 5 References


[1] ITU-T Recommendation P.800, "Methods for subjective determination of transmission quality," Geneva, 1996.

[2] ITU-T Recommendation P.830, "Subjective performance assessment of telephone-band and wideband digital codecs," Geneva, 1996.

[3] N. Johnson, S. Kotz, and A. Kemp, *Univariate discrete distributions, second edition*. New York: Wiley, 1992, ch. 3, pp. 129-133.

[4] S. Voran, "A Basic Experiment on Time-Varying Speech Quality," *Proc. 4th International MESAQIN (Measurement of Speech and Audio Quality in Networks) Conference*, Prague, Czech Republic, June 2005.

[5] L. Gros and N. Chateau, "Instantaneous and Overall Judgements for Time-Varying Speech Quality: Assessments and Relationships," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 367-377, May/June 2001.

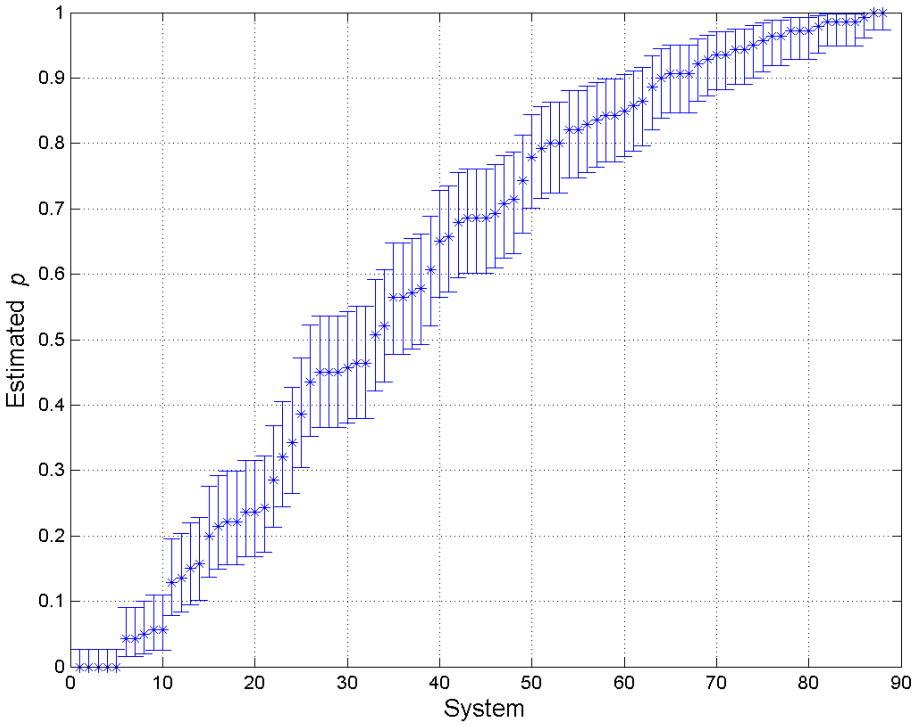


*Fig. 1 Estimated p (probability of "yes" vote) values and 95% confidence intervals for 88 systems in the experiment.*

