Ilana Volfin

Dominant Speaker Identification for Multipoint Videoconferencing

Under the supervision of
Prof. Israel Cohen

Outline

- Videoconferencing – an introduction
- Discussed realization
- Proposed method
  - Speech activity score evaluation
    - Single observation
    - Sequence of observations
  - SNR estimation
- Experimental Results

Videoconferencing

Used for :

- *Remote education*
- *Medical consulting*
- *Business meetings*
- *Personal communications*

Videoconferencing

History: [1]

1964 — 1982

AT&T
'Picturephone'
Not commercialized

Compressior
System cost: 250,00
Line cost per hour :

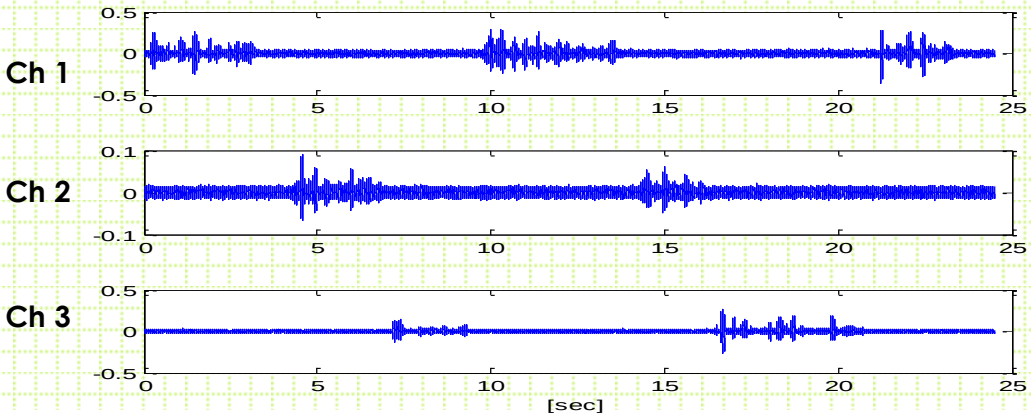
Tel
)00$
ur : 100$

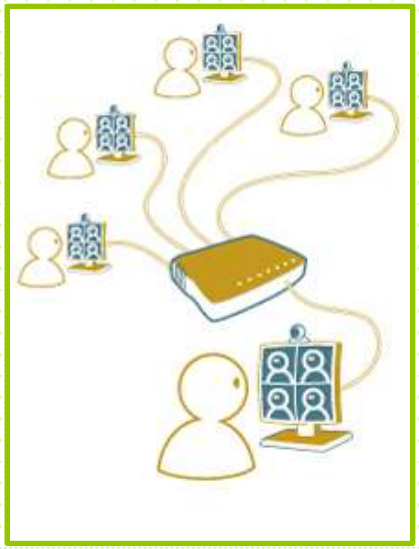
Today

[1] Sprey, "Videoconferencing as a Communication Tool",TRANSACTIONS ON PROFESSIONAL COMMUNICATION, IEEE, 1997

What is a Multipoint Conference?

- 3 or more participants
- Each at his own location
- Single microphone and video camera
- The information is administered through a central control unit

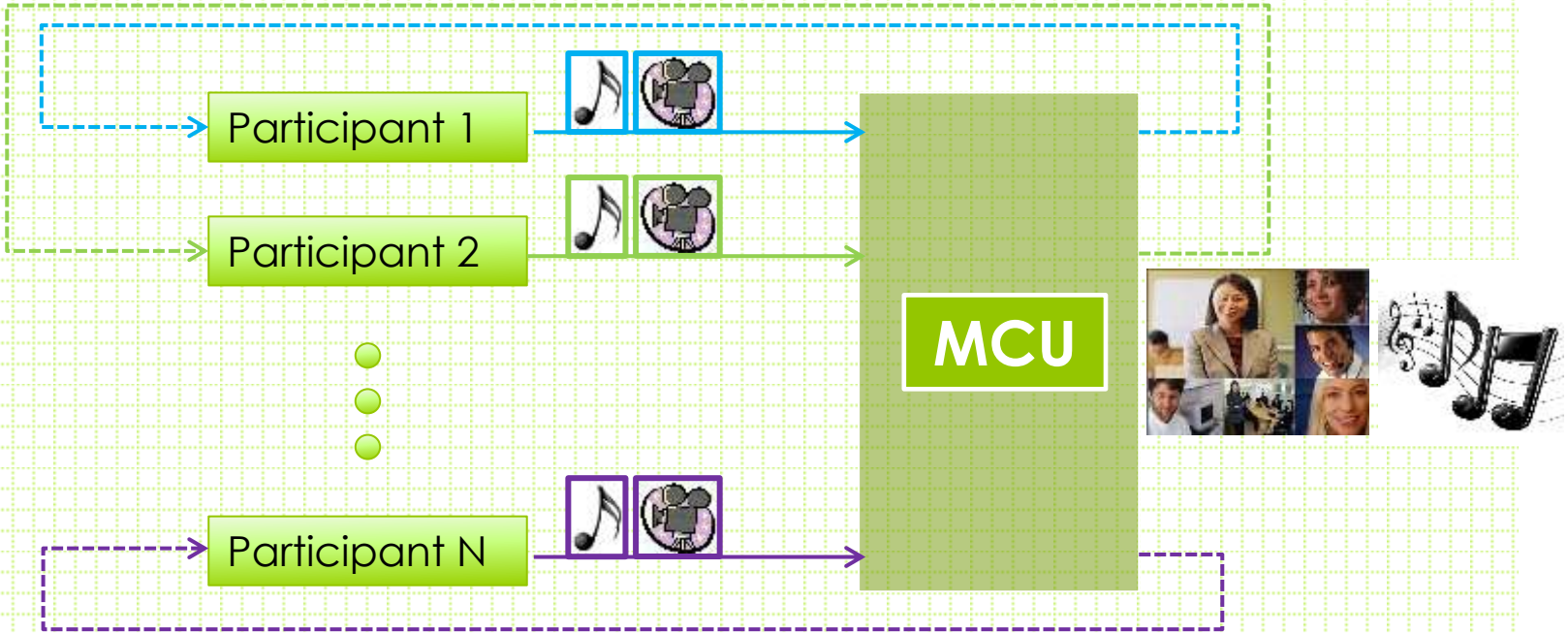




Central control unit - MCU

- **M**ultipoint **C**ontrol **U**nit

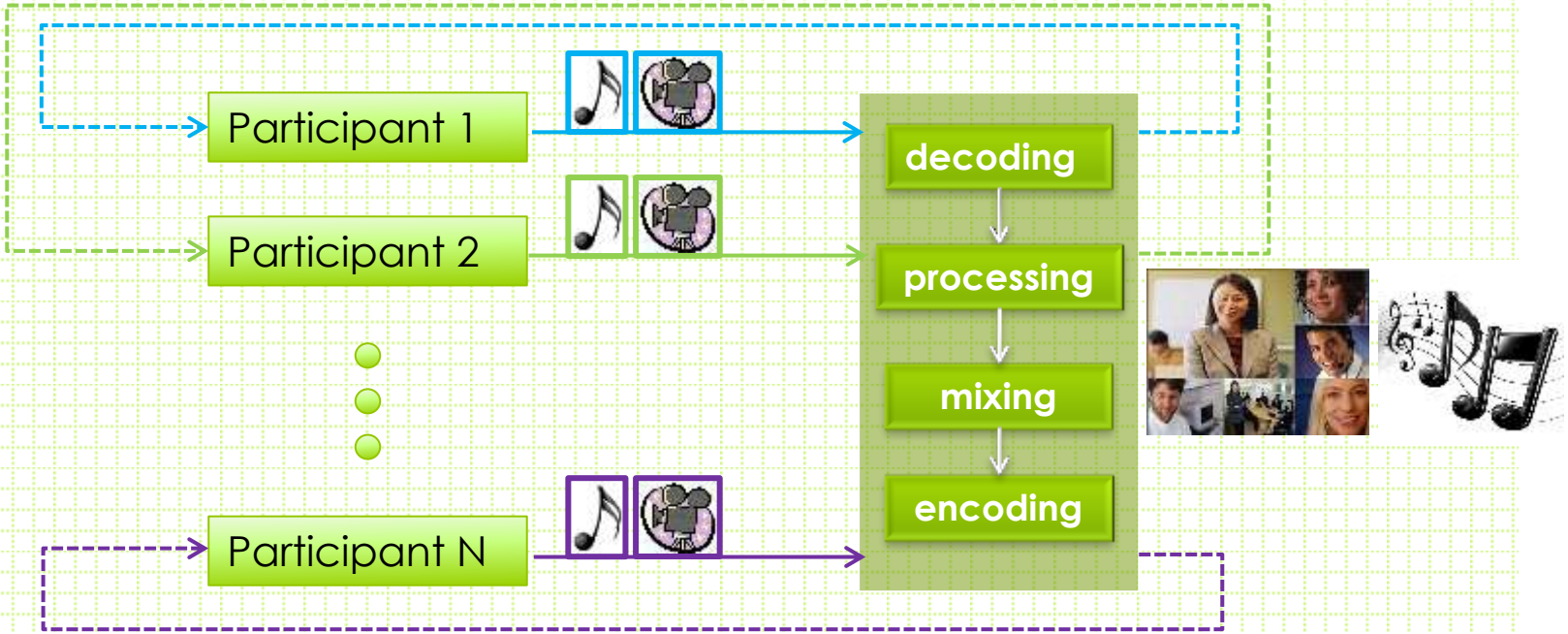
The conference is administrated through an MCU:



Central control unit - MCU

- **M**ultipoint **C**ontrol **U**nit

The conference is administrated through an MCU:

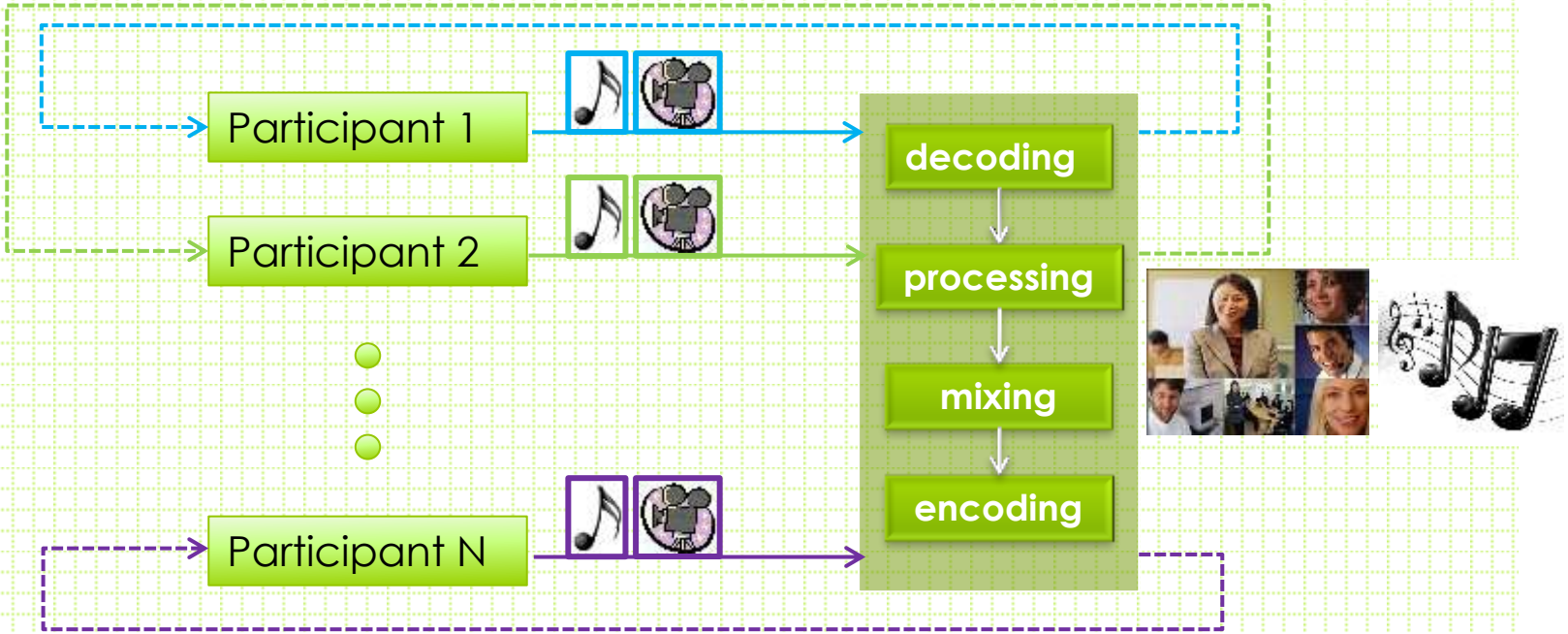


Heavy processing!

Central control unit - MCU

- **M**ultipoint **C**ontrol **U**nit

The conference is administrated through an MCU:



Reduce the amount of information to improve conference quality

Speaker selection

- Select **M** most active participants
- Discard all remaining information

Guidelines [2]

- The selection process should not introduce audio artifacts
- Transparent to the participants
- Resistance to noise
- Lack of discrimination

[2] P. J. Smith, "Voice conferencing over IP networks" , Master's thesis, Department of Electrical and Computer Engineering, McGill University, Montreal, Canada, 2002.

Speaker selection - Challenges

- Equipment
  - Low quality sensors lower SNR
  - Speakers produce crosstalk

- Surrounding
  - General noises
  - Transient noises

- Personal characteristics
  - Loud/Quiet participants

Classic Voice Activity Detection (VAD)

- Binary classification problem
  - Classify each signal frame into either *speech* or *non-speech* → 'High resolution' decision
    - Representation
      - Use a speech specific representation, to maximize discrimination ability
    - Classification method
      - Thresholding
      - Machine learning techniques

VAD and Speaker Selection

- Speaker Selection is a generalized task of VAD
  - For each signal frame in each channel :
    - Determine whether it contains prolonged speech activity or not
  - 'Lower resolution' decision is needed
    - Non-speech frames may belong to a speech burst too

Speaker Selection Methods

- Yu et al. *Computers and communications ISCC 1998*

"Linear PCM signal processing unit in multi-point video conferencing system"

  - The channel with the biggest power is determined dominant
  - To prevent frequent change of current speaker the decision is made every 1 or 2 seconds

- Chang, Lucent Technologies, 2001

"Multimedia conference call participant identification system and method "

  - Speaking participants are identified as those who pass a volume threshold

- Kwak, Verizon Laboratories, 2002

"Speaker identifier for multi-party conference"

  - Dominant speaker determined based on signal amplitude

Speaker Selection Methods

- Smith , MSc thesis, McGill University, 2002

"Voice conferencing over IP networks"

  - Participants are ranked in the order of becoming active speakers
  - A participant can move up in rankings if the smoothed power of his signal is above a *barge-in* threshold

- Xu et al. *ICME IEEE*, 2006

"Pass: Peer-aware silence suppression for internet voice conferences"

  - Very advanced VAD for ranking
  - Barge-in mechanism

Summary of existing methods

Method	Provides solution for:	Stationary Noise	Frequent Switching	Transient Noise	
Yu 1998 (Power)		-	-	-	Assume that non-dominant channels are completely silent
Chang 2001 (Volume Threshold)		-	-	-	
Kwak 2002 (Amplitude)		-	-	-	
Smith 2002 (Barge In)		√	√	-	Assume that change in signal power originates from change in speaker
Xu 2006 (Barge In & advanced VAD)		√	√	-	

Transient noise occurrences are not addressed

Discussed realization

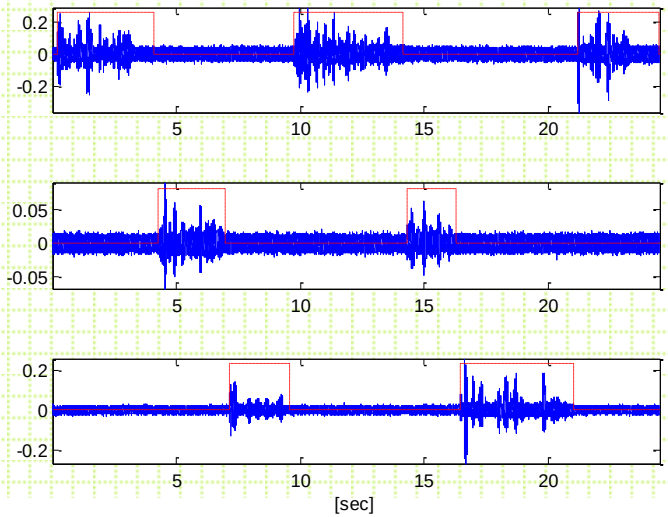
Dominant speaker identification (Speaker selection with M=1)

- Only one participant appears on each screen

- **Most video information is discarded**
- **Conversation is focused on the dominant speaker**
- **Further traffic reduction by discarding some audio**

Discussed realization

- N channels
- *Speech Burst* – a speech event composed of three sequential phases: initiation, steady state and termination.
- *Speaker Switch* – the point where a change in dominant speaker occurs
- <u>Objective</u>:
  - Follow *speech bursts*
  - Detect *speaker switches*

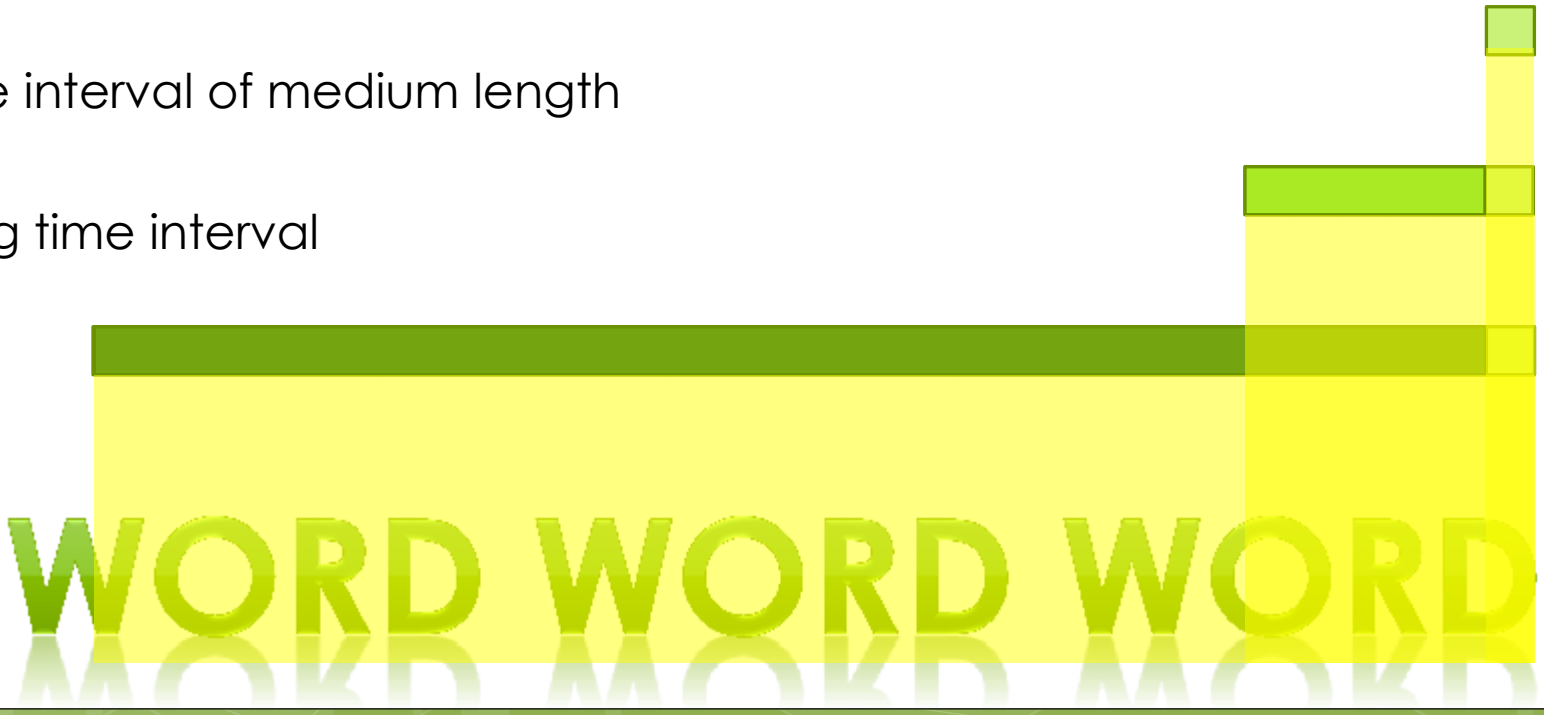


Dominant speaker identification relying on speech activity in time intervals of different length

