ECE Illinois

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN | COLLEGE OF ENGINEERING

Quality Assessment of VoIP Conversations over the Internet

Batu Sat, Zixia Huang,
and Benjamin W. Wah,

12/10/2007

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

Outline

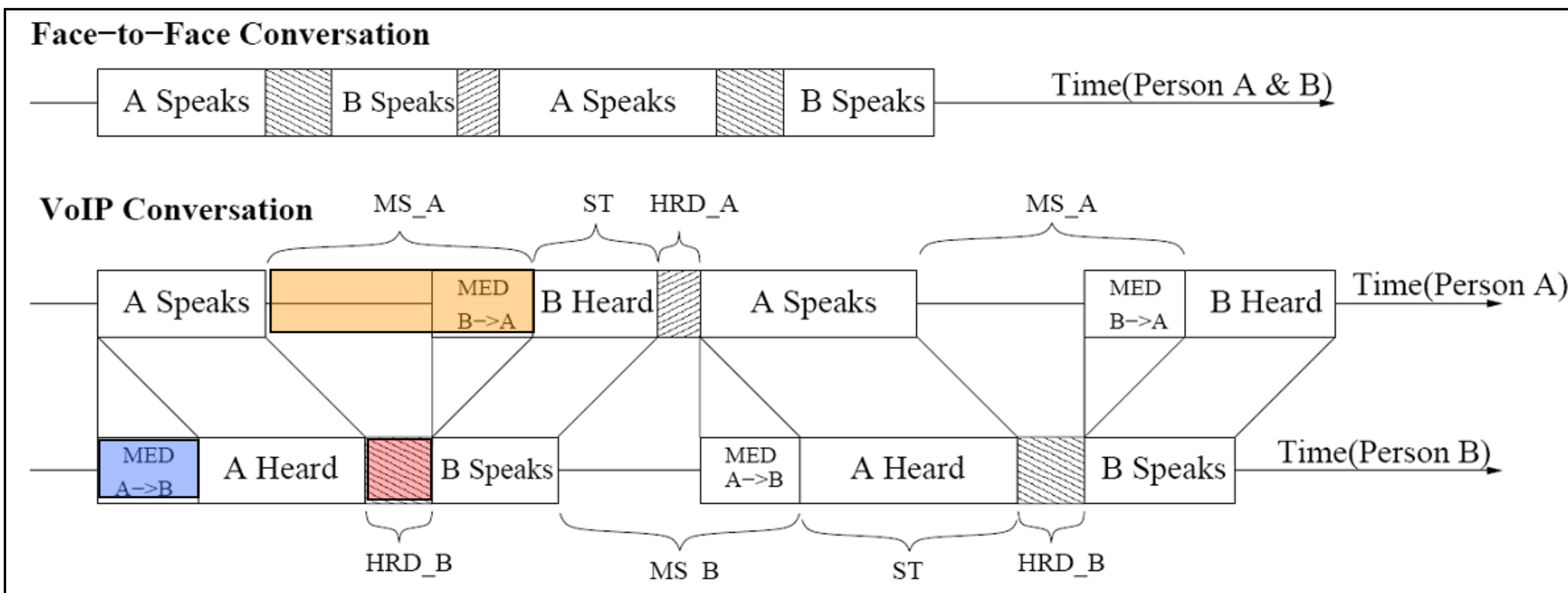
- Conversation (2-party or multi-party)
  - Conversational dynamics
  - Conversational voice communication quality (CVCQ)
- Network environment and network control
- Trade-offs in CVCQ attributes
- Multi-party transmission schemes
- Experimental results

Multi-Party Conversation

- Small group
- Turn-taking process: one person speaking at a time
  - Talk-spurts and pauses during speech
  - Mutual silences in between turns
  - Double talk (in rare cases)
- Roles
  - Previous speaker
  - Current speaker (responder)
  - Passive listener(s)

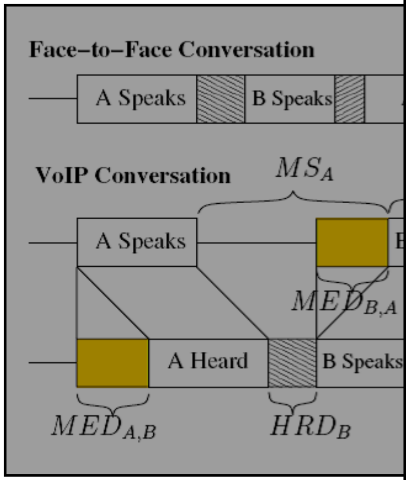
2-Party Conversational Dynamics

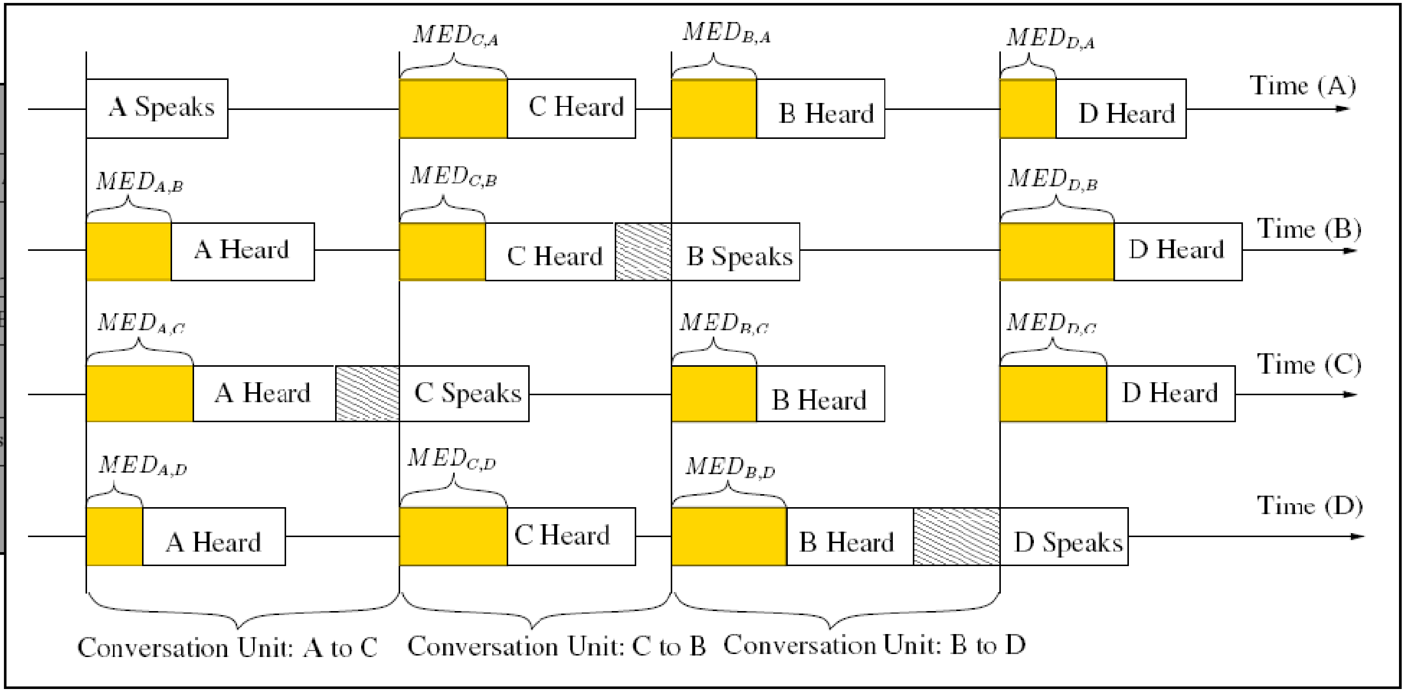
- Common perception of reality in face-to-face conversation
- Multiple realities in conversation over channel with delays
  - *Mouth-to-Ear Delay* (MED)
  - *Human Response Delay* (HRD): Duration between hearing speech and responding
  - *Mutual Silence* (MS): Perceived duration before hearing response



