Statistical Scheduling of Offline Comparative Subjective Evaluations for Real-Time Multimedia

Batu Sat, *Member, IEEE*, and Benjamin W. Wah, *Fellow, IEEE*

***Abstract*—In this paper, we study the statistical scheduling of offline subjective tests for evaluating alternative 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 various 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 select the proper control schemes that lead to the best subjective quality at run time. Since subjective tests are expensive to conduct and the number of possible control schemes and run-time conditions is prohibitively large, it is important that a minimum number of such tests be conducted offline, and that the results learned can be generalized to unseen conditions with statistical confidence. To this end, we study in this paper efficient algorithms for scheduling a sequence of subjective tests, while leaving the generalization of limited offline subjective tests to guide the operation of the control schemes at run time to a future paper. Using Monte Carlo simulations, we verify the robustness of our algorithms in terms of their accuracy and efficiency.**

***Index Terms*—Bayesian analysis, Internet, just noticeable difference, real-time multimedia, statistical scheduling, subjective tests, voice-over-IP.**

I. Introduction

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

1) *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. Many multimedia applications are characterized by multiple objective metrics, which cannot be collapsed into a single metric that captures all aspects of quality. When there are multiple metrics, quality can be denoted by a point in a multidimensional space, whose axes correspond to the individual metrics.
2) *Constrained resources.* The control schemes in these systems usually operate under limited network resources (such as constraints on bandwidth and packet rate) and computational resources.
3) *Best-effort IP network.* The IP network used may exhibit dynamic nonstationary delay and loss behavior.
4) *Communication scenario among participants.* This affects the subjective quality perceived by participants. For example, delay degradations may be more important when participants have frequent interactions.
5) *System control.* To mitigate network imperfections, the control schemes employed in these systems 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 multidimensional quality-metric space correspond to its feasible control values. This set of points form an *operating curve*.
6) *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.
7) *Multiple locally optimal operating points.* Due to the counteracting effects of the multiple quality metrics, there may be more than one locally optimal operating points of preferred subjective quality. Each point is the most preferred point among the alternatives in its neighborhood on the operating curve.

Subjective evaluations can be conducted to evaluate the quality of a control scheme. Because such evaluations cannot be performed at run time, offline tests have to be conducted during which the information learned is used to guide the operation of the control scheme(s) at run time. In general, subjective evaluations are time consuming and expensive and will require 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 in order to cover all possibilities.

A standard method for conducting subjective evaluations is to ask subjects to rank the 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 the *mean opinion score* (MOS) [1].

There are two reasons why MOS is only useful for verifying a system's performance but not suitable for designing new control schemes. Firstly, absolute scores obtained for two points on an operating curve can be used to deduce their relative

Manuscript received October 02, 2008, revised April 07, 2009. Current version published September 16, 2009. This work was supported by the National Science Foundation under Grant CNS 08-41336. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Aggelos K. Katsaggelos.

The authors are with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, 61801 USA (e-mail: batusat@illinois.edu; wah@illinois.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2009.2026097

positions. If all alternatives are mutually related under pairwise comparisons, then a total ordering can be established under ACR. In practice, two operating points may not be comparable when they involve multiple quality metrics. In this case, 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. Secondly, although each MOS score can be determined with some statistical confidence, no statistical significance can be associated with the difference of two MOS scores. For instance, 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 [1] 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 for some pairwise comparisons but can be too many for other cases.

Problem statement and approach. To address the issues described above, we study in this paper the statistical scheduling of offline comparative subjective tests for evaluating alternative operating points on an operating curve of a real-time multimedia system. Without loss of generality, we only consider an operating curve due to a single control scheme, although the approach can be easily extended to multiple control schemes. 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. A secondary goal is to efficiently schedule the subjective tests of multiple operating curves in a multimedia application. Our paper extends our previous result that only considers the existence of a single locally optimal point on an operating curve [2]. Our approach consists of the following steps.

1) *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 assessed by a measure that evaluates the relative quality of two alternatives in a *comparison category rating* (CCR) (similar to that described in Annex E of ITU P.800 [1]). By presenting two alternatives to each subject, one after another, the approach 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.
2) *Stochastic evaluations under given conditions.* To identify the best operating point at run time, we first consider the problem of determining the best operating point offline under a given set of network conditions and communication scenarios. When conducting a limited number of subjective evaluations, we use a simulator to repeat the network and communication conditions in order to eliminate variations other than the differences in the control schemes tested, We then collect the comparative subjective opinions and represent them as discrete distributions.
3) *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 that utilizes the following two principles: a) the subjective quality induced by small changes in the control scheme cannot be perceived by subjects, and b) subjective preferences between points that are in a contiguous subset of the operating curve generally point towards the locally optimal point in that subset.
4) *Learning of a classifier.* Based on the subjective preferences under a comprehensive set of test conditions, we learn a support vector machine (SVM) classifier that can generalize to unseen conditions at run time. Due to space limitation, we leave its presentation to a future paper.

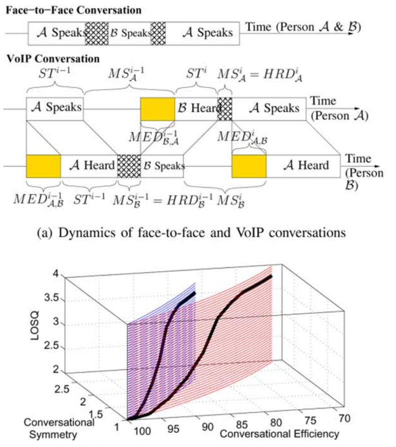
We illustrate in Section II an example POS design problem for a two-party VoIP system. In Section III, we present our notations and our model on subjective evaluations. These are followed by our search algorithms for conducting subjective tests in Section IV, both over a single operating curve as well as over multiple operating curves. Finally, we analyze in Section V the performance of our algorithms.

II. Design of POS Control in Two-Party VoIP

Our proposed approach can be used in many real-time multimedia communication applications. In this section, we describe an example application on the design of a *playout scheduler* (POS) algorithm for a real-time two-party VoIP system. This application demonstrates that subjective tests are needed to achieve high perceptual quality. It is also used as a running example to illustrate the algorithms developed for scheduling subjective tests.

A two-party VoIP conversation consists of one-way transmissions of *speech segments* (STs) in alternating directions that are separated by silence periods called *mutual silences* (MSs). When a connection has delays, the MSs perceived by a client consist of alternating short and long silence periods between turns [3]. This asymmetry is caused by the fact that after $\mathcal{A}$ speaks ST^{i-1}, the MS experienced by $\mathcal{A}$ before hearing $\mathcal{B}$'s response [$MS_{\mathcal{A}}^{i-1}$ in Fig. 1(a)] consists of the time for $\mathcal{A}$'s speech to travel to $\mathcal{B}\left(MED_{\mathcal{A},\mathcal{B}}^{i-1}\right)$, the time for $\mathcal{B}$ to construct a response (*human response delay* or $HRD_{\mathcal{B}}^{i-1}$), and the time for $\mathcal{B}$'s response to travel to $\mathcal{A}\left(MED_{\mathcal{B},\mathcal{A}}^{i-1}\right)$. Here, *mouth-to-ear delay* (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. In contrast, after $\mathcal{A}$ hears ST^i from $\mathcal{B}$, the MS experienced by $\mathcal{A}$ is only governed by her HRD $\left(MS_{\mathcal{A}}^i = HRD_{\mathcal{A}}^i\right)$. This asymmetry leads to a perception that each client is responding slowly to the other, resulting in degraded efficiency and perceptual quality [3].

Since the perceived quality of a VoIP session is subjective and not easily quantified, it is usually measured in terms of objective metrics that depend on the one-way speech quality or *listening-only speech quality* (LOSQ) and the delays of speech segments. These metrics can be based on those recommended by a standardization body, such as ITU P.862 PESQ [1], ITU G.107 E-model [4], and the ITU P.561/P.562 *call clarity index* (CCR) [1]. PESQ is a popular model for estimating LOSQ. However, it is limited when used in real-time VoIP because its computation at the receiver requires the original speech signals. The E-model has been used in VoIP system optimization [5], [6]. They do not always fit well with subjective evaluations because they employ


Face−to−Face Conversation
$\mathcal{A}$ Speaks
$\mathcal{B}$ Speaks
$\mathcal{A}$ Speaks
Time (Person $\mathcal{A}$ & $\mathcal{B}$)
VoIP Conversation
ST^{i-1}
$MS_{\mathcal{A}}^{i-1}$
ST^{i}
$MS_{\mathcal{A}}^{i} = HRD_{\mathcal{A}}^{i}$
$\mathcal{A}$ Speaks
$\mathcal{B}$ Heard
$\mathcal{A}$ Speaks
Time (Person $\mathcal{A}$)
$MED_{\mathcal{B},\mathcal{A}}^{i-1}$
$MED_{\mathcal{A},\mathcal{B}}^{i}$
$\mathcal{A}$ Heard
$\mathcal{B}$ Speaks
$\mathcal{A}$ Heard
Time (Person $\mathcal{B}$)
$MED_{\mathcal{A},\mathcal{B}}^{i-1}$
ST^{i-1}
$MS_{\mathcal{B}}^{i-1} = HRD_{\mathcal{B}}^{i-1}$
$MS_{\mathcal{B}}^{i}$
LOSQ
4
3.5
3
2.5
2
Conversational Symmetry
2.5
2
1.5
1
Conversational Efficiency
100
95
90
85
80
75
70

(a) Dynamics of face-to-face and VoIP conversations

(b) Operating curves under two conditions

Fig. 1. Representation of conversational dynamics in a 3-D objective space that involves trade-offs on CS, CE, and LOSQ.

simplistic assumptions in estimating conversational quality. For instance, the E-model assumes that degradations due to delay and LOSQ are independent and additive. They further implicitly assume that the quality of any alternative can be represented by a scalar value, which implies that any pair of alternatives can be compared and ordered. As discussed in Section I, such comparison may not always be possible.

Other objective metrics that can be computed at run time have been proposed. In this paper, we define two such metrics for measuring the quality of two-way VoIP conversations [3].

1) *Conversational efficiency* (CE). This measures the extension in time to accomplish a VoIP conversation when there are communication delays. Since a conversation over a network may be charged according to its duration, the same conversation will cost more for a network with lower CE. CE is identically perceived by both clients:

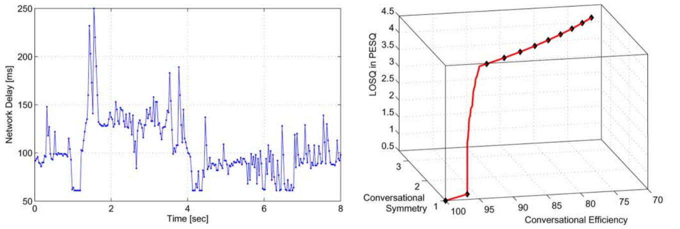
$$CE = \frac{\text{Speaking Time} + \text{Listening Time}}{\text{Total Time of Call}}. \quad (1)$$

2) *Conversational symmetry* (CS). When a client experiences highly asymmetric response times in a conversation, she tends to perceive a degradation in the naturalness of the conversation because it does not resemble a face-to-face conversation with uniform delays. One possible effect is, if $\mathcal{A}$ perceives $\mathcal{B}$ to be responding slowly, then $\mathcal{A}$ tends to respond slowly as well. To capture the asymmetry of MSs perceived by $\mathcal{A}$, we define $CS_{\mathcal{A}}$ to be the ratio of the maximum and the minimum MSs experienced by $\mathcal{A}$ in a past window:

$$CS_{\mathcal{A}} = \frac{\max_j MS_{\mathcal{A}}^{j}}{\min_j MS_{\mathcal{A}}^{j}},$$
$$j \in \{\text{speech segments in a past window}\}. \quad (2)$$

$CS_{\mathcal{A}}$ is approximately 1 in a face-to-face conversation but increases as round-trip delays increase.