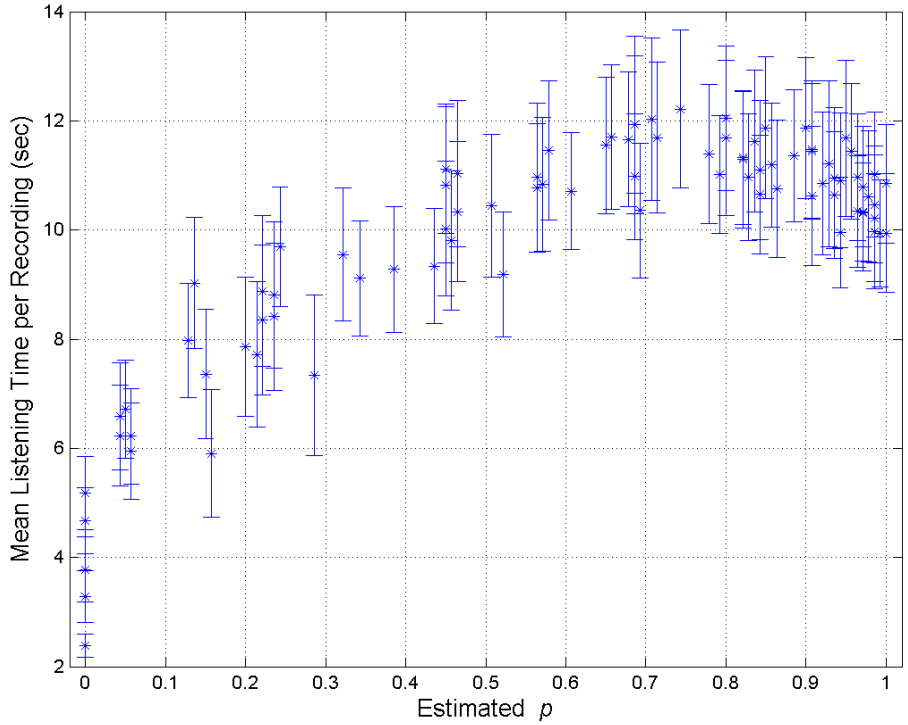


*Fig. 2 Mean listening times and 95% confidence intervals for the 88 systems in the experiment.*

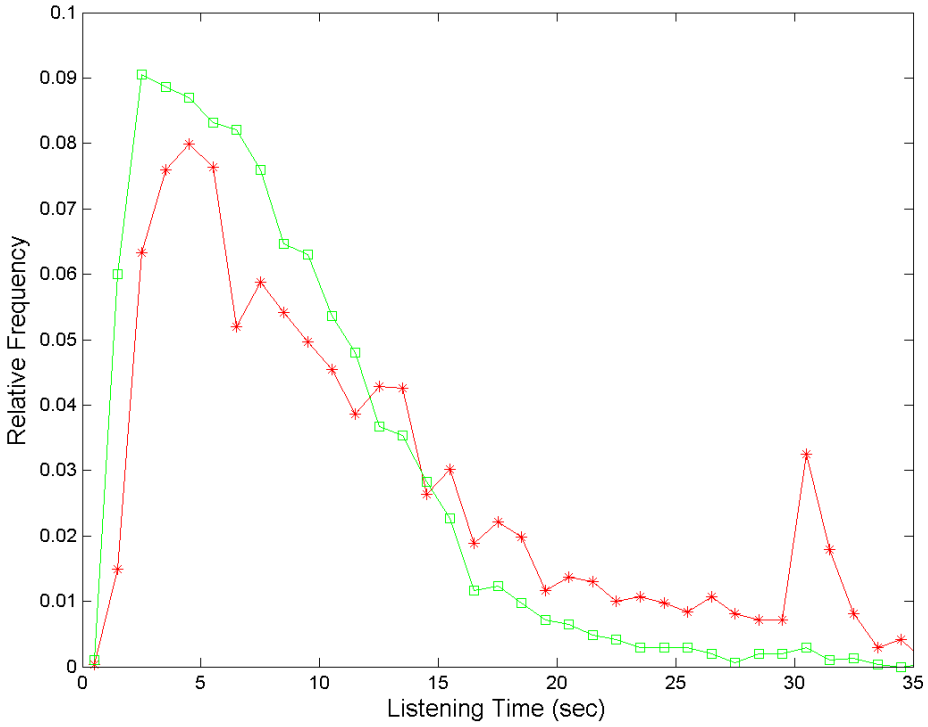


*Fig. 3 Histograms of listening times for Session 1 (asterisks) and Session 4 (squares).*

