Improving Boundary Estimation in Audiovisual Speech Activity Detection Using Bayesian Information Criterion

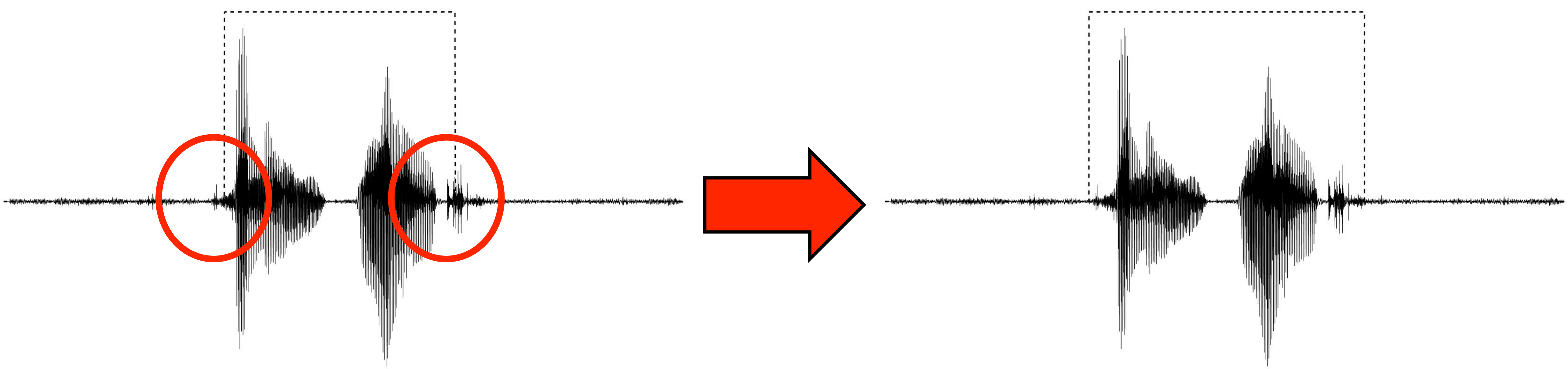
Fei Tao
John H.L. Hansen
Carlos Busso

Multimodal Signal Processing (MSP) Laboratory ,
Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering,
The University of Texas at Dallas,
Richardson TX 75080, USA

Introduction

- Speech Activity Detection (SAD) plays an important role in speech-based interfaces
- Audio-only SAD (A-SAD) may fail
  - Noise
  - Different speech mode (e.g. whisper speech)
- Introduce Visual SAD (V-SAD) to improve SAD [Aubrey et al. (2007), Joosten et al.(2013)]

- One key problem exists in V-SAD system was the precise detection of boundaries
  - Lip movement associated with non-speech event (e.g. lip smacking, laughing)
  - Anticipatory facial movements (e.g. 10 ms)
  - Low video resolution (30 fps vs. 100 fps)



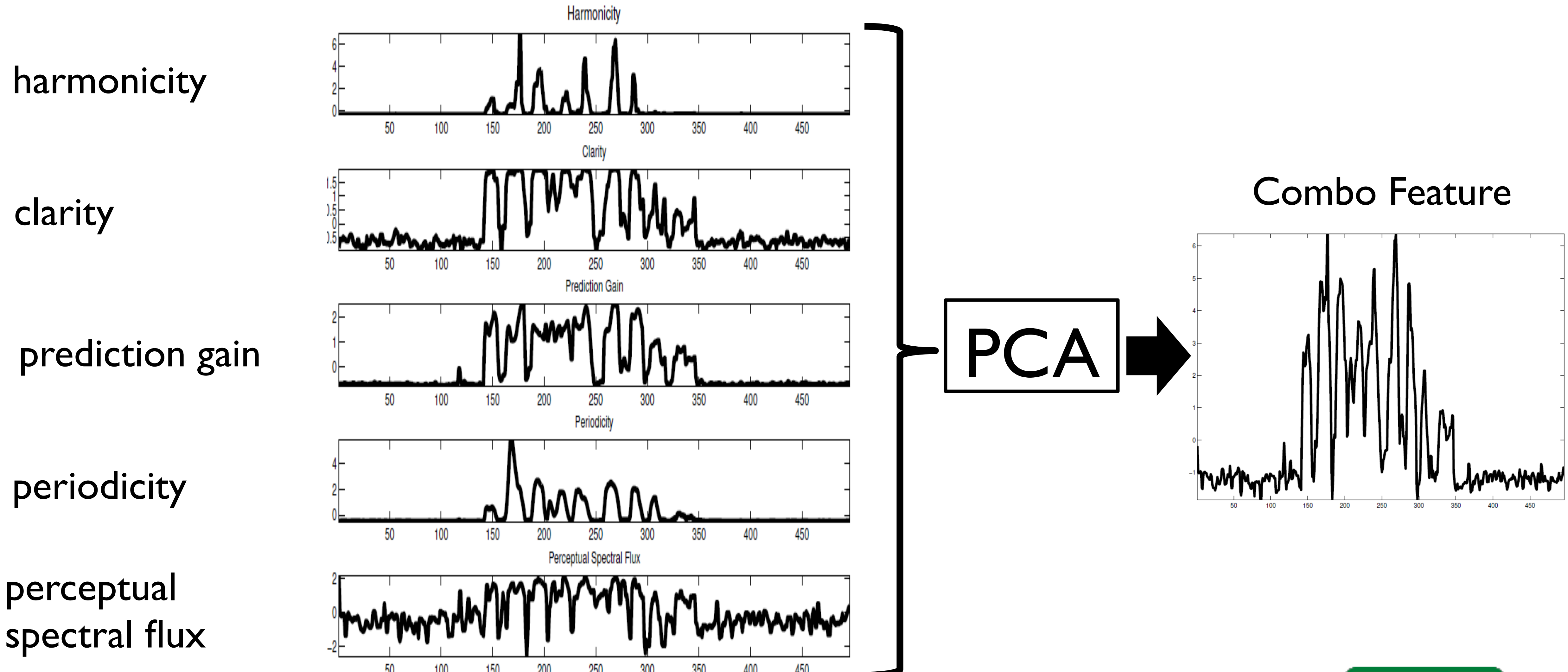
- Bayesian Information Criterion (BIC) to improve boundary detection

