

Inferring Speech Activity from Encrypted Skype Traffic

Yu-Chun Chang, Kuan-Ta Chen, Chen-Chi Wu, and
Chin-Laung Lei

Oct. 27, 2008

Outline

- Introduction
- Data description
- Proposed scheme
- Performance evaluation
- Conclusion

Introduction

- VAD (Voice Activity Detection)
 - The algorithm to extract the presence or absence of human speech in speech processing.
- Source-level VAD
 - Audio signal
 - Silence suppression
- Network-level VAD
 - Network traffic
 - Flow identification, QoS measurement

- The differences between source-level and network-level VAD

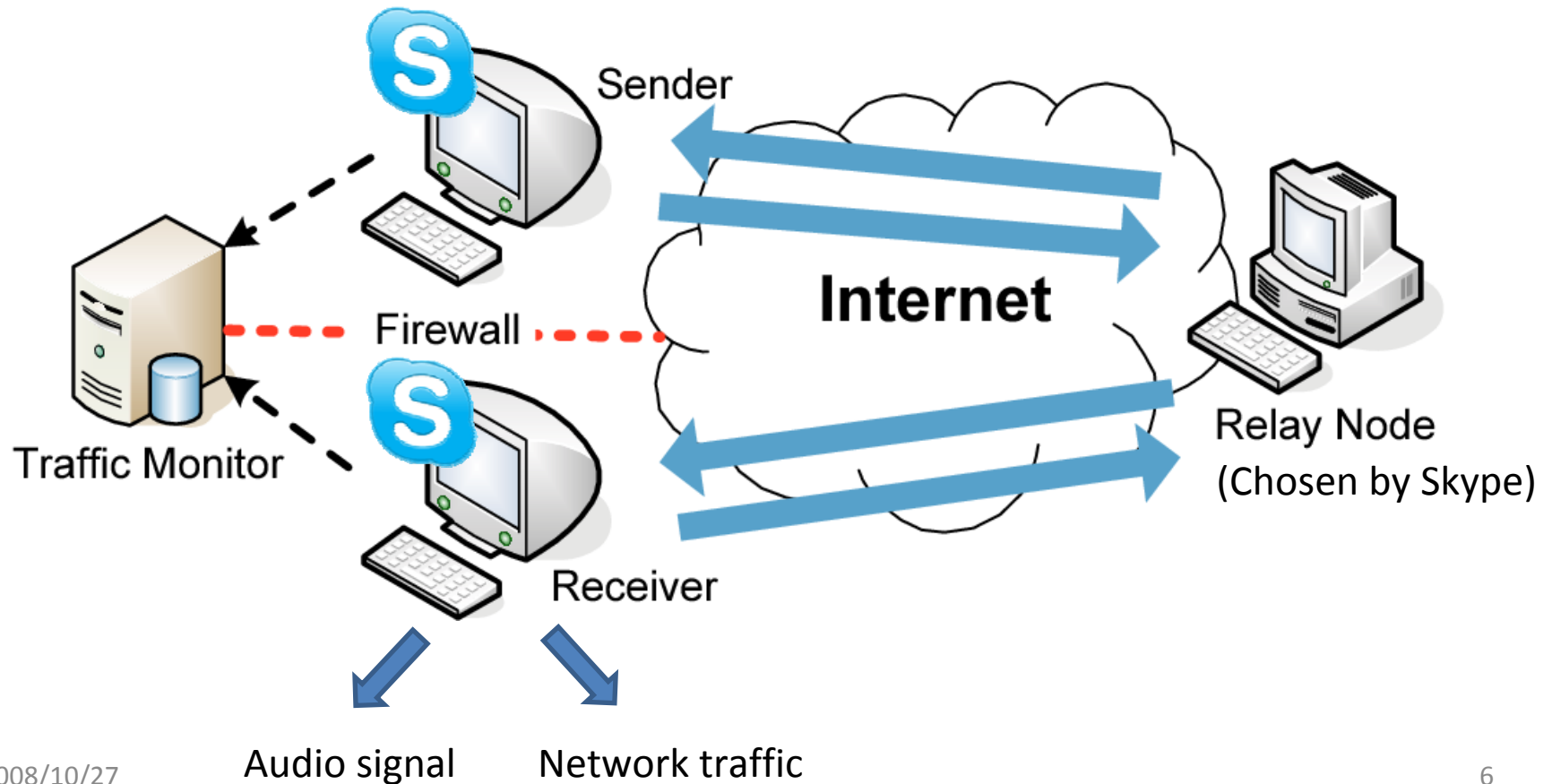
	source-level	network-level
input	audio signal	network traffic
location	speaker's host	network node
purpose	silence suppression echo cancellation	traffic management QoS measurement

Introduction (contd.)

- Challenges
 - Payload encryption
 - Skype do not support silence suppression
- Contribution
 - We propose a network-level VAD that can infer speech activity from encrypted and non-silence-suppressed VoIP traffic.

Data Description

- Experiment setup



Data Description (contd.)