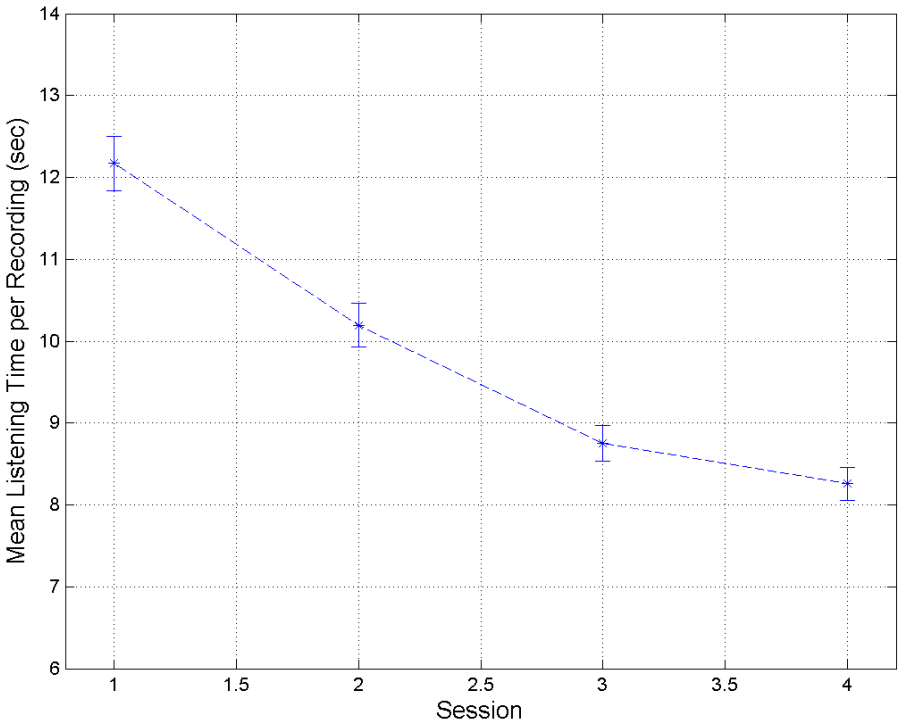


*Fig. 4 Mean listening times per recording and 95% confidence intervals, for all four sessions.*

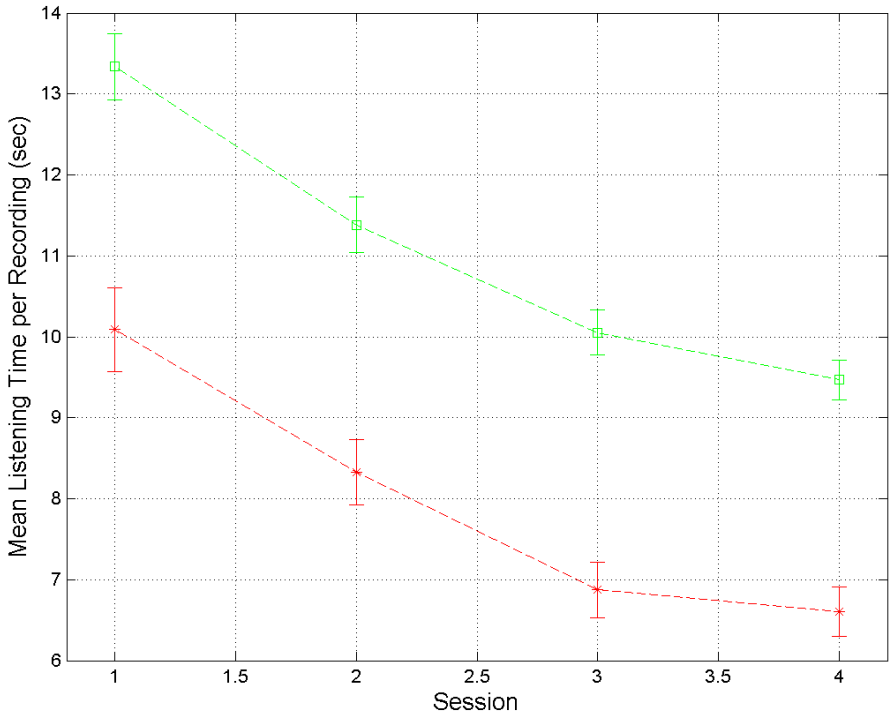


*Fig. 5 Mean listening times per recording and 95% confidence intervals for all four sessions. Listening times before "no" votes are shown with asterisks, listening times before "yes" votes are shown with squares.*