Fig. 1(b) depicts the quality of a conversational segment as a point in a 3-D space whose axes correspond to three objective metrics. Under a given conversational condition, the possible operating points are restricted to a curved plane perpendicular to the CS-CE plane, where two such conversational conditions are depicted in Fig. 1(b). By using MED as the control parameter and by using LOSQ to characterize the quality of the speech segments received, the possible operating points are further restricted to a curve on one of these planes, where the operating point shifts towards the upper right-hand corner (higher LOSQ and CS and lower CE) as MED is increased.



(a) Network delays from California to Maryland (b) Example operating curve as a function of MED

Fig. 2. Network condition and the corresponding operating curve for a conversation with medium switching frequency.

The quality of the conversation perceived by the two clients is controlled by a POS algorithm, whose goal is to find the MED under some operating condition that leads to the best subjective conversational quality. The MED that optimizes subjective quality is not at the extremes of an operating curve, but at a point where the counteracting effects on subjective quality are relatively balanced. Further, the optimal MED depends on the given network and conversational conditions. For example, for a connection with high delays and jitters, the optimal MED may need to be higher in order to improve the poor LOSQ. In contrast, for a conversation in which clients 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, MED will need to be dynamically adjusted in a closed-loop fashion in order to continually maintain high perceived conversational quality.

Finding the best MED under a given condition 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. As a result, it is infeasible to conduct indiscriminate subjective tests in order to evaluate all possible pairs of conditions that can exist at run time.

Example 1: The operating curve presented in this running example is used to illustrate the various concepts in this paper. We use a network condition with low delays, high jitters, and low losses. Fig. 2(a) depicts the packet delays observed as a function of sent times for a Maryland-Californian connection. (A comprehensive set of network conditions can be found elsewhere [7].) We use a conversational scenario with an average 24 switches/minute under no delay and whose average speech-segment and HRD durations are 1706 ms and 552 ms, respectively.

Fig. 2(b) shows a 3-D representation of the operating curve, where the curve is parametrized by MED. Starting from the lower left-hand corner, each diamond represents an increase of 100 ms in MED.

Another application for illustrating the need for subjective evaluations is in the dynamic equalization of mutual silences (MS) in multiparty VoIP [8], [9]. Here, the control parameter

TABLE I
FOUR POSSIBLE OPINIONS IN A SUBJECTIVE TEST OF A AND B

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

under given network and conversational conditions is the level of MS equalization, whereas the objective metrics are LOSQ, CE, and CS [8]. Subjective tests for guiding the selection of the best MS values are needed because there is no single objective metric that can capture all aspects of subjective conversational quality in multiparty VoIP. Yet another application that can benefit from this approach is in real-time video conferencing [10], where the controls of the encoding rate and the playout scheduler affect the delay-quality trade-offs perceived by users.

III. MODEL OF SUBJECTIVE COMPARISONS

In this section, we present the properties of comparative subjective tests, which lead to a general model for evaluating points on an operating curve. We first present the notations and basic axioms on comparative subjective evaluations. Next, we define the local optimality of an operating point and the possibility of multiple local optima on an operating curve. Based on the region of dominance of a local optimum, we present stronger axioms that are valid within the region. This is followed by a stochastic model on pairwise subjective comparison tests.

A. Comparative Subjective Tests

Comparative subjective tests are conducted by comparing two points A and B on an operating curve. Each test is conducted by asking a subject to compare A and B in a random order (to avoid any perceived bias). The alternatives are generated under the same operating conditions but under different control parameter values.

Table I shows the comparison results in one of the four opinions, where p_i is obtained by normalizing the number of subjects who responded with opinion i with respect to K, the total number of subjects. Note that the complement to the opinion "$A \approx_s B$" is "A is not about the same as B" (or $A \neq_s B$), which consists of $A >_s B$, $A <_s B$, and $A?_sB$.

The results of the subjective tests can be combined into a stationary probability distribution (or a sample distribution when K is finite) in the form of a vector called the *comparative opinion distribution* (COD):

$$\begin{aligned} COD(A,B) &= \overline{p} = (p_{-1}, p_0, p_1, p_2) \\ \text{where } \sum_i p_i(A,B) &= 1 \text{ for each } (A,B) \text{ pair.} \end{aligned} \tag{3}$$

To avoid confusion, p_i is assumed to be stationary (when K is large), and we represent a sample probability by $\hat{p}_i$ where necessary.

Example 1 (Cont'd): Using a testbed we have developed [3] and under the given network and conversational condition in Fig. 2(b), each point on the operating curve can be simulated to result in a VoIP conversation. For example, $A = 0.111$ and $B = 0.198$ result in two conversations generated using, respectively, MEDs of 111 ms and 198 ms, and represented by $(PESQ, CS, CE) = (2.60, 1.27, 0.96)$ and (4.18,1.48,0.93). After subjective tests, $COD(A,B) = (p_{-1}, p_0, p_1, p_2) = (0.75, 0.125, 0.125, 0)$, which means that B is preferred over A.

B. Monotonicity of Objective Metrics

An *operating curve* is denoted by $\mathcal{O}$, which is mapped to a real number in $[0,1]$ with extreme points $A^{\min} = 0$ and $A^{\max} = 1$. Each point on the curve is denoted by a capital letter (such as A and B) and has a one-to-one correspondence to the value of the associated control scheme that realizes the communication application. Thus, changes to the scalar control value are mapped to changes in one of the two directions along the operating curve.

Property 1. Monotonicity: Each objective metric is either monotonically non-increasing or monotonically nondecreasing with respect to increases in the corresponding control value.

Example 1 (Cont'd): Referring to Fig. 2(b), $A^{\min}$ represents the degenerate case in which POS plays each speech frame at the instant it was spoken by the remote client, whereas $A^{\max}$ represents the case in which POS waits 1 s after each speech frame is spoken before playing it.

In general, changes in MED may have no effect, or improve, or degrade the corresponding objective metrics. In this example, new packets can arrive in time for play back when MED is increased (due to either a redundant loss-concealment scheme or a higher chance for late packets to arrive), which means that there is a non-increasing relationship between MED and the rate of late packets. Because the arrivals of packets happen at discrete times, the objective metrics may have finite discontinuities when MED is increased. The relation between LOSQ and MED depends on the robustness of the speech codec on losses and the network jitter, although it is always monotonically non-decreasing. Similarly, the relation between CS (resp. CE) and MED depends on HRD (resp. ST and HRD) and is monotonically increasing (resp. decreasing) with respect to MED by definition [7].

It is possible for the monotonicity property to be violated in such a way that there are multiple local optima with respect to an objective metric when the control variable is increased. In this case, the operating curve can be divided into multiple non-overlapping segments called *regions of dominance* (Section III-D), where the objective metrics in each region satisfy the monotonicity property.

Implications on subjective preferences. The monotonicity property ensures that when perturbing from A to B, a subset of the objective metrics exhibit a non-decreasing trend, whereas the remaining metrics exhibit a non-increasing trend. However, the trade-offs among the metrics do not necessarily result in a bell-shaped subjective preference curve. These trade-offs can change at different operating points because they depend on the perceived degradation of each quality metric as well as the relative change in that perception. As a result, there can be multiple locally optimal subjectively preferred points within a region where monotonicity of the objective metrics is satisfied.

For example, if a dominant degradation is common to both A and B, then a subject may more likely prefer the point that exhibits improvement in the dominant degradation. In contrast, if A exhibits a significant improvement on the less perceived degradation, while no perceptible difference is observed between A and B with respect to the dominant degradation, then A is more likely to be preferred.

C. Basic Axioms

Axiom 1. 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 there is no difference in the objective metrics between A and itself, subjects should not perceive them to be different except for mistaken evaluations.

Axiom 2. IID: Each subject has the same level of expertise, and their responses to comparing any two points on an operating curve are independent and identically distributed (IID).

This axiom allows us to model the sample COD in (3) by a multinomial distribution. In particular, the order of a comparison does not affect $COD(A, B)$, as stated in the following axiom.

Axiom 3. Symmetry/Anti-Symmetry: Indistinguishable ($\approx_s$) and incomparable ($?_s$) opinions are symmetric: $p_i(A, B) = p_i(B, A)$ for $i \in \{0, 2\}$. Preference opinions ($>_s$ and $<_s$) are anti-symmetric: $p_{-1}(A, B) = p_1(B, A)$.

Let $B - A$ be the perturbation in the control value from a fixed A to a variable B. Since each objective metric is monotonic with respect to B and a small change in B may result in a possibly discrete but finite change in the objective metric, there will be a small fraction of subjects perceiving a difference in quality. Hence, a small change in B will result in a possibly discrete change in the probability of perceiving such a difference when the number of subjects is large. As the difference between A and B increases, the perception of the difference in their subjective quality increases. This noticeable difference is commonly used in psychophysics and is defined as follows.

Definition 1. Just Noticeable Difference (JND) of A: When comparing a fixed A and a variable B on an operating curve $\mathcal{O}$, $JND(A)$ is the $B - A$ value for which 50% of the subjects perceive a difference in their quality.

In statistical inference from a finite number of subjective evaluations, we define $JND(A)$ to be the minimum value of $|B - A|$ such that the hypothesis, $\{H_0 : p_0(A, B) < 0.5\}$, is rejected with a given statistical significance. If B is inside the JND of $A(|B - A| \leq JND(A))$, then A and B are *indistinguishable*; otherwise, they are *distinguishable*.

Definition 2. Complete Noticeable Difference (CND) of A: When comparing a fixed A and a variable B, $CND(A)$ is the minimum $|B - A|$ value such that $p_0(A, B) = 0$.

Axiom 4. 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.

When $B = A$ (thus, $|B - A| = 0$), $p_0(A, A)$ is equal to 1 due to Axiom 1. As $|B - A|$ increases, there are larger differences in their objective metrics, resulting in a non-decreasing number of subjects perceiving a difference in their quality. Eventually, all subjects perceive that A is not the same as B.

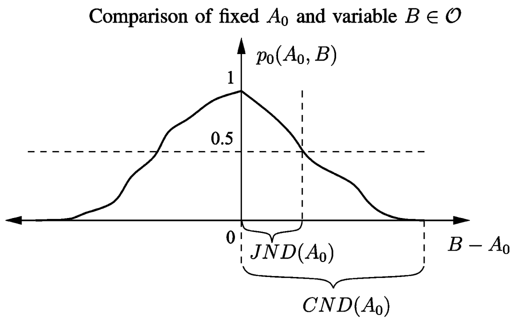


Fig. 3. Comparing a fixed A_0 with a variable B: p_0 is non-increasing as a function of $B - A_0$.

TABLE II
p_0 OF SUBJECTIVE COMPARISONS BETWEEN VARIOUS B WHEN COMPARED TO A AND A' IN FIG. 2(b)

	B					
	0.25	0.30	0.35	0.40	0.50	0.60
$A = 0.20$	0.95	0.85	0.65	0.45	–	–
$A' = 0.30$	–	–	1.00	0.90	0.60	0.40

As a result of Axiom 4, Fig. 3 shows that the $JND(A_0)$ and $CND(A_0)$ regions are single contiguous regions around A_0.

$JND(A)$ and $CND(A)$ can 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. For another A, it may require a large perturbation. Our subjective tests in two-party VoIP conversations confirm the variations in JND and CND as a function of A.