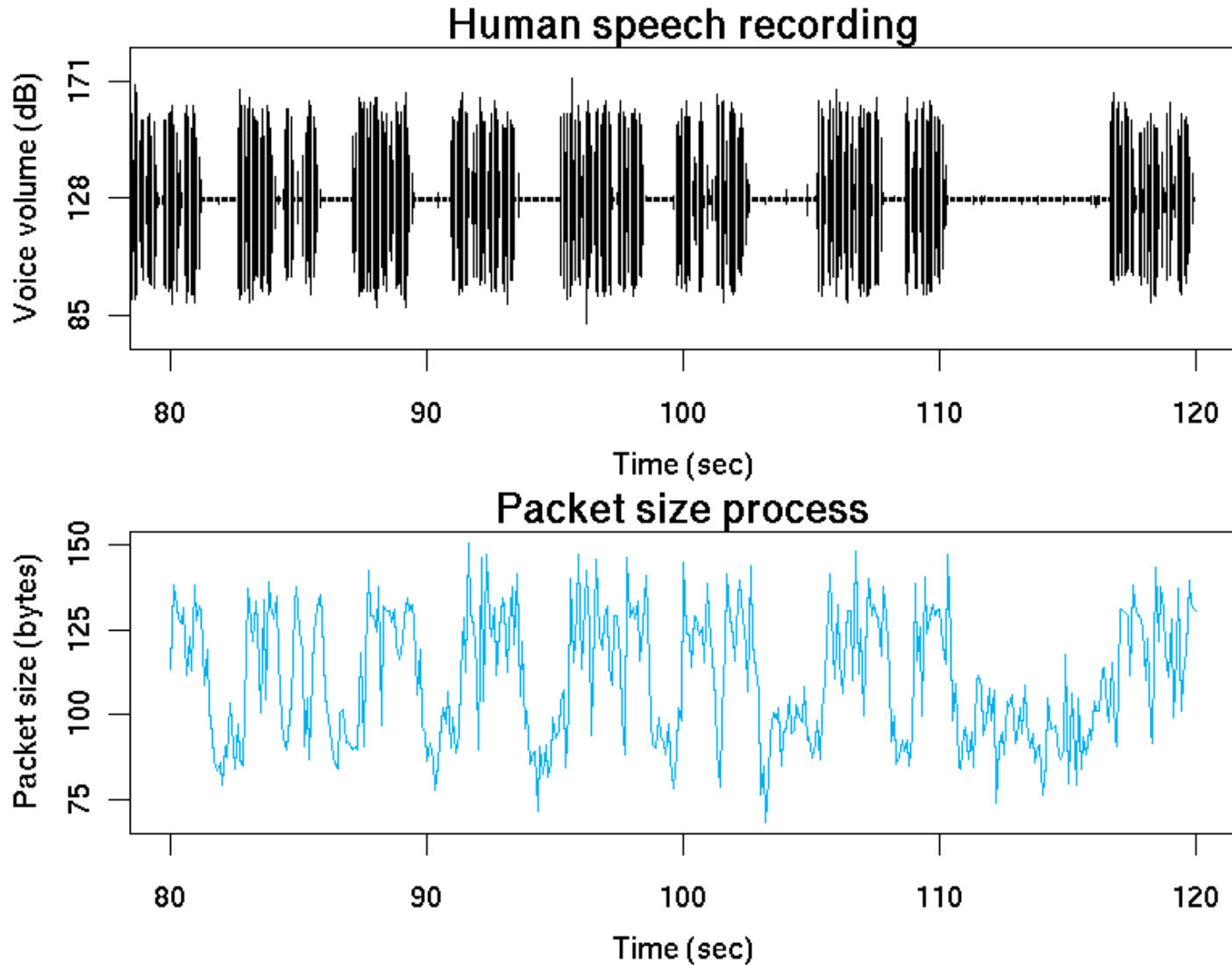
- Trace summary

Total # of traces	# TCP	# UDP
1839	1427	412
# Relay node	Mean packet size	Mean time period
1677	109.6 bytes	612.5 sec

Proposed Scheme

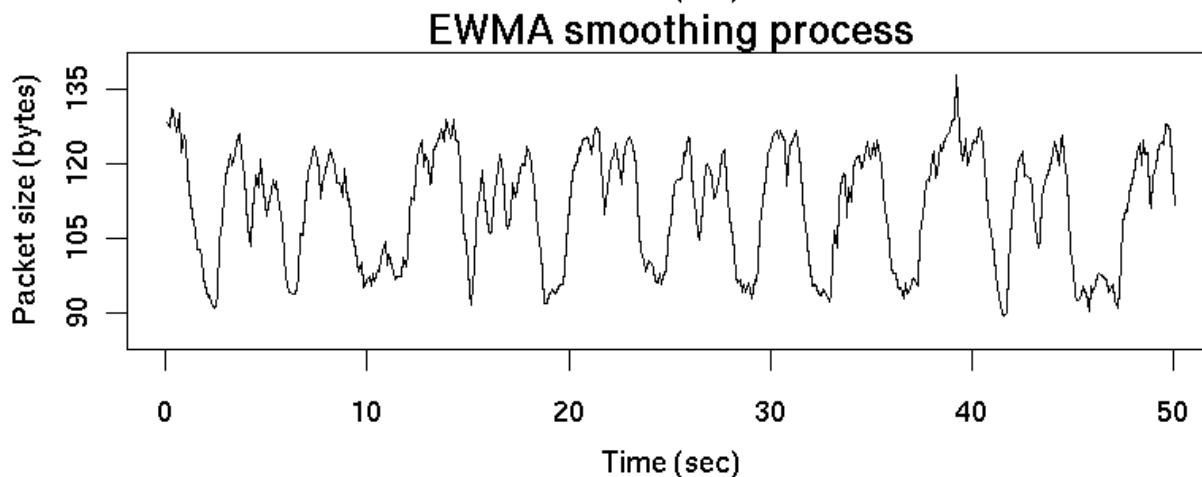
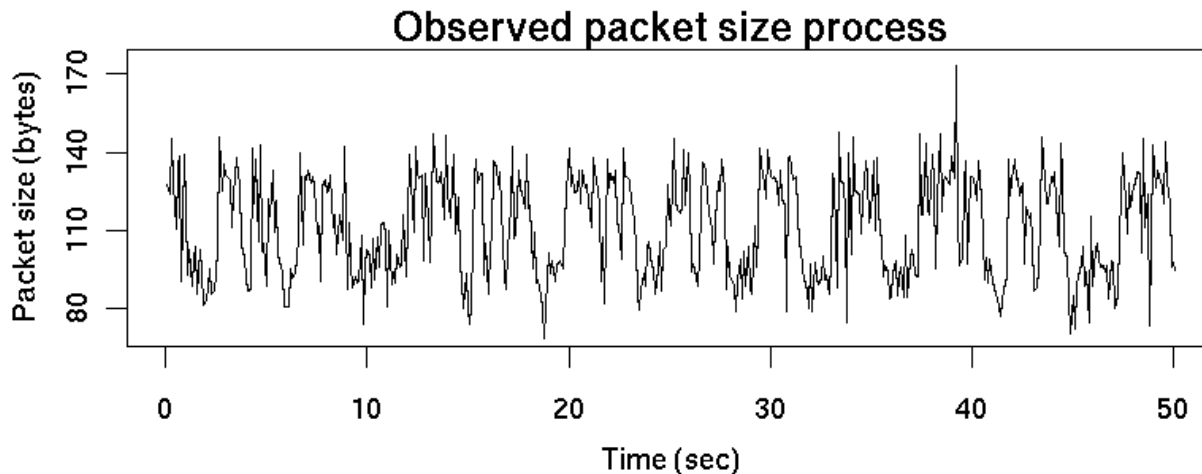
- The indicator of voice activity – packet size
- Smoothing
- Adaptive thresholding