Currently observed frame

Time interval of medium length

Long time interval

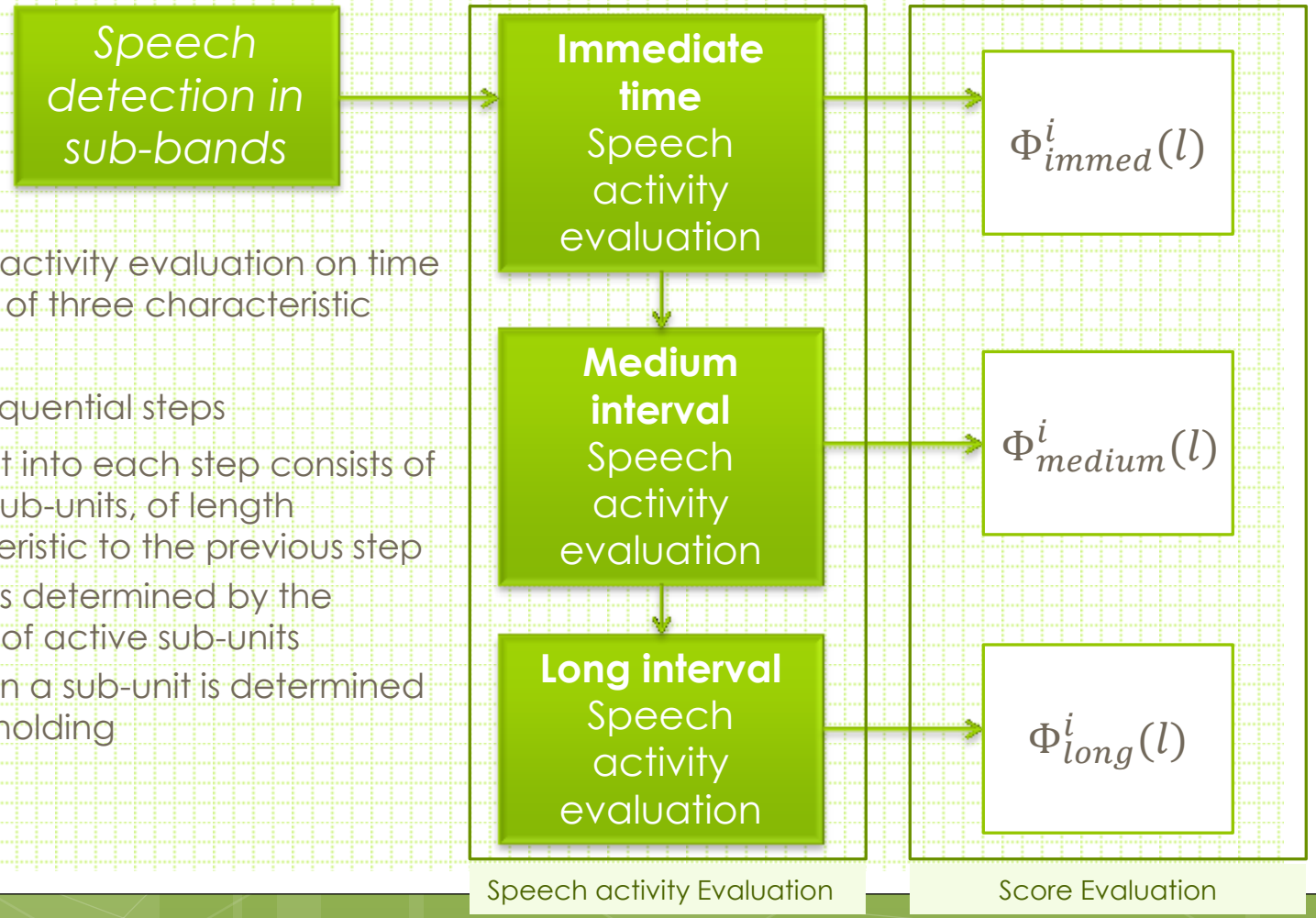


Dominant speaker identification relying on speech activity in time intervals of different length

- Two Stages:
  - **Local processing**
    - Individual processing on each channel
    - Speech activity score evaluation for the immediate, medium and long time-intervals
  - **Global decision**
    - Speech activity score comparison across channels

Local processing

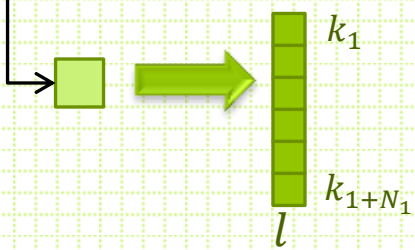
- Channel i, time frame l:



- Speech activity evaluation on time intervals of three characteristic lengths
- Three sequential steps
- The input into each step consists of smaller sub-units, of length characteristic to the previous step
- Activity is determined by the number of active sub-units
- Activity in a sub-unit is determined by thresholding

Local processing

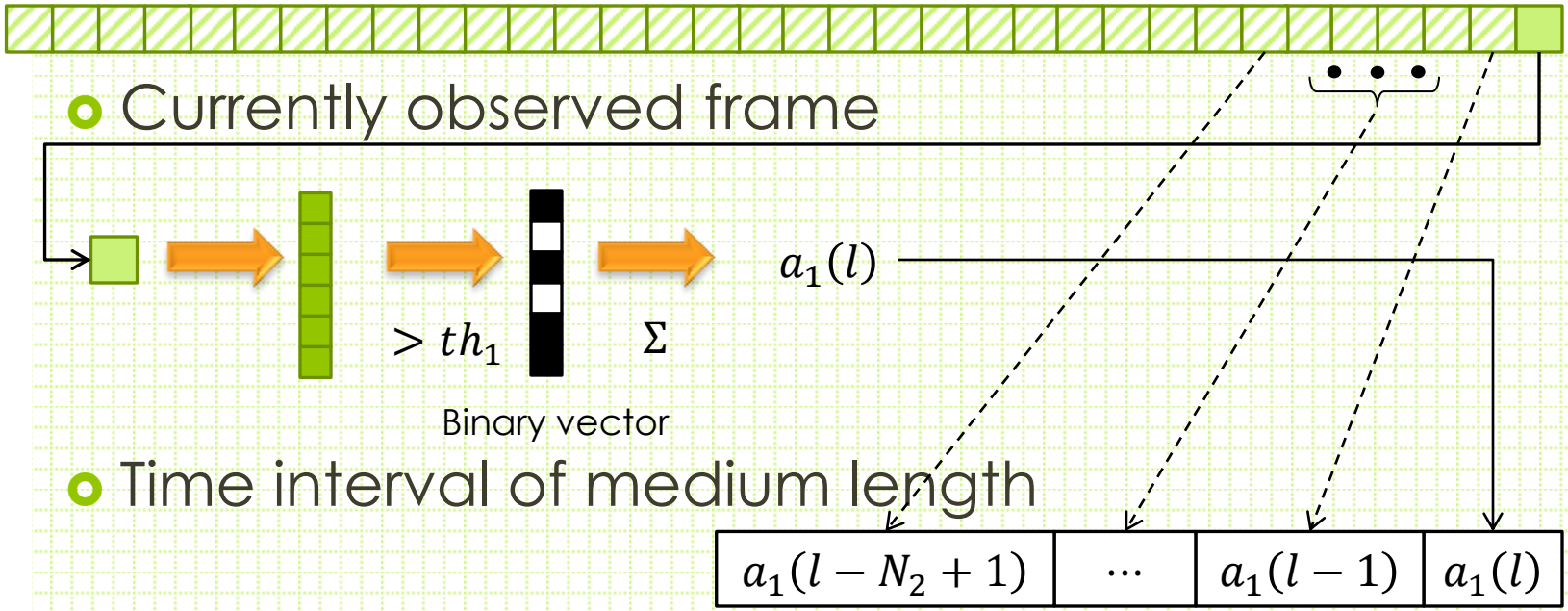
- Currently observed frame



- Time-frequency representation
- l – time index
- $[k_1, k_{1+N_1}]$ – frequency range of voiced speech

Local processing

- Currently observed frame



- Time interval of medium length

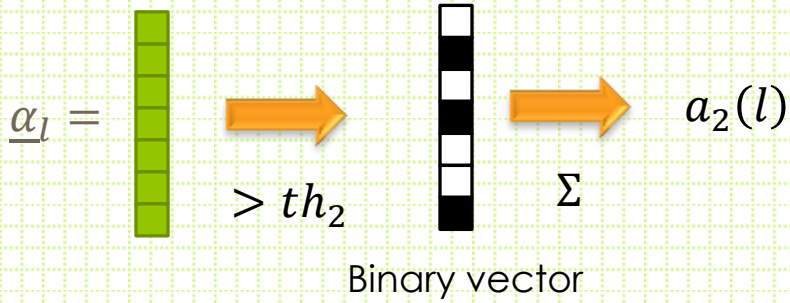
Local processing

- Currently observed frame → $a_1(l)$
- Time interval of medium length

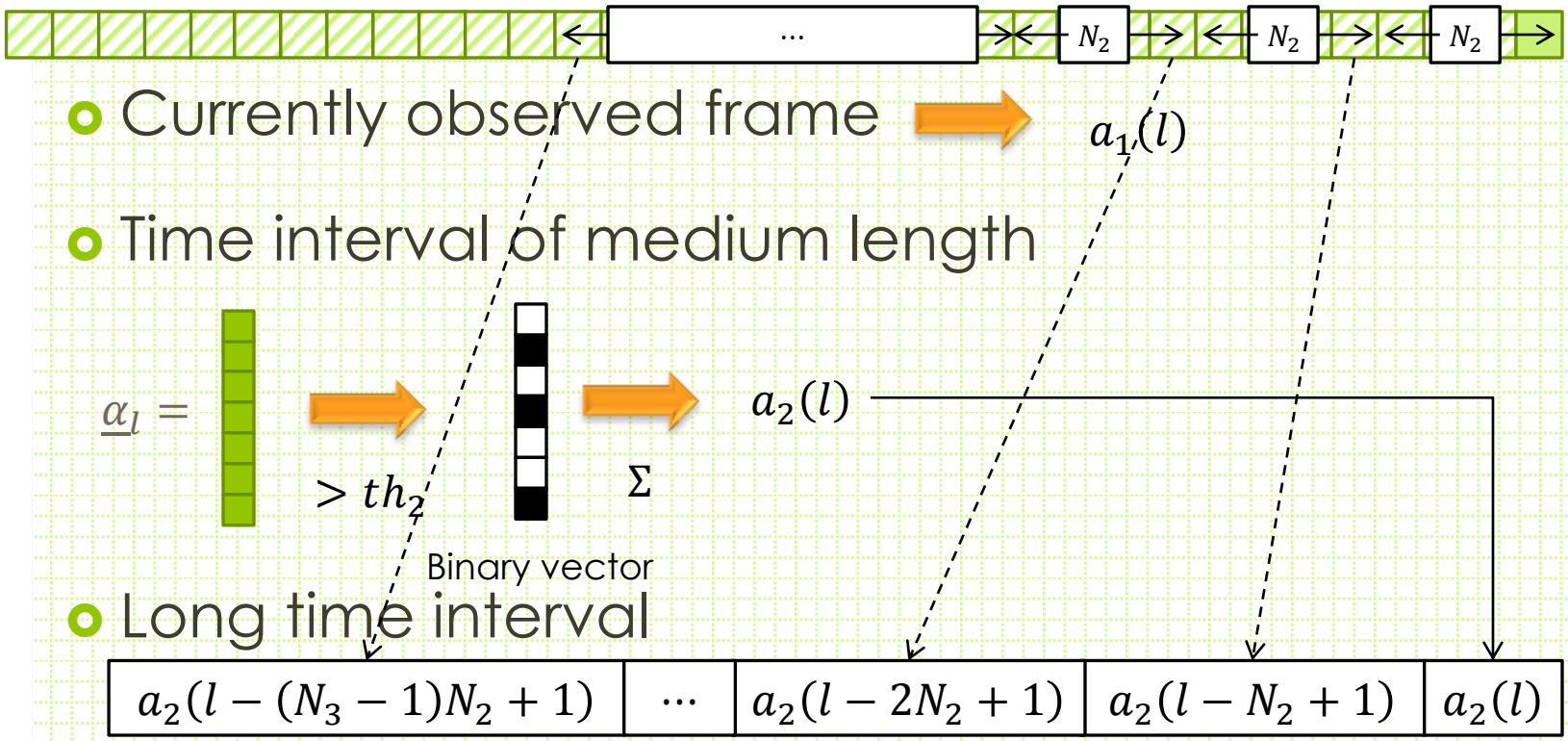
$$\underline{\alpha}_l = \begin{bmatrix} a_1(l) \\ a_1(l-1) \\ \\ \\ \\ a_1(l-N_2+1) \end{bmatrix}$$

