

Quantifying



User Satisfaction

Sheng-Wei (Kuan-Ta) Chen

Institute of Information Science, Academia Sinica

Collaborators: Chun-Ying Huang
Polly Huang
Chin-Laung Lei
(National Taiwan University)

Motivation

- Are users satisfied with our system?
 - User survey
 - Market response
 - User satisfaction metric
- To make a system **self-adaptable in real time** for better user experience
 - User satisfaction metric

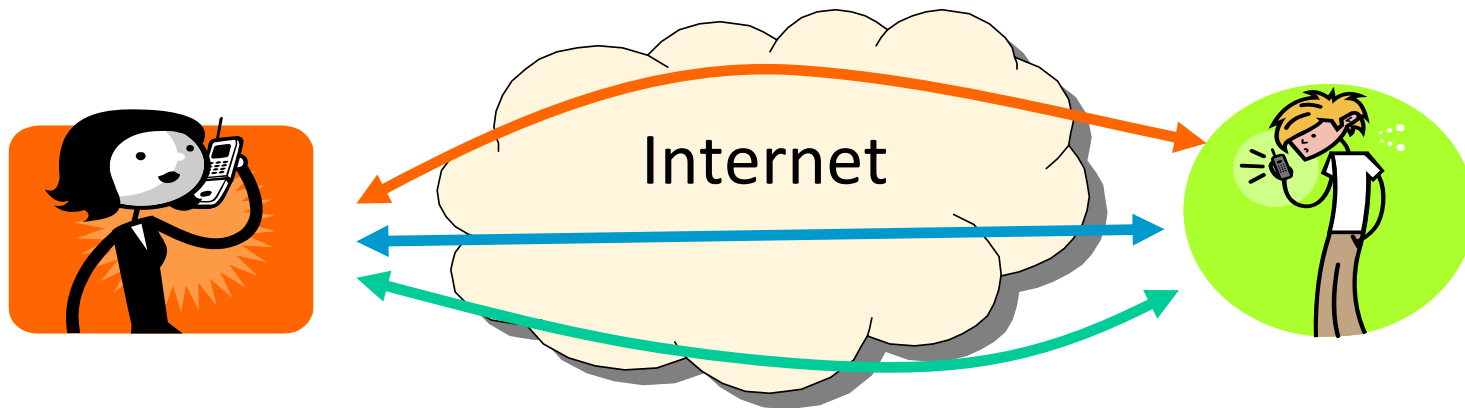
Need of a Quality-of-Experience (QoE) metric!

QoE metrics




- FTP applications: data throughput rate
- Web applications: response time and page load time
- VoIP applications: voice quality (fidelity, loudness, noise), conversational delay, echo
- Online games: interactivity, responsiveness, consistency, fairness

QoE is multi-dimensional esp. for real-time interactive applications!

What path should Skype choose?



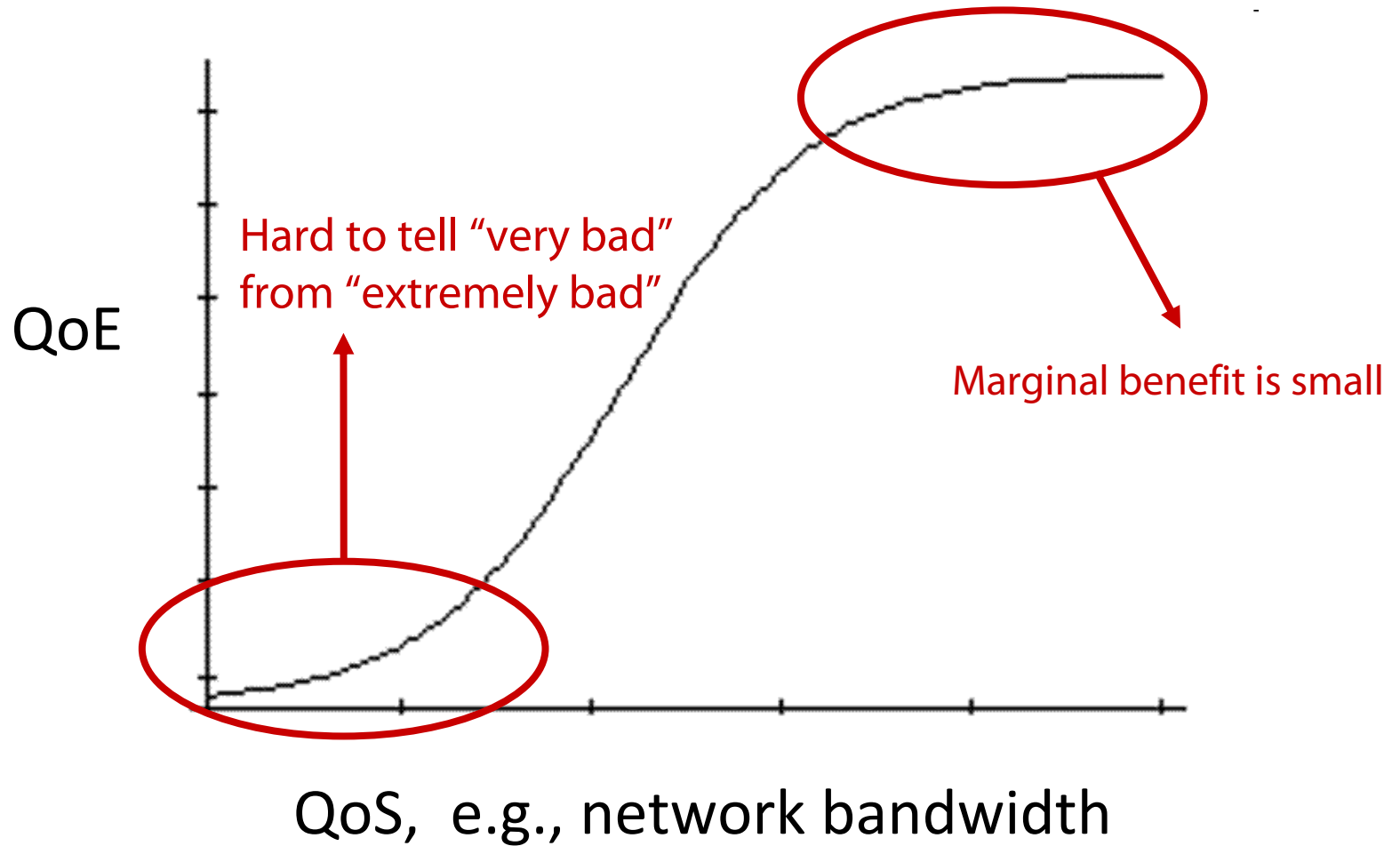
Which path is “the best”?

path	avail bandwidth	loss rate	delay
	10 Kbps	2%	100 ms
	20 Kbps	1%	300 ms
	30 Kbps	3%	500 ms

QoS and QoE

- QoS (Quality of service)
 - The quality level of “native” performance metric
 - Communication networks: delay, loss rate
 - Voice/audio codec: fidelity
 - DBMS: query completion time
- QoE (Quality of experience)
 - How users “feel” about a service
 - Usually **multi-dimensional**, and **tradeoffs** exist between different dimensions (download time vs. video quality, responsiveness vs. smoothness)
 - However, a unified (scalar) index is normally desired!

A typical relationship between QoS and QoE



Mapping between QoS and QoE

Which QoS metric is most influential on users' perceptions (QoE)?

Source rate?

Loss?

Delay?

Jitter?

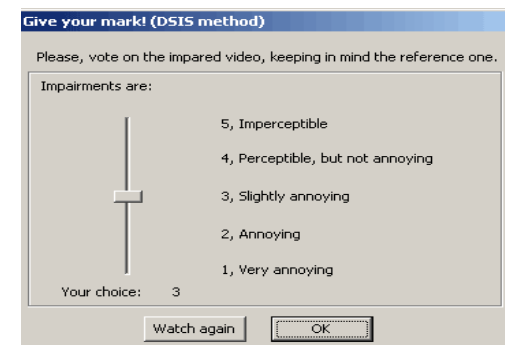
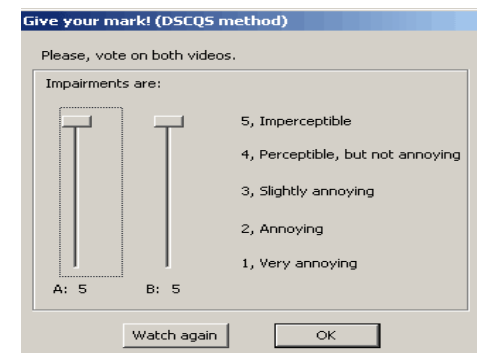
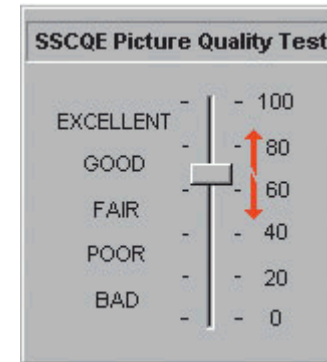
Combination of the above?