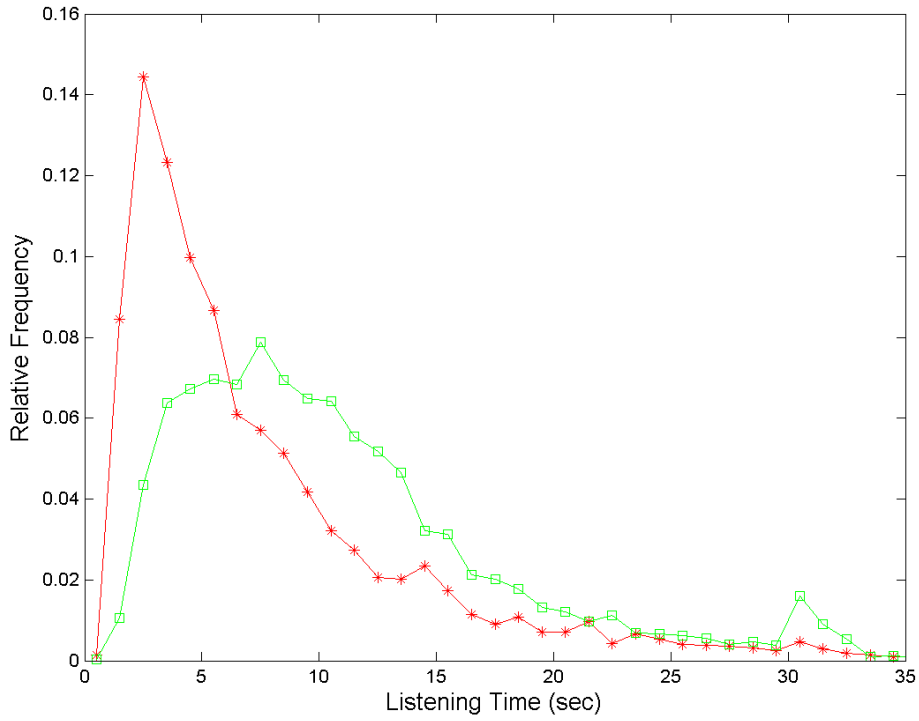


*Fig. 6 Histograms of listening times before "no" votes (asterisks) and before "yes" votes (squares).*

