Statistical Testing Of Off-line Comparative Subjective Evaluations For Optimizing Perceptual Conversational Quality In VoIP*

Batu Sat and Benjamin W. Wah
Department of Electrical and Computer Engineering
and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign
Urbana, IL 61801, USA
{batusat,wah}@uiuc.edu

Abstract

In this paper, we study the scheduling of off-line subjective tests for evaluating alternative parameter values of control schemes in real-time multimedia applications. These applications are characterized by multiple counteracting objective quality metrics (such as delay and signal quality) that can be affected by the control schemes. However, the trade-offs among these metrics with respect to the subjective preferences of users are not defined. As a result, it is difficult to use the proper control value that leads to the best subjective quality without carrying out subjective tests. Since subjective tests are expensive to conduct and the number of possible control values and run-time conditions is prohibitively large, it is important that a minimum number of such tests be conducted, and that the results learned can be generalized to unseen conditions with statistical confidence. To this end, we study in this paper optimal algorithms for scheduling a sequence of subjective tests, while leaving the generalization of limited off-line subjective tests to guide the operations of the control schemes at run time to a future paper. Using an example application in the design of the playout scheduling (POS) algorithm for a two-party VoIP system, we study the accuracy and efficiency of conducting subjective tests simultaneously.

1 Introduction

In this paper, we study methods for conducting off-line subjective tests. These tests are used to guide the operation of control schemes for real-time multimedia communication systems in order to achieve high perceptual quality. The systems involved have the following properties.

a) *Multiple objective quality metrics.* A common approach is to use some objective metrics recommended by a standardization body, such as the International Telecommunication Union (ITU) or the Internet Engineering Task Force (IETF), as well as metrics that can be computed easily. Examples include the delay incurred and the quality of the received media. In many multimedia applications, there does not exist a single objective metric that captures all aspects of quality. When using multiple metrics, the quality of a system can be denoted by a point in a multi-dimensional space, whose axes correspond to the individual metrics.

b) *Constrained resources.* The control schemes operate under limited network resources (such as constraints on bandwidth and packet rate) and computational resources.

c) *Best-effort IP network.* The IP network used exhibits dynamic non-stationary delay and loss behavior.

d) *Communication scenario among participants.* This affects the subjective quality perceived. For example, delay degradations may be more important to the participants when they have frequent interactions.

e) *System control.* To mitigate network imperfections, the control schemes employed have adjustable parameters, such as the transmission rate and the playout schedule. For a control scheme under given constraints and conditions, the set of *operating points* in the multi-dimensional quality-metric space correspond to its feasible control alternatives. This set of points form an *operating curve*.

f) *Trade-offs among objective metrics on subjective preferences.* Due to system constraints and network imperfections, trade-offs must be made among the multiple counteracting quality metrics. Since their effect on subjective user preferences is not defined, it is difficult to select the proper control parameter values in order to arrive at an operating point with the highest subjective quality.

g) *Multiple local optima.* There can be multiple optimal operating points on an operating curve around a particular operating point with the optimal subjective quality.

*Proceedings of IEEE International Symposium on Multimedia, 2008.

** Research supported by National Science Foundation Grant CNS 08-41336.

Subjective evaluations can be conducted to evaluate the quality of a control scheme. Because such evaluations cannot be performed at run time, off-line tests have to be conducted during which the information learned is used to guide the control at run time. In general, subjective evaluations are time consuming and expensive because they involve multiple subjects in order to arrive at some statistically significant results. Further, since there may be prohibitively many network conditions and communication scenarios that can be observed at run time, it is infeasible to conduct exhaustive subjective tests to cover all situations.

A standard method for conducting subjective evaluations is to ask subjects to rank the communication quality by an *absolute category rating* (ACR) and to take an algebraic mean of the opinions of the subjects in response to the same stimuli. The result obtained is denoted by a *mean opinion score* (MOS). In the following, we explain why this approach is only useful for verifying a system's performance, but not suitable for designing new control schemes.

a) *Absolute scores* obtained for two points on an operating curve can be used to deduce their relative positions. If all alternatives are mutually related under pairwise comparisons, then a total ordering can be established. In practice, two operating points may not be comparable when they involve multiple quality metrics. This happens because the perceived effects on the difference of one metric may not be consistently translated into the differences of the other metrics. Consequently, the feasible operating points of an operating curve lie on a Pareto-optimal boundary.

b) *Statistical significance.* Although MOS scores can be determined statistically, no statistical significance can be associated with the difference of two MOS scores. If the variances in the scores are large relative to their difference, then the conclusion reached on the difference is not statistically meaningful. As is stated in ITU P.800 [3] for evaluating telephone communication quality, absolute ratings are not accurate for evaluating quality when samples have high quality or their difference is barely perceptible. Hence, the number of samples required to obtain MOS with a certain level of statistical significance can be inadequate in some pairwise comparisons, but can be too many in other cases.