How to measure QoE: A quick review

- Subjective evaluation procedures
 - Human studies, not scalable
 - Costly!
- Objective evaluation procedures
 - Statistical models based on subjective evaluation results
 - **Pros: Computation without human involvement**
 - **Cons: (Over-)simplifications of model parameters**
 - E.g., use a single “loss rate” to capture the packet loss process
 - E.g., assume every voice/video packet is equally important
 - Not consider external effects such as loudness and quality of handsets

Subjective Evaluation Procedures

- Single Stimulus Method (SSM)
- Single Stimulus Continuous Quality Evaluation (SSCQE)
- Double Stimulus Continuous Quality Scale (DSCQS)
- Double Stimulus Impairment Scale (DSIS)



Objective Evaluation Methods

- Referenced models
 - speech-layer model: PESQ (ITU-T P.862)
 - Compare original and degraded signals
- Unreferenced models (no original signals required)
 - speech-layer model: P.VTQ (ITU-T P.563)
 - Detect unnatural voices, noise, mute/interruptions in degraded signals
 - network-layer model: E-model (ITU-T G.107)
 - Regression model based on delay, loss rate, and 20+ variables
 - Equations are over-complex for physical interpretation, e.g.

$$I_s = 20 \left[\left\{ 1 + \left(\frac{X_{olr}}{8} \right)^8 \right\}^{\frac{1}{8}} - \frac{X_{olr}}{8} \right]$$
$$X_{olr} = OLR + 0.2(64 + No - RLR)$$

Our goals

An objective QoE assessment framework

- passive measurement (thus scalable)
- **easy to construct models (for your own application)**
- easy to access input parameters
- easy to compute in real time

Our contributions

$$\text{USI} = 2.15 \times \log(\text{bit rate}) - 1.55 \times \log(\text{jitter}) - 0.36 \times \text{RTT}$$

bit rate: data rate of voice packets

jitter: receiving rate jitter (level of network congestion)

RTT: round-trip times between two parties

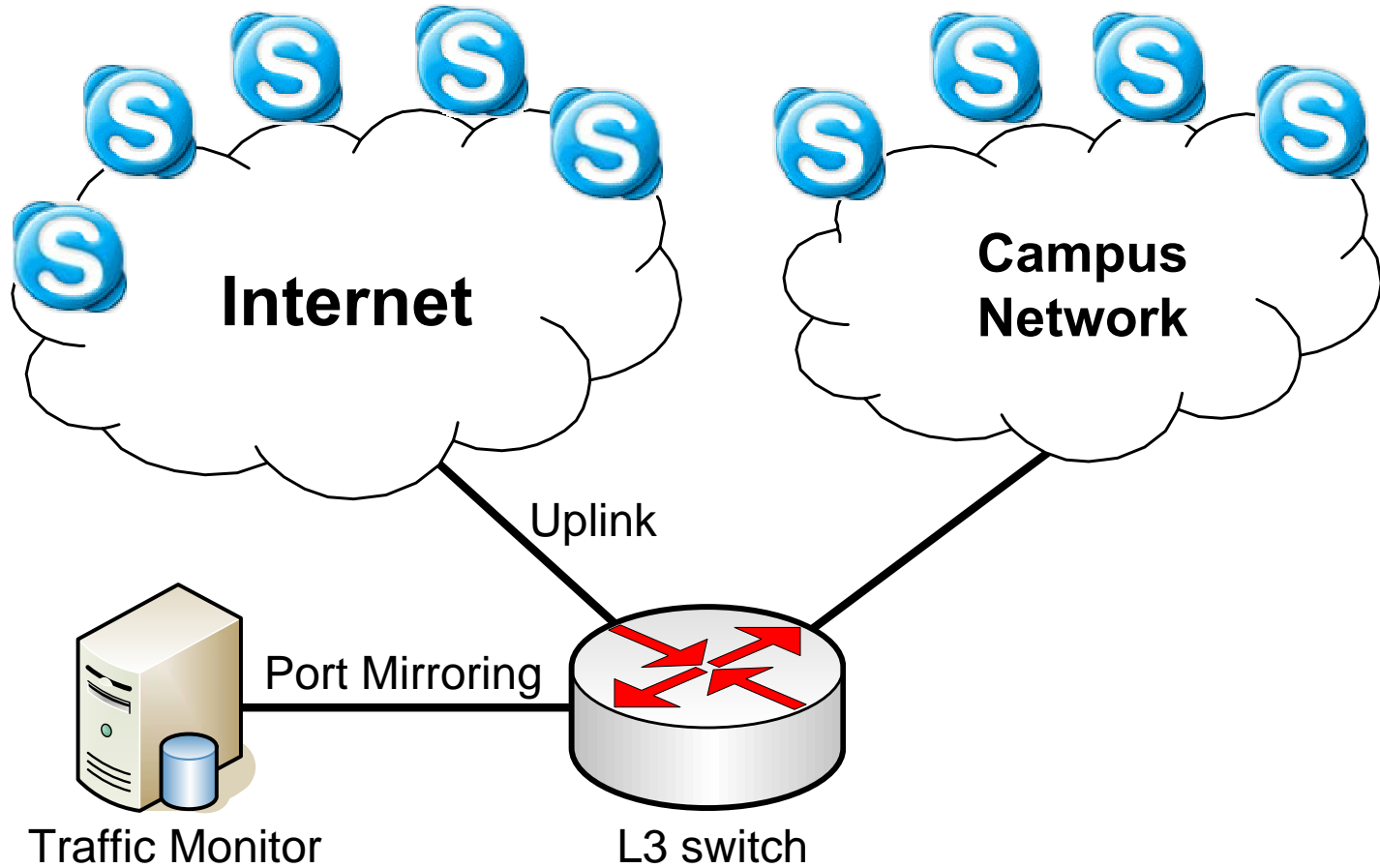
An index for Skype user satisfaction

- derived from real-life Skype call sessions
- verified by users' speech interactivities in calls
- accessible and computable in real time

Talk outline

- The Question
- ➔ ■ Measurement
- Modeling
- Validation
- Significance

Setting things up



Capturing Skype traffic

1. Identify Skype hosts and ports

- Track hosts sending http to “ui.skype.com”
- Track their ports sending UDP within 10 seconds
- → (host, port)
- Other parties which communicate with discovered host-port pairs

2. Record packets

- Whose source or destination \in these (host, port)
- Reduce the # of traced packets to 1-2%

Extracting Skype calls

1. Take these sessions

- Average packet rate within (10, 100) pkt/sec
- Average packet size within (30, 300) bytes
- For longer than 10 seconds

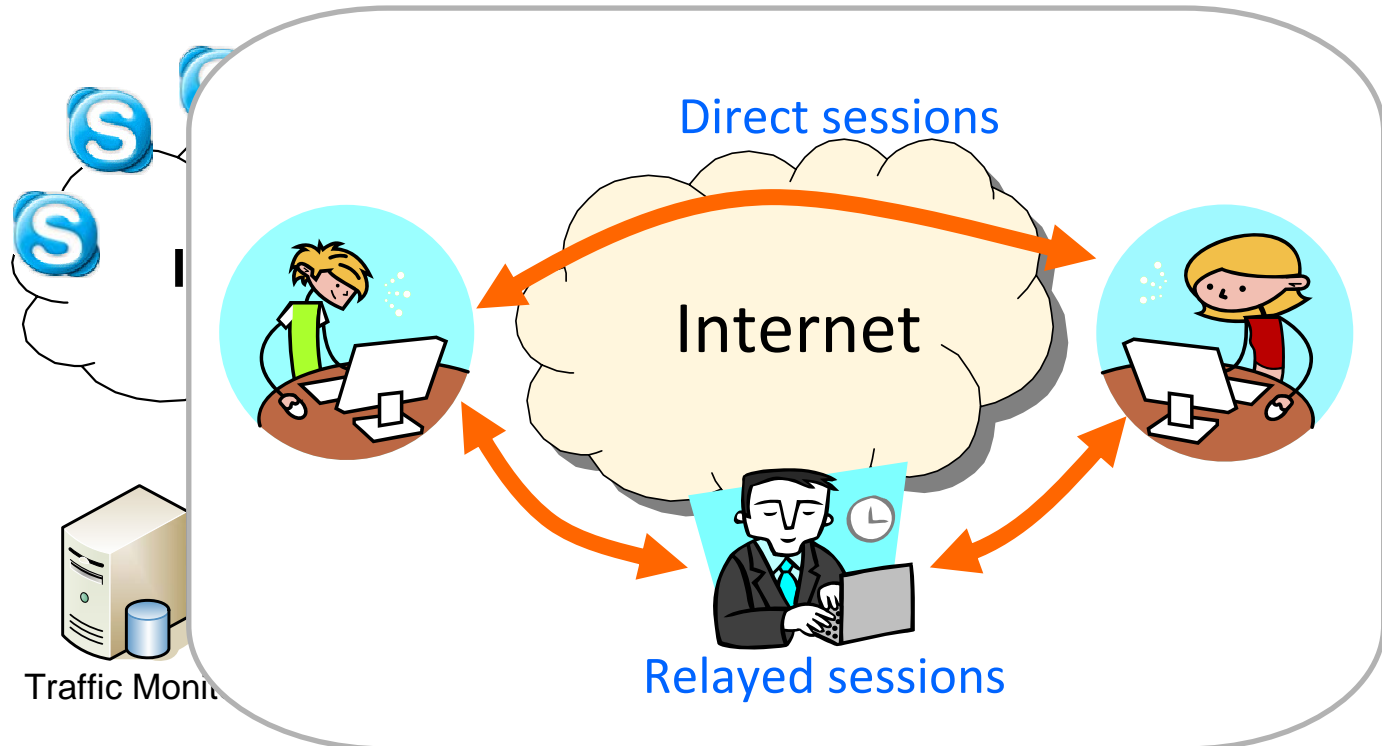
2. Merge two sessions into one relay session

- If the two sessions share a common relay node
- Their start and finish time are close to each other with 30 seconds
- And their packet rate series are correlated

Probing RTTs

- As we take traces
- Send ICMP ping, application-level ping & traceroute
 - Exponential intervals

Trace Summary



Category	Sessions	Hosts	Avg. Time
Direct	253	240	29 min
Relayed	209	369	18 min
Total	462	570	24 min