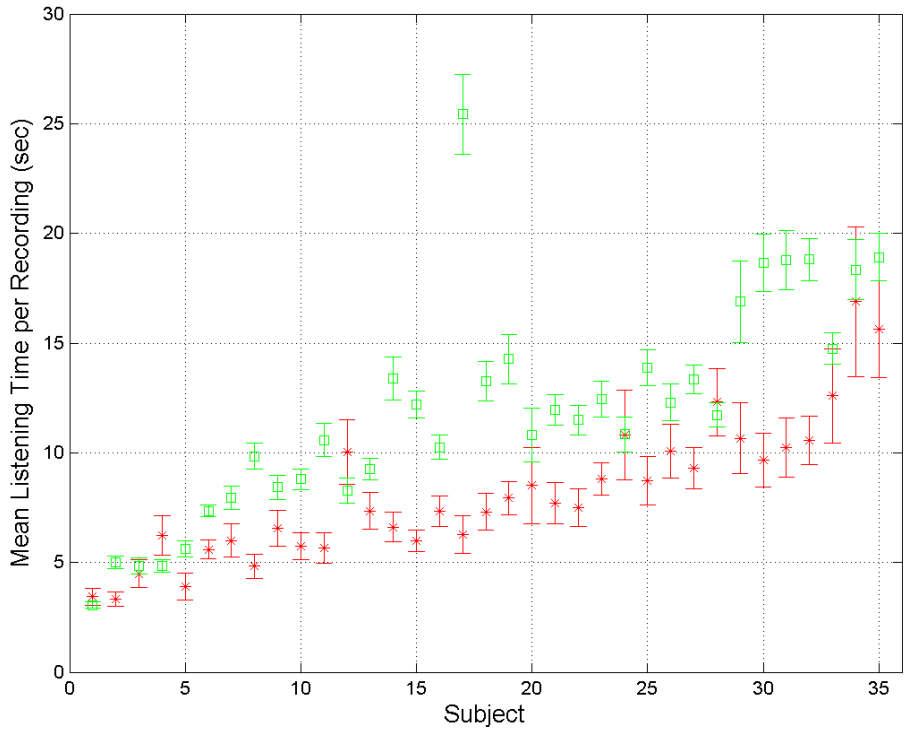


*Fig. 7 Per subject mean listening times before "no" votes (asterisks) and before "yes" votes (squares) along with 95% confidence intervals. Subjects are sorted from fastest (left) to slowest (right).*

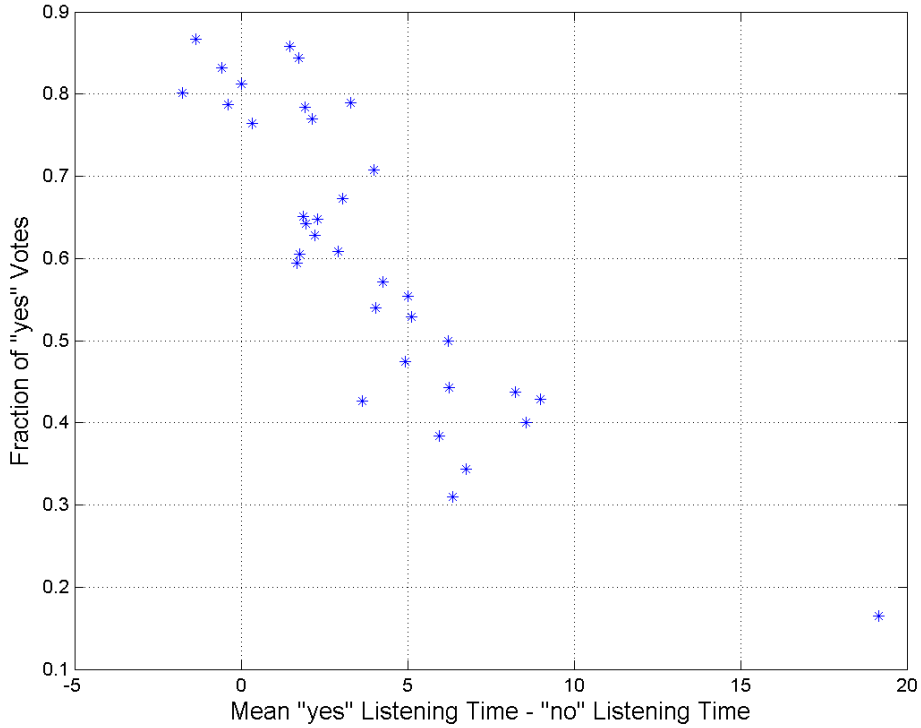


*Fig. 8 Relationship between cautious approval (x axis) and criticality (y axis) for each subject.*