Multi-Party Conversational Dynamics

- Multiple realities
  - Participants perceive different timing and duration of speech & silence events
  - MED, HRD, MS





Demonstration

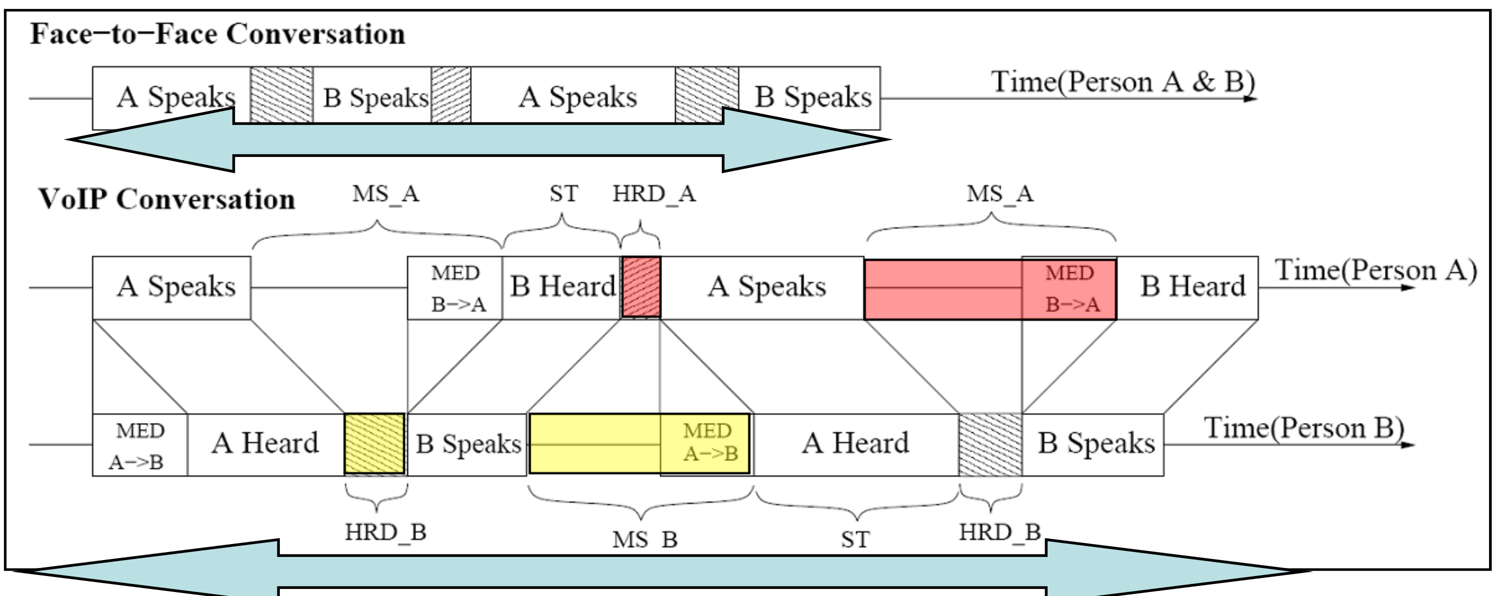
Person	Location	Face-to-Face	Our System	Skype (v3.5)
A	Berkeley (USA)			Host
B	Canada			
C	Dartmouth (USA)			
D	Hefei (China)			
E	Hong Kong			

Conversation Order (HRD=750 ms) using Trace Set 1:

A	C	A	B	C	E	D	B	C	D	C	D	E	C	D

Conversational Metrics

- *Conversational Interactivity* (**CI**) (N(N-1) values)
  - Ratio of perceived silence durations before and after person's speech.
  - Asymmetric with increasing MED
- *Conversational Symmetry* (**CS**) (N values)
  Summarize the imbalances in CI from one participant's perspective
- *Conversational Efficiency* (**CE**) (1 value)
  - Ratio of conversation durations for f2f setting over setting with channel delays
  - Decreasing with increasing MED



Conversational Voice Communication Quality

- CVCQ: Quality of an interactive conversation
  - Listening-only speech quality (LOSQ) of one-way speech (N(N-1) values)
  - Perceived degradations due to network delays
    - CI, CS, and CE can be perceived
    - MED cannot be perceived directly
- CVCQ can be represented by (LOSQ, CI, CE, CS)
  - No standard to relate these metrics to MED

Trade-offs in CVCQ

- Trade-offs among (LOSQ, CI, CS, CE) depend on MED
  - MED ⇧ LOSQ ⇧ CI, CS, CE ⇩
  - MED ⇩ LOSQ ⇩ CI, CS, CE ⇧
- Multi-dimensional on all speaker-listener pairs
- Trade-offs among
  - Individually perceived metrics (CI, CS, LOSQ)
  - Commonly perceived metric (CE)
- Improving CE by minimizing MED for each listener can cause
  - Vulnerability to unconcealable frames due to delay spikes
  - Request for repetition of an utterance → Low CE

Perceived Delay Effects (2-Party)

- Perceived effects of MED depend on **conversational conditions**
  - CI and CS depend on the **Human Response Delay (HRD)**
  - CE depends on **Switching Frequency** of conversation

Table 2: Statistics of two face-to-face conversations.

Conversation Type	Avg. single-talk duration	Avg. HRD duration	# of switches	Total Time
Social	3,737 ms.	729 ms.	7	35 sec.
Business	1,670 ms.	552 ms.	15	35 sec.

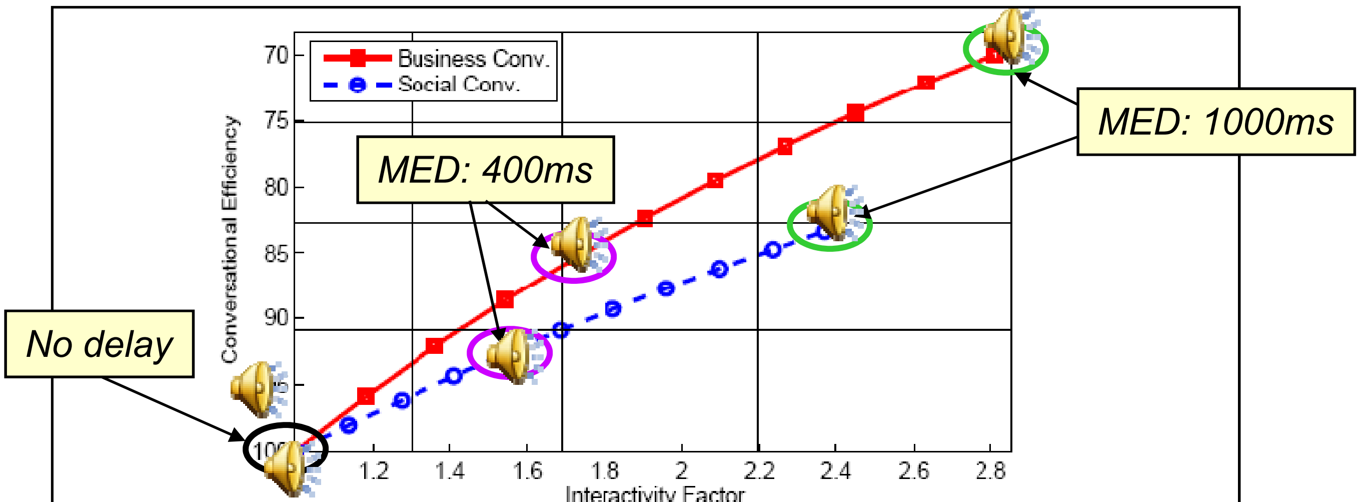


Figure 11: Effect on CI and CE when MED changes for the two conversations in Table 2.