Previous Work on SAD

- Supervised V-SAD
  - Aubrey et al (2007) applied HMM in developing V-SAD system;
  - Joosten et al (2013) applied SVM classifier
- AV-SAD Fusion
  - Takeuchi et al. (2009) combined the V-SAD and A-SAD decision boundaries using logical operators.
  - Almajai and Milner (2008) concatenated acoustic and visual features.
- No one has worked on improving the boundary detection

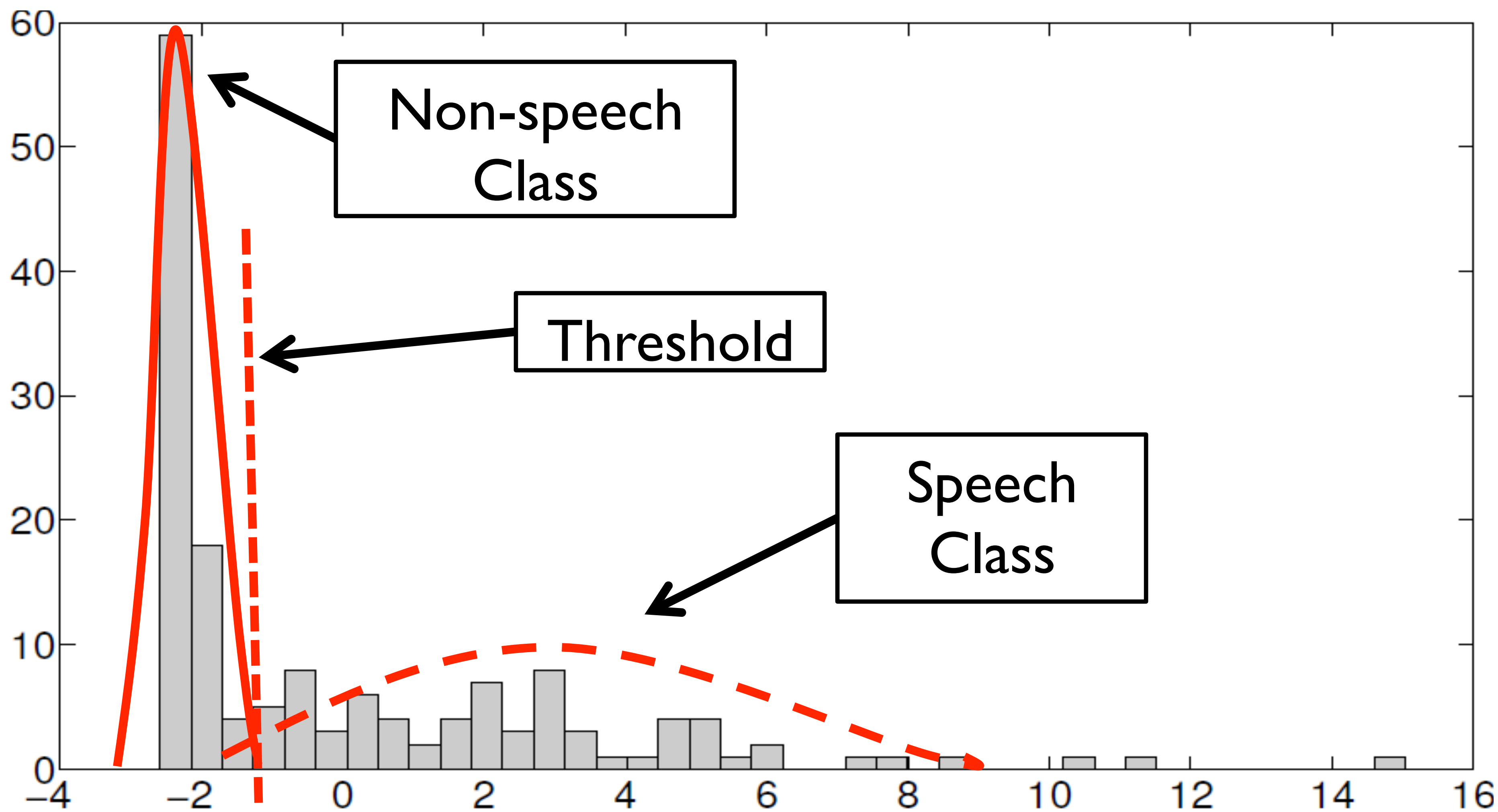
AV-SAD System: Audio Component

- Framework proposed by Sajadi and Hansen (2013)
- Audio feature (5-D)
- Principal Component Analysis (PCA) on audio feature: 1-D combo feature



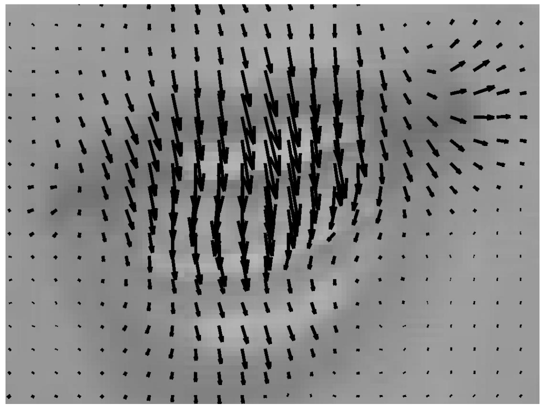
Unsupervised A-SAD

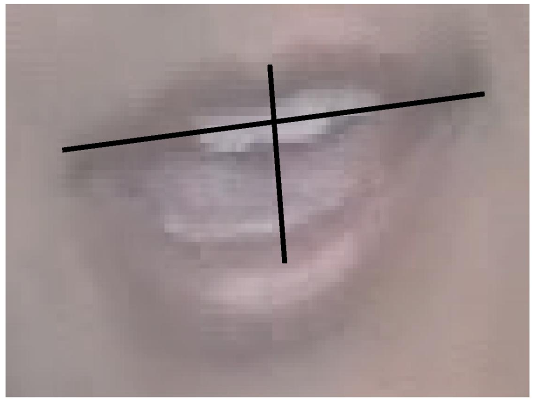
- Unsupervised clustering with EM approach



AV-SAD System: Video Component

- Video feature [Tao et al (2015)]:
  - Optical flow: OFx, OFy and OFx+OFy (OFxy)
  - Geometric feature: height (H), width (W), W x H and H+W
  - Short term statistics (0.3 s window)