Talk outline

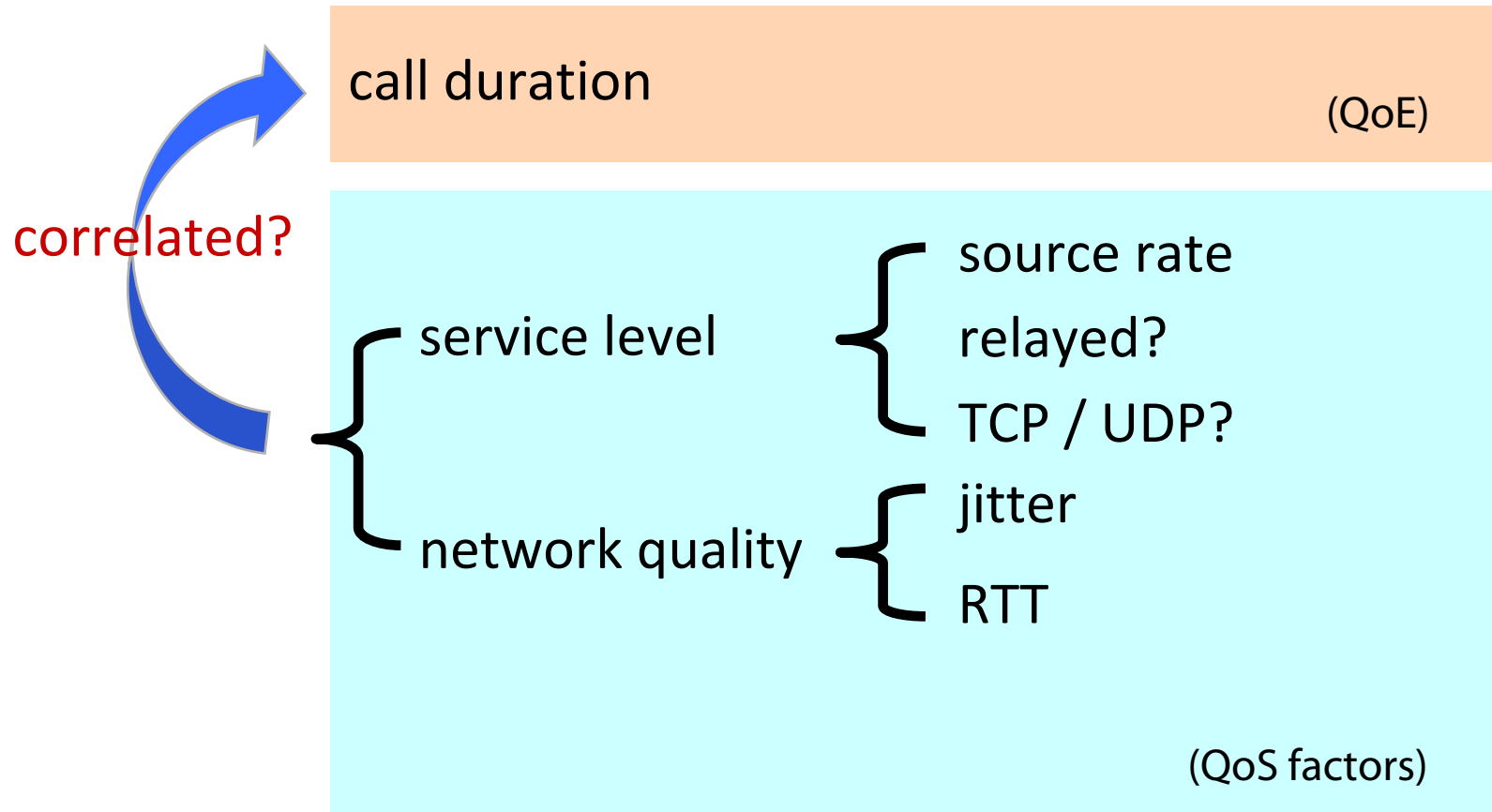
- The Question
- Measurement
- Modeling
- Validation
- Significance



The intuition behind our analysis

- The **conversation quality (i.e., QoE)** perceived by call parties is more or less related to the **call duration**
- The **network conditions** of a VoIP call are **independent** of
 - importance of talk content
 - call parties' schedule
 - call parties' talkativeness
 - other incentives to talk (e.g., free of charge)

First, getting a better sense

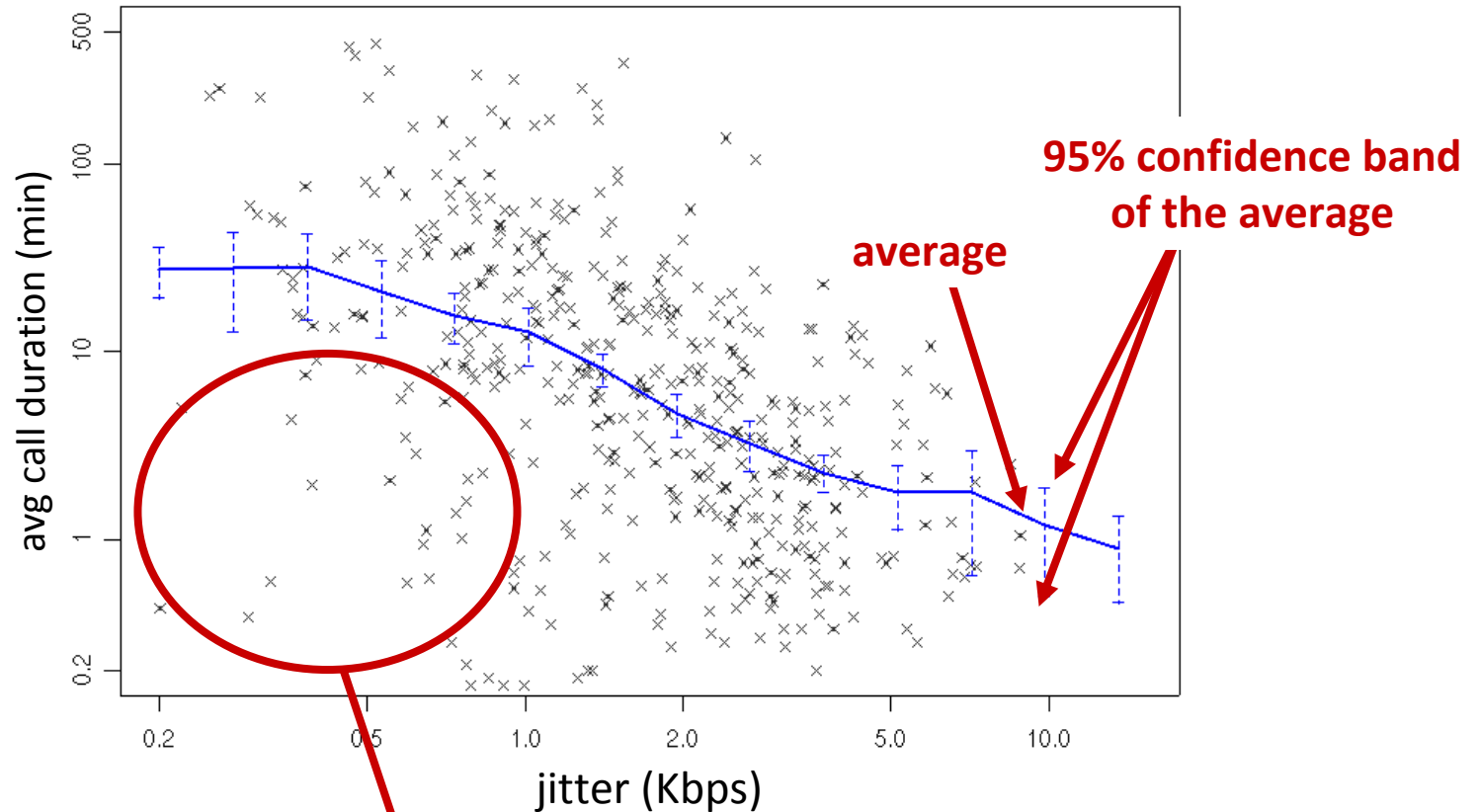


Is call duration related to each factor?

- For each factor
 - Scatter plot of the factor to the call duration
 - **See** whether they are positively, negatively, or not correlated
- Hypothesis tests
 - Confirm whether they are **indeed** positively, negatively, or not correlated