Local processing

- Currently observed frame → $a_1(l)$
- Time interval of medium length



Local processing



- Currently observed frame
- Time interval of medium length
- Long time interval

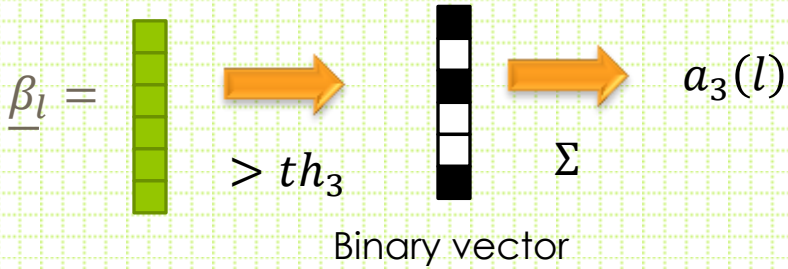
Local processing

- Currently observed frame → $a_1(l)$
- Time interval of medium length → $a_2(l)$
- Long time interval

$$\underline{\beta}_l = \begin{bmatrix} a_2(l) \\ a_2(l-N_2+1) \\ \\ \\ a_2(l-N_3N_2+1) \end{bmatrix}$$

Local processing

- Currently observed frame → $a_1(l)$
- Time interval of medium length → $a_2(l)$
- Long time interval



Local processing

- Currently observed frame → $a_1(l)$
- Time interval of medium length → $a_2(l)$
- Long time interval → $a_3(l)$

Active sub-units & Transients

- Currently observed frame → $a_1(l)$ **# of active sub-bands**
- Time interval of medium length → $a_2(l)$ **# of active frames**
- Long time interval → $a_3(l)$ **# of active blocks**

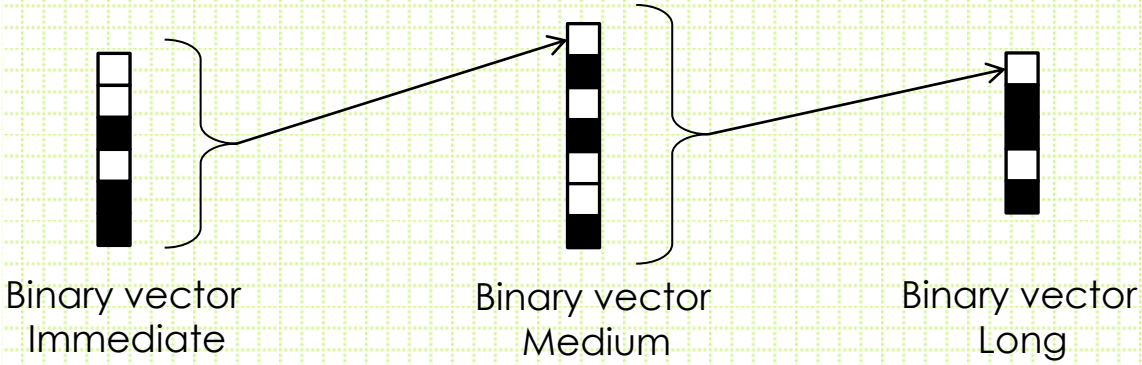
Why Thresholding and counting ?

- Smoothing
- Equalization
- Noise spikes suppression

Active sub-units & Transients

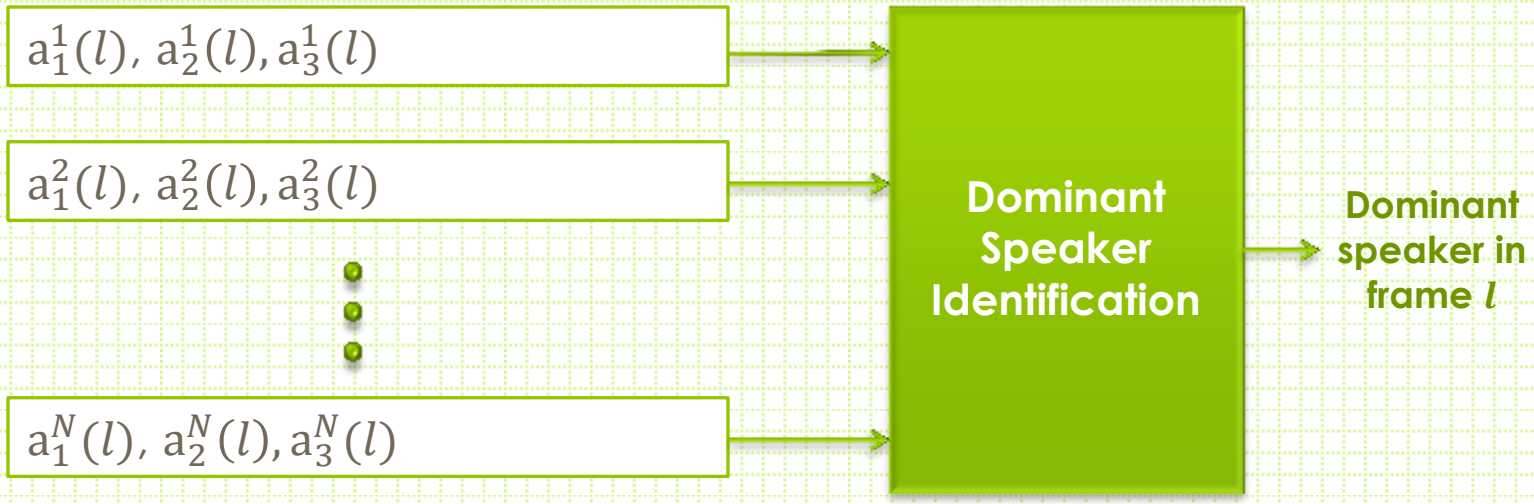
- Currently observed frame → $a_1(l)$ **# of active sub-bands**
- Time interval of medium length → $a_2(l)$ **# of active frames**
- Long time interval → $a_3(l)$ **# of active blocks**

Discriminating isolated transients from fluent audio activity:



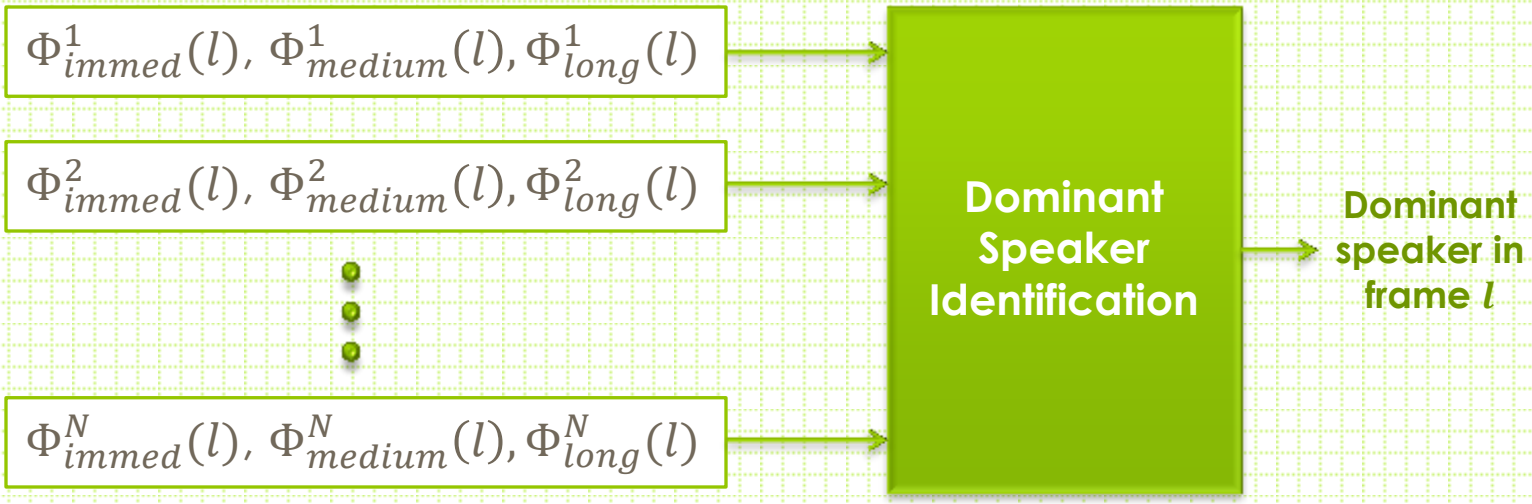
Global Decision

- Time frame l:



Global Decision

- Time frame l:



Dominant speaker identification

Activated once in a *decision-interval*

- Objective:
  - Detect *speaker switch* events
- Method:
  - Compare speech activity scores across channels
  - Each channel is evaluated in respect to:
    - the dominant channel
    - minimal activity level

Current dominant speaker remains unless activity on one of the other channels justifies a speaker switch

Score Evaluation

Single

- Score computed on a single observation
- Each observation is analyzed independently