Our Approach. Next, we describe our observations and our approach to address the issues stated above.

a) *Comparative ranking.* To determine the preferred operating point among a set of alternatives, a partial order that requires pairwise comparisons suffices. The partial order can be obtained by a measure that evaluates the relative quality of two alternatives in a *comparative category rating* (CCR) (similar to the LOSQ−*listening only speech quality*−evaluation in ITU P.800). By presenting two alternatives to each subject, one after another, it allows the incomparability of some alternatives to be identified and small differences between two to be more accurately evaluated. The disadvantage, however, is a significant increase in the number of tests because such tests will need to be conducted for each pair of alternatives instead of each alternative.

b) *Stochastic evaluation results under given conditions.* To identify the best operating point at run time, we first consider the problem of determining the best operating point off-line under a given set of network conditions and communication scenarios. In this task, we conduct a limited number of subjective evaluations. To eliminate variations other than the differences in the control schemes tested, we use simulators to repeat the network conditions and communication scenarios. We then collect the comparative subjective opinions and represent them as discrete distributions.

c) *Pruning of search space.* The idea is to systematically use the observations from past subjective tests to prune tests that have not been conducted. Our approach is based on a statistical model of subjective evaluations, which utilizes two principles that small differences cannot be perceived by subjects and that subjective preferences are uni-modal with respect to changes in the control parameter.

d) *Learning of a classifier.* Based on the subjective preferences under a comprehensive set of test conditions, we learn an SVM (support-vector-machine) classifier that can generalize to unseen conditions at run time. Due to space limitation, we leave its presentation to a future paper.

Our proposed approach can be used in many real-time multimedia communication applications. As an example, we describe in Section 2 the design of a playout scheduling (POS) algorithm in a two-party VoIP system. Another application is in the dynamic equalization of mutual silences (MS) in multi-party VoIP [4, 1] for optimizing perceptual conversational quality. Here, the control parameter under given network and conversational conditions is the level of MS equalization, whereas the objective metrics are LOSQ, conversational efficiency, and conversational symmetry [4]. Subjective tests to guide the selection of the best control value are needed because there is no single objective metric that captures all aspects of subjective conversational quality. Yet another application that can benefit from this methodology is in real-time video conferencing, where the controls of the encoding rate and the playout scheduler affect the delay-quality trade-offs perceived by users.

Problem statement. In this paper, we study the statistical scheduling of off-line comparative subjective tests for evaluating alternative operating points on an operating curve of a control scheme in a real-time multimedia system. Our goal is to minimize the number of subjective tests needed in order to determine a locally optimal operating point to within some prescribed level of statistical confidence. We assume the knowledge of the region of the operating curve where the local optimum is located, which is denoted by the *Region of Dominance* or ROD. All comparisons are, thus, conducted within the respective ROD of the local optimum.

We first describe the POS design problem for two-party VoIP in Section 2. This is followed by the model of subjective comparisons (Section 3), the Bayesian formulation for representing information learned (Section 4), the approach for conducting subjective tests (Section 5), the optimal policy (Section 6), and the evaluation results (Section 7).

2 Design of POS Control in VoIP

In this section, we illustrate an example application in the design of a POS algorithm for a two-party VoIP system. The trade-offs among the various objective metrics in this application will require subjective tests to be conducted in order to achieve high perceptual quality.

System control. A VoIP conversation consists of one-way transmissions of speech segments in alternating directions that are separated by silence periods. The one-way transmission of speech involves trade-offs between LOSQ and the perceived delay degradations. These trade-offs are controlled by a POS algorithm that adjusts the *mouth-to-ear delay* (MED). Here, MED is the total delay from the mouth of the speaker to the ear of the listener, which includes delays in encoding, packing, transmission, de-jittering, and decoding. A longer MED allows more packets to arrive in time for playout and improves LOSQ, but will degrade the interactivity and the efficiency of the conversation.

In the context of a two-party VoIP system, the POS scheme has MED as its single control variable that affects multiple objective metrics of conversational quality, such as LOSQ, conversational symmetry, and conversational efficiency [5]. Figure 1 depicts the quality of a conversational segment as a point in a multi-dimensional space whose axes correspond to these objective metrics.

Network and conversational conditions. The perceived delay effects depend on the conversational conditions, such as the human response delay, single-talk duration and switching frequency [5]. Under a given conversational condition, the feasible set of alternatives is represented by a plane in Figure 1. For given network and conversational conditions, the feasible alternatives are further restricted to an operating curve on a plane that corresponds to the conversational condition. Under these conditions, the operating point shifts towards the right on the operating curve as the control variable MED is increased.

Optimal operating point. The specific MED that optimizes subjective conversational quality depends on the given network and conversational conditions. For example, for a connection with high delays and jitters, the optimal MED can be higher in order to improve the poor LOSQ. In contrast, for a conversation in which participants take frequent turns, a lower MED may be preferred in order to reduce the annoying delay degradations. Since the network and conversational conditions may change during a conversation, the MED will need to be dynamically adjusted in a closed-loop fashion in order to maintain a high conversational quality perceived by users.

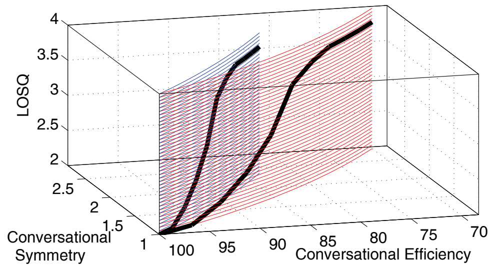

Figure 1. A 3-dimensional representation of an operating curve under two conversational conditions.