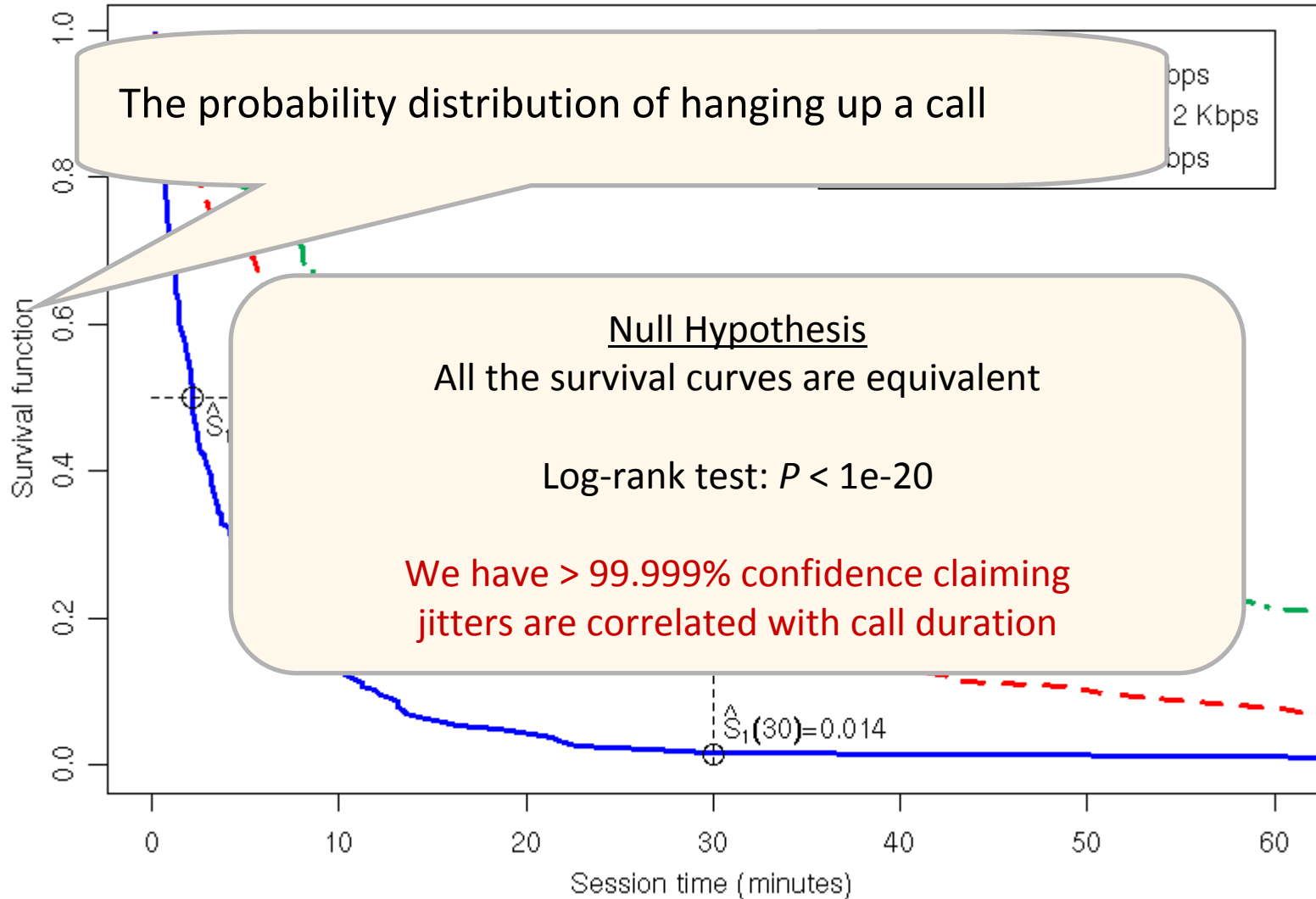
Call duration vs. jitter

(std dev of received bytes/sec)



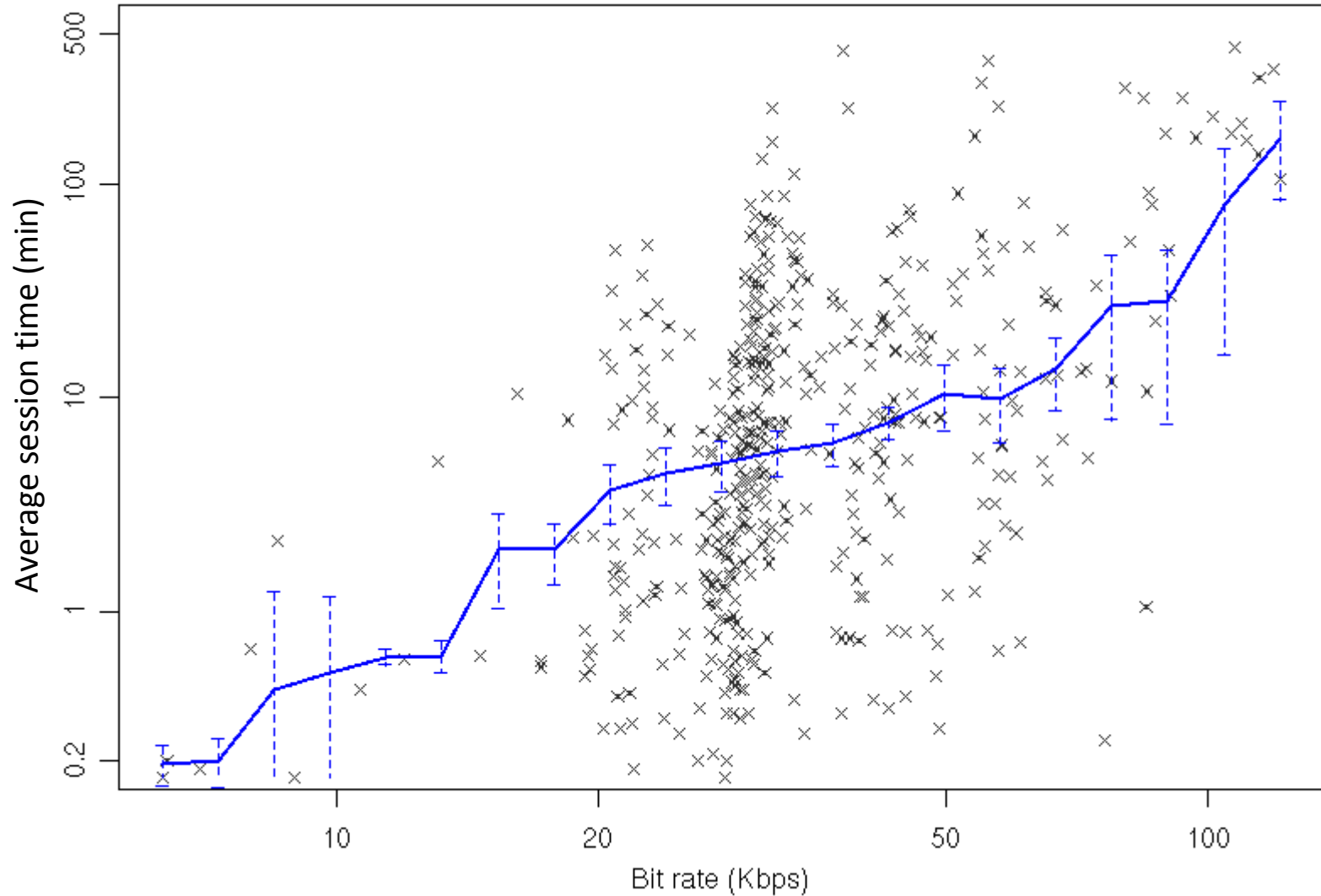
- There are short calls with low jitters
- The **average** shows a negative correlation between the 2 variables

Effect of Jitter – Hypothesis Testing

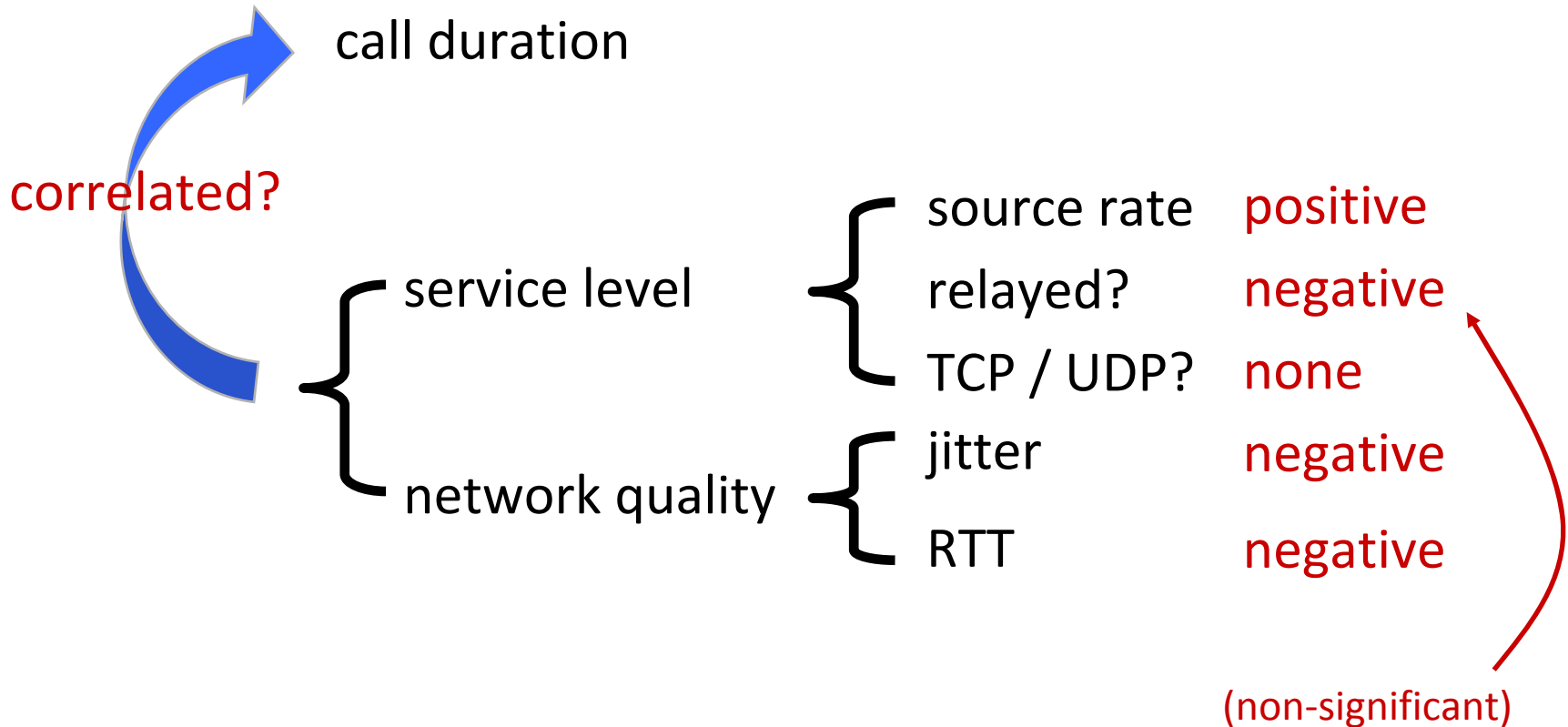


Effect of Source Rate

(the bandwidth Skype intended to use)



The better sense



Linear regression?