Example 1 (Cont'd): The example VoIP application satisfies Axioms 1–3. To illustrate JND, we have conducted pairwise subjective evaluations between alternatives on the operating curve in Fig. 2(b). Firstly, we compare $A = 0.2$ with B that has larger MED with respect to that of A. As the difference in MEDs increases, the fraction of responses indicating that the two are about the same decreases. This happens because the differences in LOSQ, CS, and CE increase at the same time, and a larger fraction of the subjects can perceive the difference between the conversations. We then repeat the experiments using $A' = 0.3$. We observe that the JND observed tends to be larger than that in the first experiment, which means that subjects are less sensitive to perceiving the changes. This is due to the fact that, as the baseline degradations due to MED are larger, the notice-ability threshold, which is related to the baseline MED, is also larger. This behavior is illustrated in Table II, which lists the values of p_0 for various B when compared to A and A'.

D. Locally Optimal Points on an Operating Curve

Intuitively, an optimum is a point that is preferred when compared to every other feasible point on an operating curve. It is preferred because it achieves the optimal trade-off among the various objective metrics and cannot perform better by operating at another point. Hence, identifying such a point is paramount in the design of adaptive system-control schemes.

Definition 3. Local Optimum: Point A_i^* is locally optimal over points in a subset of the operating curve $\mathcal{O}_i \subseteq \mathcal{O}$:

$$A_i^* = \{A \mid p_1(A,B) > 0.5 \, \forall \, B \in \mathcal{O}_i \text{ such that } |B - A| > JND(A)\}. \quad (4)$$

There can be multiple local optima on an operating curve, since changes in the multiple objective metrics along the operating curve may lead to locally optimal trade-offs.

Definition 4. Region of Dominance (ROD)*:* The ROD of a local optimum A_i^*, $ROD(A_i^*)$, is the largest contiguous region $\mathcal{O}_i$ of an operating curve $\mathcal{O}$ in which (4) is satisfied.

A local optimum is dominant (or preferred more than 50% of the time) against any point within its ROD, except for points in its JND region. However, when A_i^* is compared against B outside its ROD $(B \notin ROD(A_i^*))$, we cannot conclusively say whether A_i^* is preferred over B; that is, the hypotheses $p_0(A_i^*, B) > 0.5$, $p_1(A_i^*, B) > 0.5$, and $p_{-1}(A_i^*, B) > 0.5$ are all rejected with some statistical significance. Similarly, nothing can be concluded when comparing a point in the ROD of one local optimum with a point in the ROD of another local optimum.

Example 1 (Cont'd): The operating curve in Fig. 2(b) has a single local optimum, although it cannot be proved unless infinitely many subjective tests are conducted. In Section IV-D, we illustrate the existence of the local optimum by conducting a finite number of tests. With one local optimum, the operating curve has one ROD.

Lemma 1: There cannot be multiple local optima that are within the ROD of each other.

Proof: Assume two local optima that are within the ROD of each other (e.g., $A_1^* \in ROD(A_2^*)$ and $A_2^* \in ROD(A_1^*)$). From (4), $p_1(A_1^*, A_2^*) > 0.5$, and $p_1(A_2^*, A_1^*) > 0.5$. Due to Axiom 3, $p_{-1}(A_1^*, A_2^*) > 0.5$; thus $\sum_i p_i(A_1^*, A_2^*) > 1$. These are contradictions. ■

Definition 5. Global Optimum: A^*, is a point that dominates all points on an operating curve $(ROD(A^*) = \mathcal{O})$.

Each operating curve is not guaranteed to have a global optimum. However, if one exists, it is unique. That is, there cannot be another point outside the JND of the global optimum that satisfies the property of global optimality. This is stated formally as follows.

Lemma 2: If a global optimum exists, then it is unique.

Proof: Assume that two global optima A_1 and A_2 satisfy (4) for $\mathcal{O}_i = \mathcal{O}$ and that $|A_1 - A_2| > \max\{JND(A_1), JND(A_2)\}$. Then $p_1(A_1, A_2) > 0.5$ and $p_1(A_2, A_1) > 0.5$. Thus, $\sum_i p_i(A_1, A_2) > 1$. This is a contradiction. ■

It is, however, possible that multiple points, all within the JND of each other, satisfy the definition of global optimality. This does not cause any inconsistencies because any of the candidate points can be chosen as the global optimum and no candidate is distinguishable from another.

E. Incomparability Within the ROD of a Local Optimum

At a local optimum A_i^*, the quality metrics have an optimal trade-off. However, due to Property 1, as the point is perturbed away from A_i^* in one direction (say towards $A^{\max}$) but still within the ROD of A_i^*, a subset of the metrics exhibit more perceptible degradations that dominate the other metrics. On the other hand, if the point is perturbed in the other direction (say towards $A^{\min}$) but within the ROD of A_i^*, then a different subset of the metrics exhibit more perceptible degradations. Thus, when a subject is asked to compare the two points on different sides of A_i^*, the subject may indicate that the pair is incomparable, since different subsets of quality aspects dominate the degradation. As the distance between these points increases, the overlap between the subsets of dominant quality aspects is reduced, which causes more subjects to indicate that the pair is incomparable. As an example, in two-party VoIP, perturbations from the local optimum in one direction cause degradations due to delay to be dominant and in the other direction degradations due to speech quality to be dominant.

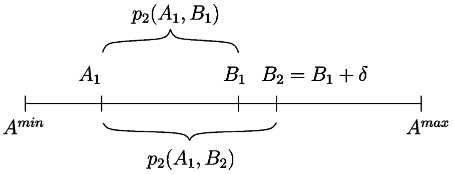


Fig. 4. Comparison of A_1 with B_1 and A_1 with B_2. p_2 increases when $\delta > 0$.

Axiom 5. Incomparability of A and B: In the stationary case when there are a large number of subjects $(K \rightarrow \infty)$, $\lim_{\delta \to 0^+} p_2(A, B + \delta) \geq p_2(A, B)$, and $\lim_{\delta \to 0^+} p_2(A - \delta, B) \geq p_2(A, B)$.

Fig. 4 illustrates the axiom. Assume that metrics 1 and 2 are monotonically non-increasing and that metrics 3 and 4 are monotonically non-decreasing. Given the comparison of A_1 and B_1, a second comparison between A_1 and B_2 is conducted, where B_2 is perturbed by an infinitesimal amount from $B_1 (B_2 = B_1 + \delta, \delta \rightarrow 0)$. Due to the monotonicity of the metrics, a perturbation from B_1 to B_2 results in some of the perceptible objective metrics to be less perceptible and some less perceptible ones to be more perceptible. This causes the subject to depend on a slightly different set of quality aspects in evaluating the quality trade-offs. This change results in a reduction in the overlap of the important metrics between A_1 and B_2 with respect to A_1 and B_1. Hence, the probability that subjects perceive A_1 and B_2 to be incomparable is monotonically non-decreasing with respect to A_1 and B_1.

Lemma 3: For $K \rightarrow \infty$ and any finite $\Delta > 0$, $p_2(A, B + \Delta) \geq p_2(A, B)$, and $p_2(A - \Delta, B) \geq p_2(A, B)$, where A, B, $B + \Delta$, $A - \Delta \in ROD(A_i^*)$.

Proof: The proof follows directly from Axiom 5 after cascading together infinitesimal changes. ■

Corollary 1: $p_2(A_2, B_2) \geq p_2(A_1, B_1)$ if $[A_1, B_1] \subseteq [A_2, B_2]$, where $A_1, B_1, A_2, B_2 \in ROD(A_i^*)$.

Corollary 2: $p_2\left(A^{\min}{}_i, A^{\max}{}_i\right) \geq p_2(A, B)$ for all $A, B \in ROD(A_i^*)$.

Example 1 (Cont'd): As the difference between the MEDs of two operating points is decreased by moving one or both of the points closer to the other, the incomparability rate among the subjects decreases as well. Table III summarizes the results of the subjective tests. Three comparisons were conducted,

TABLE III
$p_2(A,B)$ FOR THE OPERATING CURVE IN FIG. 2

(A,B)	A				B				$p_2(A,B)$
	MED	PESQ	CS	CE	MED	PESQ	CS	CE	
(0.11,1.00)	110	2.39	1.27	0.96	1000	4.18	3.42	0.72	0.60
(0.11,0.50)	110	2.39	1.27	0.96	500	4.18	2.21	0.84	0.20
(0.25,0.50)	250	4.18	1.60	0.91	500	4.18	2.21	0.84	0.0

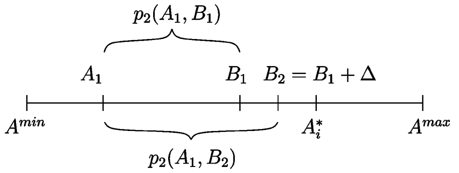


Fig. 5. Comparing A and B in the same side of the local optimum A_i^*: $|p_1 - p_{-1}|$ increases when B is perturbed towards A_i^*.

where the $[A,B]$ segment of each subsequent pair is a subset of the previous segments (e.g., $[0.25, 0.50] \subset [0.11, 0.50] \subset [A^{\min}, A^{\max}]$).

F. Subjective Preference Within the ROD of a Local Optimum

Consider a pair of operating points in the ROD of a local optimum A_i^*. In contrast to the indistinguishable (p_0) and incomparable (p_2) opinions, p_1 and p_{-1} contain information on the location of A_i^*. In this section, we present our observations and basic axiom on the preference of one point over another. These results are used later to represent the information deduced on the location of A_i^*.

Axiom 6. Subjective Preference: For $K \rightarrow \infty$ and A and B on the same side of A_i^* where $A < B$

$$|p_1(A,B) - p_{-1}(A,B)| \leq \begin{cases} \lim_{\delta \to 0^+} |p_1(A-\delta,B) - p_{-1}(A-\delta,B)| \\ \lim_{\delta \to 0^+} |p_1(A,B+\delta) - p_{-1}(A,B+\delta)|. \end{cases} \quad (5)$$

Fig. 5 explains the axiom intuitively. As B_1 moves to B_2 towards A_i^*, it will have more balance in its objective metrics and better perceived quality when compared to A_1. The difference between p_1 and p_{-1} indicates the conclusiveness of this perceptual comparison because it represents the improvement of the preferred opinion with respect to the non-preferred opinion. As B moves towards A_i^*, the conclusiveness of the comparison improves.

The POS-design problem described in Section II generally exhibits this property, where the preference towards the alternative closer to the optimal point increases as the other alternative moves away from the optimum. This perturbation makes either LOSQ or delay degradation more dominant and, thus, more perceptible.

Definition 6. Control Symmetry: For A and B on opposite sides of A_i^*, A and B are objectively symmetric, denoted by $A\|_0B$, if they are equidistant from A_i^* in terms of their control value; that is, $|A - A_i^*| = |B - A_i^*|$.

Definition 7. Subjective Symmetry: For A and B on opposite sides of A_i^*, A and B are subjectively symmetric, denoted by $A\|_sB$, if $p_1(A,B) \leq p_{-1}(A, B-\delta)$ and $p_1(A,B) \geq p_{-1}(A, B+\delta)$, where $\delta \rightarrow 0$. This means that the probabilities for one point to be preferred over another are equal from both directions.

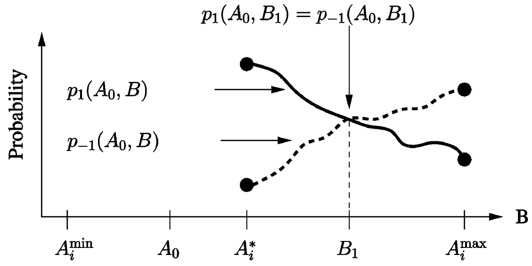


Fig. 6. Subjective symmetry of A_0 and B.