Quality metrics. The standard objective metrics for evaluating conversational quality include the E-model [2] and the *call clarity index* [3]. They do not always fit well with subjective evaluations because they employ simplistic assumptions in estimating conversational quality. For instance, the E-model assumes that degradations due to delay and LOSQ are independent and additive.

The most popular subjective metric for evaluating conversational quality is the ITU recommendation P.800 based on an ACR scale, where a pair of subjects converse over a phone system to complete a given task. As is discussed in Section 1, this approach is suitable for verification purposes, but is of limited use in designing VoIP control algorithms.

Design goal of POS control. The goal of the POS control is to find the MED under some given operating condition that leads to the best subjective conversational quality. The problem is challenging because multiple subjective evaluations are needed in each comparison in order to arrive at some statistically significant conclusions. Moreover, MED can take continuous values, which result in infinitely many realizations of the POS algorithm. Further, given infinitely many network and conversational conditions that can exist at run time, it is infeasible to conduct indiscriminate subjective tests in order to evaluate all pairs of conditions.

3 Model of Subjective Comparisons

We present in this section our model on the relative ranking of subjective evaluations.

Notation. Let $\mathcal{O}$ be the set of points on an operating curve, and $A^{\min}$ and $A^{\max}$ be the two extreme points in the set. There are four possible outcomes when comparing two points A and B on the operating curve:

Condition	Probability	Notation
A is better than B	$Pr(A > B)$	$p_1(A,B)$
A is about the same as B	$Pr(A \approx B)$	$p_0(A,B)$
A is worse than B	$Pr(A < B)$	$p_{-1}(A,B)$
A is incomparable to B	$Pr(A?B)$	$p_2(A,B)$

The distribution of the opinions can be modeled by a multi-nomial distribution, assuming that the opinions

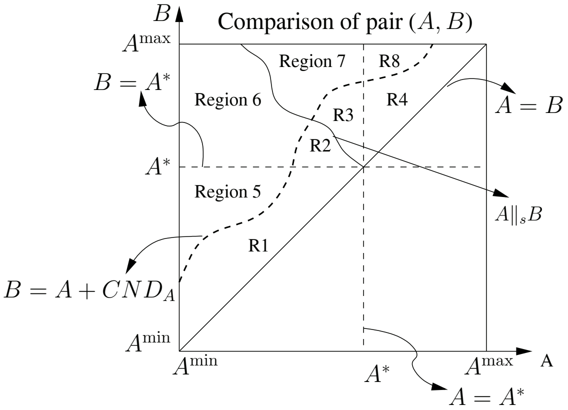

Figure 2. Model of subjective comparison of A and B: 8 regions correspond to different pairwise comparisons on the 2-D plane whose axes represent the indexes of the two points compared.

Table 1. Regions in Figure 2 with respect to the boundaries. The '$-$' symbol indicates that the region and the origin $(A^{\min}, A^{\min})$ are located in the same side of a line.

Boundary	Region in a ROD							
Line	R1	R2	R3	R4	R5	R6	R7	R8
$B - A - CND_A$	$-$	$-$	$-$	$-$	$+$	$+$	$+$	$+$
$A - A^*$	$-$	$-$	$-$	$+$	$-$	$-$	$-$	$+$
$A\|_s B$	$-$	$-$	$+$	$+$	$-$	$-$	$+$	$+$
$B - A^*$	$-$	$+$	$+$	$+$	$-$	$+$	$+$	$+$

from different subjects are independent and identically distributed. The *comparative opinion distribution* (COD) can be represented by a vector: $COD(A, B) = \overline{p} = (p_{-1}, p_0, p_1, p_2)$, where $\sum_i p_i(A, B) = 1$ for all (A, B) pairs.

3.1 Model of Pairwise Subjective Comparisons

The general model describes the probabilities of occurrence of the four possible opinions when comparing A and B. It is defined over a 2-D plane, whose axes represent the numeric values of the two points. Figure 2 depicts the model and the eight regions on the A-B plane defined in Table 1. Due to the model's anti-symmetry property (defined below), it suffices to define half of the plane where B is larger than A. When a finite number of comparisons are carried out, the result is a sampled version of the probabilities defined in the model. Let $B - A$ be the perturbation in the control value from A to B. This notation will be used when evaluating a fixed A in comparison to another variable point B.

The following are the axioms on pairwise comparisons.

Reflectivity. Comparing a point with itself results in the $A \approx_s A$ opinion from an individual perspective and $p_0(A, A) = 1$ from a collective perspective. Since no objective difference exists between the two points, subjects should not perceive the two points to have different quality.

IID. When comparing any two points on an operating curve, the responses of subjects are independent and identically distributed.

Symmetry/anti-symmetry. The order of comparison does not effect the comparative opinion between A and B. Indistinguishable ($\approx_s$) and incomparable ($?_s$) opinions are *symmetric*; thus, $p_i(A, B) = p_i(B, A)$ for $i \in \{0, 2\}$. On the other hand, preference opinions ($>_s$ and $<_s$) are *anti-symmetric*; thus, $p_{-1}(A, B) = p_1(B, A)$.