No!

Reasons

- Assumptions no longer hold
 - errors are not independent and not normally distributed
 - variance of errors are not constant
- Censorship
 - There are calls that have been going on for a while
 - There are calls that have not yet finished by the time we terminate tracing
 - We can't simply discard these calls
 - Otherwise we end up with a **biased set of calls with limited call duration**

Cox regression modeling

The Cox regression model provides a good fit

- the effect of treatment on patients' survival time
- log-hazard function is proportional to the weighted sum of factors

$$\log h(t|\mathbf{Z}) \propto \beta^t \mathbf{Z}$$

\mathbf{Z} : factors (bit rate=x, jitter=y, RTT=z, ...)

β : weights of factors

Hazard function (conditional failure rate)

The instantaneous rate at which failures occur for observations that have survived at time t

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t}$$

Functional Form Checks

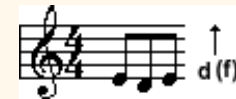
- Assumption $h(t|\mathbf{Z}) \propto \exp(\beta^t \mathbf{Z})$ must be conformed

- Example: Human beings are known sensitive to the **scale** of physical quantity rather than the **magnitude** of the quantity

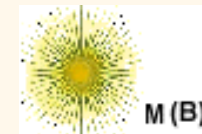
- Scale of sound (decibels vs. intensity)



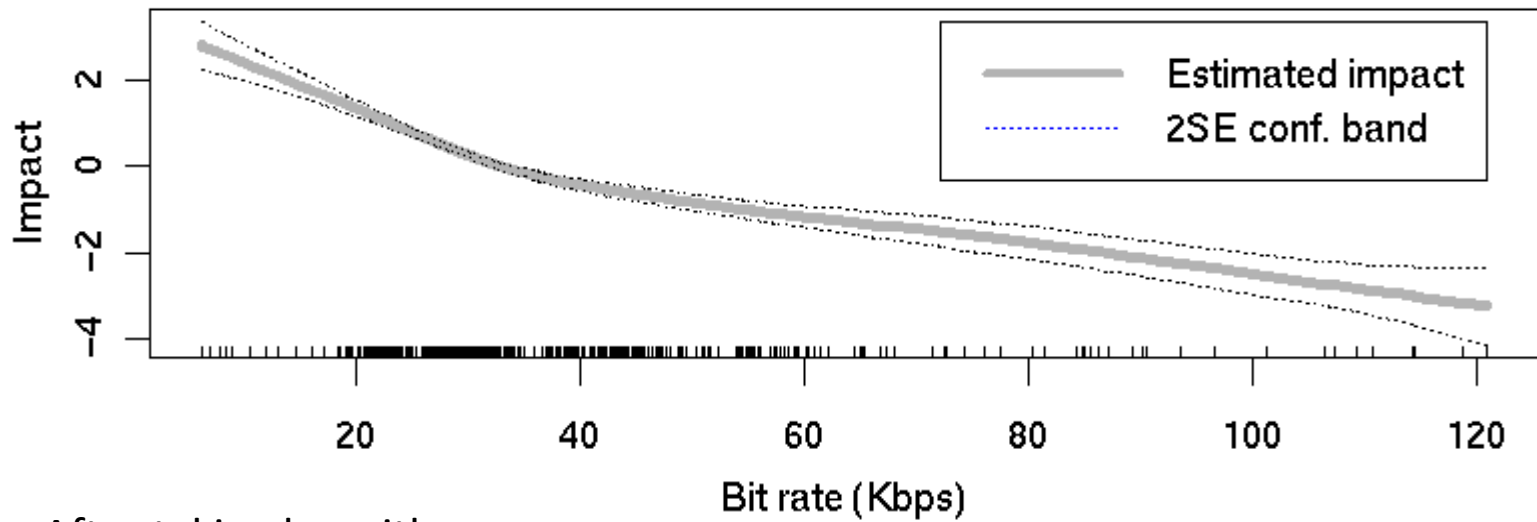
- Musical staff for notes (distance vs. frequency)



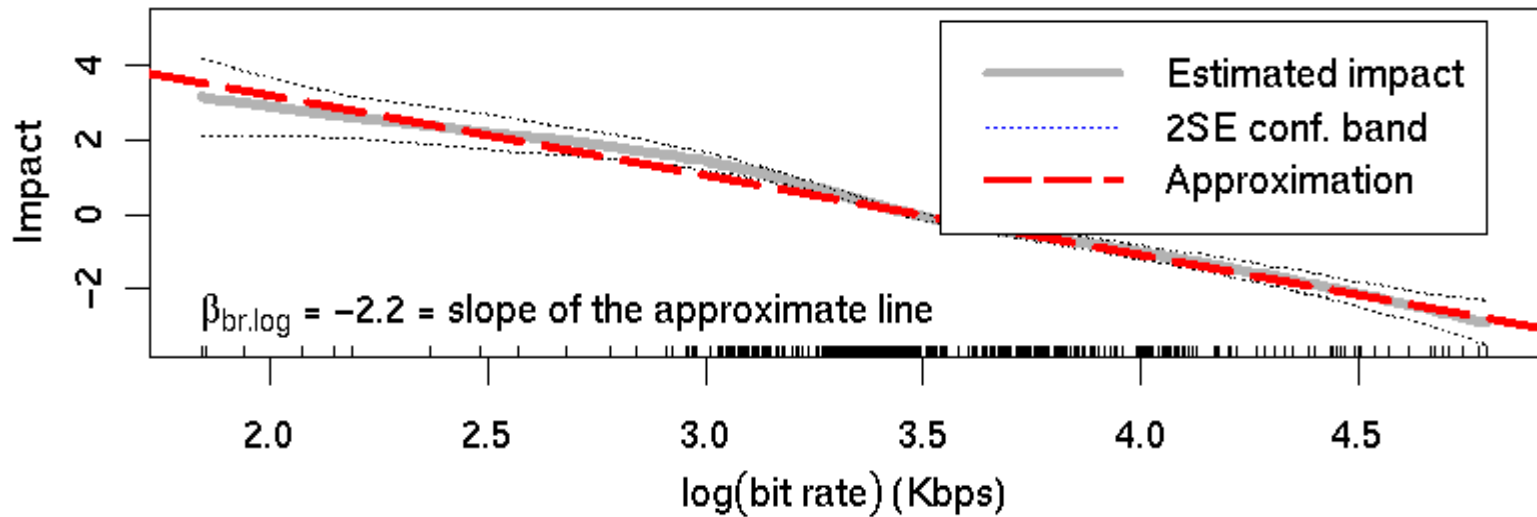
- Star magnitudes (magnitude vs. brightness)



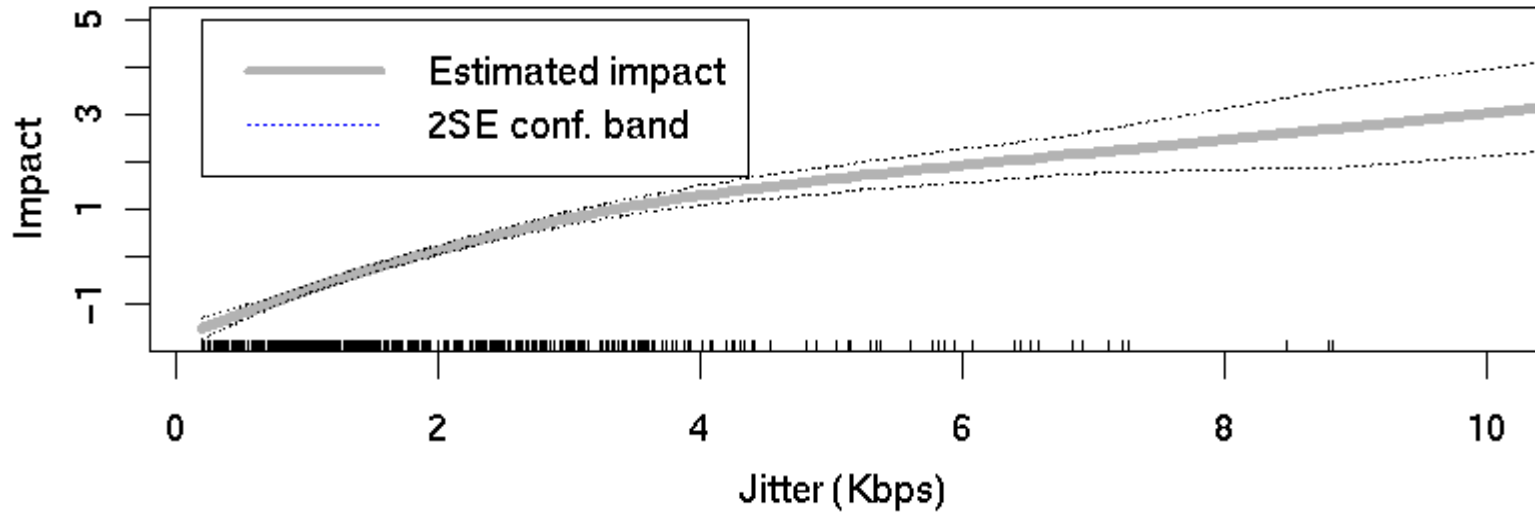
The Logarithm Fits Better (Bit rate)