More responsive to changes

Sequential

- Score computed on a sequence of observations
- Exploits natural temporal dependence in speech

More Smooth

Modeling the number of active sub-units

N – number of sub-units
a – number of active sub-units

Two Hypotheses: $\begin{cases} H_1 & speech\ is\ present \\ H_0 & speech\ is\ absent \end{cases}$

H_1 → $p(a|H_1) = \mathrm{Bin}(N,p) = \binom{N}{a} p^a (1-p)^{N-a}$

a

Likelihood Models

H_0 → $p(a|H_0) = Exp(\lambda) = \lambda e^{-\lambda a}$

→ **Single**

OR

→ **Sequential**

Score based on a single observation

Likelihood Ratio

log-Likelihood Ratio

$$\Lambda = \frac{p(a|H_1)}{p(a|H_0)}$$

$$\boldsymbol{\Phi} = log\boldsymbol{\Lambda}$$

speech activity score

$$\Phi = \log\binom{N}{a} + a\log(p) + (N-a)\log(1-p) - \log(\lambda) + \lambda a$$

Score based on a sequence of observations

- The score is computed on a *sequence* of observations

$$\underline{\boldsymbol{X}}^{l} = [a(l - M + 1), \dots, a\,(l)]$$

- Observations are *not* independent
  - A speech frame is more likely to be followed by another speech frame

    $p(q_n = H_1 | q_{n-1} = H_1) > p(q_n = H_1)$ [4]

First order Markovian dependency in the transitions between consecutive frames

[4] J. Sohn et al., "A statistical model-based voice activity detection", Signal Processing Letters, IEEE, 1999

Score based on a sequence of observations

- *HMM* of two states:
  - $\zeta_l = \begin{cases} H_{1,(l)} & speech\ is\ present\ in\ frame\ l \\ H_{0,(l)} & speech\ is\ absent\ in\ frame\ l \end{cases}$
  - State dynamics: $b_{ij} = p(\zeta_l = \mathrm{j} | \zeta_{l-1} = i)\ ,\ \ i, j \epsilon \{0,1\}$
    - $[b_{00} + b_{01}] = 1,\ [b_{10} + b_{11}] = 1$
  - Steady state probabilities

    $pH_0 = \frac{b_{10}}{b_{10}+b_{01}}, pH_1 = \frac{b_{01}}{b_{10}+b_{01}}$
  - Current observation only depends on current

    state: $p(X^l | \zeta_l, \underline{\boldsymbol{X}}^l) = p(X^l | \zeta_l)$

Score based on a sequence of observations

Likelihood Ratio

$$\Lambda_l^{seq} = \frac{p(\underline{\boldsymbol{X}}^l | H_{1,(l)})}{p(\underline{\boldsymbol{X}}^l | H_{0,(l)})} \rightarrow \frac{p(\underline{\boldsymbol{X}}^l, H_{1,(l)})pH_0}{p(\underline{\boldsymbol{X}}^l, H_{0,(l)})pH_1} \rightarrow$$

- Denote: $\alpha_l(1) \triangleq p(\underline{\boldsymbol{X}}^l, H_{1,(l)})$ and $\alpha_l(0) \triangleq p(\underline{\boldsymbol{X}}^l, H_{0,(l)})$
- Recursive formula (forward procedure) for $\alpha_l(j), j \in \{0,1\}$

$$\alpha_q(j) = \begin{cases} p(X_q = a(q)|H_{j,(q)})[\alpha_{q-1}(0)b_{j0} + \alpha_{q-1}(1)b_{j1}] & , l - M + 1 < q < l \\ p(X_q = a(q)|H_{j,(q)})p(H_j) & , q = l - M + 1 \end{cases}$$

log-Likelihood Ratio

$$\rightarrow \Lambda_l^{seq} = \frac{\alpha_l(1)}{\alpha_l(0)} \cdot \frac{pH_0}{pH_1} \rightarrow \boldsymbol{\Phi_l^{seq} = \log\frac{\alpha_l(1)}{\alpha_l(0)} + \log\frac{pH_0}{pH_1}}$$

Single Vs. Sequential

- Denote: $\Gamma(n) = \frac{\alpha_n(1)}{\alpha_n(0)}$, $\Phi^{seq} = \log\big(\Gamma(n)\big) + \log\frac{pH_0}{pH_1}$

$$\Gamma(n) = \frac{\alpha_n(1)}{\alpha_n(0)} = \frac{p(a_n|H_1)}{p(a_n|H_0)} \cdot \frac{\alpha_{n-1}(0)b_{10} + \alpha_{n-1}(1)b_{11}}{\alpha_{n-1}(0)b_{00} + \alpha_{n-1}(1)b_{01}}$$

$$= \Lambda_n \cdot \frac{b_{10} + \Gamma(n-1)b_{11}}{b_{00} + \Gamma(n-1)b_{01}}$$

$$\boxed{\Phi^{seq} = \Phi^{single} + \log\frac{b_{10} + \Gamma(n-1)b_{11}}{b_{00} + \Gamma(n-1)b_{01}} + \log\frac{pH_0}{pH_1}}$$

Single Vs. Sequential

$$\Phi^{seq} = \Phi^{single} + \log\frac{b_{10} + \Gamma(n-1)b_{11}}{b_{00} + \Gamma(n-1)b_{01}} + \log\frac{pH_0}{pH_1}$$

- In the presence of speech

$$\Gamma(n-1) \gg 0 \;\Rightarrow \log\frac{b_{10} + \Gamma(n-1)b_{11}}{b_{00} + \Gamma(n-1)b_{01}} \rightarrow \log\frac{b_{11}}{b_{10}} > 0$$

$$\Rightarrow \Phi^{seq} > \Phi^{single}$$

Single Vs. Sequential

$$\Phi^{seq} = \Phi^{single} + \log\frac{b_{10} + \Gamma(n-1)b_{11}}{b_{00} + \Gamma(n-1)b_{01}} + \log\frac{pH_0}{pH_1}$$

- In absence of speech

$$\Gamma(n-1) \ll 1 \;\Rightarrow \log\frac{b_{10} + \Gamma(n-1)b_{11}}{b_{00} + \Gamma(n-1)b_{01}} \rightarrow \log\frac{b_{10}}{b_{00}} < 0$$

$$\Rightarrow \Phi^{seq} < \Phi^{single}$$

Sequential processing improves separability

A-priori SNR Estimation

Let: $y(n) = x(n) + d(n) \xrightarrow{STFT} Y_l = X_l + D_l$

- $\lambda_l = E\{|X_l|^2\}$ – Spectral variance of speech
  - X_l is a complex vector $X_l = A_l e^{j\phi_l}$
  - $\lambda_l = E\{|X_l|^2\} = E\{|A_l|^2\}$
- $\lambda_{D_l} = E\{|D_l|^2\}$ – Spectral variance of noise

A-priori SNR : $\hat{\xi}_{l|l} = \dfrac{\hat{\lambda}_{l|l}}{\lambda_{D_l}}$

A-priori SNR Estimation

The estimator is derived in two steps: [5]

1. Propagation Step
   - Assuming we have all information up to frame $l-1$
   - Obtain one frame ahead conditional variance $\hat{\lambda}_{l|l-1}$
2. Update Step
   - Update the estimator using information from current frame, l
   - Obtain $\hat{\lambda}_{l|l}$

[5] I. Cohen, "Relaxed statistical model for speech enhancement and a priori snr estimation", Speech and Audio Processing, IEEE Transactions on, 2005.

A-priori SNR Estimation

The estimator is derived in two steps :

1. Propagation Step

- Assuming we have all information up to frame $l-1$
- The conditional variance of speech is assumed to propagate as a GARCH(1,1) model:

$$\lambda_{l|l-1} = \lambda_{min} + \mu|X_{l-1}|^2 + \delta\left(\lambda_{l-1|l-2} + \lambda_{min}\right)$$

$$\lambda_{min} > 0, \mu \geq 0, \delta \geq 0, \mu + \delta < 1$$

$$\boxed{\begin{aligned}\hat{\lambda}_{l|l-1} &= E\{\lambda_{l|l-1}|\hat{A}_{l-1}, \hat{\lambda}_{l-1|l-2}\} \\ &= \lambda_{min} + \mu\hat{A}_{l-1}^2 + \delta\left(\hat{\lambda}_{l-1|l-2} + \lambda_{min}\right)\end{aligned}}$$

A-priori SNR Estimation

The estimator is derived in two steps :

1. Propagation Step

$$\hat{\lambda}_{l|l-1} = E\{\lambda_{l|l-1} | \hat{A}_{l-1}, \hat{\lambda}_{l-1|l-2}\}$$
$$= \lambda_{min} + \mu \hat{A}_{l-1}^2 + \delta(\hat{\lambda}_{l-1|l-2} + \lambda_{min})$$

2. Update Step

- Update the estimator using information from current frame, l
- Obtain $\hat{\lambda}_{l|l}$

A-priori SNR estimation

2. Update Step:

Let: $$y(n) = x(n) + d(n) \xrightarrow{STFT} Y_l = X_l + D_l$$

- Spectral enhancement $\hat{X}_l = GY_l$
- G – spectral gain function
- G is chosen as a minimizer of a certain distortion measure

A-priori SNR estimation

2. Update Step:

- We chose G_{SP}, the minimizer of the power distortion measure

$$d_{SP} = \left(A_l^2 - \hat{A}_l^2\right)^2$$

- $\gamma_l = \frac{|Y_l|^2}{\lambda_{D_l}}$ - a posteriori SNR
- $\hat{\lambda}_{l|l-1}$ - one frame ahead conditional variance (speech)
- The resulting spectral gain function is :

$$G_{SP}\left(\hat{\lambda}_{l|l-1}, \gamma_l\right) = \sqrt{\frac{\hat{\lambda}_{l|l-1}}{\lambda_{D_l}+\hat{\lambda}_{l|l-1}}\left(\frac{1}{\gamma_l}+\frac{\hat{\lambda}_{l|l-1}}{\lambda_{D_l}+\hat{\lambda}_{l|l-1}}\right)}$$

Speech detection in sub-bands

2. Update Step
   - $\hat{X}_l = G_{SP}(\hat{\lambda}_{l|l-1}, \gamma_l) Y_l \rightarrow \hat{\lambda}_{l|l} = E\left\{\left|\hat{X}_l\right|^2\right\} = G_{SP}(\hat{\lambda}_{l|l-1}, \gamma_l)^2 |Y_l|^2$

$$\boxed{\hat{\lambda}_{l|l} = G_{SP}(\hat{\lambda}_{l|l-1}, \gamma_l)^2 |Y_l|^2}$$