Smoothness. Since each objective metric is monotonic in control value and small changes in objective metrics do not cause drastic differences in subjective preferences, $p_i(A, B)$ is continuous and piecewise differentiable with respect to B for a fixed A and variable B and for each i.

Just noticeable difference (JND) of A. When there are small changes between A and B, there will be a small percentage of subjects perceiving a difference in the subjective quality of A when compared to B. As the difference between A and B increases, the subjective perception of the difference increases as well. This noticeable difference is commonly used in psycho-physics and is defined as follows. In a comparison between a fixed point A and a variable point B, both of which belong to the same operating curve $\mathcal{O}$, JND_A is the $B - A$ value for which 50% of the subjects perceive a difference in their quality. If B is inside the JND region of A ($|B - A| \leq JND_A$), then A and B are *indistinguishable*; otherwise, they are *distinguishable*.

Complete noticeable difference (CND) of A. In a comparative evaluation between a given point A and a variable point $B \in \mathcal{O}$, CND_A is defined as the minimum $B - A$ value such that $p_0(A, B) = 0$.

Indistinguishability. The probability of an indistinguishable opinion, $p_0(A, B)$, is monotonically non-increasing with respect to $B - A$ for fixed A and variable B.

Justification. When $B - A = 0$, $p_0(A, A) = 1$ due to reflectivity. As $B - A$ increases, there are more objective differences between A and B, resulting in more subjects perceiving the difference in quality. Eventually all subjects perceive that A is not the same as B. However, JND_A may vary as a function of A. For some A, a small perturbation in the control may result in the perception of a difference in subjective quality; whereas it may require a large perturbation for subjects to perceive the difference for another A. Our subjective tests in two-party VoIP conversations confirm the variations in JND as a function of A.

Locally optimal point on an operating curve is denoted as A^* and is defined as follows:

$$A^* = \{A | p_1(A, B) > 0.5,\ \forall B \in \mathcal{O} \text{ s.t. } |B - A| > JND_A\}.$$

The locally optimal operating point A^* is, therefore, preferred among all the alternatives in its ROD, except for points that are within its JND. In general, determining A^* will require all pairwise comparisons of operating points. Since this is prohibitively expensive when the number of alternatives is large, our goal is to find A^* through a sequence of adaptively chosen pairwise comparisons of alternatives.

Incomparability of A and B. For an infinite number of subjects ($K \to \infty$), $\lim_{\delta\to0^+} p_2(A, B+\delta) \ge p_2(A,B)$ and $\lim_{\delta\to0^+} p_2(A-\delta, B) \ge p_2(A,B)$.

Justification. At a local optimum, the quality metrics have an optimal trade-off. However, due to the monotonicity property, as the point is perturbed away from A^* in one direction (say towards $A^{\max}$), a subset of the quality metrics exhibit more perceptible degradations that dominate the other metrics. On the other hand, when the point is perturbed in the other direction (say towards $A^{\min}$), a different subset of quality metrics exhibit more perceptible degradations. Thus, when a subject is asked to compare the two points on different sides of A^*, the subject may indicate that the pair is *incomparable*, since different sets of quality metrics dominate the degradation. As the distance between the points increase, the overlap between the sets of dominant quality metrics reduces, thus increasing the percentage of subjects who indicate that the pair is *incomparable*. For example, in two-party VoIP, perturbations from the local optimum causes degradations due to delay to be dominant in one direction but degradations due to speech quality to be dominant in the other direction.

Due to space limitation, we state the lemmas and corollaries in this section without proof.

Lemma. For $K \to \infty$ and for any finite $\Delta > 0$, $p_2(A, B+\Delta) \ge p_2(A,B)$ and $p_2(A-\Delta, B) \ge p_2(A,B)$.

Corollary. $p_2(A_2, B_2) \ge p_2(A_1, B_1)$ if $[A_1, B_1] \subseteq [A_2, B_2]$.

Corollary. $p_2(A^{\min}, A^{\max}) \ge p_2(A,B)\ \forall A, B \in \mathcal{O}$.

Subjective preference. For $K \to \infty$ and A and B on the same side of A^*,

$$|p_1(A,B) - p_{-1}(A,B)| \le \begin{cases} \lim_{\delta\to0^+} |p_1(A-\delta,B) - p_{-1}(A-\delta,B)| \\ \lim_{\delta\to0^+} |p_1(A,B+\delta) - p_{-1}(A,B+\delta)|. \end{cases}$$

Justification. To explain the concept intuitively, as A is fixed and B is perturbed towards A^*, the point closer to A^* will have more balance in their objective metrics and better perceived quality. The difference between p_1 and p_{-1} is an indication of the conclusiveness of the comparison in directing the most likely location of A^* in relation to A and B. Thus, as B, the point closest to A^*, is perturbed towards A^*, the conclusiveness of the comparison improves.

Control symmetry. For A and B on opposite sides of A^*, A and B are objectively symmetric, denoted by $A\|_0 B$, if they are equi-distant from A^* in terms of their control value; that is, $|A - A^*| = |B - A^*|$ or $A + B - 2A^* = 0$.

Subjective symmetry. For A and B on opposite sides of A^*, A and B are subjectively symmetric, denoted by $A\|_s B$, if $p_1(A,B) = p_{-1}(A,B)$. This means that the votes that indicate one point can be preferred over the other is equal from both directions.