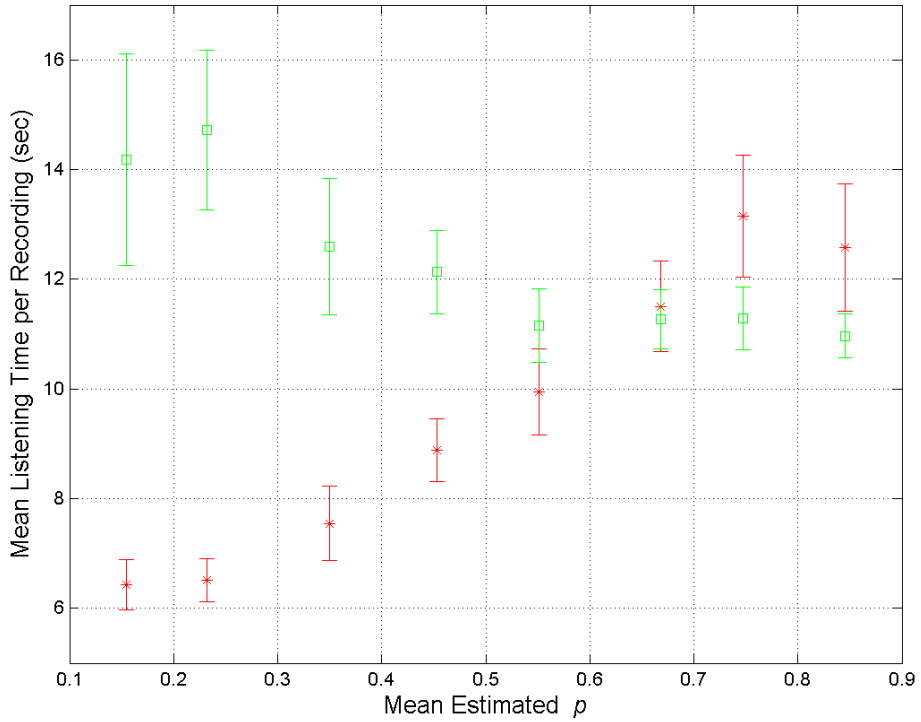


*Fig. 9 Mean listening times before "no" votes (asterisks) and before "yes" votes (squares) and 95% confidence intervals versus mean system quality.*

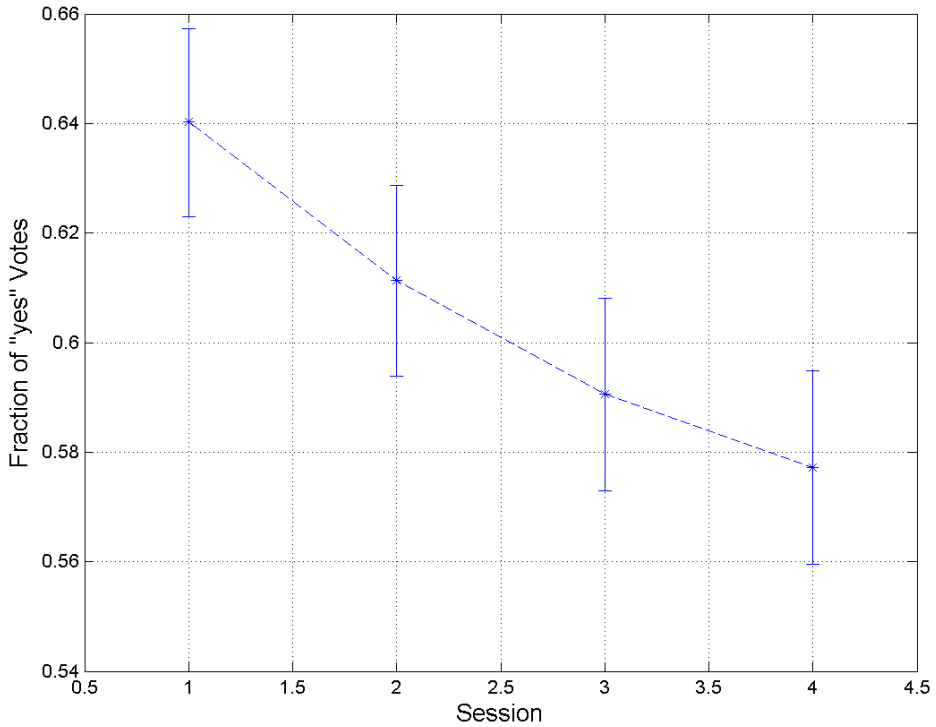


*Fig. 10 Fraction of "yes" votes in each session along with 95% confidence intervals.*