In the special case in which the objective metrics are continuous at A and B with respect to the control value, subjective symmetry results in $p_1(A,B) = p_{-1}(A,B)$. To account for possibly finite discontinuities in the objective metrics at A and B, we need to define subjective symmetry with respect to $\delta \rightarrow 0$.

Lemma 4: A subjectively symmetric point B on opposite side of A_i^* with respect to A exists if $p_1(A, A_i^*) \geq p_{-1}(A, A_i^*)$ and $p_1(A, {A^{\max}}_i) \leq p_{-1}(A, {A^{\max}}_i)$, where $A < A_i^* < B < {A^{\max}}_i$. Such a point B, if exists, is unique.

Proof: Existence. We know that p_1 and p_{-1} are either non-increasing or non-decreasing between A_i^* and ${A^{\max}}_i$ when A is fixed and B is between A_i^* and ${A^{\max}}_i$. Assuming $p_1(A_0, A_i^*) \geq p_{-1}(A_0, A_i^*)$ and $p_1(A_0, {A^{\max}}_i) \leq p_{-1}(A_0, {A^{\max}}_i)$, then there exist at least one B_1 at the cross-over point in Fig. 6 between the two curves that satisfy the condition in Definition 7 with respect to A_0 and B, namely, $A_0\|_sB$.

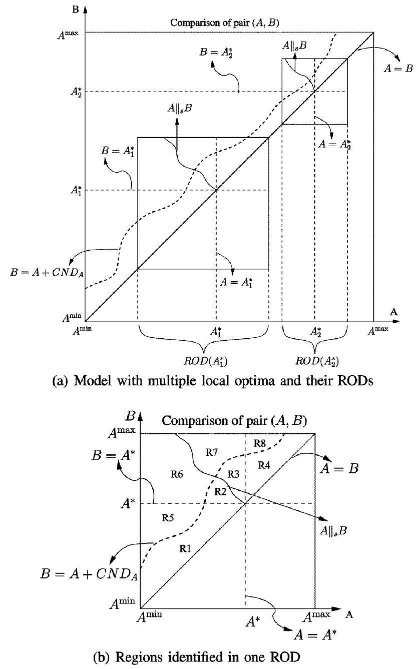
Uniqueness. Since the functions $p_1(A_0,B)$ and $p_{-1}(A_0,B)$ are monotonic (non-increasing or non-decreasing), point B that satisfies the condition must be unique. In case where both functions are constant, namely, $p_1(A_0,B) = p_{-1}(A_0,B)$, then there is a region in which the condition is satisfied. Since all points in such a region satisfy the condition, the region is unique. ∎

The comparison of subjectively symmetric points does not result in any new information on the location of A_i^*. However, when comparing A with any point that is larger than B where $A\|_sB$, then A is more preferred. This observation will be useful for deducing the location of A_i^* from the result of subjective tests.

G. General Model of Subjective Comparisons

In this section, we use the axioms presented to develop a general model of subjective comparisons. Fig. 7(a) depicts the general case with multiple local optima, where the axes represent the two points compared. Due to Axiom 3, it suffices to assume $B \geq A$. We focus on a detailed description of the COD for $ROD(A_i^*)$.

A more restricted model describes the probabilities of occurrence of the four possible opinions when comparing A and B in one ROD. Fig. 7(b) depicts the model and the eight regions, whose properties on COD are summarized in Table IV.



(a) Model with multiple local optima and their RODs

(b) Regions identified in one ROD

Fig. 7. Model of subjective comparison of two operating points A and B on an operating curve.

TABLE IV
PROPERTIES ON COD OF THE EIGHT REGIONS IN FIG. 7(b) DEFINED WITH RESPECT TO THE FOUR BOUNDARY LINES WITH $B > A$: $B - A - CND(A)$, $A - A^*$, $A\|_s B$, AND $B - A^*$

	Regions in a ROD							
	R1	R2	R3	R4	R5	R6	R7	R8
p_0	> 0	> 0	> 0	> 0	0	0	0	0
p_{-1} vs. p_1	$p_{-1} > p_1$	?	?	$p_{-1} < p_1$	$p_{-1} > p_1$	?	?	$p_{-1} < p_1$
p_2	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0

The eight regions defined with respect to the four boundary lines have the following properties.

- *Regions R1 and R5:* A and B are on the same side of A_i^*, where $A < B < A_i^*$. If a pair compared belongs to these regions, then the result satisfies $p_{-1} < p_1$, according to Axiom 6. Without knowing A_i^*, such a result indicates that A_i^* is more likely to be larger than B. This is consistent with the actual location of A_i^* and guides the search in the right direction.
- *Regions R4 and R8:* A and B are on the same side of A_i^*, where $A_i^* < A < B$. If a pair compared belongs to these regions, then the result satisfies $p_1 < p_{-1}$, according to Axiom 6. Without knowing A_i^*, such a result indicates that A_i^* is more likely to be smaller than A. This is consistent with the actual location of A_i^* and guides the search in the right direction.
- *Regions R2, R3, R6, and R7:* A and B are on opposite sides of A_i^*, where $A < A_i^* < B$. If a pair compared belongs to these regions, then p_1 or p_{-1} can be larger. Such a result is inconclusive for guiding the search.

The model does not specify the result when comparing one point in an ROD against another point outside of the ROD. Such comparisons do not provide information for identifying a local optimum and should be avoided.

Lemma 5: The RODs corresponding to different local optima do not overlap.

Proof: Assume that $ROD(A_1^*)$ and $ROD(A_2^*)$ overlap and that A and B are chosen in the overlapped region. Without loss of generality, assume $A_1^* < A_2^*$. By applying Axiom 6 on $A_1^*, p_1(A, B) > p_{-1}(A, B)$. On the other hand, when applying Axiom 6 on $A_2^*, p_1(A, B) < p_{-1}(A, B)$. This is a contradiction. ■

IV. EFFICIENT AND ACCURATE SEARCH ALGORITHMS FOR FINDING LOCAL OPTIMA

Based on the fundamental understanding of subjective tests in the last section, we develop in this section a systematic approach for conducting pair-wise comparative subjective tests among points on one or more operating curves. There are two counteracting metrics of success for this task, the most important being the accuracy of the local optimum estimated. Since there are infinitely many points on a continuous operating curve, it is impossible to identify the optimum via a finite number of tests. However, it suffices to estimate the local optimum to within the JND of its actual location, since both are indistinguishable in this region. The second metric of success is the number of subjective comparisons conducted. Although more comparisons would lead to a better estimate, it is important to develop a method that achieves the desired level of accuracy using the minimum number of tests.

The development of an efficient search strategy is based on the following observations. Firstly, the COD obtained by comparing two points on an operating curve provides information on the preferred trade-offs among the associated objective metrics, which indicates a direction in which the local optimum is likely to be located. As more evidence is collected, the collective information leads to an estimate of the ROD and its local optimum with higher statistical confidence. Using the information on the estimated location of a local optimum, our strategy chooses the next pair of points to be compared in order to minimize the total number of comparisons to achieve a given level of accuracy.

A. Conducting Subjective Evaluations of a Single Operating Curve

There are several alternatives for conducting subjective evaluations of points on an operating curve. A general approach is to divide the sequence of tests into *batches* with M tests each, ask all subjects to conduct the tests in a batch, update the estimation of the local-optimum candidates, and adaptively choose a new sequence of tests in the next batch. Because the test results in one batch are used to optimize the tests in the next batch, the tests must be synchronized so that those in the current batch are completed before beginning those in the next batch.

Sequential evaluations $(M = 1)$. In one extreme, each batch consists of one pair of points to be tested. This approach results in the least number of tests before updating the estimate of the local-optimum candidate. Hence, it leads to a better choice of the next test to be conducted and a lower bound on the total number of tests. However, it also results in an upper bound on the number of batches, making it inconvenient for subjects because they have to combine their results at the end of each test before the next test can be determined.

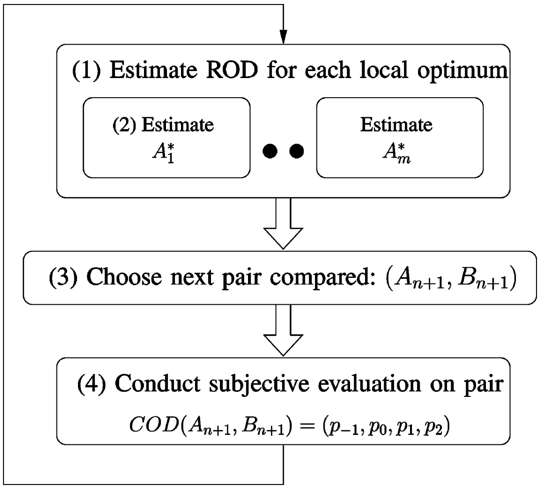

(1) Estimate ROD for each local optimum
(2) Estimate A_1^*
Estimate A_m^*
(3) Choose next pair compared: (A_{n+1}, B_{n+1})
(4) Conduct subjective evaluation on pair
$COD(A_{n+1}, B_{n+1}) = (p_{-1}, p_0, p_1, p_2)$

Fig. 8. Method for identifying a local optimum through subjective comparisons.

Batch-parallel evaluations $(M > 1)$. To avoid frequently synchronizing the subjects in their tests, M pairs of tests can be evaluated by all subjects in each batch, before updating the local-optimum estimate. Its disadvantage is that, when M is large, most of the test results do not provide new information on the estimate.

Fully parallel evaluations $(M \gg 1)$. In the other extreme, all evaluations are conducted in a single batch. In this case, the estimate on the local optimum can only be obtained after all the subjects have completed a predefined set of comparisons. A trivial solution is to select $N \doteq \left(A^{\max}{}_i - A^{\min}{}_i / JND\right)$, 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_i^* to within JND of the actual A_i^*. This approach gives an upper bound on the number of tests. However, since $A^{\min}{}_i$, $A^{\max}{}_i$, and JND are unknown, a separate set of tests is needed to first find these values. Such tests can be as expensive as conducting tests to find the local optima.

Because sequential evaluations are more effective for reducing the total number of tests in identifying a local optimum of an operating curve, we study this method in detail in this section. Fig. 8 depicts the three steps in our method, which involve estimating the values of $A^{\min}{}_i$, $A^{\max}{}_i$, and JND, as well as the local optimum in an ROD.

- Step 1: Given the evidence collected, estimate the ROD of each local optimum.
- Step 2: Given the ROD of a local optimum and the evidence collected so far, estimate the local optimum.
- Step 3: Given an estimate of the local optimum, chose the next pair of points to be evaluated.

In the rest of this section, we first present the second step for estimating the local optimum in an ROD, since the first and the last steps utilize this step. Based on our proposed method, we present at the end of the section a heuristic for batch-parallel evaluations and the approach for the subjective evaluation of multiple operating curves.

B. Step 2: Finding a Local Optimum in a Given ROD

In this section, we develop the second step of our method, using an estimate of the ROD and the previous comparison results. Our goal is to refine the estimate of the local optimum in order to get a better confidence.

The model in Section III-G for comparing points in $ROD(A_i^*)$ allows us to determine a likely direction on the location of A_i^*. However, its non-parametric nature makes it difficult to combine the result of a test with the prior information obtained. Hence, we cannot calculate the statistical likelihood on the probable locations of A_i^*.