The indicator of voice activity – Packet size



Smoothing

- EWMA (Exponentially Weighted Moving Average)



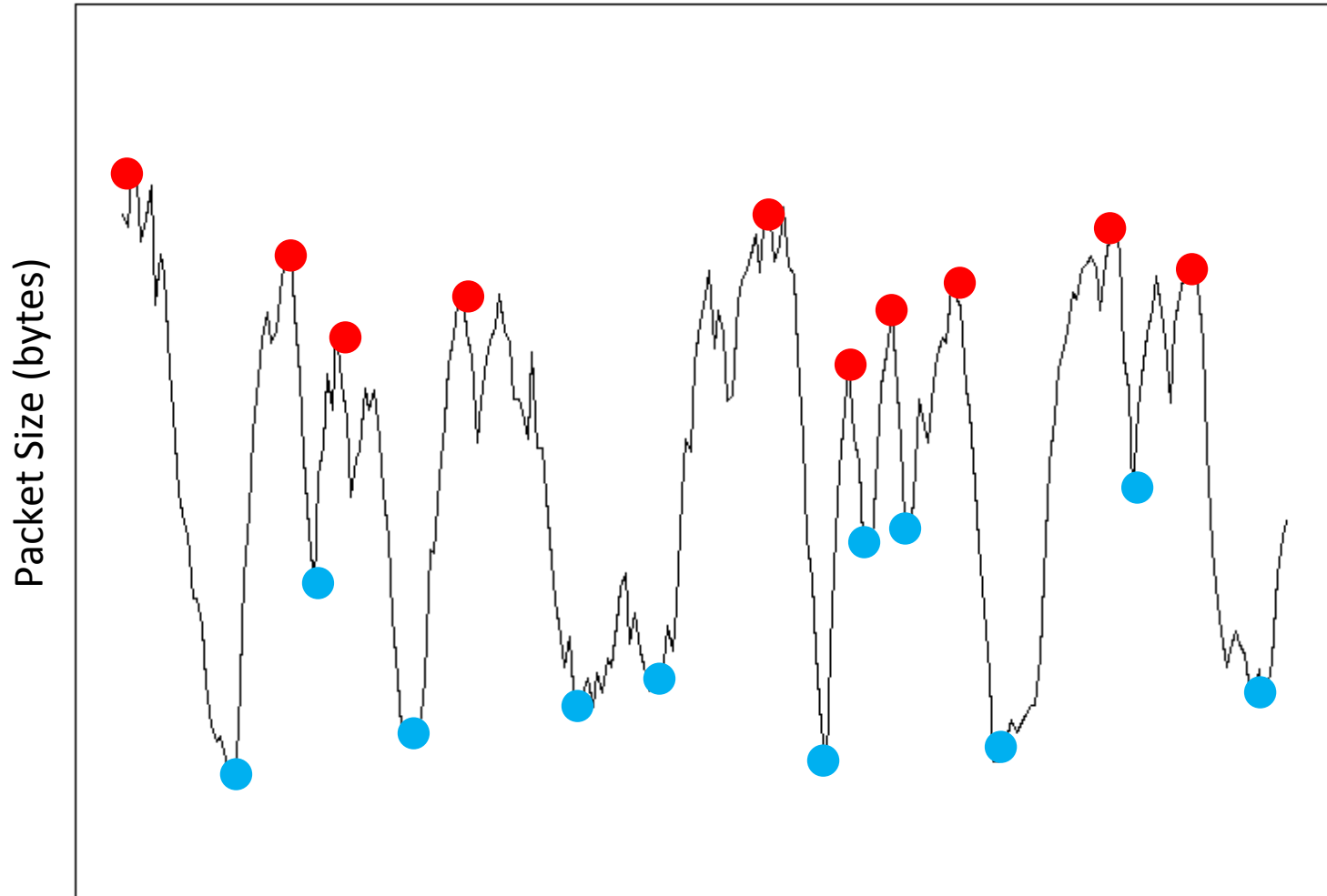
EWMA :