Feature Set

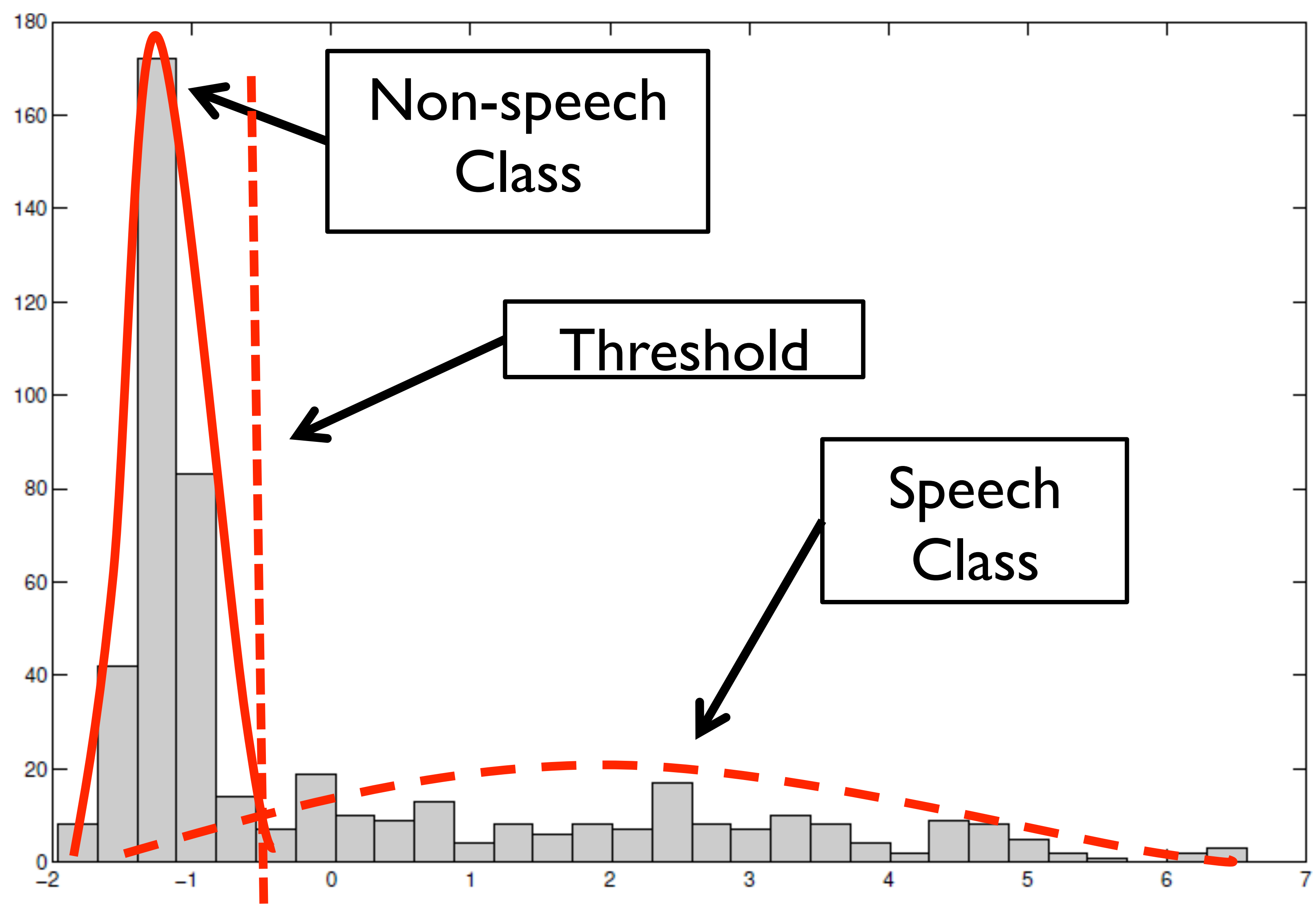
Set	OFx	OFy	OFxy	H	W	W+H	WxH
Temporal Variance	✓	✓	✓	✓	✓	✓	✓
Zero Crossing Rate	✓	✓	✓	✓	✓	✓	✓
Speech Periodic Characteristic	✓	✓	✓	✓	✓	✓	✓
First Order Derivative				✓	✓	✓	✓