To address this issue, we develop in this section a parametric model of subjective comparisons in $ROD(A_i^*)$ after simplifying the general model. The simple model allows a probabilistic representation of our knowledge on the location of A_i^* and a way to statistically combine the deductions from multiple comparisons. It also allows us to develop an adaptive search algorithm (Section IV-D) that significantly reduces the number of comparisons needed for identifying A_i^*. In addition, an estimate on the confidence of the result provides a consistent stopping condition for our algorithm. We evaluate the effect of our simplifications using Monte Carlo simulations in Section V.

Our simplified parametric model on $ROD(A_i^*)$ is derived with the following assumptions.

Assumption 1: $CND(A_i)$ and $JND(A_i)$ are constant in $ROD(A_i^*)$. Further, p_0 is linear with respect to $B - A$.

We know intuitively and from subjective experiments that $JND(A_i)$ and $CND(A_i)$ depend on A_i and can vary in $ROD(A_i^*)$. However, the task of estimating a continuous function is as hard as estimating the optimum itself. For tractability, we make a simplification that $JND(A_i)$ and $CND(A_i)$ do not change with A_i.

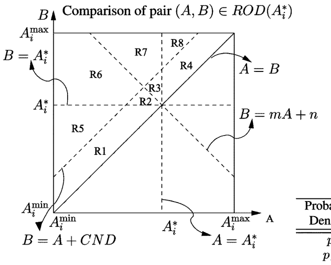
Assumption 2: The boundary line representing subjectively symmetric pairs, $A\|_s B$, is a straight line of the form $B = mA + n$ on the A-B plane, where $m = (-\gamma/\Delta - \gamma)$ and $n = (\Delta/\Delta - \gamma)A_i^*$.

This approximation is justified because the preferred trade-offs among objective metrics are slowly changing around a point. Hence, it is reasonable within the ROD of a local optimum.

Assumption 3: For tractability in derivations, we specify the parameters m and n of the $A\|_s B$ line as a probability distribution. By symmetry, $A_i^*\|_s A_i^*$; thus, (A_i^*, A_i^*) is on the $A\|_s B$ line. It suffices to specify another point on the $A\|_s B$ line to uniquely identify it. Since control symmetry is defined for points that satisfy $A < A_i^* < B$, the line has to pass through $B - A = \Delta$ (where $\Delta > 0$) between $(A_i^* - \Delta, A_i^*)$ and $(A_i^*, A_i^* + \Delta)$. For simplicity, we assume that the cross-over point is uniformly distributed on this line segment. The cross-over point is represented by $(A_i^* - \Delta + \gamma, A_i^* + \gamma)$, where γ is a random variable uniformly distributed in $[0, \Delta]$.

This assumption results in a piecewise linear likelihood function derived later to represent the information learned on the location of A_i^*.

Assumption 4: In the general model, A is more preferred than B $(p_1(A, B) > p_{-1}(A, B))$ if $A < A_i^* < B$ and $B > B'$ where $A\|_s B'$. In the simplified model, we assume that $p_{-1}(A, B) = 0$ when deducing the likely direction of A_i^*



(a) The simplified parametric model

Probability Densities	Regions in $ROD(A_i^*)$							
	R1	R2	R3	R4	R5	R6	R7	R8
p_0	>0	>0	>0	>0	0	0	0	0
p_{-1}	>0	>0	0	0	>0	>0	0	0
p_1	0	0	>0	>0	0	0	>0	>0
p_2	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0	≥ 0

(b) COD for the simplified model (assuming $B > A$)

Fig. 9. Eight regions corresponding to different pairwise comparisons on the A-B plane. The boundary lines separating the regions are similar to those in Table IV, except that the boundary line $A\|_s B$ becomes $B - mA - n$.

after obtaining an $A >_s B$ opinion. Similarly, when $B < B'$ where $A\|_s B'$ or when subjective symmetry with respect to A does not exist, we assume $p_1 = 0$ and use this property in our derivations when an $A <_s B$ opinion is obtained.

Our model describes the probabilities of occurrence of the four possible opinions when comparing two operating points in $ROD(A_i^*)$. For a constant $CND(A_i)$ independent of A_i, Fig. 9 depicts the 2-D model and the eight regions with respect to the four boundary lines.

Bayesian Formulation. We assume an estimate of the ROD of a local optimum in which we are fairly certain that a local optimum exists. Since there may be evidence to suggest multiple local optima, we apply the following procedure to each ROD individually.

The information deduced on the location of A_i^* can be represented by a *belief function*, which is a probability density function (PDF) defined over the set of operating points in $ROD(A_i^*)$. It is denoted by $f_{A_i^*}(a)$ 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 PDF by its probability mass function and integration by summation.

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

$$f_{A_i^*}^0(a) = 1; \quad a \in \left[A^{\min}{}_i, A^{\max}{}_i\right]. \tag{6}$$

Deductions from a single pairwise comparison. Based on the distribution of the opinions obtained by comparing A and B, we can improve our knowledge on the location of A_i^* by a Bayesian formulation. The analysis allows us to obtain the posterior probability from the prior probability and the new evidence:

$$\begin{aligned} f_{A_i^*}(a|COD(A,B)) &= \overline{p}) \\ &= \frac{L(a|COD(A,B) = \overline{p}) \times f_{A_i^*}(a)}{\int_0^1 L(\eta|COD(A,B) = \overline{p}) \times f_{A_i^*}(\eta) d\eta}. \end{aligned} \tag{7}$$

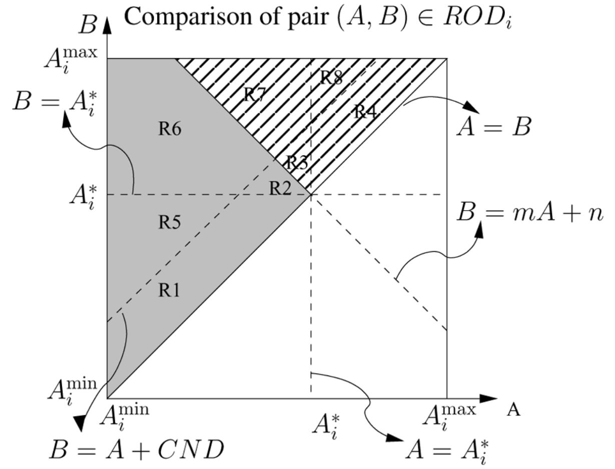


Fig. 10. Deductions on which regions the (A, B) pair may be located, based on the $A >_s B$ and $A <_s B$ responses. Cross-shaded regions indicate that the subject perceives A_i^* to be in $[A + \gamma, A^{\max}{}_i]$. Solid-shaded regions indicate that the subject perceives A_i^* to be in $[A^{\min}{}_i, A + \gamma]$.

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

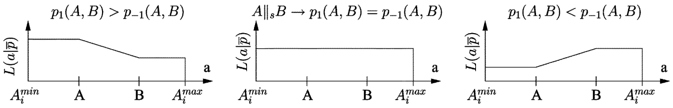
Based on Assumptions 2 and 3, the $A\|_s B$ line satisfies $B = mA + n = (-\gamma/\Delta - \gamma)A + (\Delta/\Delta - \gamma)A^*$, where $B - A = \Delta$ and γ is uniform in $[0, \Delta]$. Next, we analyze the deductions of the four responses.

a) Implications of $A >_s B$. If a subject prefers A over B, this means $A_i^* \notin [A + \gamma, A^{\max}{}_i]$, since $p_1 = 0$ in Regions 1, 2, 5, and 6 [Fig. 9(b)]. Thus, $A_i^* \in \left[A^{\min}{}_i, A + \gamma\right]$ (solid-shaded regions in Fig. 10).
b) Implications of $A <_s B$. If a subject prefers B over A, this means $A_i^* \notin \left[A^{\min}{}_i, A + \gamma\right]$, since $p_{-1} = 0$ in Regions 3, 4, 7, and 8 [Fig. 9(b)]. Thus, $A_i^* \in [A + \gamma, A^{\max}{}_i]$ (cross-shaded regions).
c) Implications of $A \approx_s B$. If a subject indicates that A is about the same as B, A_i^* can be in any of Regions 1, 2, 3, and 4. This does not provide any information on the location of A_i^*, and $A_i^* \in \left[A^{\min}{}_i, A^{\max}{}_i\right]$.
d) Implications of $A?_s B$. If a subject indicates that A is incomparable to B, A_i^* can be in any of the eight regions. This does not provide any information on the location of A_i^*, and $A_i^* \in \left[A^{\min}{}_i, A^{\max}{}_i\right]$.

The *likelihood function* $L(a|\overline{p})$ is a function of $a \in [A^{\min}{}_i, A^{\max}{}_i]$ and indicates the likelihood of obtaining $\overline{p}$ as a result of subjective comparison of A and B if $A_i^* = a$. Using Axiom 2, the likelihood of a to be the optimum can be evaluated by the occurrence frequencies of the four outcomes analyzed above. Conditioned on the value of γ and the result of the subjective comparison, we can represent the likelihood function as

$$L(a|\overline{p}, \gamma) = \begin{cases} p_1 + p_0 + p_2, & \text{if } A^{\min}{}_i < a < A + \gamma \\ p_{-1} + p_0 + p_2, & \text{if } A + \gamma < a < A^{\max}{}_i. \end{cases} \tag{8}$$

Since γ is uniformly distributed over $[0, \Delta]$, where $\Delta = B - A$, the expectation taken over γ results in a likelihood function



(a) A is more preferred than B (b) A and B are preferred equally (c) B is more preferred than A

Fig. 11. Likelihood functions based on the subjective comparison of A and B.

that is only conditioned on $COD(A,B) = \overline{p}$, the result of the subjective evaluation. $L(a|\overline{p})$ is defined as

$$\begin{aligned} &L(a|\overline{p}) \\ &\quad = E_\gamma[L(a|\overline{p},\gamma)] = \int_0^\Delta L(a|\overline{p},\gamma)Pr(\gamma)d\gamma \\ &\quad = \begin{cases} p_0 + p_2 + p_1, & \text{if } A^{\min}{}_i < a < A \\ p_0 + p_2 + \frac{p_1(B-a)+p_{-1}(a-A)}{B-A}, & \text{if } A \le a \le B \\ p_0 + p_2 + p_{-1}, & \text{if } B < a < A^{\max}{}_i. \end{cases} \end{aligned} \tag{9}$$

Fig. 11 depicts the three possible cases of the likelihood function defined in (9) for 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_i^*.

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

$$\begin{aligned} f_{A_i^*}^n(a) &= \frac{f_{A_i^*}^{n-1}(a) \times L(a|COD(A_n,B_n) = \overline{p})}{\int_{A^{\min}{}_i}^{A^{\max}{}_i} f_{A_i^*}^{n-1}(\eta) \times L(\eta|COD(A_n,B_n) = \overline{p})d\eta} \\ &= \frac{\prod_{i=1}^{n} L(a|COD(A_n,B_n) = \overline{p})}{\int_{A^{\min}{}_i}^{A^{\max}{}_i} \prod_{i=1}^{n} L(\eta|COD(A_n,B_n) = \overline{p})d\eta}. \end{aligned} \tag{10}$$

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 written in a closed form as is shown in (10).

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

$$\hat{A}_i^*(f) = \arg\max_a \left\{ \int_{a-JND/2}^{a+JND/2} f(\xi)d\xi \right\}. \tag{11}$$

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

$$U(f) = Pr(|\hat{A}_i^* - A_i^*| \le JND) = \int_{\hat{A}_i^*-JND/2}^{\hat{A}_i^*+JND/2} f(\xi)d\xi. \tag{12}$$

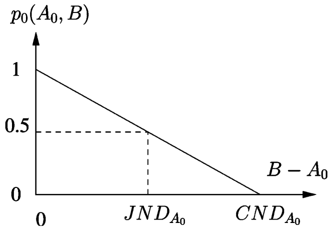


Fig. 12. p_0 decreases linearly as a function of $B - A$.