$$P_i = \lambda Y_i + (1 - \lambda) P_{i-1}$$

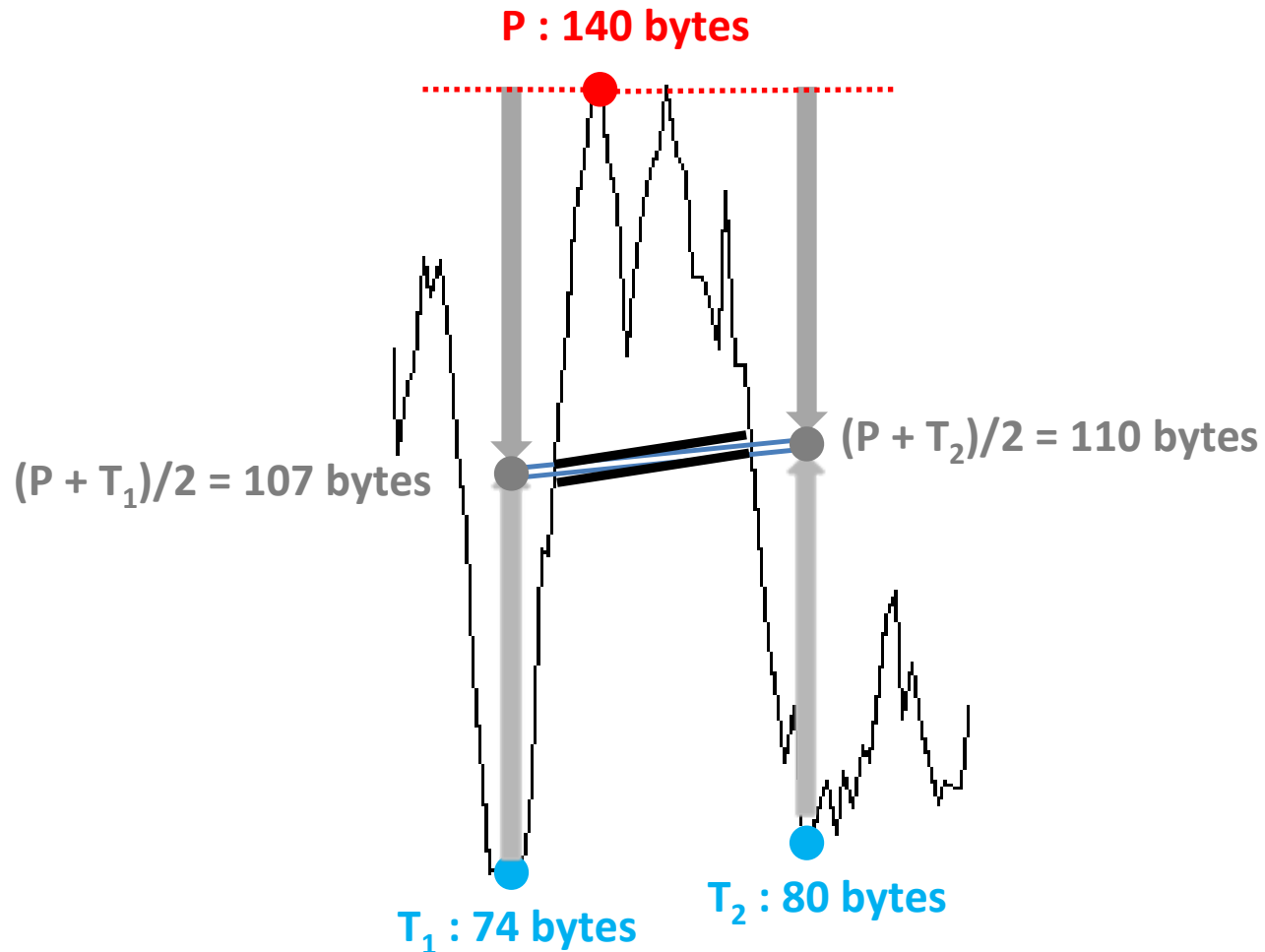
($\lambda = 0.2$)