25-D feature in total

Unsupervised V-SAD

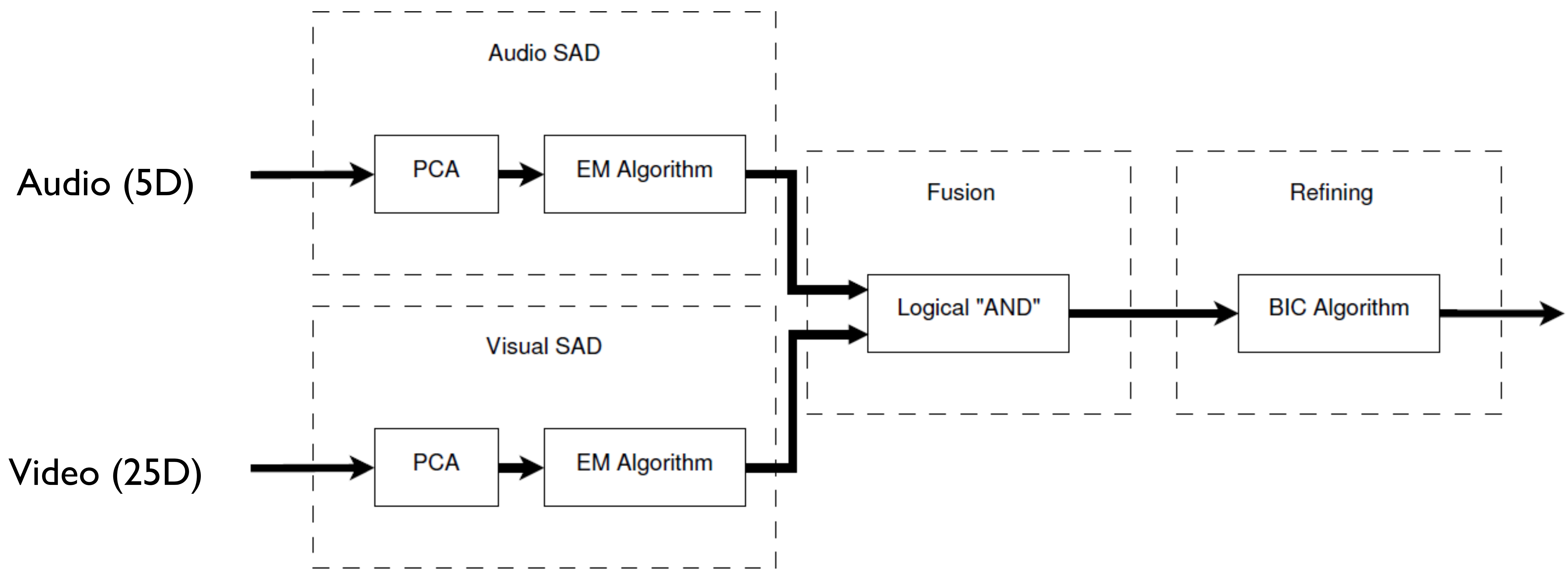
- Similar approach to unsupervised A-SAD
  - PCA on 25-D feature
  - EM to form two classes on “combo” feature