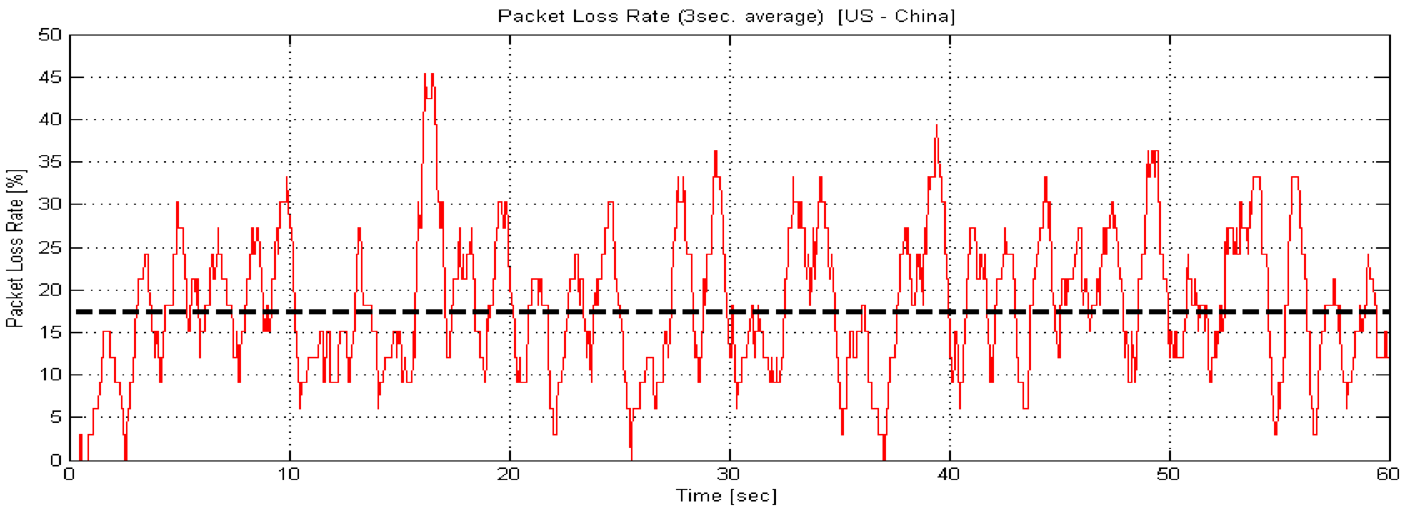
IEEE Int'l Symposium on Multimedia, 2007

Outline

- Conversation
- Network environment and network control
  - Network conditions
  - Network control: POS and LC
  - Trade-offs in system-controllable metrics
- Trade-offs in CVCQ attributes
- Multi-party transmission schemes
- Experimental results

Network Conditions: Packet Loss

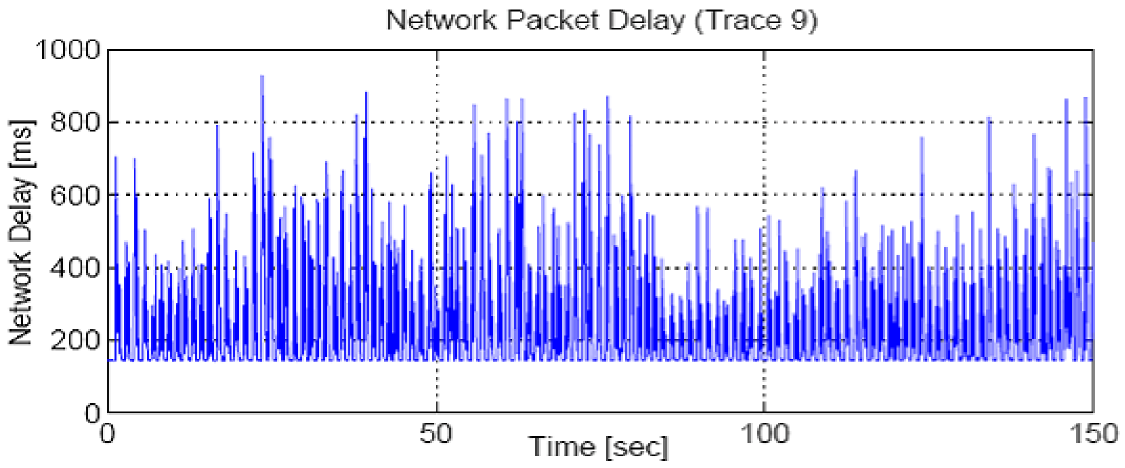
- Network-loss conditions change in a matter of seconds
  - Stationary models cannot track fast changing conditions



- Retransmission of lost speech packets not feasible in real-time VoIP
  - Redundancy needed to conceal lost packets at receiver
  - Piggybacking previously sent frame(s) in current packet
  - Require receiver to wait for redundant packet
  - Too frequent packet transmissions cause congestion

Network Conditions: Packet Delay

- Packets experience network delays with jitters and spikes



- Speech needs to be played smoothly in the presence of jitters and delays
  - Employ jitter-buffers and adaptive play-out scheduling (POS)

Multi-Party Internet Traffic Behavior

- Large variations in delay, jitter and loss
  - Disparities across destinations from a single source
  - Correlations in delay spikes from one source

