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

Batu Sat and Benjamin W. Wah

12/15/2008

Department of Electrical and Computer Engineering and the Coordinated Science Laboratory
University of Illinois at Urbana-Champaign, USA

Outline

• Introduction
• Approach & Problem Statement
• Model and Deductions of Subjective Comparisons
• Subjective Evaluations Methods
• Strategy for Simultaneous Evaluations
• Experimental Results
General Problem Studied

- Design the operation of control schemes
  - Real-time multi-media communication systems
  - Achieves high perceptual conversational quality
  - Robust to dynamic network conditions & communication scenarios

- Systems with common properties
  - Trade-offs among objective metrics on subjective preferences
  - Constrained resources on best-effort IP network
  - Communication scenario among participants

Subjective Evaluations

- On-line subjective evaluations are infeasible
  - Offline subjective tests are expensive and require multiple subjects

- Absolute Category Rating (e.g. ITU P.800 MOS)
  - Two operating points with multiple quality metrics may not be comparable
  - Not very accurate for small difference or high quality
  - Statistical significance cannot be associated with MOS differences
  - Suitable for verification of system performance
Outline

- Introduction
- Approach & Problem Statement
- Model and Deductions of Subjective Comparisons
- Subjective Evaluations Methods
- Strategy for Simultaneous Evaluations
- Experimental Results

Play-out Scheduling Design

- System control
  - MED: Trade-off between LOSQ and delay degradations
  - Goal: Choose optimal MED at run time – optimal operating point
- Network conditions: non-stationary & connection dependent
- Conversational scenarios
  - Frequency of conversation turns
  - Speech and silence durations
- Multiple quality metrics
  - Operating curve in multi-D space
Our Approach

- Comparative ranking leading to partial order
- Dividing the problem into two stages
  - Identify best operating point off-line given operating curve
  - Learn and generalize from limited number of conditions at run-time
- Simulation & evaluation of results under given conditions
  - Repeatability of subjective tests relating results to control parameter
- Pruning of search space
  - Small changes in objective space may not be subjectively perceptible
  - Systematically use previous subjective preference results to reduce future tests
- Combining of multiple pair-wise comparisons using Bayesian framework
- Learning of a classifier to generalize to similar but unseen conditions

Problem Statement

- Statistical scheduling of off-line comparative subjective tests for evaluating alternative operating points on an operating curve of a control scheme in real-time multimedia systems
- Assumptions:
  - Domain knowledge on problem: identify monotonic quality metrics
  - Region of Dominance (ROD) is known on operating curve
- Not studied in this paper:
  - Multiple operating curves corresponding to different conditions
  - Learning and generalization of a classifier
Outline

- Introduction
- Approach & Problem Statement
- Model and Deductions of Subjective Comparisons
- Subjective Evaluations Methods
- Strategy for Simultaneous Evaluations
- Experimental Results

Model of Subjective Comparisons

Notation:
- Operating curve: $O$, set of feasible points
- $A_{\text{min}}$, $A_{\text{max}}$: two extreme points on $O$
- Comparative Opinion Distribution when comparing $A$ and $B$
  - COD($A,B$) = $(p_1, p_0, p_1, p_2)$

<table>
<thead>
<tr>
<th>Condition</th>
<th>Probability</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A$ is better than $B$</td>
<td>$Pr(A &gt; B)$</td>
<td>$p_1(A,B)$</td>
</tr>
<tr>
<td>$A$ is about the same as $B$</td>
<td>$Pr(A \approx B)$</td>
<td>$p_0(A,B)$</td>
</tr>
<tr>
<td>$A$ is worse than $B$</td>
<td>$Pr(A &lt; B)$</td>
<td>$p_1(A,B)$</td>
</tr>
<tr>
<td>$A$ is incomparable to $B$</td>
<td>$Pr(A \not\approx B)$</td>
<td>$p_2(A,B)$</td>
</tr>
</tbody>
</table>
Model of Subjective Comparisons

2-D representation of comparing A and B
- 4 boundary lines and 8 regions: based on relative location to A*

Axioms of Subjective Comparisons (1/3)

Reflectivity
- Comparing a point with itself: \( p_0(A,A) = 1 \)

Independent and Identically Distributed
- Finite no. of IID samples

Symmetry/anti-symmetry
- Order of comparison does not affect outcome
  - \( p_0(A,B) = p_0(B,A) \)
  - \( p_2(A,B) = p_2(B,A) \)
  - \( p_1(A,B) = p_1(B,A) \)
Axioms of Subjective Comparisons (2/3)

**Just Noticeable Difference (JND)**
- 50% of subjects perceiving a difference in quality with respect to A
- Complete Noticeable Difference (CND)

**Indistinguishability**
- \(|B-A| \leq JND_A\)
- \(p_0(A,B)\) is monotonically non-increasing w.r.t. \(B-A\)

**Locally optimal point**
- \(A^* = \{ A \mid p_1(A,B) > 0.5 \text{ for all } B \text{ in } O \text{ such that } |B-A| > JND_A\}\)
- \(A^*\) is preferred among all points within the ROD, except within JND