25-D → PCA → Visual combo feature



Proposed Approach

- Unsupervised A-SAD and V-SAD [Sajadi and Hansen (2013), Tao et al (2015)]:
- Audio-visual fusion
  - Logical fusion: "AND" and "OR"
- BIC refine



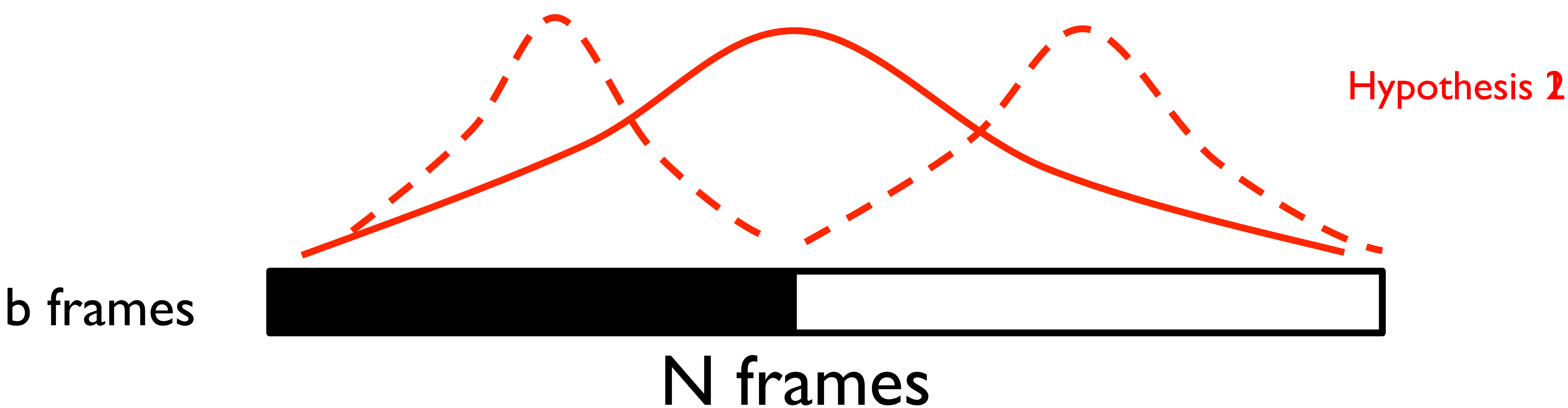
Bayesian Information Criterion (BIC) Refine

- The BIC is a criterion used to select a model among potential candidate models [Zhou and Hansen (2005)]
  - Hypothesis 1 (H1): one single distribution
  - Hypothesis 2 (H2): bimodal distribution
  - $\Delta BIC = BIC(H2) - BIC(H1)$

$$\mathrm{BIC}(H_2) = -\frac{b}{2}\log|\hat{\Sigma}_1| - \frac{N-b}{2}\log|\hat{\Sigma}_2| - \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\cdot\log N$$