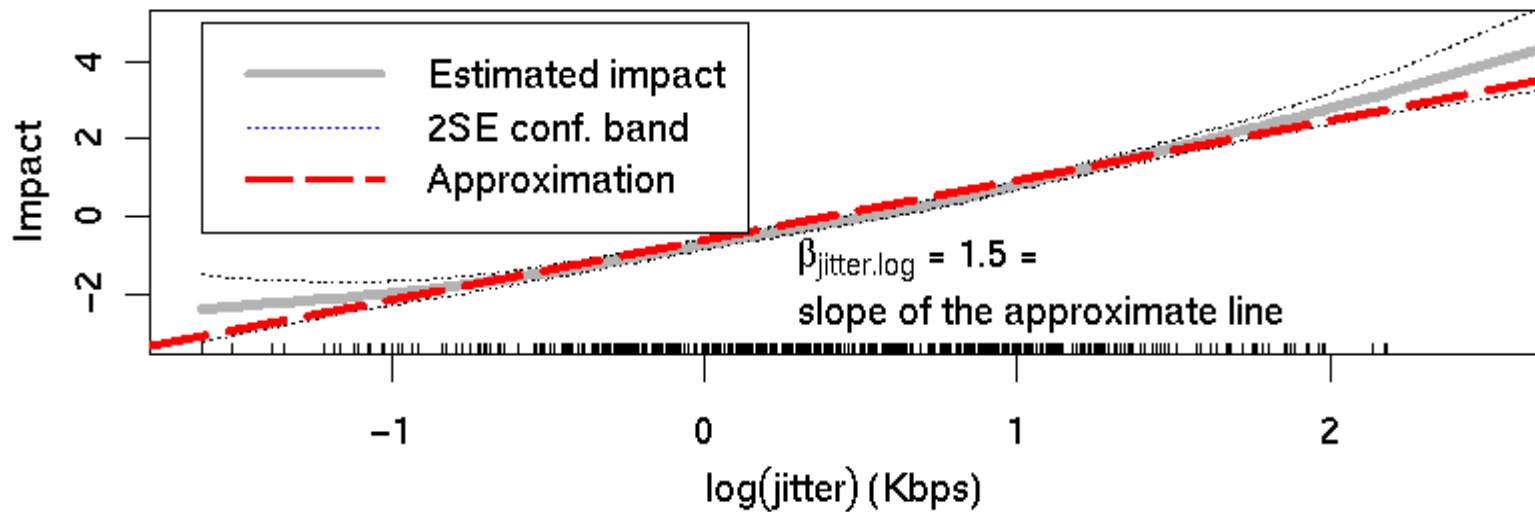
After taking logarithm ...



The Logarithm Fits Better (Jitter)



After taking logarithm ...



Final model & interpretation

variable	coef	std. err.	signif.
log(bit rate)	-2.15	0.13	< 1e-20
log(jitter)	1.55	0.09	< 1e-20
RTT	0.36	0.18	4.3e-02

Interpretation

A: bit rate = 20 Kbps

B: bit rate = 15 Kbps, other factors same as A

The hazard ratio between A and B can be computed by

$$\exp((\log(15) - \log(20)) \times -2.15) \approx 1.86$$

➔ The probability B will hang up is 1.86 times the probability A will do so at any instant.

Hang-up rate and USI

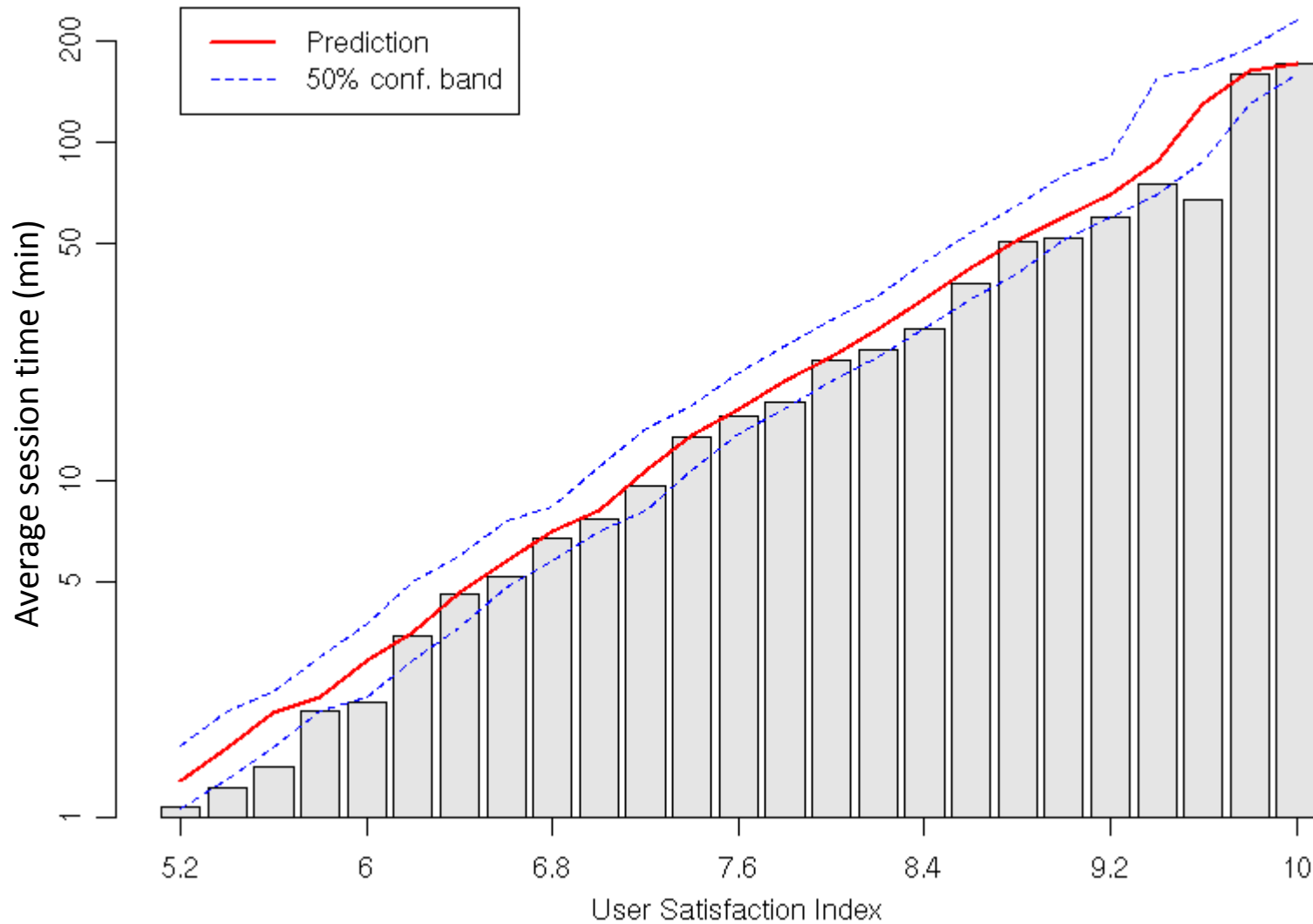
Hang-up rate =

$$2.15 \times \log(\text{bit rate}) - 1.55 \times \log(\text{jitter}) - 0.36 \times \text{RTT}$$

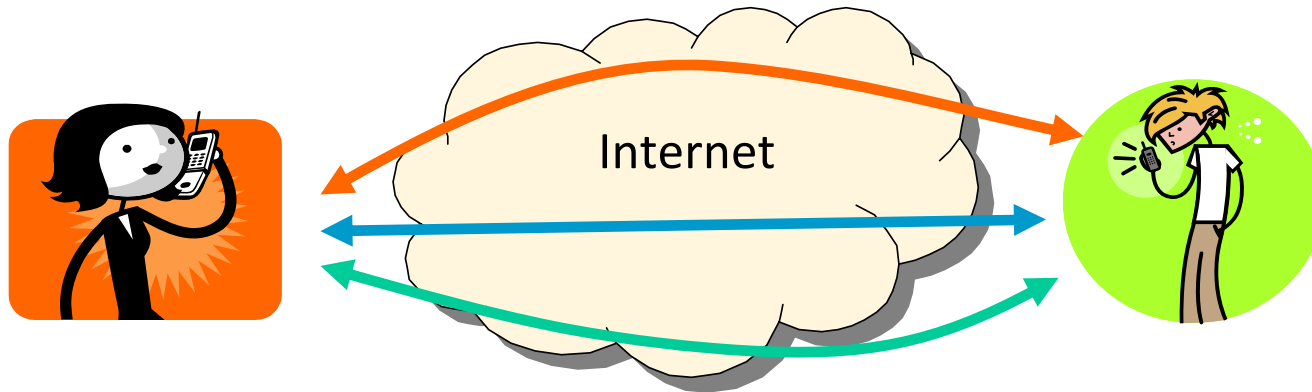
User satisfaction index (USI) =


–Hang-Up Rate

Actual and Predicted Time vs. USI



The multi-path scenario



path	avail bandwidth	jitter	RTT	USI
	10 Kbps	2 Kbps	100 ms	3.84
	20 Kbps	1 Kbps	300 ms	6.33
	30 Kbps	3 Kbps	500 ms	5.43



BUT,

is call hang-up rate a good indication of user satisfaction?

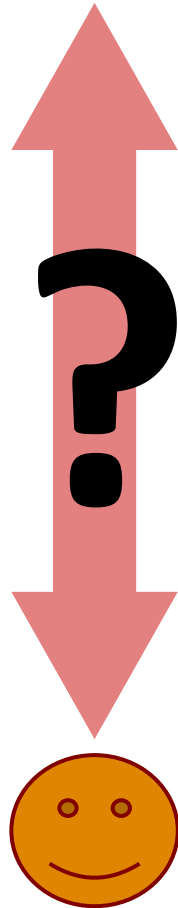
Talk outline

- The Question
- Measurement
- Modeling
- Validation
- Significance



User satisfaction: Validation

Call duration



intuition: call duration \leftrightarrow satisfaction
not confirmed yet

User satisfaction: One step further

Call duration



now we're going to check!