The *a-priori* SNR estimator:

$$\boxed{\hat{\xi}_{l|l} = \frac{\hat{\lambda}_{l|l}}{\lambda_{D_l}} = \frac{\hat{\lambda}_{l|l-1}}{\lambda_{D_l} + \hat{\lambda}_{l|l-1}} \left(1 + \frac{\hat{\lambda}_{l|l-1}\gamma_l}{\lambda_{D_l} + \hat{\lambda}_{l|l-1}}\right)}$$

Experimental results

Three Experiments:

1. Identification of the dominant speaker
2. Robustness to transient audio occurrences
3. Results on a real multipoint conference

Experimental results

The proposed method was compared to:

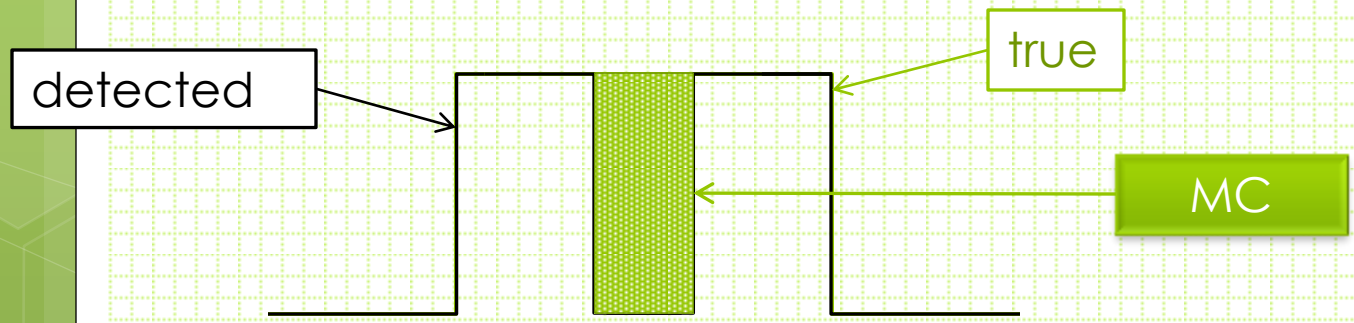
- The speaker with the highest VAD score in the *decision-interval* is selected
  1. ***RAMIREZ*** [6]
  2. ***SOHN*** [4]
  3. ***GARCH*** [7]

4. The speaker with the highest ***SNR*** is selected
5. The speaker with the highest signal ***POWER*** is selected

[6] Ramirez et al. "Statistical voice activity detection using a multiple observation likelihood ratio test", Signal Processing Letters, IEEE,2005
[7] S. Mousazadeh and I. Cohen, "Ar-garch in presence of noise: Parameter estimation and its application to voice activity detection", Audio, Speech, and Language Processing, IEEE Transactions on, 2011

Performance Evaluation

- Quantitative Evaluation:
  - False Speaker Switches #
  - Mid Sentence Clipping (MC) – percent of undetected mid section of a speech burst



Experiment #1

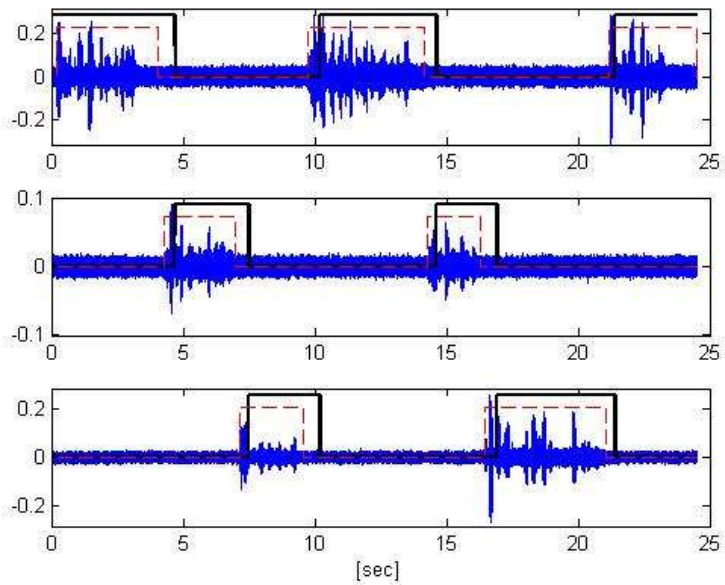
Signals:

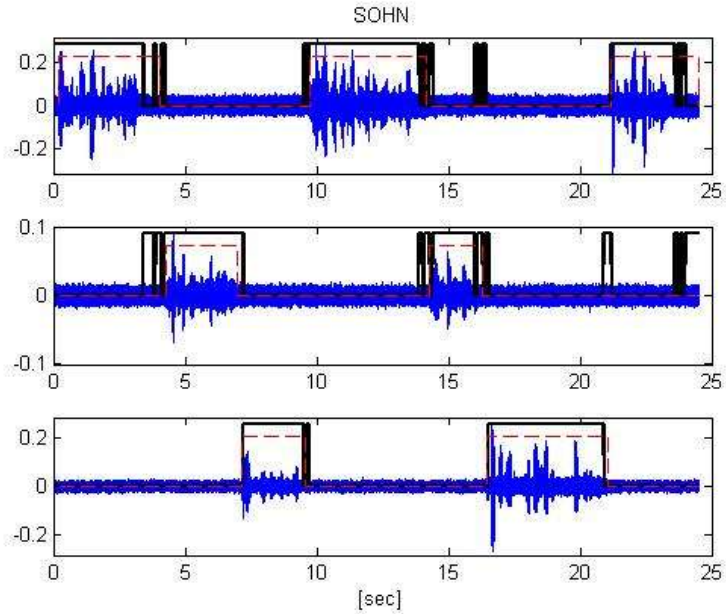
- One or more concatenated TiMit sentences of same speaker (with added white noise)
- The non silent part of the sentence is expected to be detected as a continuous speech burst