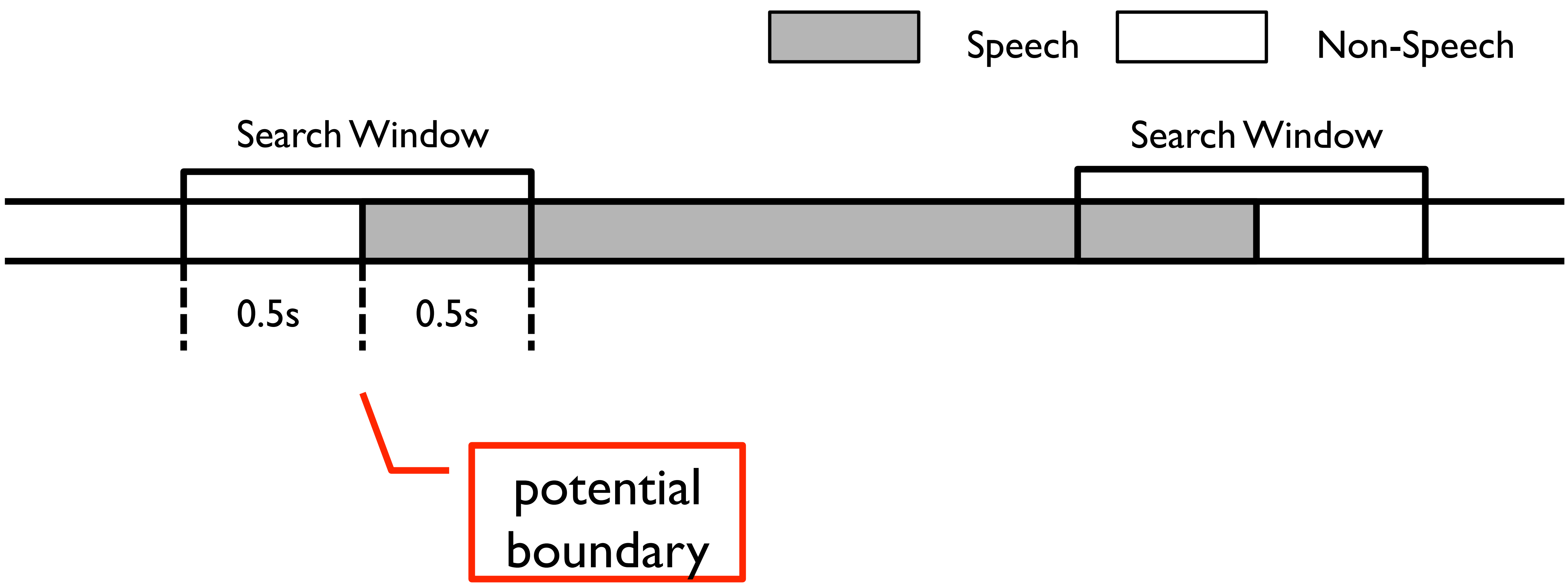
d is the feature dimension

$\hat{\Sigma}$ is covariance of N frames,



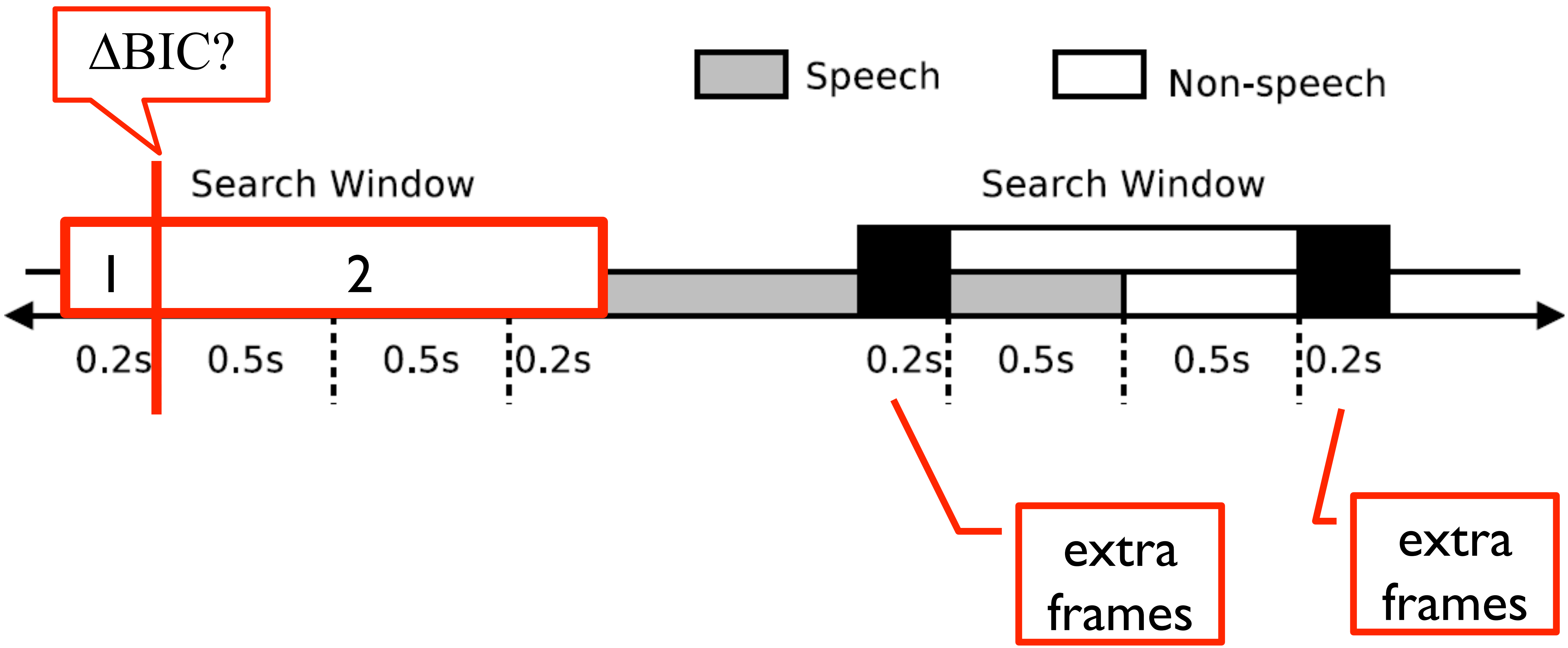
Bayesian Information Criterion (BIC) Refine