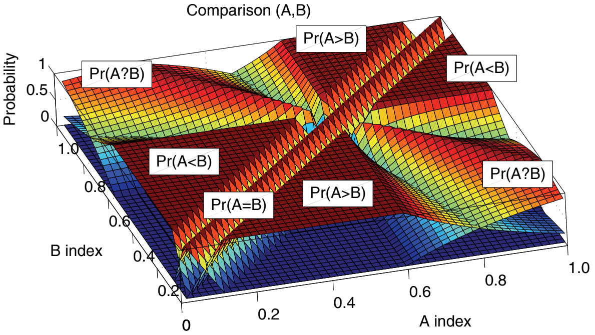

Figure 3. The PDF of the four opinions on the 2-D plane whose axes represent the indexes of the two points compared. A complete evaluation of the 50 discrete operating points is illustrated, where the regions corresponding to the four possible dominating opinions are shown. Assume that $A, B \in [0,1]$, $A^* = 0.5837$, and $JND = 0.05$.

Lemma. A subjectively symmetric point B with respect to A around A^* exists if $p_1(A, A^*) \ge p_{-1}(A, A^*)$ and $p_1(A, A_i^{\max}) \le p_{-1}(A, A_i^{\max})$. Such B, if exists, is unique.

3.2 Simplified Parametric Model

To derive properties of the model and generate efficient search strategies, simplifications are needed on the comparison model. These simplifications allow us to represent the information learned about the location of the local optimum and combine the information on multiple comparisons in a way that improves the search.

Figure 3 depicts the four probabilities as surfaces on the 2-D plane for the 8 regions in Table 2. It also illustrates the regions in which one of the four opinions is dominant. We make the following assumptions in deriving the simplified model for point within the ROD of a local optimum.

Assumption 1. CND and JND are constant and do not vary with respect to A within the ROD of a local optimum. Further, p_0 is linear as a function of $B - A$.

Assumption 2. The boundary line representing subjectively symmetric pairs, $A\|_s B$, is assumed to be a straight line on the A-B plane and is represented by $B = mA + n$, where $m = \frac{-\gamma}{\Delta-\gamma}$ and $n = \frac{\Delta}{\Delta-\gamma}A^*$. This approximation is justified because the preferred trade-offs among objective metrics is slowly changing around a point and is reasonable within the ROD of a local optimum.

Assumption 3. For generalizability of the simplified model, we specify the parameters m and n of the $A\|_s B$ line stochastically. By symmetry, $A^*\|_s A^*$; hence, (A^*, A^*) is on the $A\|_s B$ line. Thus, it suffices to specify another point on the $A\|_s B$ line to uniquely identify the line. Since control symmetry is defined for a pair of points on different sides of A^*, the line has to pass through the line $B - A = \Delta$ (where $\Delta > 0$) between $(A^* - \Delta, A^*)$ and $(A^*, A^* + \Delta)$.

For simplicity of the derivation, we assume that the crossing point is uniformly distributed on the line segment. This assumption results in a piecewise linear shape of the likelihood function that represents the information learned on the location of a local optimum. This cross-over point can be represented by $(A^* - \Delta + \gamma, A^* + \gamma)$, where γ is a random variable uniformly distributed in $[0, \Delta]$.

Assumption 4. In the general model, A is more preferred than B ($p_1(A, B) > p_{-1}(A, B)$) if $A < A^* < B$ and $B > \{B|A\|_s B\}$. In our simplified model, we further assume that $p_{-1} = 0$. This assumption will be used in our derivation when deducing the more likely direction of A^* when an $A >_s B$ opinion is obtained. Similarly, for the opposite case when $B < \{B|A\|_s B\}$ or when subjective symmetry of A does not exist, we assume that $p_1 = 0$. This property will be used when an $A <_s B$ opinion is obtained.

4 Deductions of the Optimal Alternative

Bayesian Formulation. The information deduced about the location of A^* can be represented in a *belief function*. This is a probability density function (PDF) defined over the set of operating points in ROD. It is denoted by $f_{A^*}(a)$ when the operating curve is continuous and by a probability mass function when the curve is discrete. In the rest of this paper, we use belief functions defined over a 1-D continuous space to represent the likelihood of each operating point to be optimal. It is understood that, for a discrete operating curve, the notation can be converted by replacing PDFs with its probability mass function and integration with summation. Without loss of generality, we map the operating curve $[A^{\min}, A^{\max}]$ to $[0, 1]$.

Initial knowledge on the location of A^*. Before any subjective test is conducted, the location of A^* is assumed to be uniformly likely at any operating point on the operating curve. This initial belief function is

$$f^0_{A^*}(a) = 1,\ a \in [A^{\min}, A^{\max}] \tag{1}$$

Deductions from a single pairwise comparison. Based on the distribution of opinions when comparing A and B, we can improve our knowledge on the location of A^*. A Bayesian formulation can be used to obtain the posterior probability from the prior probability and new evidence.

$$f_{A^*}(a|\overline{p}) = \frac{L(a|COD(A, B) = \overline{p}) * f_{A^*}(a)}{\int_0^1 L(\eta|COD(A, B) = \overline{p}) * f_{A^*}(\eta)d\eta}. \tag{2}$$

The formulation requires the prior belief function on the location of A^* and the likelihood function $L(a|\overline{p})$. Before we present the derivation of the likelihood function, we show the deductions using the subjects' responses.