However, since JND can only be estimated using the results of the comparisons already conducted, its estimation has an inherent error. The calculation of this error and the confidence bounds are discussed later in this section. Using the confidence bounds on the estimated JND (which increases as a function of the number of tests conducted), we calculate the confidence bounds for $\hat{A}_i^*$.

Estimation of JND and CND. As is discussed above, JND is needed in (12) when evaluating the confidence of $\hat{A}_i^*$. The estimated CND is also used in Step 3 when choosing the next pair of points to be compared. To this end, we use the p_0 and $B - A$ values of those previously conducted subjective results for estimating CND and JND. Fig. 12 depicts the linear relation between p_0 and $B - A$, according to Assumption 1 of our simplified model. The model further assumes that JND and CND (which is twice of JND) are constant and do not vary with A:

$$p_0 = 1 - \frac{B-A}{CND}. \tag{13}$$

For K subjects in the evaluation, let $\hat{p_0}$ be the empirical distribution of a binomial random variable with p_0:

$$\Pr(\hat{p_0} = x) = \binom{K}{Kx}(p_0)^{Kx}(1-p_0)^{K(1-x)}. \tag{14}$$

Given the model that assumes a linear function of p_0 with respect to $B - A$ and that passes through $(B - A, p_0) = (0,1)$ and $(CND, 0)$, CND is the only unknown to uniquely specify the line. Based on a single comparison, using the empirical $\hat{p_0}$ value and $B - A$, the estimated CND is

$$\hat{CND} = \frac{B-A}{1-\hat{p_0}}. \tag{15}$$

The error in $\hat{CND}$ is due to variations in the empirical distribution $\hat{p_0}$ around the actual p_0. It can be calculated using the binomial distribution and the likelihood that $\hat{p_0}$ can be obtained when p_0 is equal to a particular value:

$$L(p_0 = y|\hat{p_0} = x) = \binom{K}{Kx} y^{Kx}(1-y)^{K(1-x)}. \tag{16}$$

Furthermore, using (15), the likelihood can be defined for cases when CND is equal to a particular value:

$$\begin{aligned} &L(CND = z \mid \hat{p_0} = x, B - A = \Delta) \\ &\qquad = \frac{\Delta}{z^2}\binom{K}{Kx}\left(1-\frac{\Delta}{z}\right)^{Kx}\left(\frac{\Delta}{z}\right)^{K(1-x)}. \end{aligned} \tag{17}$$

Similar to the estimation of the belief function in Section IV-B, $\hat{CND}$ can be obtained by a Bayesian formulation. Since the actual p_0 needs to be in $[0, 1]$, we normalize the belief function as (18) at the bottom of the page.

The distribution of $\hat{CND}$ can then be calculated using (15), whose confidence improves with the number of comparisons conducted. Once the distribution is obtained, the 90th percentile confidence intervals are calculated and used in estimating A_i^* and its utility in (12).

Initially, no evidence suggests that there will be more then one local optimum. Since the optimum is equally likely to be anywhere on the operating curve, we choose the initial $\hat{A}_i^*$ to be 0.5. Further, the initial $\hat{CND}$ is arbitrarily chosen to be 0.1.

C. Step 1: Estimating the ROD of an Unknown Local Optimum

In this section, we develop a method for estimating the boundaries of the ROD for one or more local optima. An *evidence* of comparing A and B is denoted by e and consists of the tuple $(A, B, COD(A, B))$. An *evidence set* is denoted by $\mathcal{E}$ and is a collection of evidences obtained by subjective evaluations. The *complete evidence set*, containing the results of all the past n comparisons conducted so far, is denoted by $\mathcal{E}_{all}$; see (19) at the bottom of the page.

As is discussed above, there may be multiple local optima on an operating curve, where each local optimum dominates over its corresponding ROD. These multiple RODs on an operating curve, if they exist, do not overlap with each other (Lemma 5). Further, only evidence for which both comparison points are within the ROD of a local optimum provides a reliable direction on the location of that local optimum (Section III-G).

In case of multiple local optima on an operating curve, the result obtained by comparing a pair of points in one ROD cannot be combined with that of comparing a pair in another ROD. As a result, when comparing points on an operating curve with multiple local optima, some of the evidences would give conflicting (or inconsistent) directions on the location of the local optima. If this situation cannot be explained by the noise in the finite number of subjective tests, then it indicates the existence of multiple local optima (and multiple ROD regions), where one evidence belongs to one ROD and another to the other ROD.

Let $\mathcal{E}_i \subseteq \mathcal{E}_{all}$ be the subset of evidences that correspond to $ROD(A_i^*)$. Based on the possibly inconsistent evidences found, we discriminate them into different subsets that correspond to different local-optimum candidates.

Definition 8. Inconsistent Evidence: For $\hat{A}_i^* < A$, an evidence is inconsistent if the hypothesis $\{H_0 : p_1 \geq p_{-1}\}$ can be rejected with some statistical significance. Similarly, for $B < \hat{A}_i^*$, an evidence is inconsistent if the hypothesis $\{H_0 : p_{-1} \geq p_1\}$ can be rejected with the same statistical significance.

When two sets of evidences are inconsistent, it means that each is pointing to a different local optimum. As a result, the operating curve should be divided into two RODs, each corresponding to one local optimum.

Procedure for identifying multiple RODs on an operating curve. This consists of three steps.

a) Initially, all evidences that are mutually consistent with other evidences in the set are used to determine an ROD. Although this step provides a superset of the actual ROD region that can potentially overlap with each other, it ensures that the RODs are not over-pruned due to noisy evidences. This step is described in detail as follows.

Initial condition. After the first comparison, since it has no inconsistent evidence, we assume that there is one local optimum, and its ROD is the entire operating curve. As subsequent comparisons are done, we determine whether the new evidence on the current estimate of the local optimum is consistent with existing evidences.

Existence of multiple ROD regions. If the new evidence is found to be inconsistent, a new set of evidences is formed, say $\mathcal{E}_2$, that corresponds to a new local-optimum candidate. The new evidence will be taken from $\mathcal{E}_1$ and placed in $\mathcal{E}_2$. Further, all evidences that are consistent with the new evidence will be duplicated from existing sets to $\mathcal{E}_2$. The procedure will be repeated for all existing sets until each set has evidences that are mutually consistent with each other. The procedure results in the largest set of mutually consistent evidences in each set.

Initial ROD estimation. Next, we map each evidence set to its corresponding ROD. We identify the minimum and the maximum of the points compared in each evidence set in order to determine its bounds:

$$\begin{aligned} A_i^{\hat{\min}} &= \min\{A_j \mid (A_j, \bullet, \bullet) \in \mathcal{E}_i\} \\ A_i^{\hat{\max}} &= \max\{B_j \mid (\bullet, B_j, \bullet) \in \mathcal{E}_i\}. \end{aligned} \tag{20}$$

We repeat the estimation of the ROD for each local-optimum candidate. At this point, the RODs estimated may overlap, since some evidences can be members of multiple sets. As RODs do not overlap (Lemma 5), it is necessary to update the initial RODs estimated in order to ar-

$$\Pr{}^{post}(CND = z) = \frac{\Pr^{prior}(CND = z) \times L(CND = z|\hat{p_0} = x, B - A = \Delta)}{\int_0^1 \Pr^{prior}(CND = \eta) \times L(CND = \eta|\hat{p_0} = x, B - A = \Delta) d\eta} \tag{18}$$

$$\mathcal{E}_{all} = \left\{ \underbrace{(A_1, B_1, COD(A_1, B_1))}_{e_1}, \underbrace{(A_2, B_2, COD(A_2, B_2))}_{e_2}, \ldots, \underbrace{(A_n, B_n, COD(A_n, B_n))}_{e_n} \right\} \tag{19}$$

rive to non-overlapping RODs. Although we can simply construct evidence sets that do not overlap, this condition is not sufficient. For example, one of the evidences in the first set can have one of its points compared in the ROD of the second set, which causes the two RODs to overlap.

b) In the second step, the local optima are estimated based on the initial RODs found. It uses the set of evidences for which both points compared are within a single ROD to obtain a belief function and a corresponding estimate of a local optimum. To obtain a subset of the initial ROD estimates that do not overlap, we first estimate the belief functions of different local-optimum candidates individually over their possibly overlapping RODs. We then estimate the local optimum using the procedure in Section IV-B.

c) Lastly, the RODs are refined again to ensure that each is a contiguous region with a local optimum and that they do not overlap with each other.

Updated estimation of RODs. Starting from the local optimum estimate, the corresponding ROD is a contiguous region until an inconsistent evidence is found. This step ensures that the RODs are non-overlapping. In case there is no inconsistent evidence, the entire operating curve is taken to be a single ROD:

$$\begin{aligned}\hat{A}_i^{\min} &= \max\Big\{ \min\{A_j \mid (A_j, \bullet, \bullet) \in \mathcal{E}_i\} \\ &\qquad \max\{B_j \mid (\bullet, B_j, \bullet) \notin \mathcal{E}_i\}\Big\} \\ \hat{A}_i^{\max} &= \min\Big\{ \max\{B_j \mid (\bullet, B_j, \bullet) \in \mathcal{E}_i\} \\ &\qquad \min\{A_j \mid (A_j, \bullet, \bullet) \notin \mathcal{E}_i\}\Big\}. \end{aligned} \tag{21}$$

Eliminating noisy evidence and merging RODs. Due to noise in the subjective evaluations, it is possible to have inversions of preference directions in some comparisons (such as $p_1 > p_{-1}$ instead of $p_1 < p_{-1}$). This may cause the locations of the inconsistent evidences to interleave with each other and result in overlapping RODs. In this case, most of the evidences would point to one direction, whereas a few in the same vicinity would point to another. Eliminating such noisy evidences requires merging the divided ROD regions into one contiguous region. This task can be achieved by increasing the statistical significance level when identifying inconsistent evidence pairs.

Once the RODs are estimated, the information is passed to Step 2 (Section IV-B), and a local optimum is identified in one of the RODs.

D. Step 3: Identifying the Next Pair of Points to be Compared

Based on the procedures in Sections IV-B and C, we present our search algorithm for choosing the next pair of points to be compared. Our goal is to minimize the number of comparisons before A_i^* is identified with high confidence. We first describe our observations on the optimal sequence of comparisons and reduce the problem to choosing the optimal pair in each step. We then derive the optimal pair of points to be compared in the next step.

Sequence of comparisons. As more pairwise evaluations are conducted, the combined belief function evolves from uniform to a shape that is centered around $\hat{A}_i^*$. Since it is not feasible to exactly identify A_i^* in a continuous search space, we stop the search once a certain level of confidence is reached. The confidence level chosen will affect the efficiency of the algorithm and the accuracy of the result.

At the beginning of the nth comparison, when given the utility $U(f^{n-1})$, the expected number of comparisons left to reach the stopping condition if the optimal pair is chosen in the nth test is denoted by

$$\begin{aligned} S(U(f^{n-1})) &= 1 + S(U(f^n)) \\ &= 1 + \min_{A_n, B_n} S(U(f^n \mid A_n, B_n)). \end{aligned} \tag{22}$$

The following are the arguments leading to the evaluation of $\min_{A_n,B_n} S(U(f^n \mid A_n, B_n))$. For any A and B, $L(a \mid A, B)$ is uni-modal. Let $mode(L)$ be the set of points satisfying the modality. It is clear that $A_i^* \in mode(L(a \mid A, B))$. Since any comparison conducted over the same ROD is consistent and A_i^* is common to all the comparisons, it is clear that $mode(L(a \mid A_1, B_1)) \cap mode(L(a \mid A_2, B_2)) \neq \emptyset$. For any sequence of A-B pairs, $f^n(a)$ is uni-modal and $A_i^* \in mode(f^n)$. Hence, $U(f^n)$ is monotonically non-decreasing with respect to n for any sequence of comparisons, 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 of the current belief function.