- Focus on transition area
  - Potential boundary given by previous steps
  - ΔBIC computed for each frame in search window
  - Extra frames before and after search window



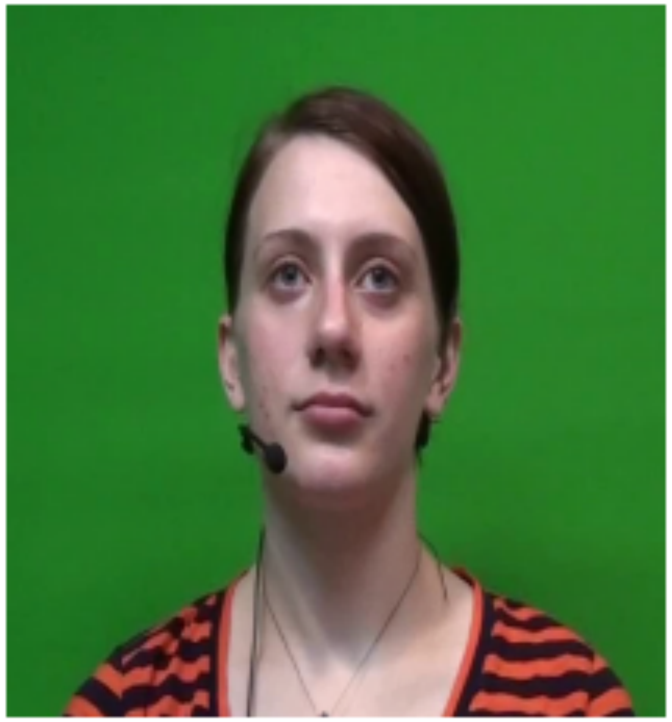
Bayesian Information Criterion (BIC) Refine

- Focus on transition area
  - Potential boundary given by previous steps
  - ΔBIC computed for each frame in search window
  - Extra frames before and after search window



Corpus Description

- MSP Audio-visual Whisper (MSP-AVW) corpus
  - 20 males and 20 females
  - 120 TIMIT sentences per speaker (60 in neutral, 60 in whisper)
  - Audio: SHURE 48 KHz close-talk microphone
  - Video: high definition SONY cameras (1440 × 1080) at 29.97 fps



Experiment and Result

- Performance without BIC
  - Whisper decreases performance by ~20%
  - V-SAD is robust to different modes
  - Under neutral condition, the fusion decreases the performance by ~5%
    - The ground truth of the labels was annotated based only on audio
    - Original sampling frequency is low (29.97 fps)
  - Under whisper condition, the fusion improves the performance by ~8%

Modality	Set	Acc [%]	Pre [%]	Rec [%]	F [%]
A-SAD	Nsen	94.05	97.15	89.85	93.35
	Wsen	67.96	61.02	88.65	72.28
V-SAD	Nsen	78.06	75.11	89.45	80.40
	Wsen	78.20	72.69	89.10	80.06
AV-SAD	Nsen	89.47	97.90	79.93	88.00
	Wsen	81.28	81.73	79.21	80.45