Based on the simplified model, the $A\|_s B$ line satisfies the following criteria:

$$B = mA + n = \frac{-\gamma}{\Delta - \gamma} A + \frac{\Delta}{\Delta - \gamma} A^*, \tag{3}$$

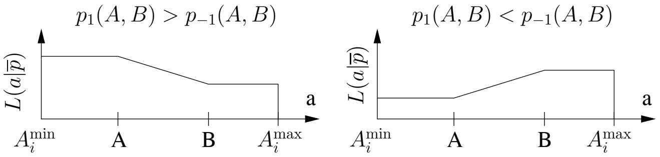

Figure 4. Likelihood functions based on the comparison of the (A, B) pair for 2 cases where $p_1(A, B) > p_{-1}(A, B)$ and $p_1(A, B) < p_{-1}(A, B)$.

where $B - A = \Delta$ and γ is uniform in $[0, \Delta]$. Next, we consider the four responses and analyze their deductions.

- Implication of the $A >_s B$ opinion: $A^* \notin [A + \gamma, A^{\max}]$, since $p_1 = 0$ in Regions 1, 2, 5, and 6. Thus, any $a \in [A^{\min}, A + \gamma]$ can be A^*.
- Implication of the $A <_s B$ opinion: $A^* \notin [A^{\min}, A + \gamma]$, since $p_{-1} = 0$ in Regions 3, 4, 7, and 8. Thus, any $a \in [A + \gamma, A^{\max}]$ can be A^*.
- Implication of the $A \approx_s B$ opinion: A^* can be in any of Regions 1 through 4. This opinion does not provide any information on the location of A^*. Thus, any $a \in [A^{\min}, A^{\max}]$ can be A^*.
- Implication of the $A ?_s B$ opinion: A^* can be in any of the 8 regions. This opinion does not provide any information on the location of A^*. Thus, any $a \in [A^{\min}, A^{\max}]$ can be A^*.

The *likelihood function* $L(a|\overline{p})$ is a function of $a \in [A^{\min}, A^{\max}]$ and indicates the likelihood of obtaining $\overline{p}$ as the result of a subjective comparison of A and B if $A^* = a$. The likelihood of a to be the optimum can be evaluated using the occurrence frequencies of the 4 outcomes analyzed above. This formulation assumes that the subjects have equal expertise, and their responses are independent and identically distributed (*Axiom of IID*). Conditioned on the value of γ and the result of the subjective comparison, we can represent the likelihood as a function of a as follows:

$$L(a|\overline{p}, \gamma) = \begin{cases} p_1 + p_0 + p_2 & \text{if } A^{\min} < a < A + \gamma \\ p_{-1} + p_0 + p_2 & \text{if } A + \gamma < a < A^{\max}. \end{cases} \tag{4}$$

However, the value of γ is unknown and assumed to be uniformly distributed over $[0, \Delta]$, where $\Delta = B - A$. Thus, the expectation taken over γ results in the likelihood function that is only conditioned on $COD(A, B) = \overline{p}$, the result of the subjective evaluation. This likelihood function $L(a|\overline{p})$ is defined as

$$\begin{aligned} L(a|\overline{p}) = E_\gamma[L(a|\overline{p}, \gamma)] &= \int_0^\Delta L(a|\overline{p}, \gamma) Pr(\gamma) d\gamma \\ &= \begin{cases} p_0 + p_2 + p_1 & \text{if } A^{\min} < a < A \\ p_0 + p_2 + \frac{p_1(B-a) + p_{-1}(a-A)}{B-A} & \text{if } A \le a \le B \\ p_0 + p_2 + p_{-1} & \text{if } B < a < A^{\max}. \end{cases} \end{aligned}$$

Figure 4 depicts the two possible cases of the likelihood function defined above in a subjective comparison.

Deductions on subsequent evaluations. The belief function (posterior density) obtained from the Bayesian formulation can be used as the prior knowledge in a subsequent application of the formulation. We assume that the COD results from comparing different pairs are independent in terms of the information on the location of A^*.

For $a \in [A^{\max}, A^{\max}]$, the combined belief function after the n^{th}, $n \geq 1$, comparison is

$$f^n_{A^*}(a) = \frac{f^{n-1}_{A^*}(a) * L(a|COD(A_n, B_n) = \overline{p})}{\int_{A^{\min}}^{A^{\max}} f^{n-1}_{A^*}(\eta) * L(\eta|COD(A_n, B_n) = \overline{p}) d\eta}. \tag{5}$$

The combination process is associative, meaning that the order of the combination does not affect the combined belief function. Further, based on the independence property, the combined belief function found by cascading the Bayesian formulation can be obtained in a closed form:

$$f^n_{A^*}(a) = \frac{\prod_{i=1}^{n} L(a|COD(A_n, B_n) = \overline{p})}{\int_{A^{\min}}^{A^{\max}} \prod_{i=1}^{n} L(\eta|COD(A_n, B_n) = \overline{p}) d\eta}. \tag{6}$$