Y : Observed packet size
P : Smoothed packet size

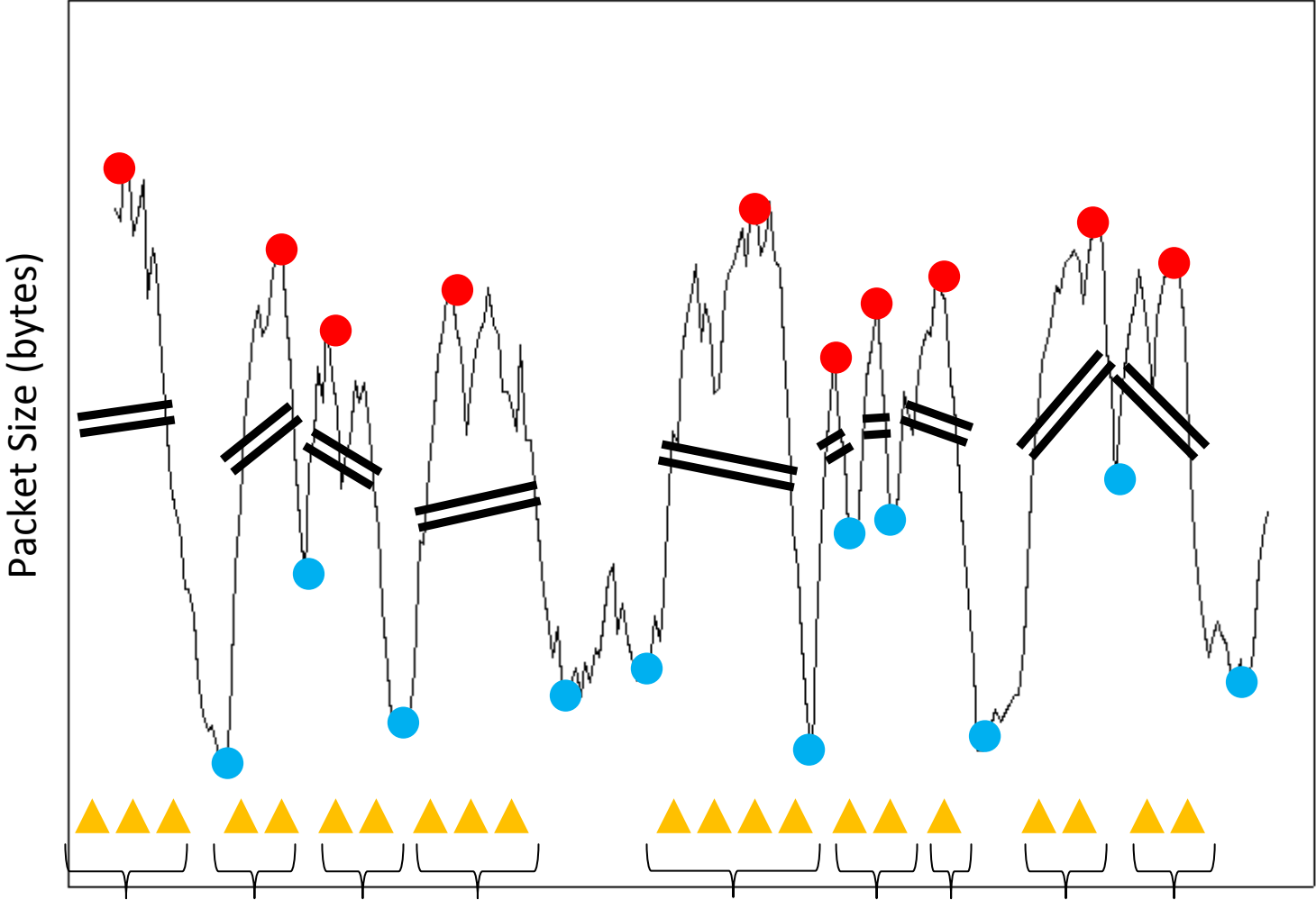
Adaptive thresholding



Adaptive thresholding (contd.)



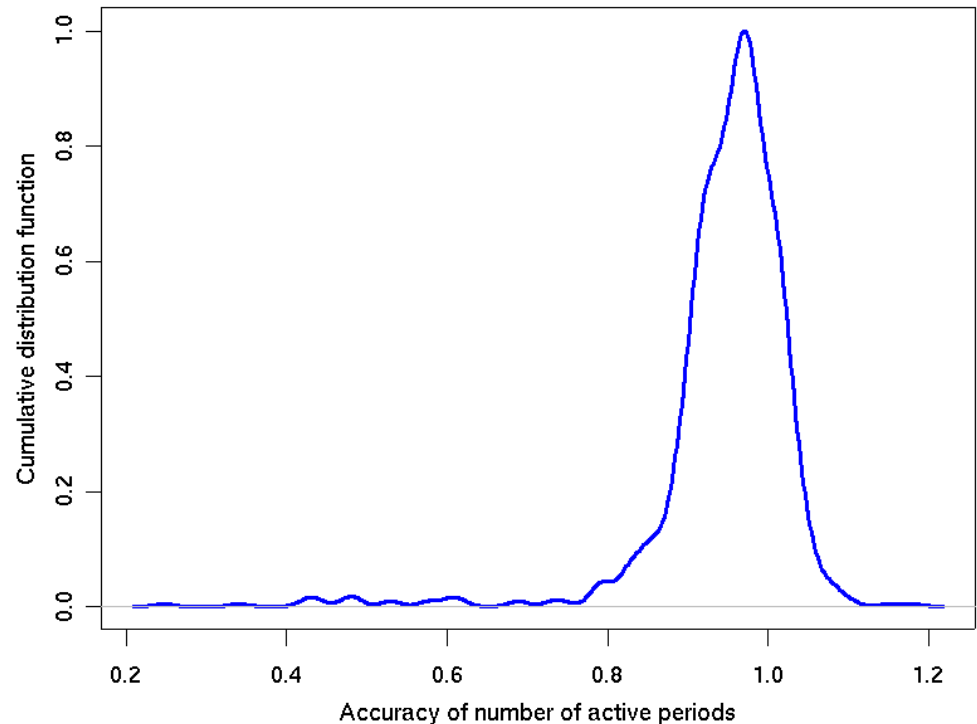
Adaptive thresholding (contd.)



Performance Evaluation

- Number of ON periods

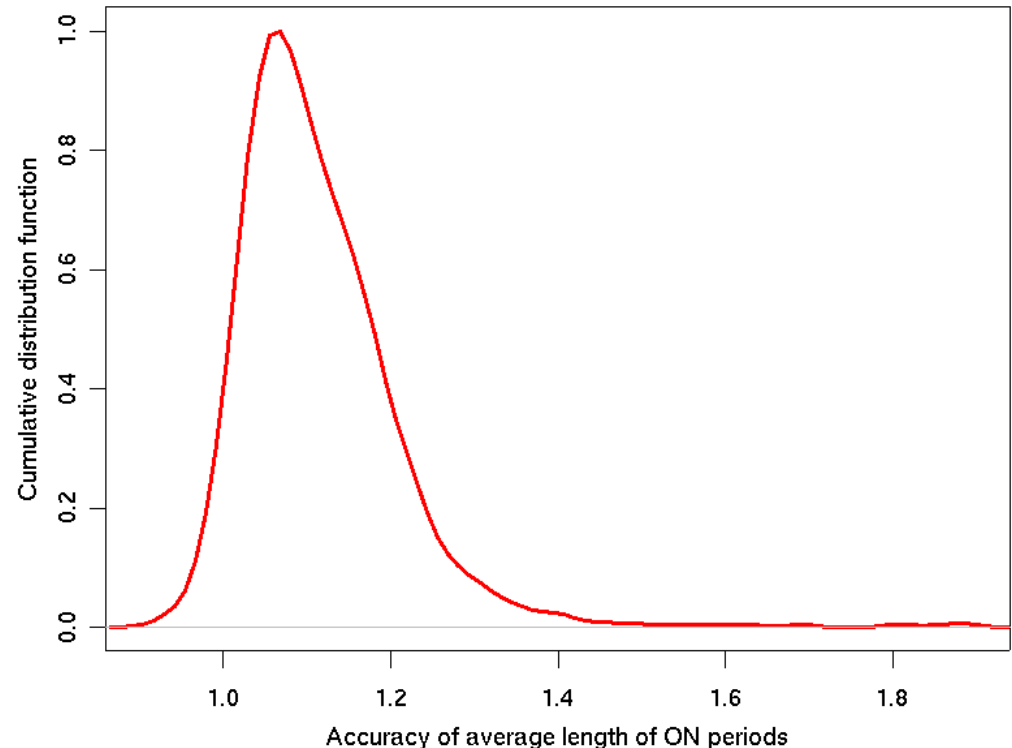
$$\frac{\text{Number_of_estimated_ON_periods}}{\text{Number_of_true_ON_periods}}$$



Performance Evaluation (contd.)