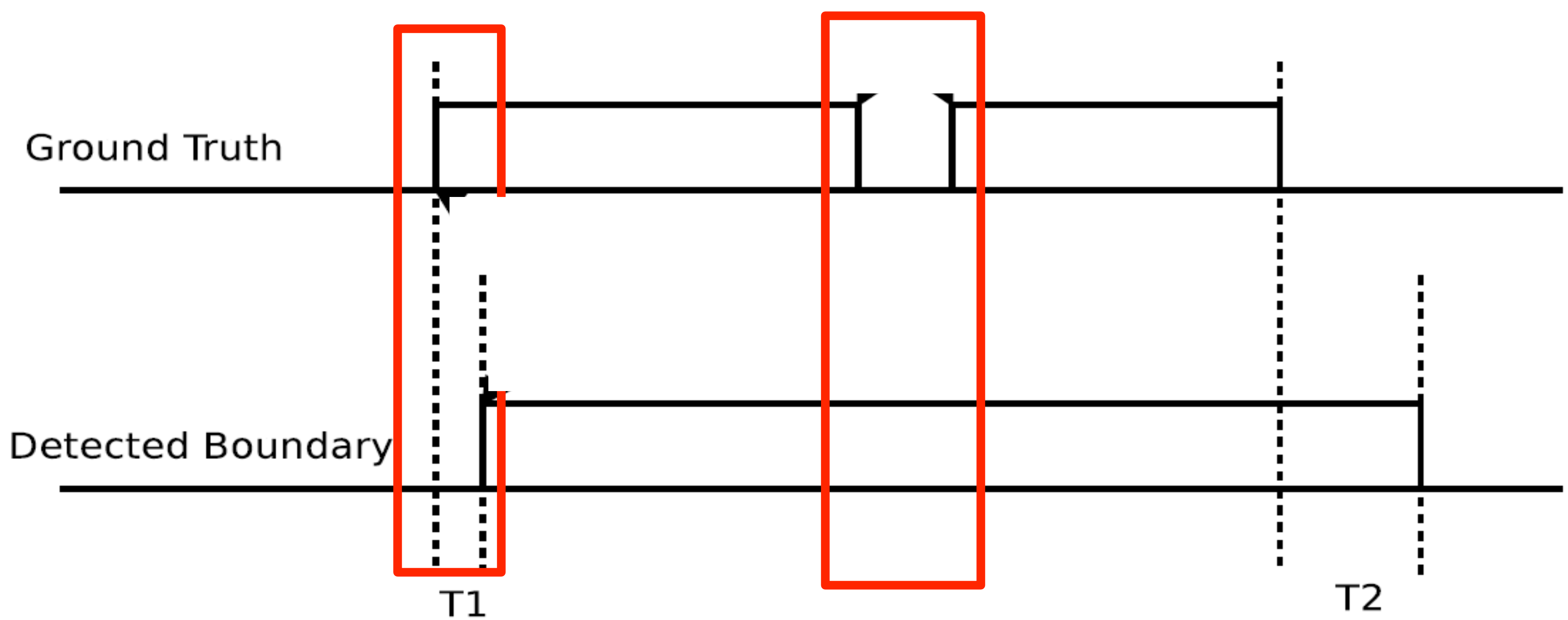
- Performance with BIC:
  - Apply BIC on detected boundary from AV-SAD

	Set	ACC [%]	Pre [%]	Rec [%]	F [%]
AV-SAD	Nsen	89.47	97.90	79.93	88.00
	Wsen	81.28	81.73	79.21	80.45
AV-SAD + A-BIC	Nsen	91.11	97.47	83.77	90.10
	Wsen	82.91	84.47	79.48	81.90
AV-SAD + V-BIC	Nsen	88.53	92.22	83.18	87.47
	Wsen	78.67	76.63	80.54	78.53
AV-SAD + AV-BIC	Nsen	91.25	97.49	84.05	90.27
	Wsen	82.87	83.76	80.37	82.03

- A-BIC improves the system:
  - For speech detection, ~2% absolute improvement
- V-BIC impairs the system
  - Modalities mismatch
- AV-BIC achieves best performance on speech detection

Median Local Boundary Mismatch

- Local Boundary Mismatch (LBM)
  - the mismatch frames between the detected boundary and ground truth in local regions



- Median Local Boundary Mismatch (MLBM)
  - Represents the boundary detection performance
  - Lower is better

- Boundary detection performance:
  - Up-sampling to 100 fps for MLBM comparison

	Set	MLBM [fps]
AV-SAD	Nsen	35.00
	Wsen	64.00
AV-SAD + A-BIC	Nsen	25.00
	Wsen	56.00
AV-SAD + V-BIC	Nsen	42.00
	Wsen	71.00
AV-SAD + AV-BIC	Nsen	25.00
	Wsen	53.00

- A-BIC improves the system:
  - For MLBM, relatively improve 28.5% under neutral and 12.5% under whisper
- V-BIC impairs the system
  - Modalities mismatch
- AV-BIC achieves best performance on boundary detection

Conclusion and Future Work

- Conclusion
  - AV-SAD is explored showing that visual modality will improve robustness under whisper condition
  - Proposed a approach to improve boundary detection in SAD by BIC
  - AV-BIC achieves best performance
- Future Work
  - Better fusion approach need be explored

THANK YOU !

QUESTION?