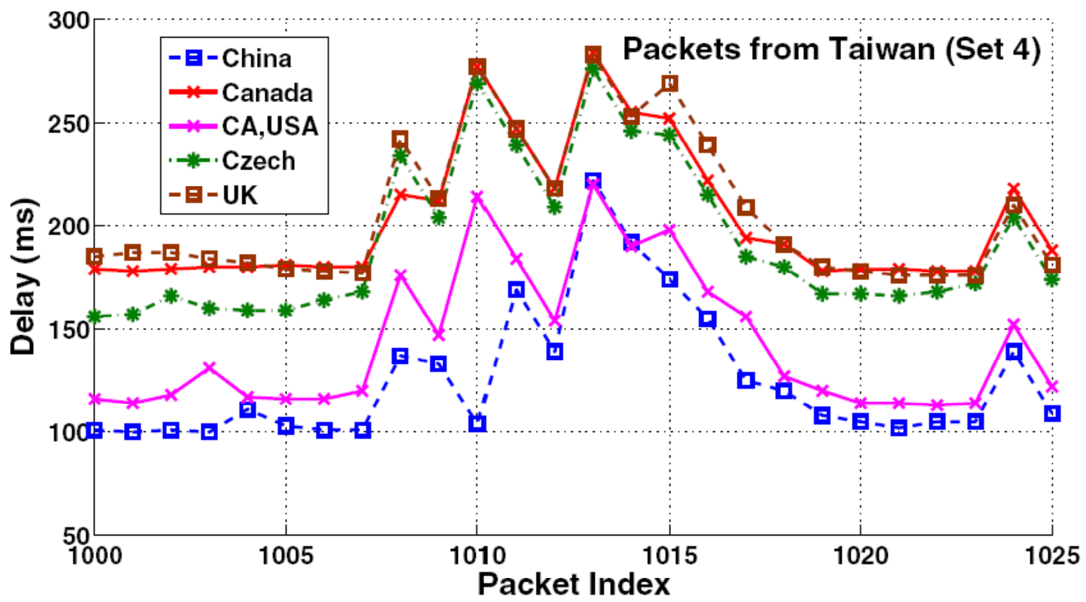


Figure 3. Delay behavior of packets collected from Taiwan to Xian (China), Canada, California (USA) and Czech at 1:00 CST in August 2007 (Trace 4).

Outline

- Conversation
- Network environment and network control
- Trade-offs in CVCQ attributes
  - Trade-offs via system controllable metrics
  - CVCQ representation
  - Subjective tests
- Multi-party transmission schemes
- Experimental results

Design Goals

- Observations on multi-party VoIP conferencing:
  - Multiple perspectives over delayed channels
  - Disparities in network conditions across speaker-listener pairs → asymmetry
  - Trade-offs among multiple quality metrics (LOSQ, CI, CE, CS)
  - Not scalable by P2P extension of two-party VoIP system
- Design a VoIP conferencing system with high conversational quality that is consistent across time and participants.
  - Multi-party transmission topology
  - Loss concealment schemes
  - Play-out scheduling schemes

VoIP System Architecture

- Speaker/Listener
  - Perceptual metric: CVCQ
  - Measurable metric: LOSQ, CI, CE, CS
  - Conversational conditions: SF, HRD
  - Delays introduced/perceived
- VoIP client
  - Speech encoding/decoding
    - Mixing of received streams
  - Network Control
    - Loss Concealments
    - Play-out scheduling
  - Measurable metrics (MED_{est}, UCFP)
- Transmission scheme
  - Receiver: single or multiple
  - Sender: single or multiple
  - Overlay network control (if present)

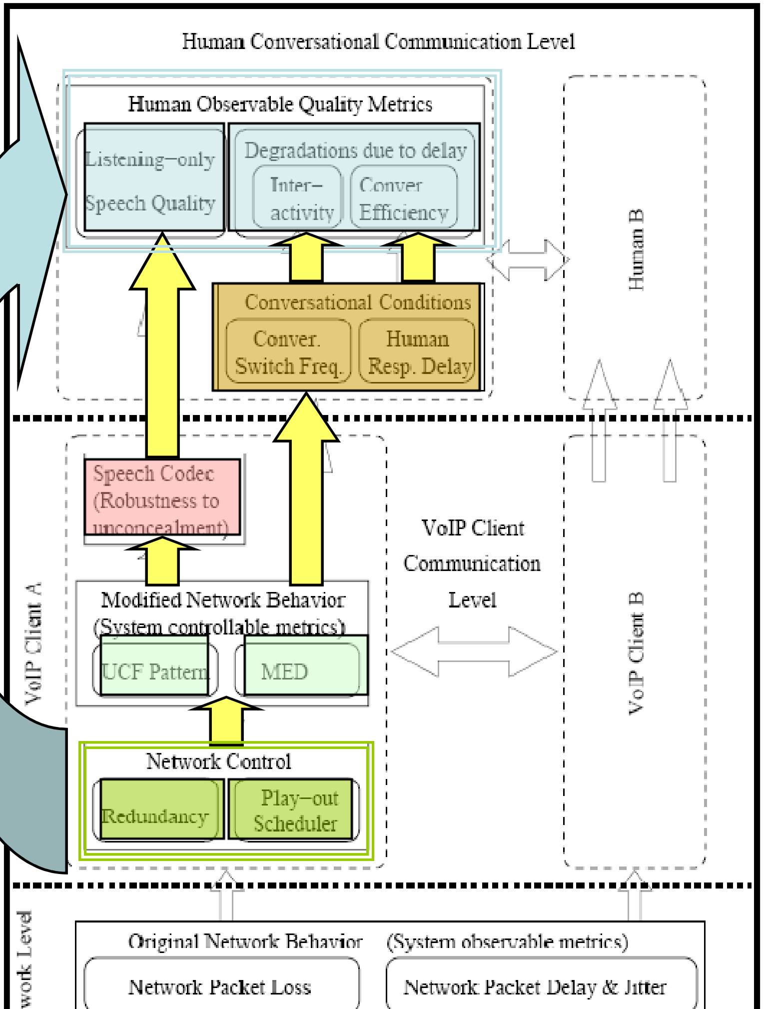


Figure 1: Architecture of a VoIP system.

Network Control via POS & LC

- Goal: Mitigate network imperfections
- System-observables
  - Network-loss rate & burstiness
  - Network delays & jitters
- System controls
  - Redundancy rate (degree of piggybacking)
  - Play-out schedule of speech segments
- Intermediate quality metrics (system-controllables)
  - Un-concealable Frame Rate (UCFR)
  - Un-concealable Frame Pattern (UCFP)
  - Mouth-to-ear delay (MED)

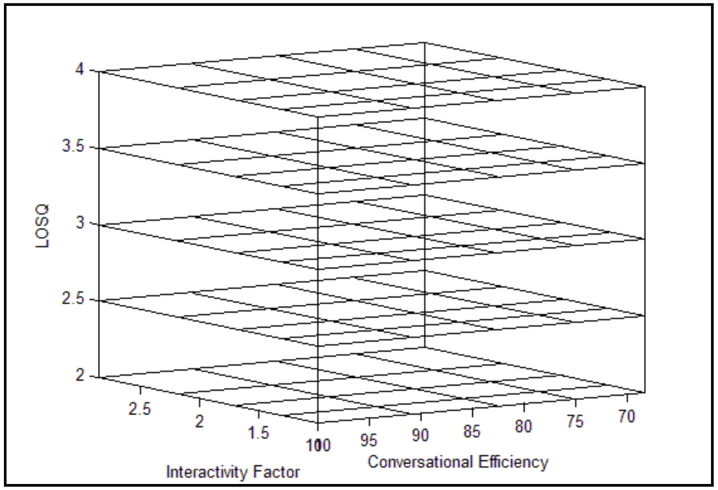
Trade-offs in System-Controllable Metrics