Speech interactivity



intuition: interactive and tight speech activities in a cheerful conversation



Identifying talk bursts

The problem

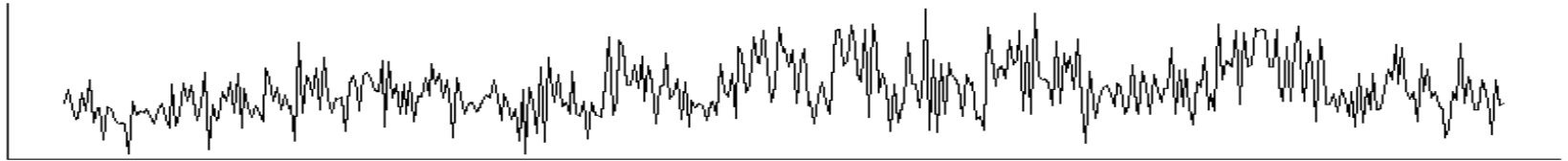
- Every voice packet is **encrypted** with 256-bit AES (Advanced Encryption Standard)

Possible solutions

- packet rate: **no silence suppression** in Skype
- **packet size: our choice**

What we need to achieve

- Input: a time series of packet sizes



- Output: estimated ON/OFF periods (ON = talk / OFF= silence)



Speech activity detection

1. Wavelet de-noising

Removing high-frequency fluctuations

2. Detect peaks and dips

3. Dynamic thresholding

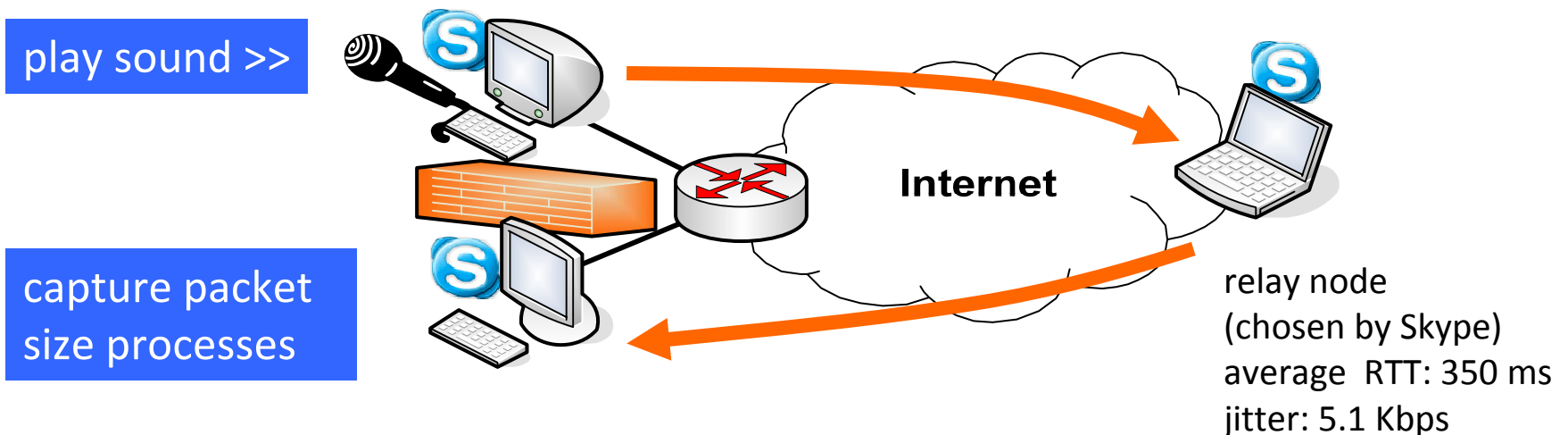
Deciding the beginning/end of a talk burst

Speech detection algorithm: Validation

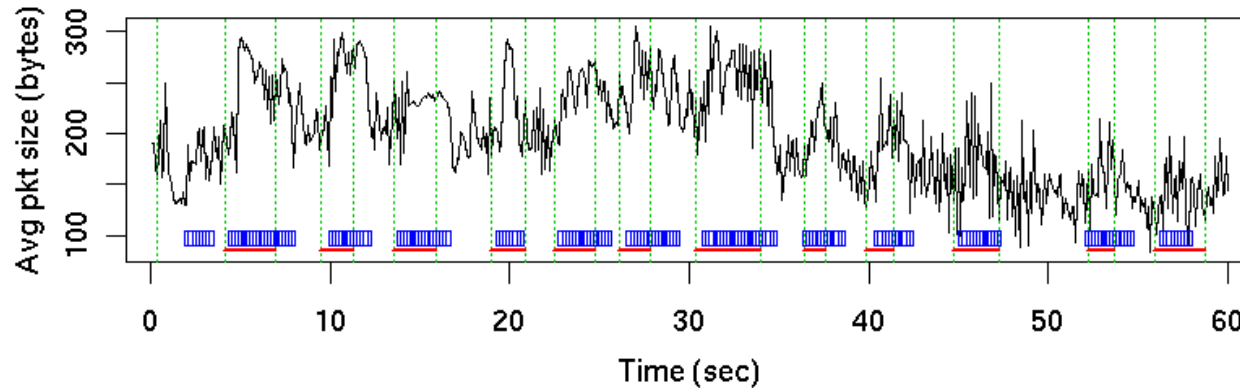
The speech detection algorithm is validated with:

- synthesized sin waves (500 Hz – 2000 Hz)
- real speech recordings

Force packet size processes contaminated by serious network impairment (delay and loss)

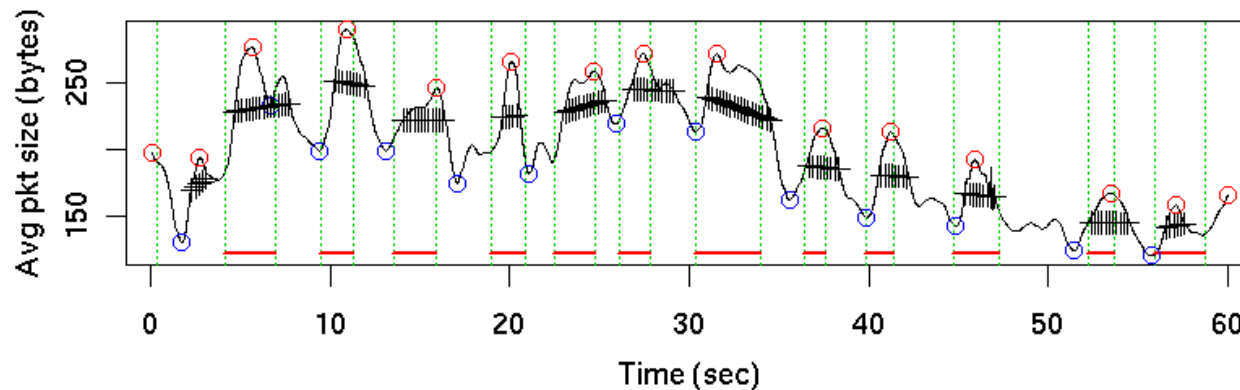


Validation with synthesized sin waves



— true ON periods
— estimated ON periods

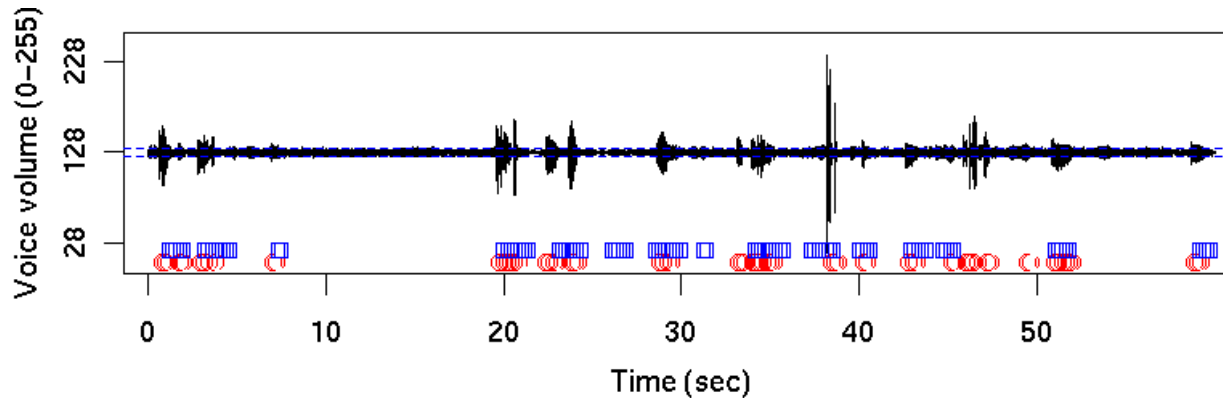
(a) Original packet size process



(b) Wavelet denoised process with estimated ON periods

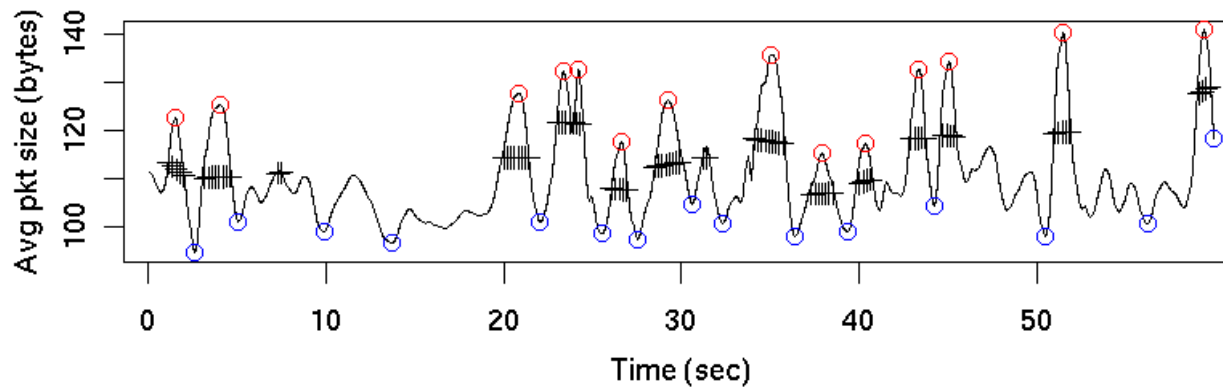
- 3 times for each of 10 test cases
- correctness (ratio of matched 0.1-second periods): 0.73 – 0.92