Axioms of Subjective Comparisons (3/3)

**Incomparability**
- Perceptible degradations are different between A and B

**Subjective Preference**
- \(|p_1(A,B) - p_1(A,B)|\) increases as point closer to \(A^*\) is perturbed towards \(A^*\)

**Control Symmetry**
- \(|A-A^*| = |B-A^*|\)

**Subjective Symmetry**
- \(p_1(A,B) = p_1(A,B)\)
Deductions on Optimal Alternative

- Simplified parametric model needed
  - Allow information learned on multiple comparisons to be combined
- Belief function
  - Representing knowledge on location of $A^*$
  - $f_{A^*}(a)$, where $a$ on $O$
- Initial Knowledge
  - Assuming uniformly distributed $f_{A^*}^0(a) = 1, \ a \in [A^\min, A^\max]$

Simplified Parametric Model (1/2)

**Assumption 1**: CND and JND are constant and do not vary with respect to $A$ within ROD of local optimum

**Assumption 2**: Boundary line representing subjective symmetric pairs $A || B$ is a straight line
**Simplified Parametric Model (2/2)**

**Assumption 3:**

\[ B = mA + n = \frac{-\gamma}{\Delta - \gamma} A + \frac{\Delta}{\Delta - \gamma} A^* \]

\( m \) and \( n \) are stochastic:

\( A \parallel_B B \) is defined by red point \((A^*, A^*)\) and one of green points

Green point is equally likely on \( B - A = \Delta \) line thus, \( \gamma \) is uniform in \([0, \Delta]\)

**Assumption 4:**

If \( A < A^* < B \) and \( B > \{B \parallel_B A\} \),
then \( p_1(A, B) = 0 \)

If \( B < \{B \parallel_B A\} \) or \( \{B \parallel_B A\} \) does not exist, then \( p_1(A, B) = 0 \)

---

**Deductions on Single Pairwise Comparison**

**Bayesian Formulation**

- Posterior probability from prior probability and new evidence

\[ f_{A^*}(a | \overline{p}) = \frac{L(a)COD(A, B) - \overline{p}) \times f_{A^*}(a)}{\int_{0}^{1} L(\eta)COD(A, B) - \overline{p}) \times f_{A^*}(\eta)d\eta} \]

- Likelihood function \( L(a | p) \) indicates the likelihood of obtaining \( p \) as the result of a subjective comparison of \((A, B)\) if \( A^* = a \)

- Likelihood is obtained using occurrence frequencies of 4 outcomes
  - \( A > s B \) \( \rightarrow \) \((A, B)\) in regions 1, 2, 5 or 6 \( \rightarrow \) any \( a \in [A^{\text{min}}, A + \gamma] \) can be \( A^* \)
  - \( A < s B \) \( \rightarrow \) \((A, B)\) in regions 3, 4, 7 or 8 \( \rightarrow \) any \( a \in [A - \gamma, A^{\text{max}}] \) can be \( A^* \)
  - \( A = s B \) \( \rightarrow \) \((A, B)\) in regions 1, 2, 3 or 4 \( \rightarrow \) No deduction
  - \( A \not\in s B \) \( \rightarrow \) \((A, B)\) in regions 1 through 8 \( \rightarrow \) No deduction
Deductions on Single Pairwise Comparison

- Conditioned on the value of $\gamma$ Likelihood function is
  
  $L(a|\gamma) = \begin{cases} 
  p_1 + p_2 + p_3 & \text{if } A_{\text{min}} < a < A + \gamma \\
  p_2 + p_2 + p_2 & \text{if } A + \gamma < a < A_{\text{max}}.
  \end{cases}$

- Unconditioned likelihood function is
  
  $L(a|\bar{p}) = E_{\gamma}[L(a|\gamma)] = \int_{A_{\text{min}}}^{A_{\text{max}}} L(a|\gamma) Pr(\gamma)d\gamma$

  $= \begin{cases} 
  p_0 + p_2 + p_1 & \text{if } A_{\text{min}} < a < A \\
  p_0 + p_2 + \frac{p_0(p_n-a)+p_{n-1}(a-A)}{b-A} & \text{if } A \leq a \leq B \\
  p_0 + p_2 + p_{n-1} & \text{if } B < a < A_{\text{max}}.
  \end{cases}$

Deductions on Subsequent Comparisons

- Combined belief function after $n^{th}$ comparison;
  
  $f_{A^n}(a) = \frac{\int_{A_{\text{min}}}^{A_{\text{max}}} f_{A^n-1}(a) \times L(a|COD(A_n, B_n) = \bar{p}) \, d\eta}{\int_{A_{\text{min}}}^{A_{\text{max}}} \int_{A_{\text{min}}}^{A_{\text{max}}} f_{A^n-1}(a) \times L(\eta|COD(A_n, B_n) = \bar{p}) \, d\eta \, d\eta}.$

- Combination is associative and in closed form
  
  $f_{A^n}(a) = \frac{\prod_{i=1}^{n} L(a|COD(A_n, B_n) = \bar{p})}{\int_{A_{\text{min}}}^{A_{\text{max}}} \prod_{i=1}^{n} L(\eta|COD(A_n, B_n) = \bar{p}) \, d\eta}$
**Estimation of Optimal Alternative & Utility**