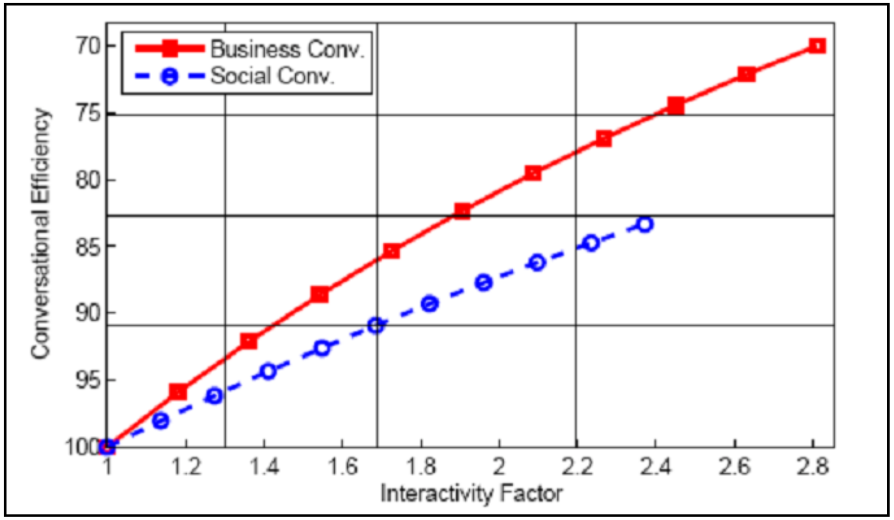
- Trade-offs between UCFR and MED
  - Depending on network conditions
  - Must be adaptive

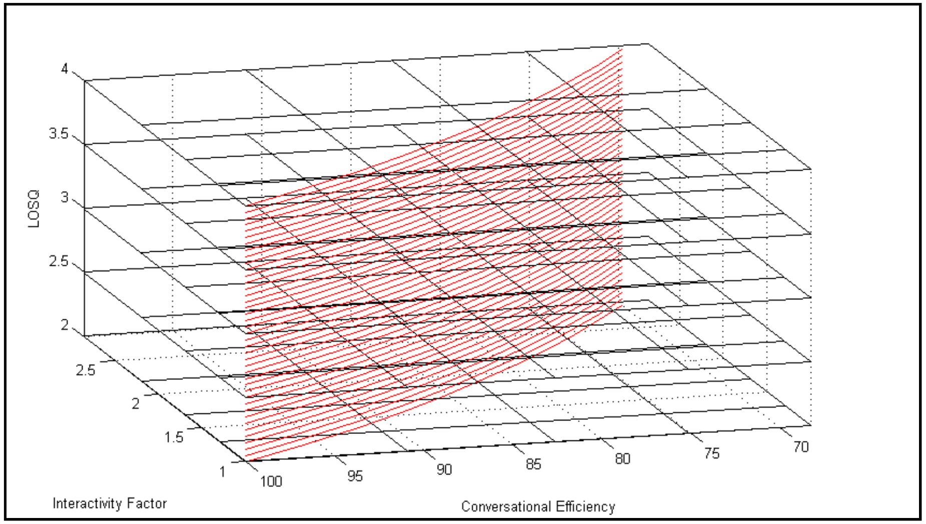
Network Control used under conditions		Network Delay Condition: Low Jitter	Network Delay Condition: High Jitter
Network Loss Condition	Low Loss	No-redundancy Short & slow changing MED	No-redundancy UCFR improves gracefully with MED
Network Loss Condition	High Loss (Bursty)	Redundant Piggybacking MED to allow receipt of redundant packets	Redundant Piggybacking High MED to reduce UCFR

CVCQ Representation

- CVCQ = (LOSQ,CI,CE)
  - Point in 3-D space
- CI and CE depend on MED and conversational conditions
  - Given conversational condition
    - Restricted to curve (e.g. C_business, C_social) on (CI,CE) plane
    - Restricted to plane (e.g. P_business) on (LOSQ,CI,CE) space

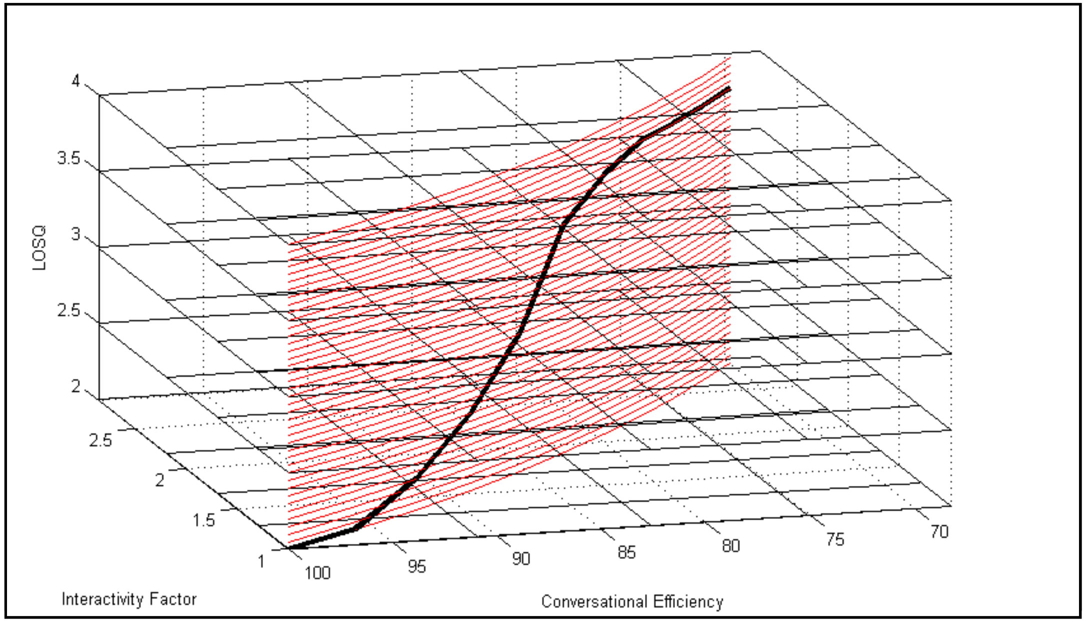






Trade-offs in CVCQ Attributes

- LOSQ depends on MED, redundancy, codec, and network conditions
- For given codec, network conditions + POS/LC policy
  - Restricted to a curve on the P_business plane in (LOSQ,CI,CE) space



- Different planes for different conversational conditions
- Different curves for different network conditions + POS/LC policy

Subjective Tests: Comparative MOS

- Comparing perceived quality of two conversations (ITU P.800)
  - Subjects asked to compare A against B
- Illustration of user preference in 2-D
  - Direction of arrow represents preference
  - Length of arrow represents strength of preference

Table 3: Comparison MOS tests: User responses.

User response	CMOS score
A is strongly preferred against B	−2
A is preferred against B	−1
A and B are preferred equally	0
B is preferred against A	1
B is strongly preferred against A	2

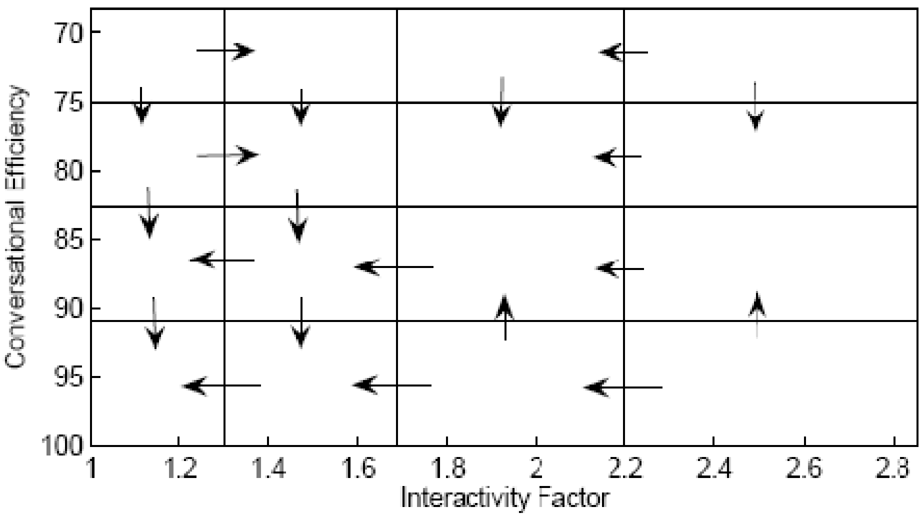


Figure 13: Relative user opinion on the CI-CE plane. The size of an arrow indicates the degree of preference of one alternative over another.