Validation with speech recordings



— true ON periods
— estimated ON periods

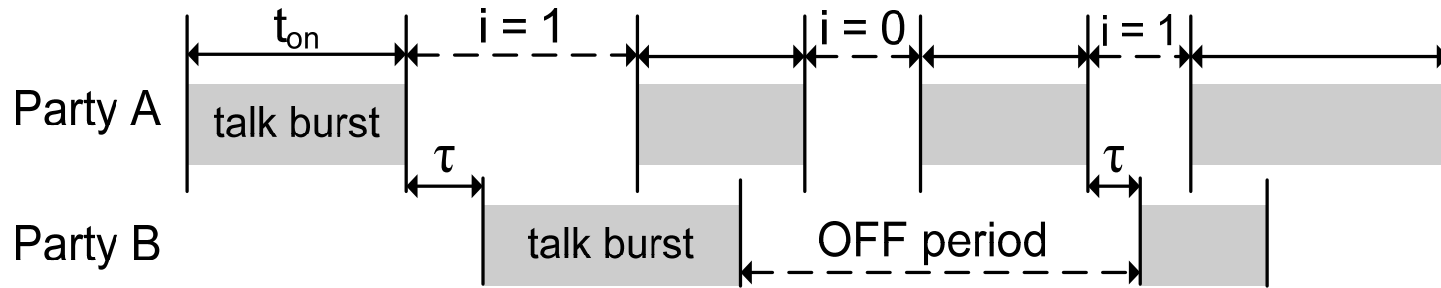
(a) Human speech recording



(b) Wavelet denoised process with estimated ON periods

- 3 times for each of 3 test cases
- correctness (ratio of matched 0.1-second periods): 0.71 – 0.85

Speech interactivity analysis



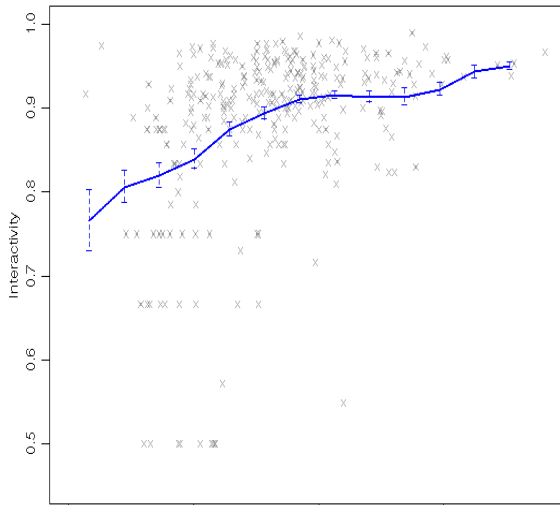
Responsiveness:	$\text{count}(i = 1) / \text{count}(i = 1 \text{ or } i = 0)$
Avg. Response Delay:	$\text{mean}(\tau)$
Avg. Burst Length:	$\text{mean}(t_{on})$

Responsiveness: whether the other party responds

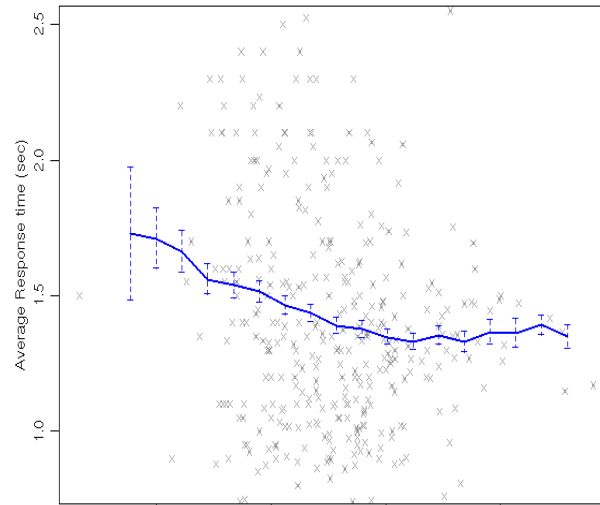
Response delay: how long before the other party responds

Burst length: how long does a speech burst last

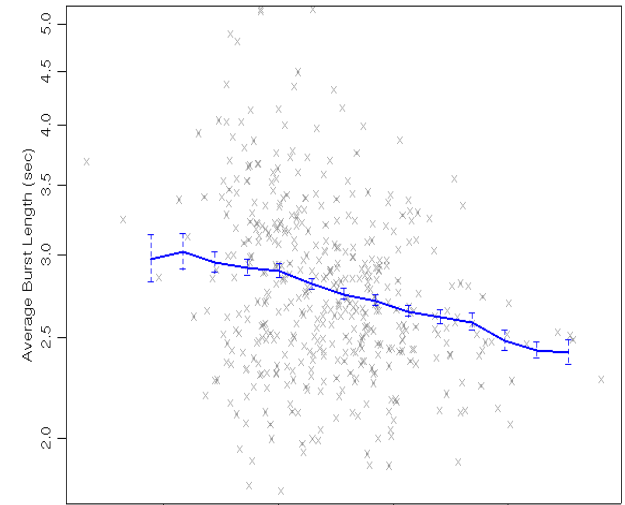
USI vs. Speech interactivity



higher USI →
higher responsiveness



higher USI →
shorter response delay



higher USI →
shorter burst length

All are statistically significant (at 0.01 significance level)

Speech interactivity in conversation supports the proposed
USI

Talk outline

- The Question
- Measurement
- Modeling
- Validation

 ■ Significance

Implications

- should put more attention to delay jitters (rather than focus on network delay only)
- and the encoding bit rate!

Significance

QoE-aware systems that can optimize user experience **in run time**

- Is it worth to sacrifice 20 ms latency for reducing 10 ms jitters (say, with a de-jitter buffer)?
- Pick the most appropriate parameters **in run time**
 - playout scheduling (buffer time)
 - coding scheme (& rate)
 - source rate
 - data path (overlay routing)
 - transmission scheme (redundancy, erasure coding, ...)

Future work (1)

Measurement

- larger data sets (p2p traffic is hard to collect)
- diverse locations

Validation

- user studies
- comparison with existing models (PESQ, etc)

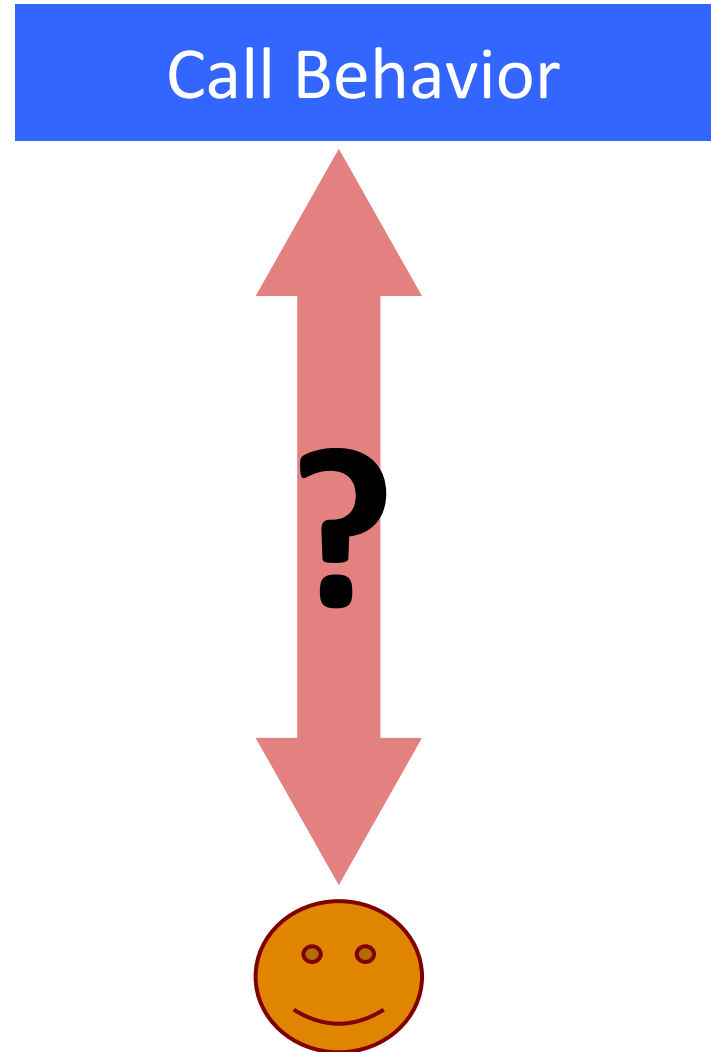
Future work (2)

Beyond “call duration”

- Who hangs up a call?
- Call disconnect-n-connect behavior

More sophisticated modeling

- Voice codec
- Pricing effect
- Time-of-day effect
- Time-dependent impact



Thank You!

Sheng-Wei (Kuan-Ta) Chen

<http://www.iis.sinica.edu.tw/~swc>