Utility. The aim of the subjective tests is to obtain $\hat{A}^*$, an estimate of A^*, with high confidence. Thus, the utility of a belief function is the confidence or the probability that $\hat{A}^*$ is within JND_{A^*} of A^*. The estimation error of less than JND_{A^*} is insignificant, since operating points within the JND of A^* is indistinguishable when compared to A^*. Given belief function f, $\hat{A}^*$ is defined to be the point that maximizes the probability of a successful estimation:

$$\hat{A}^*(f) = \arg\max_{a} \left\{ \int_{a-JND}^{a+JND} f(\xi) d\xi \right\}. \tag{7}$$

Given f and $\hat{A}^*$, the utility is defined as follows:

$$U(f) = Pr(|\hat{A}^* - A^*| \leq JND) = \int_{\hat{A}^*-JND}^{\hat{A}^*+JND} f(\xi) d\xi. \tag{8}$$

Stopping criteria. As more pairwise evaluations are conducted, the combined belief function evolves from uniform to a shape that is centered around $\hat{A}^*$. Since it is not feasible to identify A^* with 100% confidence in a continuous search space, we stop the evaluations once the 95% confidence is reached. This means that the actual A^* is within JND of $\hat{A}^*$ with 95% probability.

5 Subjective Evaluation Methods

In this section, we present the general problem formulation and consider several alternatives for conducting subjective evaluations. As is described above, the idea is to systematically use past observations and the updated belief function to choose tests that are likely to produce good deductions and prune the search space. One possibility is to divide the sequence of tests into subsets (called *batches*), ask all subjects to conduct the evaluations in a batch, and update the belief function, before choosing a new sequence of tests adaptively in the next batch.

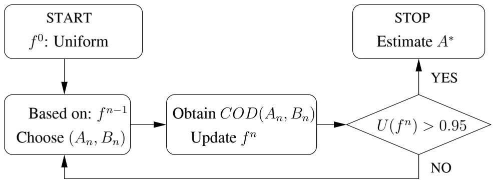

Figure 5. Adaptive choice of comparison points in simultaneous and batch-based evaluations.

Problem formulation. Find a set of M comparison pairs, $\overline{A_n} = [A_n^1, \ldots, A_n^M]$ and $\overline{B_n} = [B_n^1, \ldots, B_n^M]$ in each batch, based on the current belief function, so that the stopping criterion is achieved with the minimum number of total subjective pairwise evaluations:

$$\text{Choose } (\overline{A_n}, \overline{B_n}) \text{ to min } n^* \doteq \min\{n \mid U(f^n) \geq 0.95\}.$$

Simultaneous evaluations. In one extreme, the evaluations are conducted on one pair of operating points in each batch ($M = 1$), before updating the belief function (Figure 5). Since the choice of the pairs compared in the next batch is optimized based on updated information, this results in a lower bound on the number of pairs evaluated. However, the tests need to be synchronized and is inconvenient when multiple batches of tests have to be conducted.

Batch-based evaluations. To avoid synchronizing the subjects in their tests, multiple pairs can be evaluated by all the subjects in each batch ($M > 1$), before updating the belief function (Figure 5). Its disadvantage is that all but a few of the comparisons in a batch may be useful, and the remaining comparisons do not provide any new information for updating the belief function.

Independent evaluations. In the other extreme case, all evaluations are conducted in a single batch. In this case, the information on A^* can only be obtained after all the subjects have completed a predefined set of comparisons. A trivial solution is to select $N \doteq \frac{A^{\max} - A^{\min}}{JND}$, which represents a finite number of operating points that are JND from each other. A complete evaluation of the $N(N-1)/2$ pairs allows us to estimate A^* to within JND of the actual A^*. This approach gives an upper bound on the number of tests.

6 Strategy for Simultaneous Evaluations

At the beginning of the n^{th} comparison, given $U(f^{n-1})$, the expected number of comparisons to reach the stopping criterion if the optimal pair is chosen in each batch is

$$\begin{aligned} S(U(f^{n-1})) &= 1 + S(U(f^n)) \\ &= 1 + \min_{A_n, B_n} S(U(f^n|A_n, B_n)). \end{aligned} \tag{9}$$

The following are the arguments leading to the evaluation of $\min_{A_n, B_n} S(U(f^n|A_n, B_n))$. For any A, B pair, $L(a|A, B)$ is uni-modal. Let $mode(L)$ be the set of points satisfying the modality. It is clear that $A^* \in$

Table 2. Expected number of comparisons.

Algorithm	JND			
	0.1	0.03	0.01	0.003
1. Independent Eval.	45	≈ 500	≈ 5000	≈ 50000
2. Random (any M)	31.1	192	> 300	> 300
3. Optimal (M=1)	6.4	9.9	18.3	49.6
4. Heuristic (M=2)	6.7	11.3	21.4	56.5
Heuristic (M=3)	9.6	15.6	30.4	78.7
Heuristic (M=4)	14.0	19.6	34.2	81.2

$mode(L(a|A,B))$ for any A-B pair. Since any comparison conducted over the same operating curve is consistent, where A^* is common to all the comparisons, it is clear that $mode(L(a|A_1,B_1)) \cap mode(L(a|A_2,B_2)) \neq \emptyset$. Further, for any sequence of A-B pairs, the combined belief function $f^n(a)$ is uni-modal and $A^* \in mode(f^n)$. Hence, $U(f^n)$ is a monotonically non-decreasing function of n for any sequence of comparison pairs, and $S(U)$ is a non-increasing function of U. Thus, minimizing the expected number of steps left is equivalent to maximizing the expected utility under the current belief function in each step.