Outline

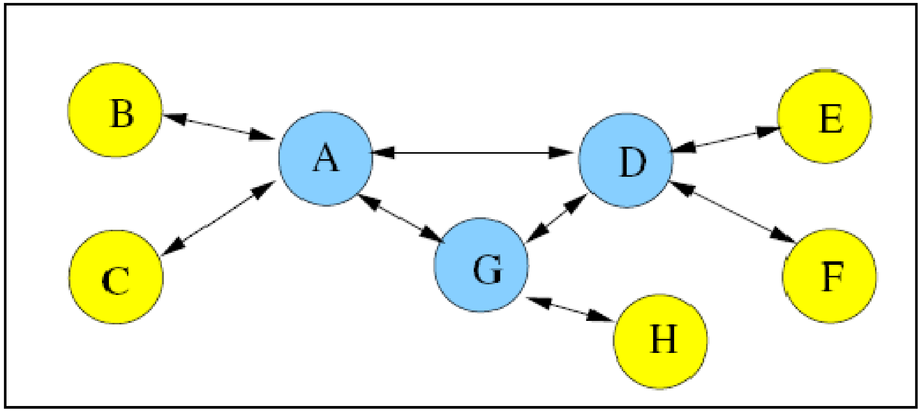
- Conversation
- Network environment and network control
- Trade-offs in CVCQ attributes
- Multi-party transmission schemes
  - Overlay network topology
  - Previous work on coding, POS, and LC
  - Proposed POS and LC
- Experimental results

Multi-party Transmission Schemes

- Decentralized: P2P unicasts from speaker to listeners
  - Not scalable in the number of participants
  - Near-end congestion at senders
- Centralized: Single “Host” relays to all participants
  - Multiple simultaneous speakers
    1. Decode → Mix → Re-encode → Send new packet to listeners
    2. Relay packets (combining packets to the same destination if possible)
  - Host may be overloaded
- Hybrid
  - IP Multicast: Not fully adopted by ISPs to allow global availability
  - Overlay Network: Virtual network over VoIP clients
    - Only a subset of the P2P links are used, less strain on any one node

Overlay Network

- Trade-off between
  - Maximum packet transmission rate of any node in topology
  - Maximum end-to-end delay (ME2ED) between any node pair
  - Adaptive to changing network conditions
- Proposed topology structure
  - Parent nodes (blue): Fully connected with other parent nodes
    - Relay packets received after combining received packets
  - Children nodes (yellow): Only connected to one parent node



Greedy Design of Overlay Network

- Collect P2P topology network conditions at call establishment
- Find max E2E delay for P2P topology
- Find single parent topology with the best ME2ED
- Add new parent nodes until
  - Improvement in ME2ED is small
  - OR bottleneck pair (one with ME2ED) is already directly connected (no hope for improvement of ME2ED)

Speech Codecs: Desirable Properties

- High perceptual speech quality under no loss
- Wide-band (16 KHz) encoding of speech
- Low-bit-rate
- Low algorithmic delay and computational complexity
- Robustness to bursty losses
- Multi-mode operation to allow graceful adaptation to network conditions

Speech Codecs

- ITU standardized speech codecs
  - G.722.2: Wideband, ADPCM, [6.6 – 24 Kbps], 20ms frames
  - G.729.1: Wideband CELP-TDBWE, [8-32 Kbps], 20ms frames
- Propriety speech codecs
  - iSAC: Wideband, Hybrid, [10-32 Kbps], [30-60ms framing options], GIPS
  - Skype claims to use iSAC in two-party system, codec used in multi-party system is unknown
- Speech Codec used in our system
  - G.722.2, 24 Kbps mode, 40 ms packet period (2 frames)
  - High quality under both lossy and non-lossy conditions,
  - Low algorithmic delay, low computational complexity

Play-Out Scheduling

- Decentralized approach:
  - Trade-offs between MED & LOSQ for each speaker-listener pair
  - May cause asymmetry in conversation perceptions
- Fully centralized approach:
  - Trade-offs between MED & LOSQ for all speaker-listener pairs
  - Overhead in collecting network info & disseminating decision

Observations:

- Limiting factor in CE: Max MED across listeners from source
  - Non-bottleneck listener's MED does not affect efficiency
- Disparities in MEDs from different speakers to one listener
  - Bottleneck listener can be different for each speaker

Our Adaptive POS Scheme

- Decentralized approach with side information
  - Bottleneck listener for current speaker: determined by previous network statistics
  - Each listener finds the individually optimal MED (w/ high LOSQ)
  - Non-bottleneck listeners relaxes their MED according to bottleneck info
    - More robust to possible delay spikes
    - More symmetric conversational dynamics across listeners
    - Minimal effect on overall conversational efficiency
  - Bottleneck listener uses individually optimal MED, updates clients for future

Loss Concealments

Observations:

- Speech quality needs to be consistently high across listeners
- Play-out scheduling cannot conceal packets lost
- Internal error concealment schemes of speech codecs are inadequate
- Disparities in network loss conditions across links

Loss concealment scheme:

- Dynamic loss conditions
- Link-specific (overlay network) loss concealment, rather than end-to-end
- Redundancy decision should be made available for POS schemes at clients
  - To wait adequately to allow for redundant information to arrive in time for play-out

POS/LC Schemes (Bottleneck Path)

- Loss concealment
  - Control redundancy degree

$$R_{i+FBD} = \min\{R \mid UCFR_i^W(\bar{p}, \bar{R}) \leq 2\%\}.$$

- POS
  - Estimate CVCQ curve by conversational and network conditions
  - Adjust system-controllable metrics to maximize user preference along curve

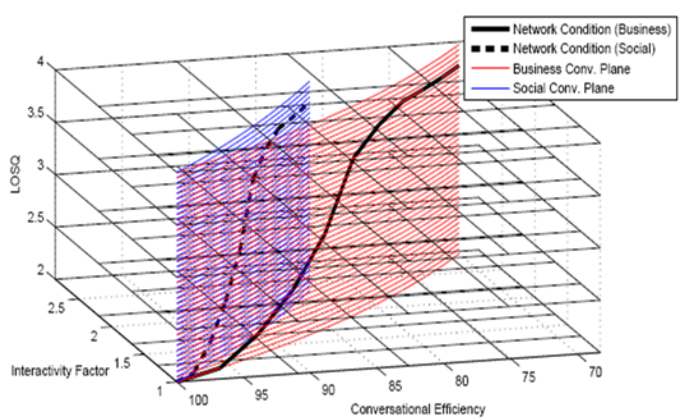


Figure 12: The planes represent the conditions due to the conversational type. The curve on each plane represents the conditions imposed by the network.