Individual comparisons. Based on Section IV-B, when both points compared are in an ROD of a local-optimum candidate, their subjective comparison would provide a correct direction on the location of A_i^* with respect to the points compared. However, if one or both points are not in the same ROD or they are in RODs of different local optima, then we cannot guarantee that a correct direction can be found. The comparisons that do not point to the intended local optimum introduce inconsistencies and reduce the confidence of the A_i^* estimated. As is described in Section IV-C, such inconsistencies are eliminated from the set of evidences during the estimation of a particular local optimum. Of course, such comparisons are wasted and do not improve our knowledge on A_i^*. In short, based on the updated estimation of each ROD, we should identify points to be compared that are in the same ROD where the local optimum is to be located.

The following are the observations for identifying the optimal pair of points to be compared next.

1) Indistinguishable and incomparable opinions do not lead to any deductions on A_i^*. Further, when $p_1 = p_{-1}$, the two points compared are subjectively symmetric and do not lead to new information on A_i^*. Hence, for a comparison to be useful, p_0 and p_2 should be small and the difference between p_1 and p_{-1} should be large. Depending on which of p_1 or p_{-1} is larger, the likely direction of the search can then be determined.

2) Due to Axiom 4, p_0 is monotonically non-increasing with $B - A$ and reaches 0 at $B - A = CND(A)$. Hence, choosing points that are very close to each other does not provide any evidence on A_i^* because p_0 is high, and thus, the difference between p_1 and p_{-1} is low. In contrast, choosing A far away from B reduces p_0, although choosing them beyond $B - A = CND$ does not reduce p_0 further (since it is already equal to zero).

TABLE V
COD OF THE SUBJECTIVE COMPARISONS CONDUCTED ON PAIRS OF POINTS ON THE OPERATING CURVE IN FIG. 2

Comparisons	A				B				COD(A,B)			
Made	MED	PESQ	CS	CE	MED	PESQ	CS	CE	p_{-1}	p_0	p_1	p_2
1	250	4.18	1.60	0.91	500	4.18	2.21	0.84	0.125	0.500	0.375	0.000
2	110	2.39	1.27	0.96	180	4.18	1.44	0.93	0.750	0.250	0.000	0.000
3	215	4.18	1.52	0.92	539	4.18	2.30	0.82	0.125	0.625	0.250	0.000
4	111	2.60	1.27	0.96	198	4.18	1.48	0.93	0.750	0.125	0.125	0.000

3) Due to Axiom 5, $p_2 = 0$ at $B - A = 0$ and is monotonically non-decreasing with respect to $B - A$ and reaches its maximum at $B - A = A^{\max} - A^{\min}$. Thus, choosing two points that are far apart increases p_2, which indirectly reduces the difference between p_1 and p_{-1} and the conclusiveness of the comparison.

4) Given an unknown A_i^* and any pair A and B, p_1 and p_{-1} are uncorrelated. Thus, maximizing $p_1 + p_{-1}$ (which minimizes $p_0 + p_2$) is equivalent to maximizing the expected value of $|p_1 - p_{-1}|$. Note that $\arg\min\{p_0 + p_2\}$ is achieved at $B - A = CND$.

5) The utility is maximized when the disparity between the two horizontal levels of the likelihood function in Fig. 11 [or the first and last cases in (9) and represented by $|p_1 - p_{-1}|$] is maximized. Due to Axiom 6, the difference between p_1 and p_{-1} increases when one of the points is close to or equal to A_i^*.

6) Thus, given $\hat{A}_i^*$ and $C\hat{N}D$ after the nth comparison, the optimal choice of the $n+1$st comparison should include $\hat{A}_i^*$ as one of the points and the other CND away from it in either direction on the operating curve:

$$(A_{n+1}, B_{n+1}) = \begin{cases} (\hat{A}_i^* - C\hat{N}D, \hat{A}_i^*), & \text{if } n \text{ is odd} \\ (\hat{A}_i^*, \hat{A}_i^* + C\hat{N}D), & \text{if } n \text{ is even.} \end{cases} \quad (23)$$

Note that since A and $B \in \mathcal{O}_i$, the selection made by (23) needs to be augmented to keep A_{n+1} and B_{n+1} within their corresponding RODs.

As more comparisons are conducted, the estimated local optimum $\hat{A}_i^*$ improves, which increases the disparity between p_1 and p_{-1} (Axiom 6). Since the region with a higher likelihood contains $\hat{A}_i^*$, the corresponding utility increases as well. Note that utility improves at a rate related to n.

Example 1 (Cont'd): Based on the method in this section, we have conducted a sequence of subjective comparisons between pairs of points on the operating curve in Fig. 2(b). Table V shows the objective metrics of each pair of points compared and the COD of the subjective comparisons on the four comparisons made.

As described above, the belief function, $\hat{A}^*$, and $C\hat{N}D$ are updated based on the latest result after each comparison. Initially, A^* is equally likely to be anywhere on the operating curve. Thus, the initial $\hat{A}^*$ is 500 ms (or 0.5), and the initial $C\hat{N}D$ is 0.25. Since there is no inconsistent evidence, ROD is equal to the entire operating curve, which happens to be valid throughout the comparisons for this example. After the first comparison, the belief function (Fig. 13) indicates that A^* is more likely to be less than 250 ms.

However, for this operating curve, any MED less than 110 ms results in PESQ less than 2.0. As our previous experience shows that such conversations are not likely to be preferred against conversations with higher PESQ, we prune the operating curve below 110 ms in order to avoid unnecessary comparisons with conversations of inferior quality. Thus, after the first comparison, our $\hat{A}^*$ is 180 ms (midway between 110 ms and 250 ms).

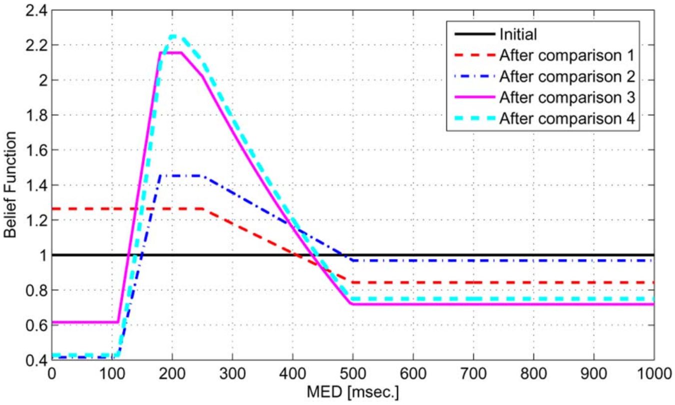


Fig. 13. Initial belief function and its evolution after each subjective comparison.

The second comparison selects $A = 0.110$ and $B = 0.180$, based on (23), since any A less than 0.110 would not provide information on the optimal point. The result indicates that $B = 0.180$ is strongly preferred against A. The updated belief function (Fig. 13) leads to the selection of the third pair compared as $(A, B) = (0.215, 0.539)$, since $\hat{A}^* = 0.215$ (midway between 0.180 and 0.250) and $C\hat{N}D = 0.324$.

In the third comparison, subjects slightly prefer the 215-ms alternative against the 539-ms alternative, since there is no difference in speech quality and the difference in degradation due to delay is not very perceptible (with a low switching frequency on this conversation).

Based on the previous results, the fourth comparison is selected to be (0.111, 0.198), and subjects prefer 198-ms alternative significantly against the 111-ms alternative due to significant improvement in LOSQ.

After four comparisons, $\hat{A}^*$ is 208 ms and the utility is 64%. The operating point identified has very high LOSQ and little delay degradations. We observe in Fig. 13 that the belief function evolves with each comparison and centers around $\hat{A}^*$.

E. Batch-Parallel Evaluations

For batch-parallel evaluations $(M > 1)$, 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 pairs compared in the next batch, based on the current belief function. We identify $M - 1$ equally spaced points $\{C^j, j = 1, \ldots, M-1\}$ in the search space for which one of them is $\hat{A}_i^*$:

$$C^j = \text{mod}\left(\frac{j-1}{M-1} + \hat{A}_i^*, 1\right), \; j = 1, \ldots, M-1. \quad (24)$$

For equal spacing, points are wrapped around the operating curve via the modulo operation. We conduct two comparisons involving $\hat{A}_i^*$, with points CND away from it in either direction, which correspond to the optimal pair for the even and odd cases in (23). For each of the remaining $M-2$ points identified, it is compared with a point CND away in the direction opposite to that of $\hat{A}_i^*$:

$$(A_n^j, B_n^j) = \begin{cases} (C^j - C\hat{N}D, C^j), & \text{if } C^j < \hat{A}_i^* \\ (C^j, C^j + C\hat{N}D), & \text{if } C^j > \hat{A}_i^* \\ \text{Both pairs above}, & \text{if } C^j = \hat{A}_i^*. \end{cases} \quad (25)$$

F. Putting Everything Together: Conducting Subjective Evaluations of Multiple Operating Curves

Recall from Section I that the goal of our subjective tests is to identify the most preferred point for each of a set of operating curves that model a comprehensive set of operating conditions in a multimedia system. The results of the subjective tests lead to a mapping between the objective metrics representing the two alternatives on an operating curve and the pair-wise subjective preference among those alternatives. We then learn these mappings for the multitude of operating curves using a pair-wise-preference SVM classifier. We utilize this classifier and the Bayesian formulation described above to combine individual preferences in order to identify the appropriate control at run time in response to an unseen operating condition.

Also recall from Section IV-A that sequential evaluations of a single operating curve are the most effective in terms of minimizing the number of tests performed for that curve, when identifying a local optimum to within some statistical confidence. However, they are inconvenient because subjects have to synchronize their test results with each other in order to estimate the local-optimum candidate before the next test can be carried out.

Based on these observations, the optimal strategy to minimize the total number of subjective tests for a set of operating curves is to test each curve sequentially and all the curves in parallel. In this approach, each subject is presented with a set of operating points to be compared, one from each operating curve to be tested. The tests in each set can be performed in any order and independent of other subjects because the result of comparisons from one operating curve does not depend on that of another curve. At the end of the tests, the results from all the subjects are combined in order to generate a local optimum estimate and identify the next pair of operating points to be compared for each of the operating curves. As the number of operating curves to be tested is large, this approach allows subjects to independently carry out a batch of independent tests, without having to synchronize their results in a locked-step fashion with other subjects. The number of iterations is bounded by the typically small number of iterations to identify a local-optimum candidate of an operating curve.

V. Performance Analysis by Monte Carlo Simulations

In evaluating our approach, we use Monte Carlo simulations to generate the probabilities of the four opinions in Table I for our general model in Fig. 7. Since these probabilities are functions of A and B, each will exist as a surface in the A-B plane. We then apply our search algorithm in Section IV, which is initialized by $\hat{A}_i^* = 0.5$ and $C\hat{N}D = 0.1$ (Section IV-B). Based on the A and B selected and a multinomial distribution, we generate the corresponding sample $COD(A, B)$ for K subjects. We then update the estimates of the ROD and the local optimum using our Bayesian procedures and repeat the search until a local optimum is found. By verifying the accuracy of our estimate with respect to the local optimum in the reference general model, we can verify the robustness of our simplified parametric model used in deriving the algorithm.