- Average length of ON periods

$$\frac{\text{Mean_length_of_estimated_ON_periods}}{\text{Mean_length_of_true_ON_periods}}$$



Performance Evaluation (contd.)

- State correctness

$$\frac{|M \text{ _and _} N|}{|M \text{ _or _} N|}$$

ON period -> 1
OFF period -> 0

True speech activity (M) : 0 0 1 1 1 0 0 0 1 1 0 1 1 0 0 0 0 0 1 1 1 1 1 0

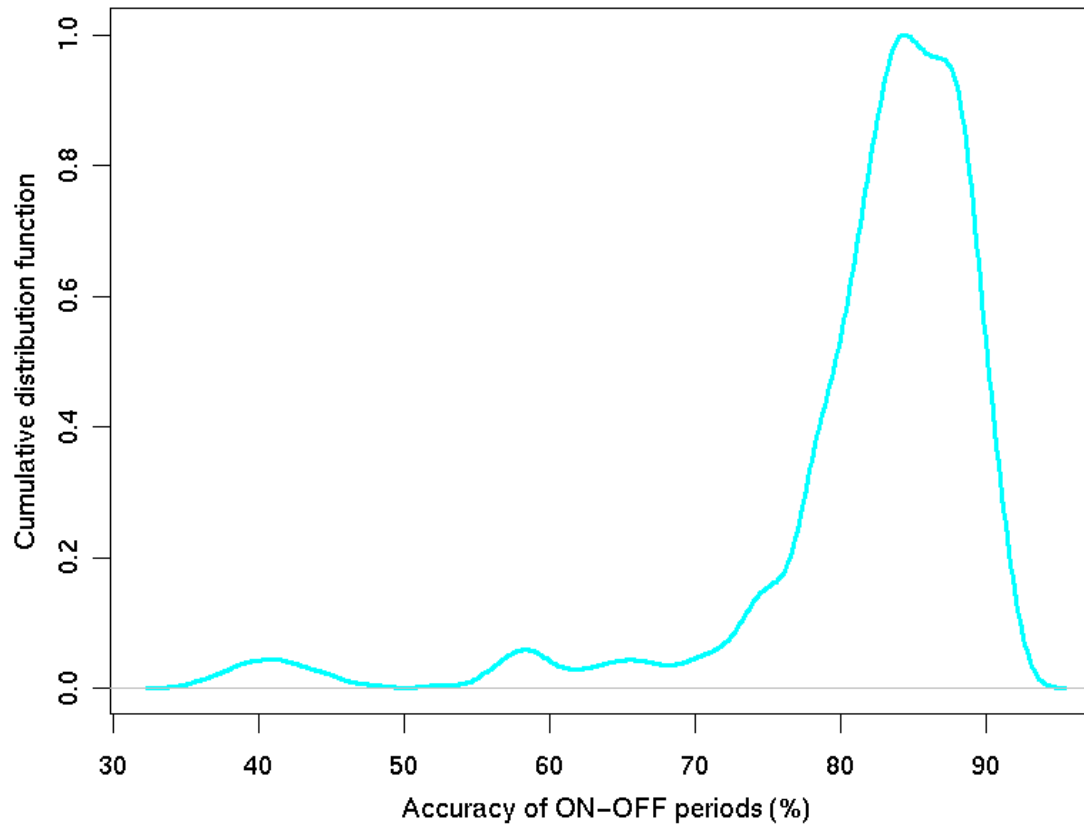
Estimated speech activity (N): 0 1 1 1 1 0 0 0 1 0 0 1 1 0 0 0 0 1 1 1 0 1 1 1

M and N: 0 0 1 1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 1 0 1 1 0

M or N: 0 1 1 1 1 0 0 0 1 1 0 1 1 0 0 0 0 1 1 1 1 1 1 1

Performance Evaluation (contd.)

- State correctness



Conclusion

- We propose the *network-level VAD* which infers speech activity from network traffic instead of audio signal.
- We propose a *VAD* algorithm that can extract voice activity from encrypted and non-silence-suppressed VoIP network traffic.