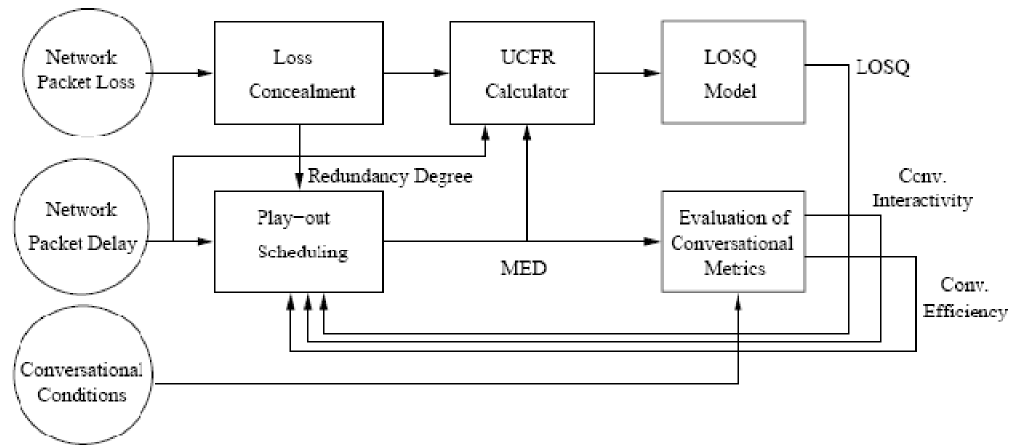
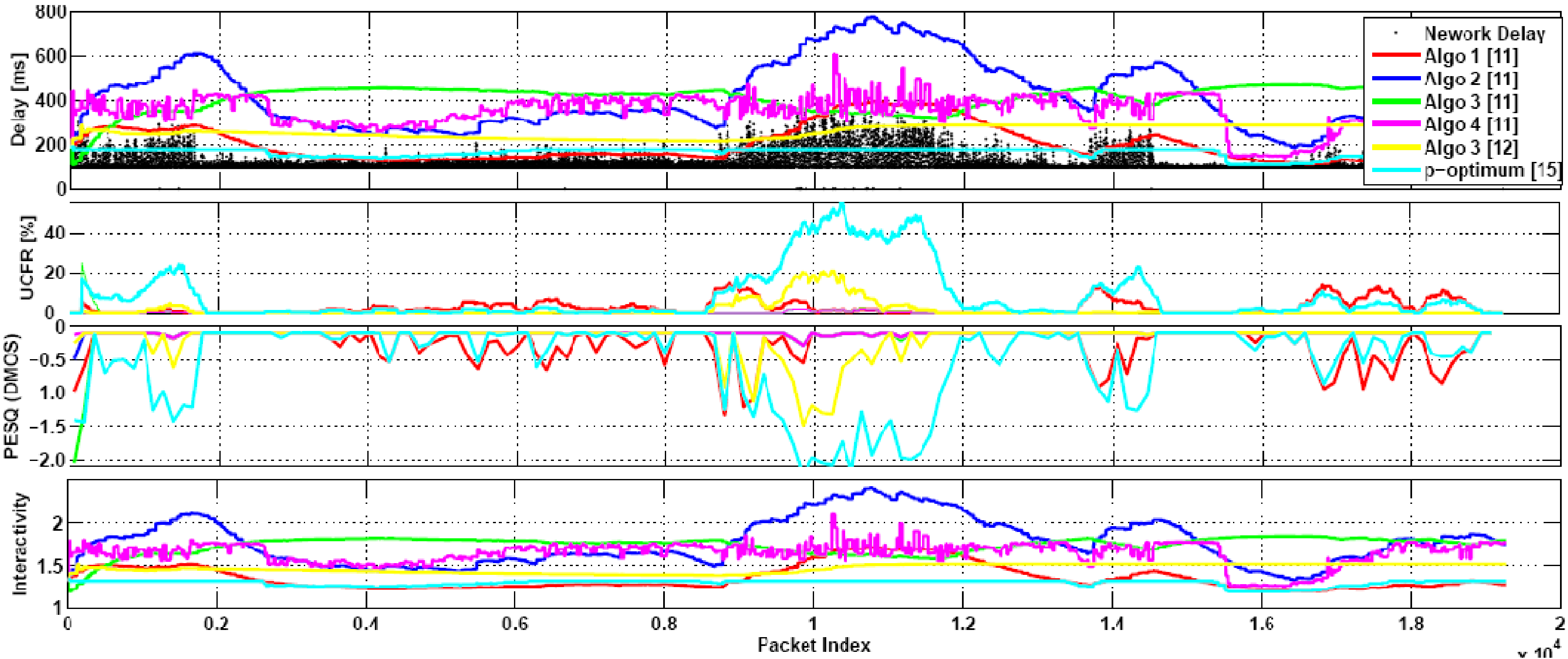


Figure 14: Proposed network control by POS/LC.

Adaptive POS: Previous Work

- None of previous schemes provides consistent balance between CVCQ attributes under changing network conditions
  - [11]: open loop scheme calculates running estimates of mean and variations in network delays and choose play-out delay at the beginning of talk spurt
  - [13]: closed loop scheme adapts LC based on the late loss rate collected in window
  - [15]: trained regression model based on Bernoulli loss models to estimate PESQ

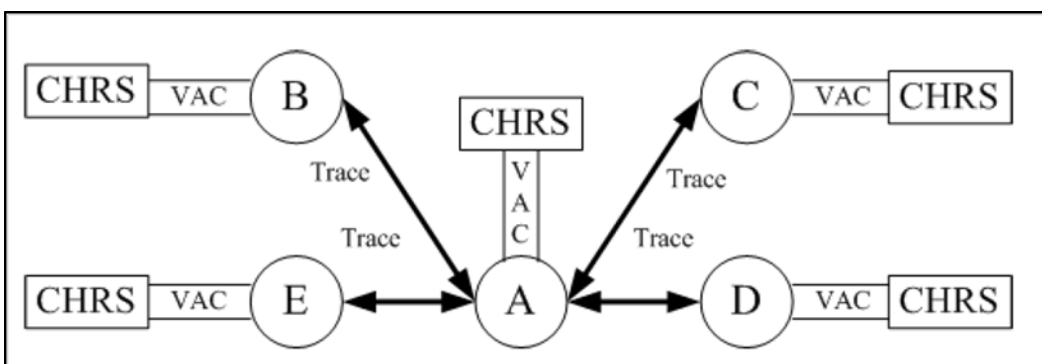


Outline

- Conversation
- Network environment and network control
- Trade-offs in CVCQ attributes
- Multi-party transmission schemes
- Experimental results
  – Experimental setup
  – Demo of multi-way conversation/Skype

Experimental Setup

- 5 client machines (Windows), connected through a router
- VoIP client software (Skype 3.5.0.214) running on each
- Router machine (Linux) drops and delays packets
  - For each connection based on PlanetLab traces
- Conversational Human Response Simulator (CHRS) on each
  - Listens to output waveform of VoIP software (other parties' speeches)
  - Waits for 750 ms after other's speech before playing speech
  - Each CHRS knows the pre-determined order for smooth turn switching
  - Conversations are recorded at each client by CHRS for off-line analysis