Experiment #1

Proposed method

Sohn VAD based method

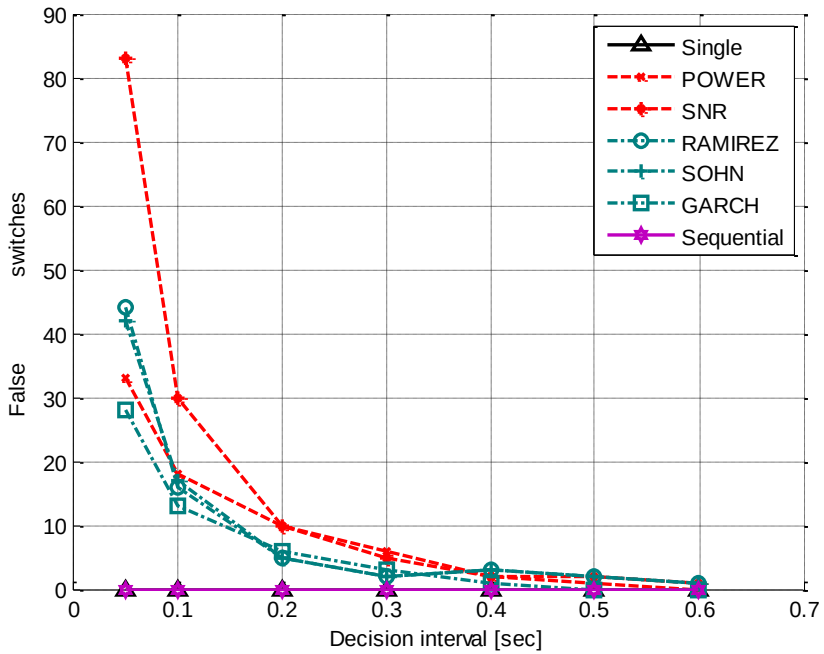




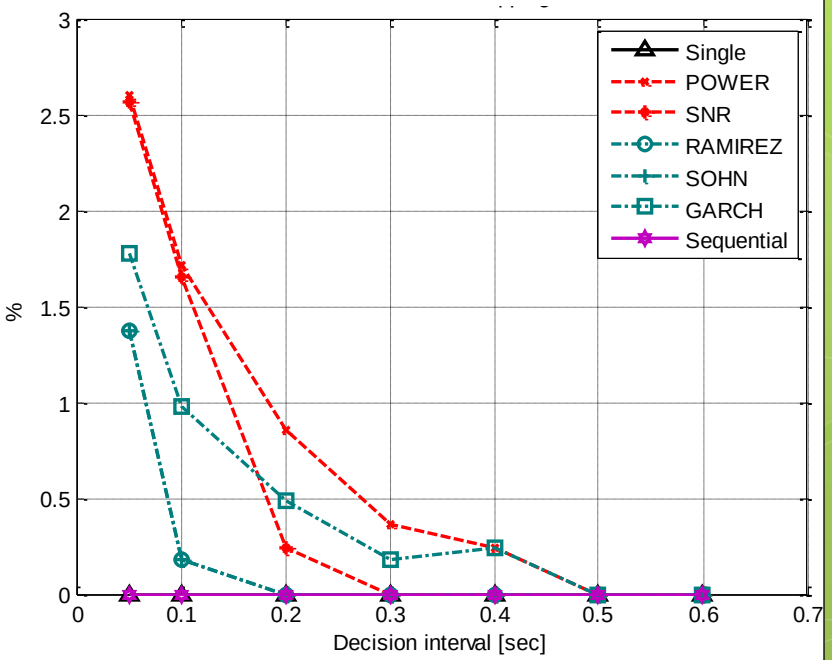
Decision interval = 0.1 sec

Experiment #1

False Speaker Switches



Mid Sentence Clipping

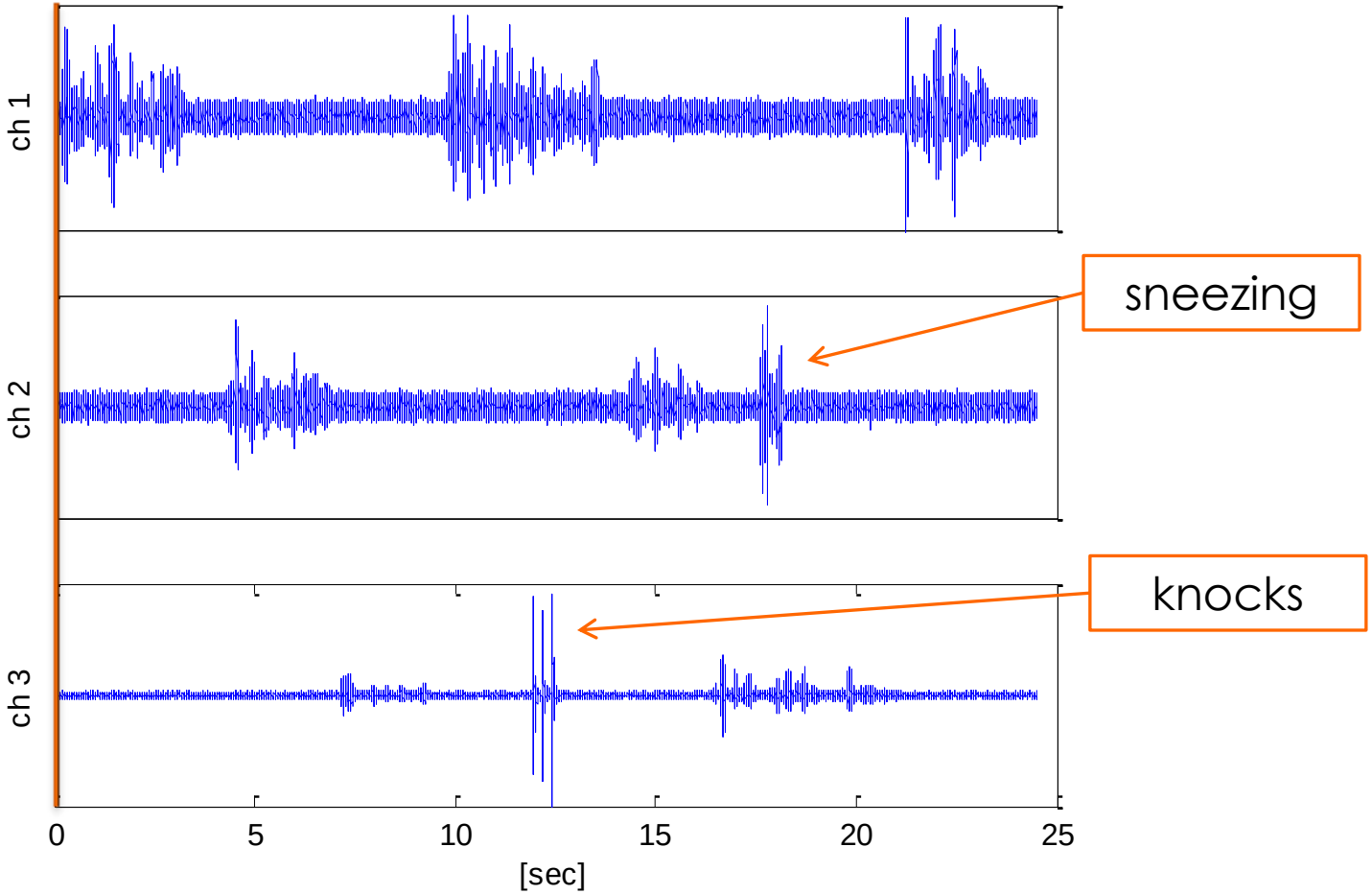


Experiment #2

Signals:

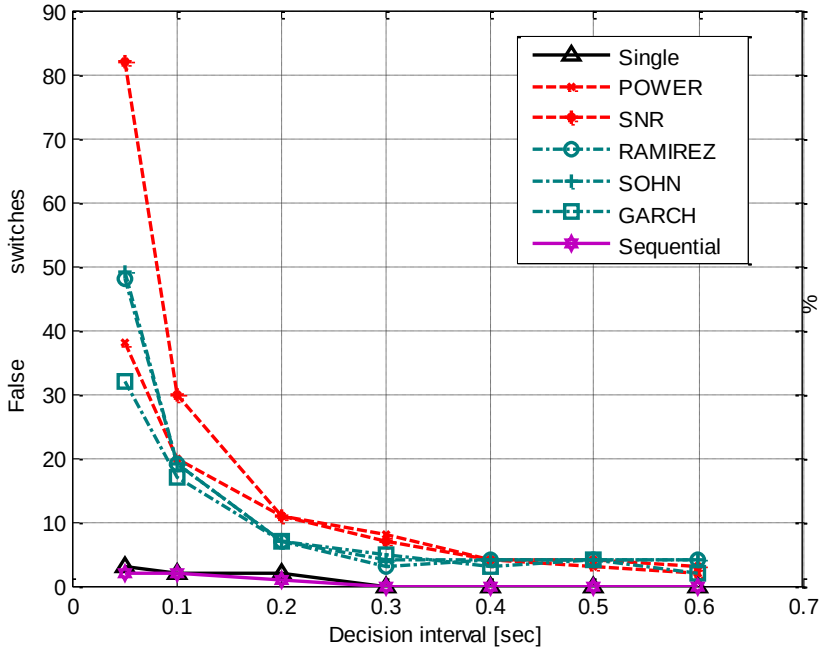
- One or more concatenated TiMit sentences of same speaker (with added white noise)
- A sound of sneezing was added to channel 2
- A sound of door knocks was added to channel 3