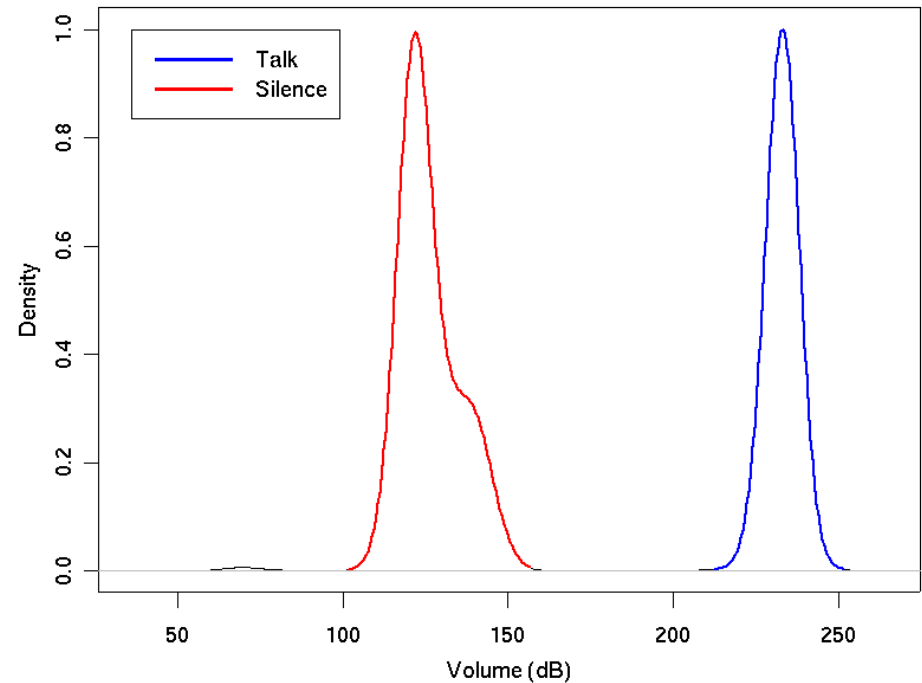
- Thanks

Backup slides

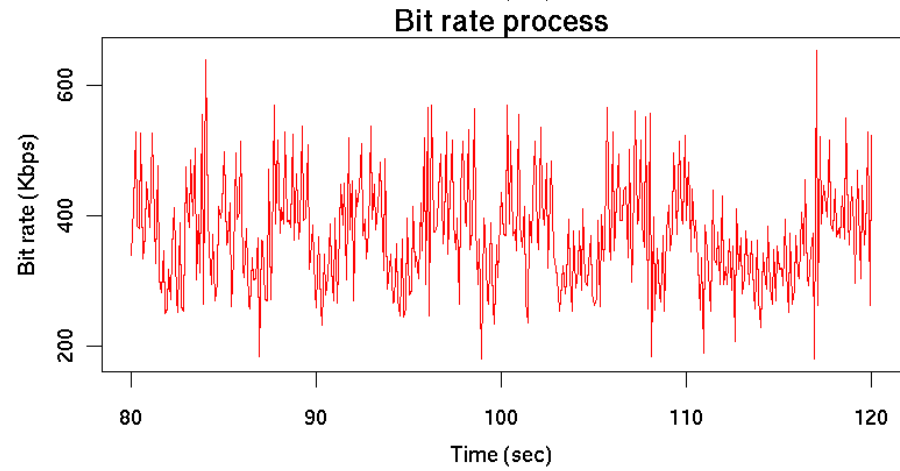
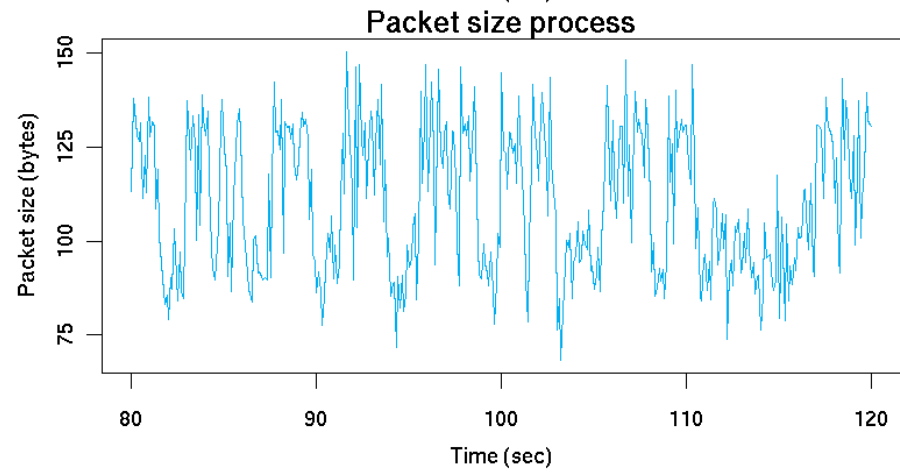
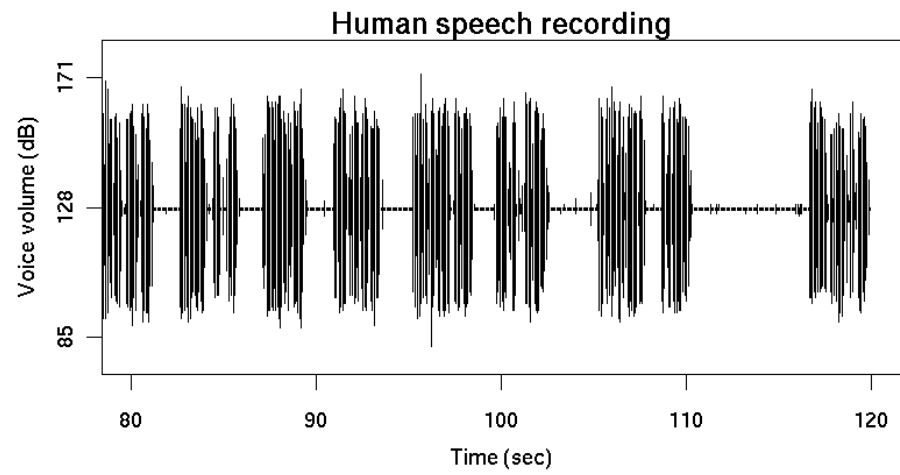
VAD on audio signaling

$$volume = 10 * \log\left(\sum_i S_i^2\right)$$

Static threshold : 183 db



J.-S. R. Jang, "Audio signal processing and recognition,"
<http://www.cs.nthu.edu.tw/jang>