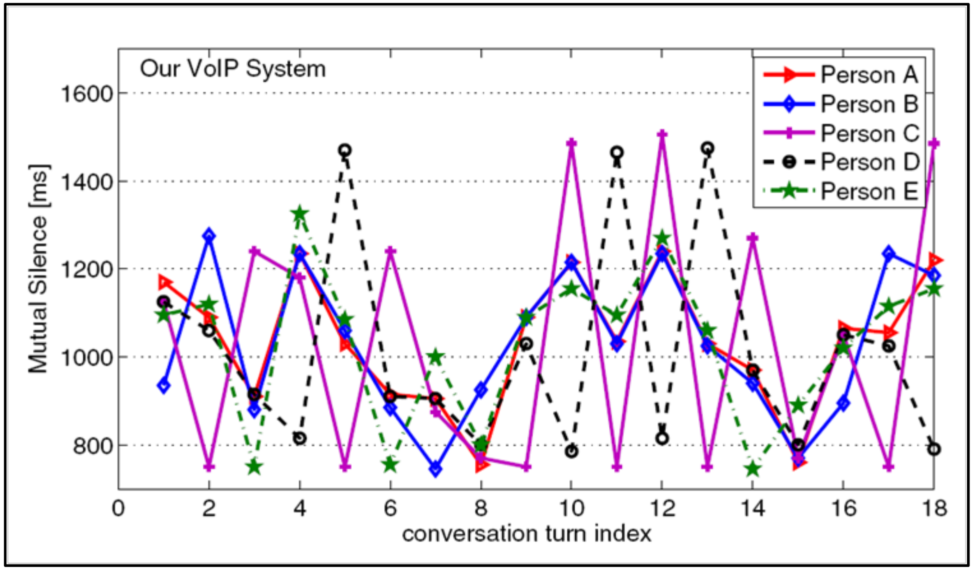
Person	Trace Set 1	Trace Set 2	Trace Set 3
A	CA, USA	UK	NH, USA
B	Canada	Hong Kong	UK
C	NH, USA	Finland	Canada
D	Hefei, China	NH, USA	Hungary
E	Hong Kong	Hungary	Hefei, China

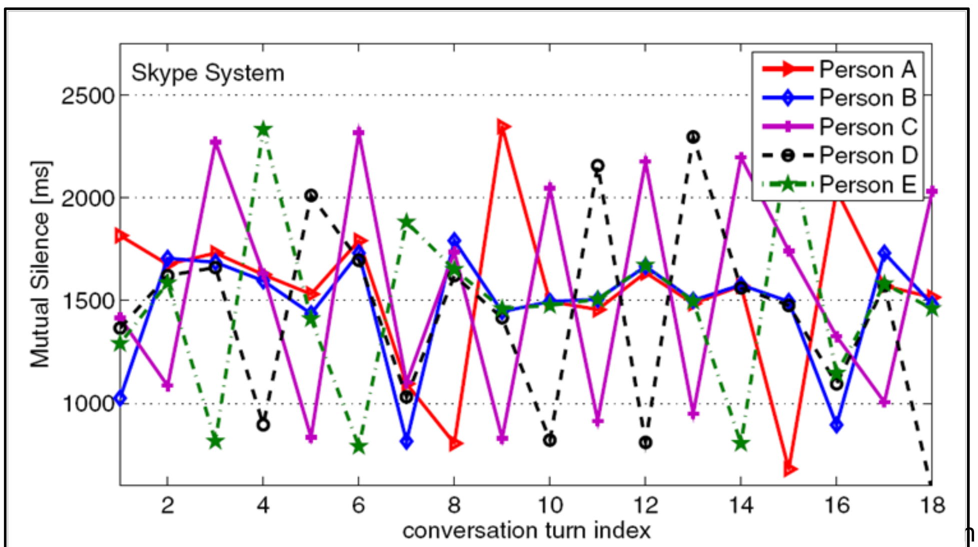
Experimental Results

- Offline analysis of conversational recordings
  - Calculation of LOSQ (PESQ – ITU P.862), CI, CE, CS
  - Subjective Comparative MOS tests for conversations
    - Two conversations (Our system & Skype): played in random order
    - Opinions in Comparative Category Rating (CCR) scale {-3,-2, -1, 0, 1, 2, 3}

Set	System	MS [ms] Rsp.	PrSpk.	Lst.	CI	CS	CE	PESQ	CMOS
1	Ours	1256	780	1029	1.62	1.68	70	3.477	+0.87
	Skype	2078	853	1510	2.44	1.80	62	2.754	
2	Ours	1072	780	925	1.35	1.40	73	3.741	+0.80
	Skype	1975	866	1462	2.32	2.11	63	2.916	
3	Ours	1071	780	928	1.36	1.35	72	3.735	+1.13
	Skype	1983	898	1463	2.29	2.40	62	2.995	

Experimental Results (cont'd)





- Shorter MS, thus better CI & CE
  - Our adaptive overlay network scheme chooses transmission topology that leads to a shorter end-to-end delay
  - Skype uses a single-parent topology based on the client initiating the conference call
- More symmetry (better CS)
  - POS schemes at each client utilizes a centralized side information to relax MEDs
  - Effect of relaxed MEDs to CE is minimal.

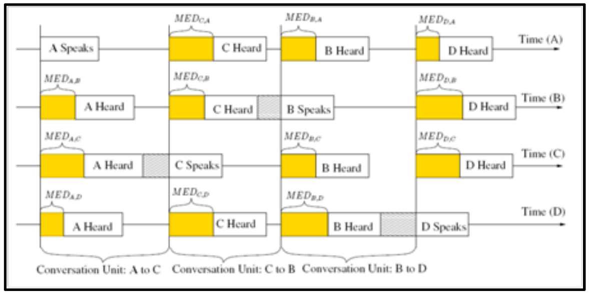
Demonstration

Person	Location	Face-to-Face	Our System	Skype (v3.5)
A	Berkeley (USA)			Host
B	Canada			
C	Dartmouth (USA)			
D	Hefei (China)			
E	Hong Kong			

Conversation Order (HRD=750 ms) using Trace Set 1:

A	C	A	B	C	E	D	B	C	D	C	D	E	C	D

Conclusions



- Conversational quality
  - LOSQ, MED
  - CI, CS, and CE
  - Subjective tests to guide control
  - Optimized via intermediate quality metrics

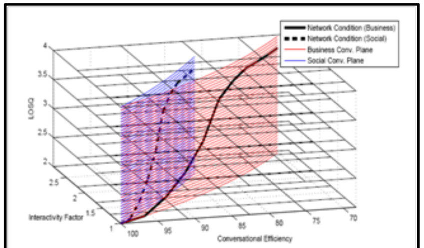
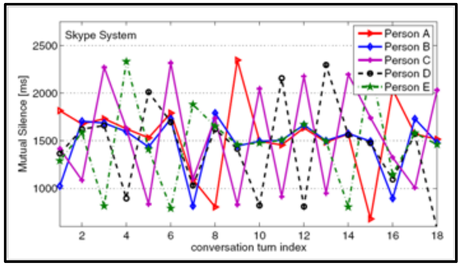


Figure 12: The planes represent the conditions due to the conversational type. The curve on each plane represents the conditions imposed by the network.

- Trade-offs achieved via network controls
  - LC via redundant piggybacking
  - Suitable mouth-to-ear delays via POS
  - Overlay topology
  - Solution different from 2-party conversation



Questions?