Experiment #2

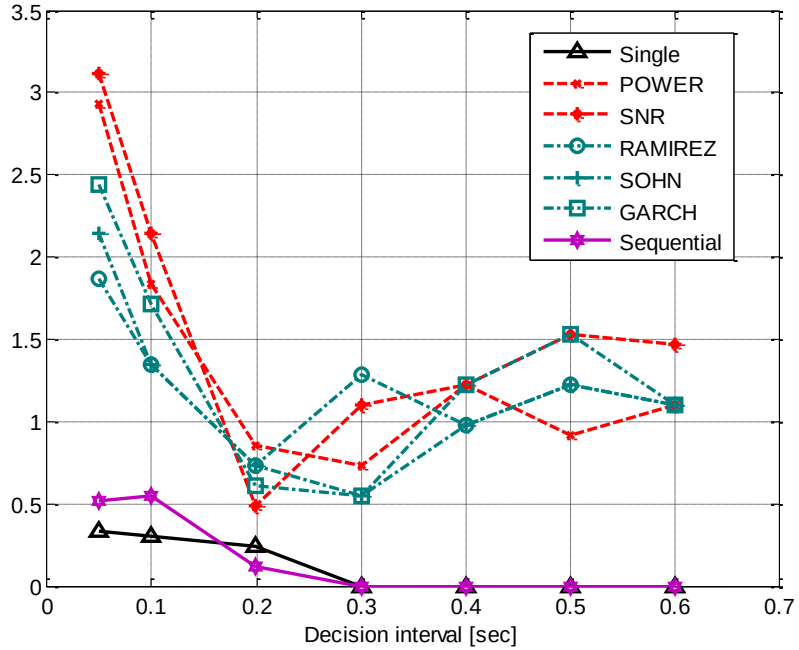


Experiment #2

False Speaker Switches

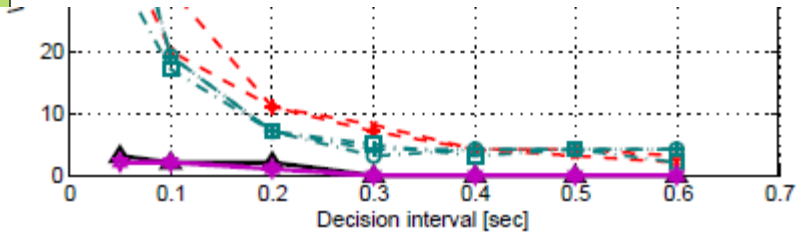


Mid Sentence Clipping

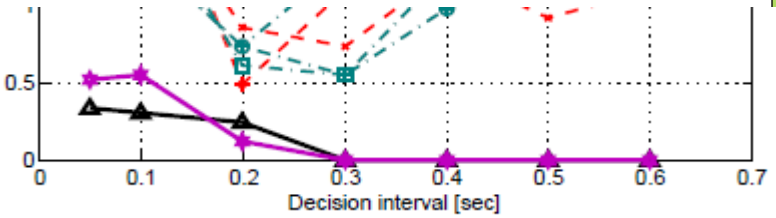


Experiment #2

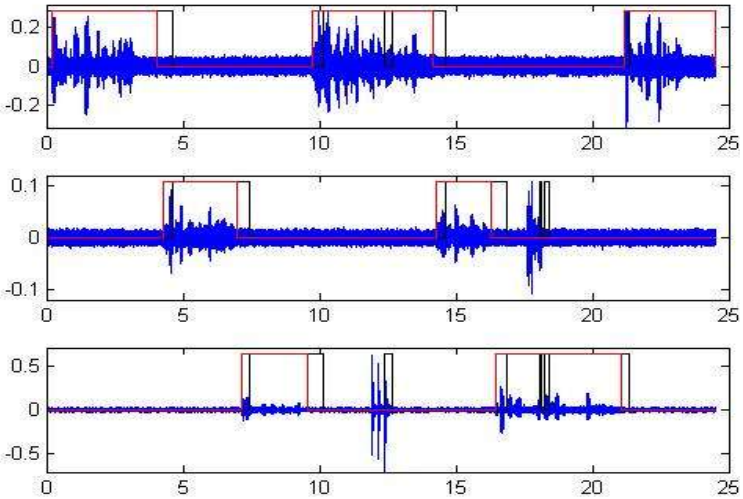
False Speaker Switches



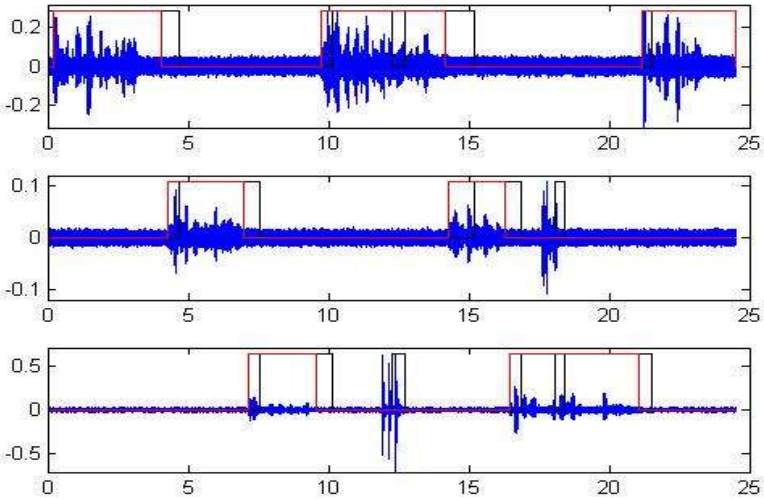
Mid Sentence Clipping



Single



Sequential

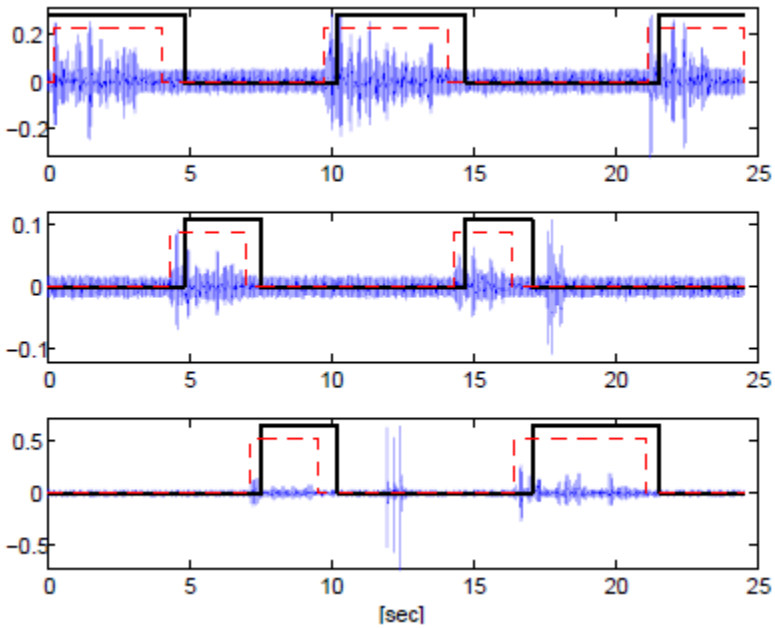


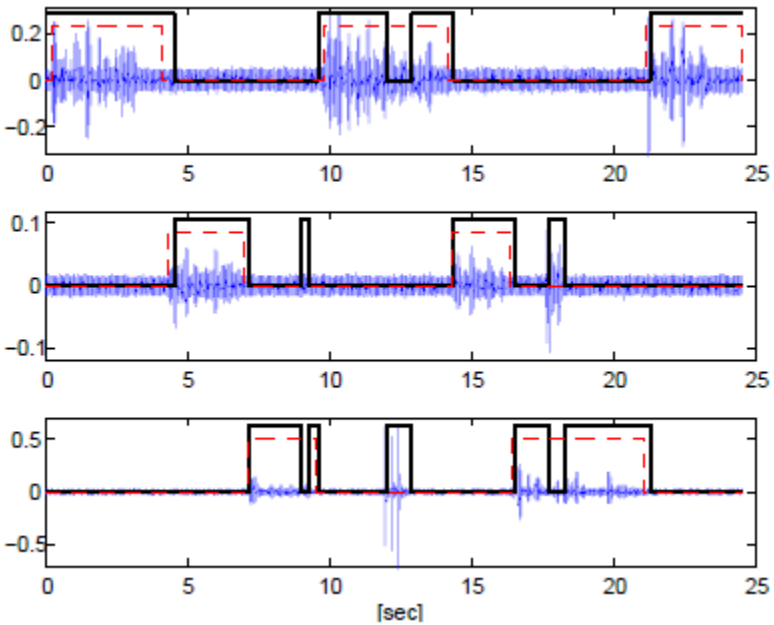
Decision interval = 0.05 sec

Experiment #2

Proposed Method

GARCH VAD based method





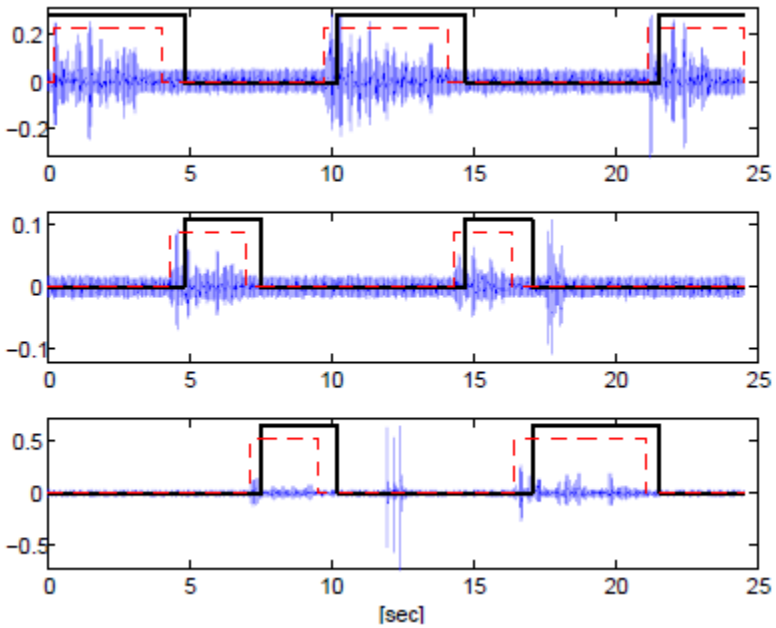
Red – hand labeled

Black – Algorithm's result

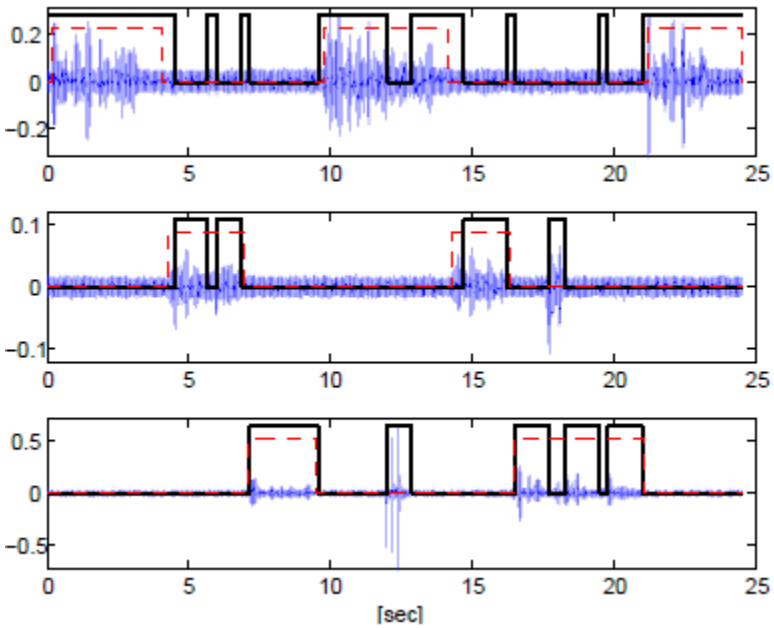
Decision interval = 0.3 sec

Experiment #2

Proposed Method



POWER based method



Red – hand labeled
Black – Algorithm's result

Decision interval = 0.3 sec

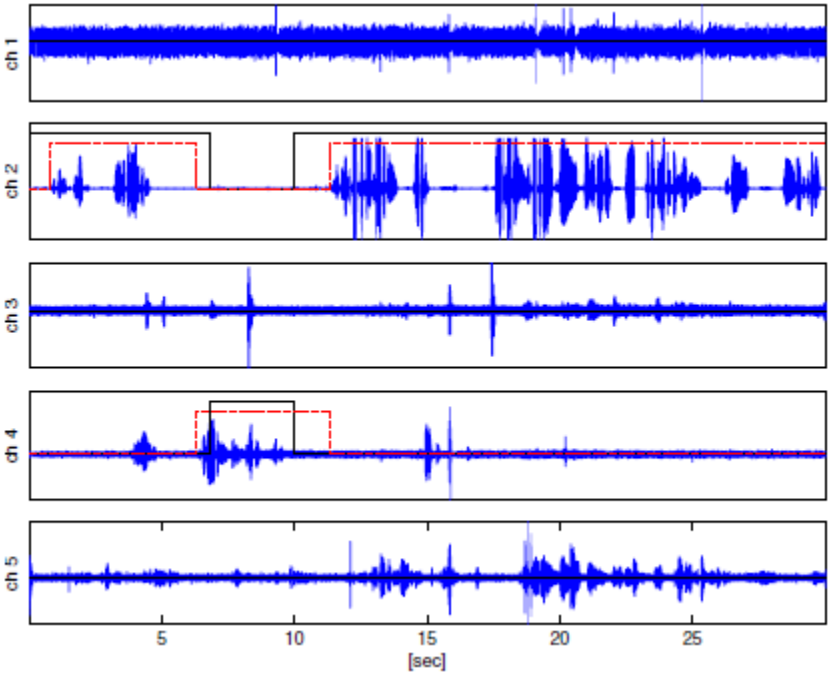
Experiment #3

Signals:

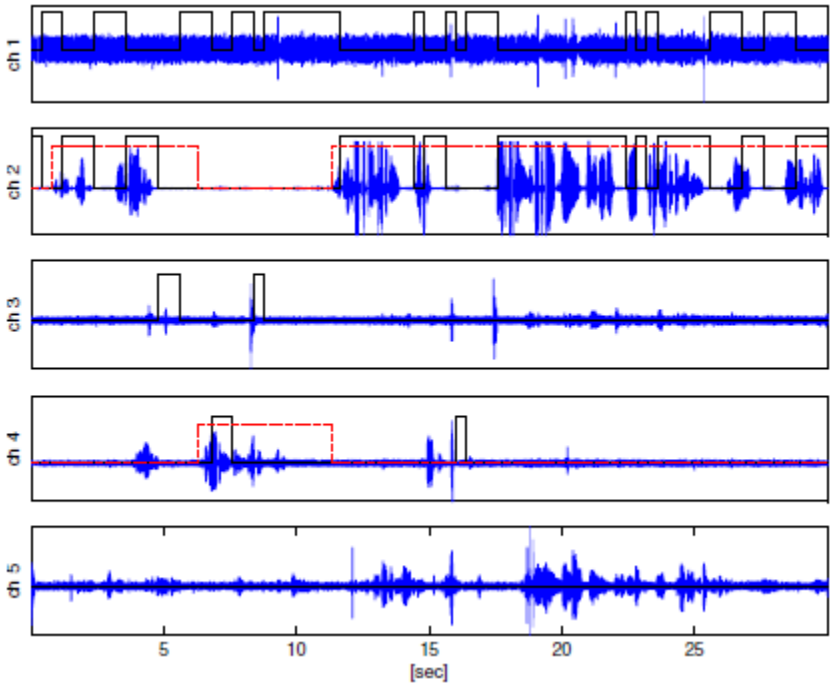
- Real Multi-channel conversation
- 5 channels
- Channels 2 & 4 – clean speech
- Channel 1 – mostly noise
- Channels 3 and 5 - crosstalk

Experiment #3

Proposed Method



POWER based method



Red – hand labeled
Black – Algorithm's result

Decision interval = 0.4 sec

Summary

- Novel method for dominant speaker identification
- Two approaches to speech activity score evaluation
- Experimental framework
- Less false speaker switches
- Robustness to transient audio occurrences

Future Research

- Non causal processing
- Temporally variable thresholds
- Preprocessing
  - Speech enhancement
  - Echo detection (or cancellation)

Thank You