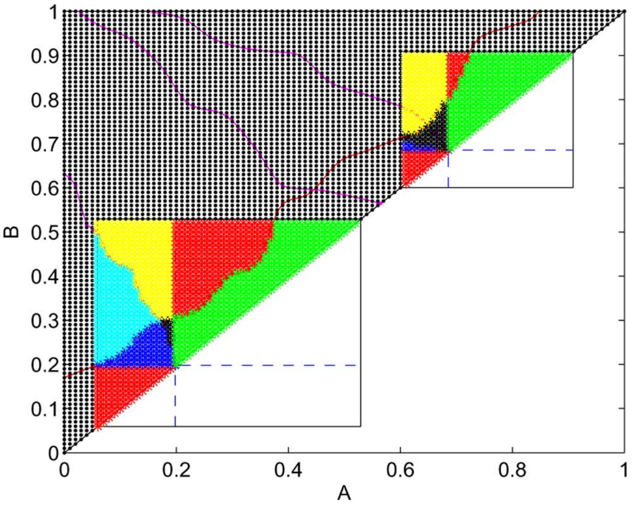


Fig. 14. Example of the regions for two local optima on an operating curve.

Since the generation of the general model in Fig. 7(a) is rather involved, we summarize its details as follows. Given the number of local optima on an operating curve, we first randomly determine the boundaries of each ROD and the position of the local optimum in it. We then generate the CND line as a continuous random walk around a given average CND value. Similarly, we generate the subjective symmetry line as a continuous random walk, when given the standard deviation of the subjective symmetry line with respect to a straight line. Finally, we generate the p_i values for a finite number of A-B pairs, specifically, 100 steps in A and 100 steps in B, and using cubic interpolations in between. In particular, p_2 is monotonically non-decreasing with $B-A$ (${p_2}^{\max}$ at $B-A = A^{\max} - A^{\min}$ to 0 at $B = A$); and p_0 is monotonically non-increasing with $B-A$ [1 at $B = A$ to 0 at $B - A = CND(A)$]. Since $p_1 > p_{-1}$ when A and B are on the same side of A_i^* and within $ROD(A_i^*)$, we set p_1 to be proportional to $|B-A_i^*|/(|B-A_i^*|+|A-A_i^*|)$ and p_{-1} proportional to $|A-A_i^*|/(|B-A_i^*|+|A-A_i^*|)$. We also normalize p_1 and p_{-1} so that their sum is $1-(p_0+p_2)$. A similar procedure is applied when A and B are on different sides of A_i^*. Fig. 14 illustrates the boundary lines with two local optima generated.

We compare our method with two other procedures, one randomly choosing the next pair on an operating curve and the other based on the fully parallel approach that chooses $N(N-1)/2$ pairs of points (Section IV-A).

Assuming a known JND, Table VI compares the average performance of the four algorithms. It shows that conducting fully parallel evaluations and random comparisons are very expensive, and that choosing pairs sequentially based on our procedure reduces the number of comparisons by five folds for simpler problems ($JND = 0.1$) and 1000 folds for harder problems ($JND = 0.003$).

TABLE VI

EXPECTED NUMBER OF COMPARISONS FOR AN OPERATING CURVE WITH A SINGLE LOCAL OPTIMUM WHEN JND IS KNOWN. ALL COMPARISONS RESULT IN A SUCCESSFUL ESTIMATION OF A^* (TO WITHIN THE JND OF THE ACTUAL VALUE)

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

TABLE VII

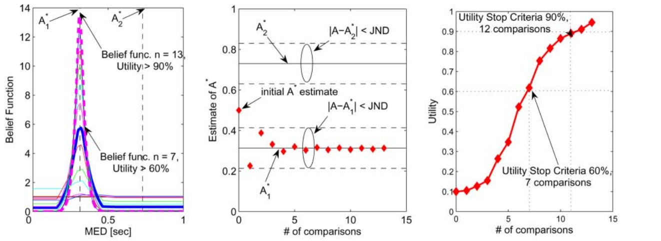
SEQUENTIAL SCHEME: EXPECTED NUMBER OF COMPARISONS AND THE PERCENT OF SUCCESSFUL ESTIMATION FOR SINGLE, DOUBLE, AND TRIPLE LOCAL OPTIMA, WHERE JND IS UNKNOWN AND ESTIMATED AFTER EACH COMPARISON

JND	Utility Stop %	Single Optimum		2 Optima		3 Optima	
		Acc %	E[n]	Acc %	E[n]	Acc %	E[n]
0.1	15	85	2.7	80	2.8	90	2.1
	20	95	3.9	90	3.8	90	3.5
	30	95	5.7	90	5.5	100	6.2
	40	95	7.4	95	7.1	100	8.3
	50	100	8.7	100	8.9	100	10.6
	60	100	10.3	100	9.9	100	12.4
	90	100	17.5	100	18.8	100	20.4
0.03	15	75	5.8	75	5.8	45	3.3
	20	95	7.4	85	7.6	85	6.1
	30	95	8.5	85	8.6	100	8.2
	40	95	9.7	90	10.3	100	9.8
	50	95	10.5	95	12.2	100	11.1
	60	95	11.6	100	13.9	100	12.7
	90	100	16.7	100	19.1	100	20.7

Also shown are the results of batch-parallel comparisons, which give the trade-offs between the number of batches and the number of tests in each. For instance, with $JND = 0.03$, it takes an average of 4.9 batches, each with $M = 4$ tests, in a batch-parallel algorithm, as compared to an average of 9.9 batches, each with one test, in a sequential algorithm. Hence, tests should be designed to balance the overhead of synchronizing test results in each batch and the benefit of sequential algorithms that minimize the number of comparisons.

Table VII summarizes the performance of our scheme for operating curves with single and multiple local optima. It also shows the trade-off between the expected number of comparisons and the accuracy of estimating the local optima (Acc %), using a stopping criterion defined by the utility (Utility Stop %) in (12). For all the cases studied, it suffices to stop the search when the utility reaches 50%, which leads to at least 95% success rate in predicting one of the local optima and requires approximately half of the number of comparisons. For those 5% of the cases in which the algorithm fails to find a point within the JND of the local optimum, the estimation errors are very small. In short, there is a substantial reduction in the number of comparisons by using a stopping criterion based on a smaller utility value, while incurring a negligible error in the estimation.

Fig. 15 illustrates a typical application of our sequential algorithm using a simulated model with two local optima. We observe that the belief function focuses around one of the local optima, and the utility of the estimation increases with each comparison. Since our algorithm does not know the number of local optima, their ROD boundaries, and the JND values, it estimates them after each comparison. Fig. 15 also depicts the convergence of the optimum estimate in the simulations. When compared to the results in Table VI, the unavailability of JND causes an increase in the expected number of comparisons needed in order to find an operating point within the JND of a local optimum.



(a) Evolution of belief function

(b) Convergence of the optimum estimated around one local optimum

(c) Improvement of utility (confidence) with number of comparisons

Fig. 15. Illustration of the application of our sequential search algorithm on a simulated model with two local optima.

VI. CONCLUSION

In this paper, we have studied algorithms for the statistical scheduling of offline subjective tests for evaluating alternative control schemes in real-time multimedia applications. Our goal is to identify the most preferred point for each of a set of operating curves that model a comprehensive set of operating conditions in these systems.

Our results show that sequential evaluations of a single operating curve are the most effective in terms of minimizing the number of tests performed for that curve when identifying a local optimum to within some statistical confidence. We have developed in this paper the axioms for characterizing subjective tests of a single operating curve, a general model of subjective tests, and a simplified parametric model for developing efficient search algorithms. Our simulation results show a substantial reduction in the number of comparisons by using a stopping criterion based on a lower confidence level, while incurring a negligible error in the estimation.

Our future work involves applying our search algorithms to identify preferred operating points in two-party [3] and multiparty [8] voice-over-IP applications. The results of the subjective tests lead to a mapping between the operating condition of an operating curve and the associated control that achieves the highest subjective quality for that operating condition. These mappings for the multitude of operating curves are then learned using an SVM classifier, which is used to generate the appropriate control at run time in response to unseen operating conditions.

REFERENCES

[1] ITU-T P-Series Recommendations, International Telecommunication Union. [Online]. Available: http://www.itu.int/rec/T-REC-P/en.

[2] B. Sat and B. W. Wah, "Statistical testing of off-line comparative subjective evaluations for optimizing perceptual conversational quality in VoIP," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2008, pp. 424–431.

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

[4] ITU-T G-Series Recommendations, International Telecommunication Union. [Online]. Available: http://www.itu.int/rec/T-REC-G/en.

[5] C. Boutremans and J.-Y. Le Boudec, "Adaptive joint playout buffer and FEC adjustment for Internet telephony," in *Proc. IEEE INFOCOM*, 2003, vol. 1, pp. 652–662.

[6] L. Sun and E. Ifeachor, "Voice quality prediction models and their applications in VoIP networks," *IEEE Trans. Multimedia*, vol. 8, no. 4, pp. 809–820, Aug. 2006.

[7] B. Sat and B. W. Wah, "Evaluating the conversational voice quality of the Skype, Google-Talk, Windows Live, and Yahoo Messenger VoIP systems," *IEEE Multimedia*, vol. 16, no. 1, pp. 46–58, Jan.–Mar. 2009.

[8] 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. Symp. Multimedia*, Taichung, Taiwan, Dec. 2007, pp. 3–10.

[9] 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. Conf. Multimedia and Expo*, Jul. 2008, pp. 493–496.

[10] D. Wu, Y. T. Hou, and Y.-Q. Zhang, "Transporting real-time video over the Internet: Challenges and approaches," *Proc. IEEE*, vol. 88, no. 12, pp. 1855–1875, Dec. 2000.

Batu Sat (S'01–M'08) received the B.S. degree in electronics and telecommunications engineering from Istanbul Technical University, Istanbul, Turkey, in 2001, and the M.S. degree in electrical engineering from the Electrical and Computer Engineering Department at the University of Illinois at Urbana-Champaign (UIUC), Urbana, IL, in 2003. He is pursuing the Ph.D. degree at UIUC.

His current research interests are in the development of evaluation and design methods of real-time multimedia communication systems.

Mr. Sat is a member of the ACM.

Benjamin W. Wah (SM'85–F'91) received the Ph.D. degree in computer science from the University of California, Berkeley, CA, in 1979.

He is currently the Franklin W. Woeltge Endowed Professor of Electrical and Computer Engineering and Professor of the Coordinated Science Laboratory of the University of Illinois at Urbana-Champaign, Urbana, IL. He also serves as the Director of the Advanced Digital Sciences Center, a large research center of the University of Illinois located in Singapore and funded by Singapore's Agency for Science, Technology, and Research (A*STAR). His current research interests are in the areas of nonlinear search and optimization, multimedia signal processing, and computer networks.

Dr. Wah has received a number of awards, including the University Scholar of the University of Illinois (1989), the IEEE Computer Society Technical Achievement Award (1998), the IEEE Millennium Medal (2000), the Raymond T. Yeh Lifetime Achievement Award from the Society for Design and Process Science (2003), the IEEE Computer Society W. Wallace-McDowell Award (2006), the Pan Wen-Yuan Outstanding Research Award (2006), the IEEE Computer Society Richard E. Merwin Award (2007), the IEEE-CS Technical Committee on Distributed Processing Outstanding Achievement Award (2007), and the IEEE-CS Tsutomu Kanai Award (2008). He cofounded the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING in 1988 and served as its Editor-in-Chief between 1993 and 1996. He has served the IEEE Computer Society in various capacities, including Vice President for Publications (1998 and 1999) and President (2001). He is a Fellow of the AAAS and ACM.