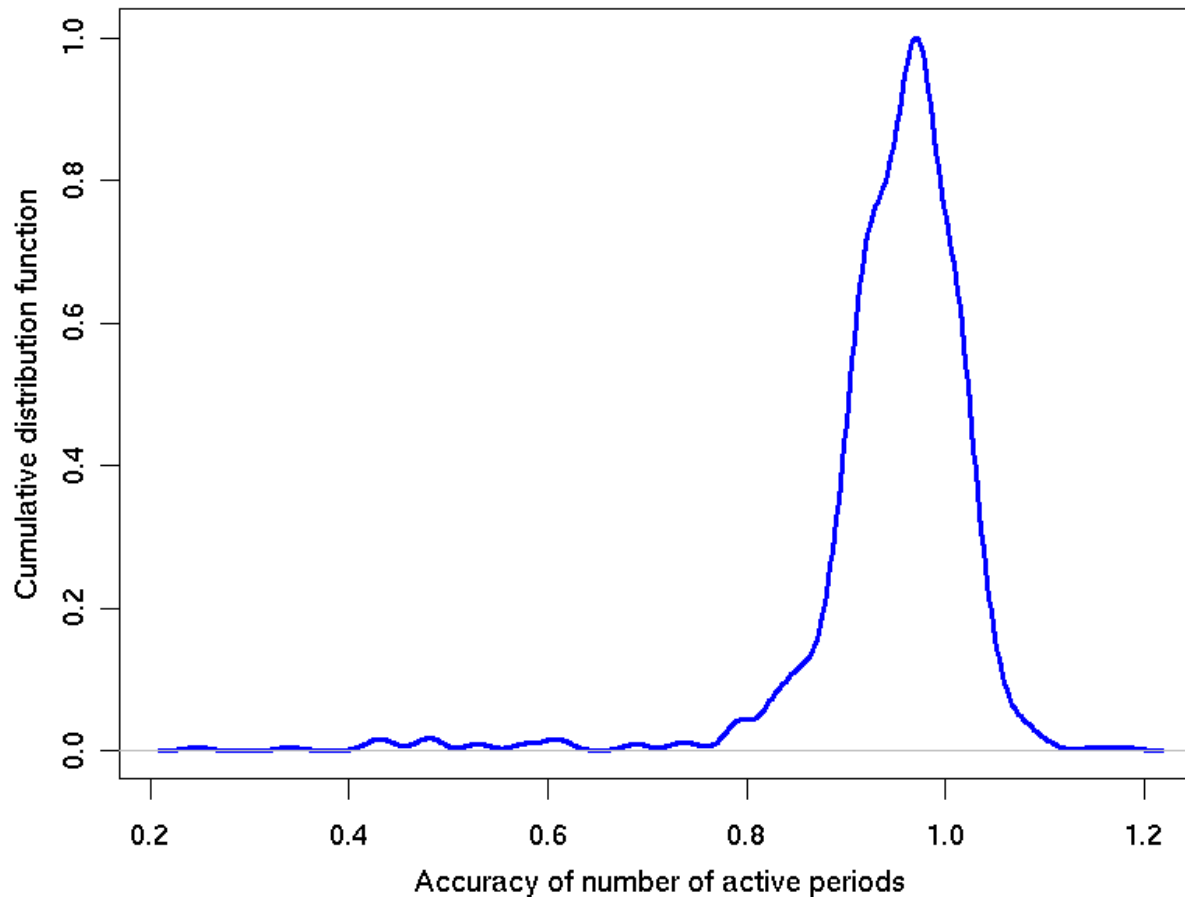


I am a student of National Taiwan University.



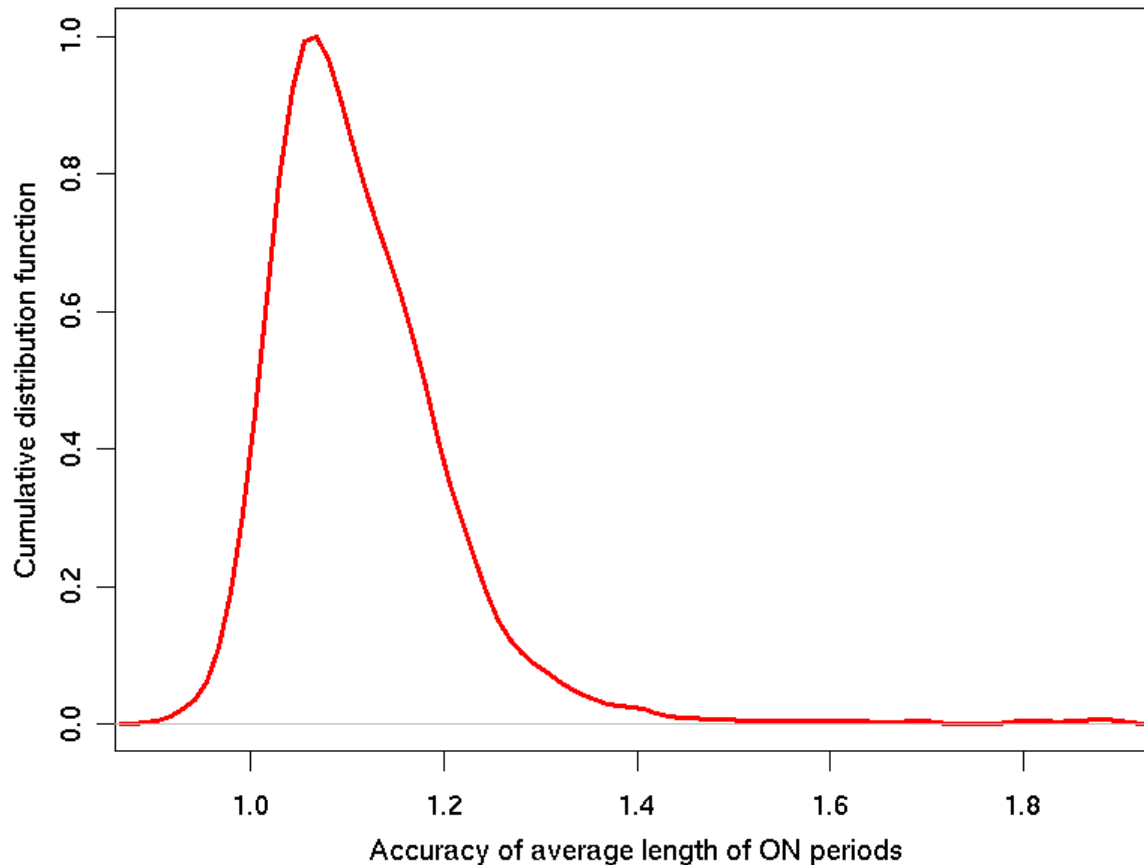
Performance Evaluation (contd.)

- Number of ON periods



Performance Evaluation (contd.)

- Average length of ON periods



Performance Evaluation (contd.)

- State correctness

$$\frac{|M \cap N|}{|M \cup N|}$$

True speech activity (M) : 0 0 1 1 1 0 0 0 1 1 0 1 1 0 0 0 0 0 0 1 1 1 1 1 0

Estimated speech activity (N): 0 1 1 1 1 0 0 0 1 0 0 1 1 0 0 0 0 0 1 1 1 0 1 1 1