Our analysis of the likelihood function leads to the following arguments in finding an optimal strategy in each step. The difference between p_1 and p_{-1} indicates the conclusiveness of the comparison result; thus, the expected utility increases monotonically with $|p_1 - p_{-1}|$. Minimizing $\{p_0 + p_2\}$ maximizes $\{p_1 + p_{-1}\}$, which in turn maximizes the expected value of $|p_1 - p_{-1}|$. Given that p_0 decreases with $B-A$ and achieves 0 at $B-A = CND$, and that p_2 increases with $B-A$ and achieves its maximum $(\max\{p_2\} \leq 1)$ at $B - A = A^{\max} - A^{\min}$, $\arg\min\{p_0+p_2\}$ is achieved at $B - A = CND$. For any given $B - A$, the highest $|p_1 - p_{-1}|$ value is achieved when either A or B is equal to A^*. Based on the current estimation of A^* and CND, the optimal pair for the next comparison is

$$(A_n, B_n) = \begin{cases} (\hat{A}^* - C\hat{N}D, \hat{A}^*) & \text{if } n \text{ is even} \\ (\hat{A}^*, \hat{A}^* + C\hat{N}D) & \text{if } n \text{ is odd,} \end{cases} \qquad (10)$$

where one of the points is chosen as the current estimate of A^* and the other point CND away (in either direction). This approach minimizes the possibility of obtaining opinions that do not lead to any deductions on the location of A^* and maximizes the conclusiveness of the comparison.

Batch-based evaluations. The derivation of the optimal sequence of pairs is intractable, since $2M$ variables have to be optimized simultaneously. Further, a numeric solution is too expensive when the number of operating points or M is large. Thus, we use a heuristic to find the set of comparison pairs in the next batch, based on the current belief function. We identify $M - 1$ equally spaced points C^i in the search space for which one of them is equal to $\hat{A}^*$.

$$C^i \quad = \quad \bmod\left(\frac{i-1}{M-1} + \hat{A}^*, 1\right). \qquad (11)$$

For equal spacing, points are wrapped around the operating curve via the modulo operation. We conduct two comparisons involving $\hat{A}^*$, with points JND away from it in either direction, which correspond to the optimal pair for the even and odd cases. For each of the remaining $M - 2$ points identified, it is compared with a point JND away in the direction opposite to that of $\hat{A}^*$:

$$(A_n^i, B_n^i) \quad = \quad \begin{cases} (C^i - C\hat{N}D, C^i) & \text{if } C^i < \hat{A}^* \\ (C^i, C^i + C\hat{N}D) & \text{if } C^i > \hat{A}^* \\ \text{Both pairs above} & \text{if } C^i = \hat{A}^* \end{cases} \quad (12)$$

7 Performance Analysis

We have conducted Monte-Carlo simulations to analyze the performance of the following four algorithms:

- Algo. 1: Complete pairwise comparisons among JND spaced points in the search space.
- Algo. 2: Pairs are chosen in each batch randomly by a uniform distribution in the search space (any M).
- Algo. 3: A single pair ($M = 1$) in each batch is chosen optimally using (10).
- Algo. 4: Multiple pairs ($M > 1$) are chosen in each batch using heuristic (12).

Table 2 compares the performance of simultaneous evaluations and batch-based evaluations for the four algorithms. It shows that conducting independent evaluations and random comparisons are very expensive, and that choosing pairs optimally reduces the number of comparisons needed by 5 folds for simpler problems ($JND = 0.1$) and by 1,000 folds for harder problems ($JND = 0.003$).

The results also indicate that multiple comparisons in a batch reduces the number of batches but increases the total number of comparisons. Therefore, subjective experiments should be designed to balance the overhead of synchronization and the benefit of updating the belief function in order to reduce the total number of comparisons.

References

[1] Z. X. Huang, B. Sat, and B. W. Wah. Automated learning of play-out scheduling algorithms for improving the perceptual conversational quality in multi-party VoIP. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, pages 493–496, July 2008.

[2] Int'l Telecommunication Union. ITU-T G-Series recommendations. `http://www.itu.int/rec/T-REC-G/en`.

[3] Int'l Telecommunication Union. ITU-T P-Series recommendations. `http://www.itu.int/rec/T-REC-P/en`.

[4] B. Sat, Z. X. Huang, and B. W. Wah. The design of a multi-party VoIP conferencing system over the Internet. In *Proc. IEEE Int'l Symposium on Multimedia*, pages 3–10, Taichung, Taiwan, Dec. 2007.

[5] B. Sat and B. W. Wah. Playout scheduling and loss-concealments in VoIP for optimizing conversational voice communication quality. In *Proc. ACM Multimedia*, pages 137–146, Augsburg, Germany, Sept. 2007.