- *Utility* of a belief function is the probability that $A^*$ estimate is within JND of $A^*$
  - Estimation error of less than JND is insignificant

- **$A^*$ estimate**: operating point with maximum utility on $O$
  \[
  A^*(f) = \arg\max_a \left\{ \int_{-JND}^{a+JND} f(\xi)d\xi \right\}
  \]

- **Utility**
  \[
  U(f) = Pr(|\hat{A}^* - A^*| \leq JND) = \int_{\hat{A}^*-JND}^{\hat{A}^*+JND} f(\xi)d\xi
  \]

---

**Outline**

- Introduction
- Approach & Problem Statement
- Model of Subjective Comparisons
- Subjective Evaluations Methods
- Strategy for Simultaneous Evaluations
- Experimental Results
Subjective Evaluation Methods

Problem Formulation
• Find a set of M comparisons $\overline{A}_n = \{A^1_n, \ldots, A^M_n\}$ and $\overline{B}_n = \{B^1_n, \ldots, B^M_n\}$ in each batch, based on current belief function
• Choose $(\overline{A}_n, \overline{B}_n)$ to min $n^* = \min\{n \mid U(f^n) \geq 0.95\}$.

Strategy for Simultaneous Evaluations
• Minimize total comparisons by maximizing utility in each step:
  – $S(U)$: expected number of comparisons left if $U$ is the current utility
    \[
    S(U(f^{n-1})) = 1 + S(U(f^n)) = 1 + \min_{A_n, B_n} S(U(f^n | A_n, B_n))
    \]
• Observations on Bayesian Formulation
  – Likelihood function is unimodal with mode of $A^*$
  – $S(U)$ is a non-increasing function of $U$
  – $U(F^n)$ is a non-decreasing function of $n$
• Comparison between $(A, B)$ is most conclusive when $|p_1 - p_{-1}|$ is large
  – $\min(p_0 + p_2) \Rightarrow \max|p_1 - p_{-1}|$
  – If $A$ or $B = A^*$ and $B - A = \text{CND} \Rightarrow \{p_0 + p_2\}$ is minimized
Simultaneous and Batch Evaluations

- Optimal choice of next comparison pair (simultaneous)

\[ (A_n, B_n) = \begin{cases} 
(A^* - C\tilde{N}D, \hat{A}^*) & \text{if } n \in \text{Even} \\
(\hat{A}^*, A^* + C\tilde{N}D) & \text{if } n \in \text{Odd.} 
\end{cases} \]

- Batch Based Evaluations (heuristic) – M per batch:

\[ C_i = \text{Mod}\left( \frac{i - 1}{M} + \hat{A}^*, 1 \right) \]

\[ (A_{i,n}^i, B_{i,n}^i) = \begin{cases} 
(C_i^i - C\tilde{N}D, C_i^i) & \text{if } C_i^i < \hat{A}^* \\
(C_i^i, C_i^i + C\tilde{N}D) & \text{if } C_i^i > \hat{A}^* \\
\text{Both pairs above} & \text{if } C_i^i = \hat{A}^* 
\end{cases} \]

Performance Analysis

Alg. 1: Complete pair-wise comparison among JND spaced points on \( O \)
Alg. 2: Choose pairs randomly and uniformly in search space (any M)
Alg. 3: Choose a single pair (M=1) in each batch optimally
Alg. 4: Choose multiple (M>1) pairs in each batch using heuristic

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>0.1</th>
<th>0.03</th>
<th>0.01</th>
<th>0.003</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Independent Eval.</td>
<td>45</td>
<td>≈ 500</td>
<td>≈ 5000</td>
<td>≈ 50000</td>
</tr>
<tr>
<td>2. Random (any M)</td>
<td>31.1</td>
<td>192</td>
<td>&gt; 300</td>
<td>&gt; 300</td>
</tr>
<tr>
<td>3. Optimal (M=1)</td>
<td>6.4</td>
<td>9.9</td>
<td>18.3</td>
<td>49.6</td>
</tr>
<tr>
<td>4. Heuristic (M=2)</td>
<td>6.7</td>
<td>11.3</td>
<td>21.4</td>
<td>56.5</td>
</tr>
<tr>
<td>Heuristic (M=3)</td>
<td>9.6</td>
<td>15.6</td>
<td>30.4</td>
<td>78.7</td>
</tr>
<tr>
<td>Heuristic (M=4)</td>
<td>14.0</td>
<td>19.6</td>
<td>34.2</td>
<td>81.2</td>
</tr>
</tbody>
</table>
Conclusion

- Divide offline problem into two stages:
  - 1st given an operating curve find optimal operating point
  - 2nd learn/generalize multiple operating curves for run-time
- Bayesian framework to represent and combine information learned
- Adaptation of next comparison to reduce total comparisons
- Applicable to other real-time multimedia communication problems

Future Work

- Multiple local optima on one operating curve
- Multiple operating curves
- Learning/generalizing of classifiers