INTERSPEECH 2015
September 6 - 10
Dresden, Germany

An Unsupervised Visual-only Voice Activity Detection Approach Using Temporal Orofacial Features

Fei Tao
John H.L. Hansen
Carlos Busso

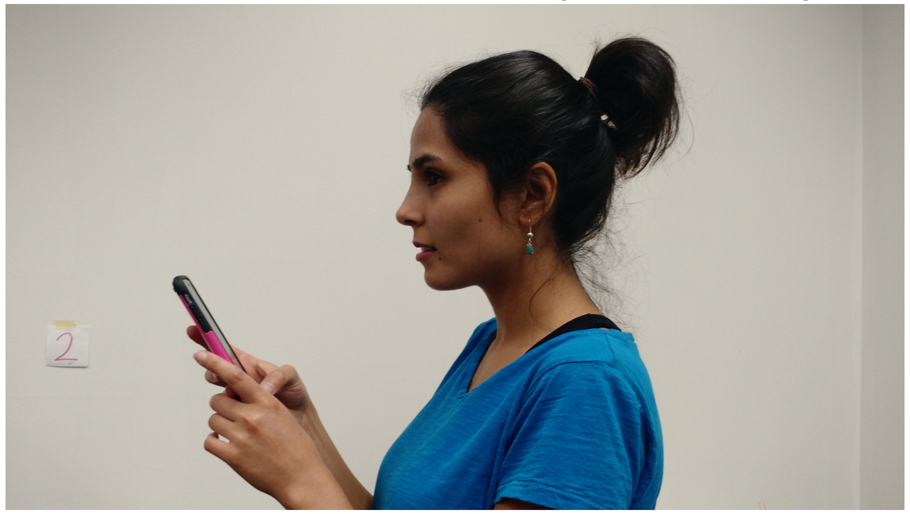
Multimodal Signal Processing (MSP) Laboratory ,
Center for Robust Speech Systems (CRSS),
Department of Electrical Engineering,
The University of Texas at Dallas,
Richardson TX 75080, USA

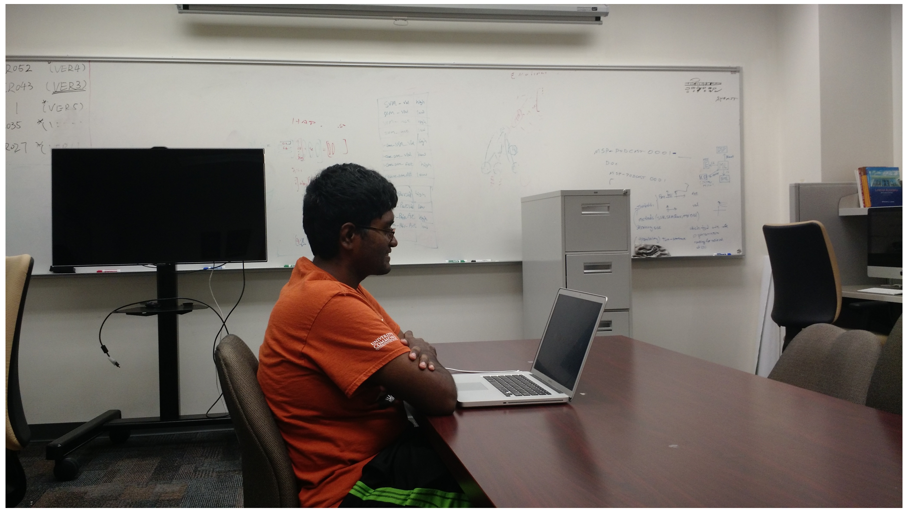
Outline

- Introduction
- Related Work
- Corpus Description
- Proposed Approach
- Experiment and Result
- Conclusions and Future Work

Introduction

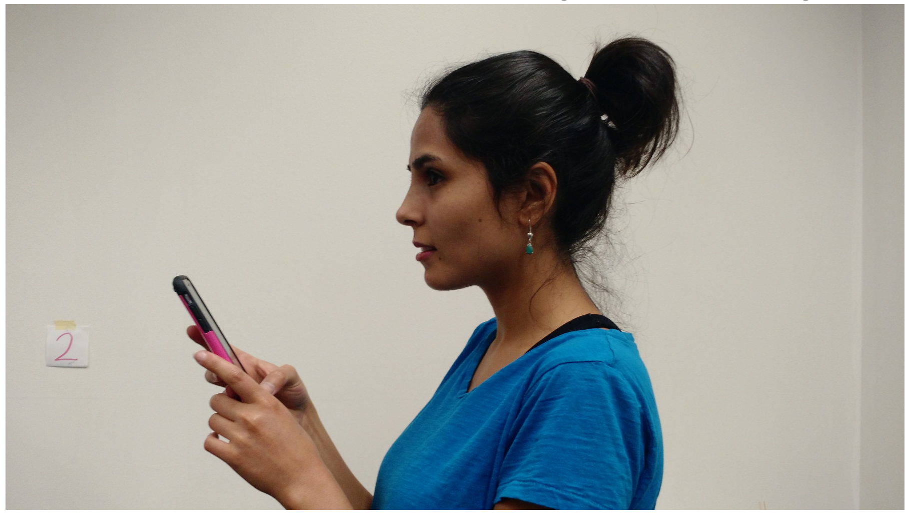
- Voice Activity Detection (VAD) plays an important role in speech-based interfaces
- Audio based VAD (AVAD) has challenges:
  - Background noise
  - Different speech modes (e.g. emotion, soft speech , whisper)
- Visual VAD (VVAD) becomes an alternative

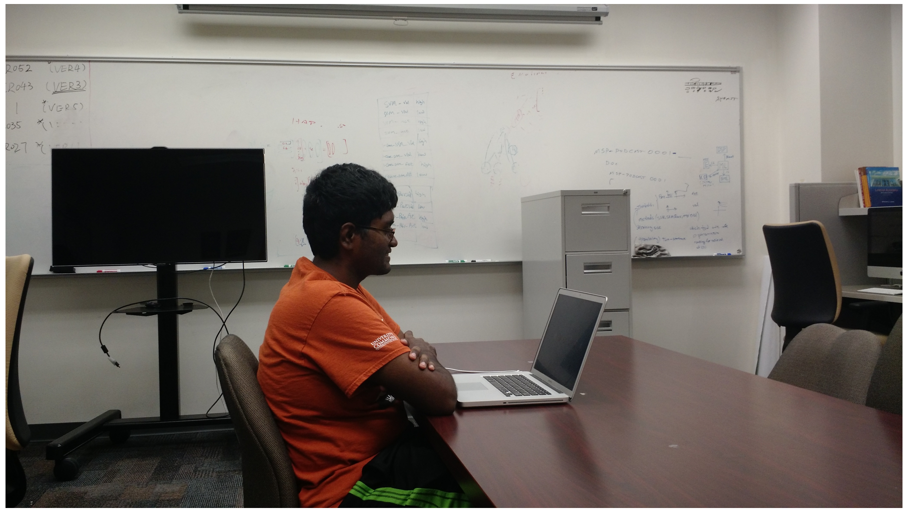




Introduction

- Voice Activity Detection (VAD) plays an important role in speech-based interfaces
- Audio based VAD (AVAD) has challenges:
  - Background noise
  - Different speech modes (e.g. emotion, soft speech , **whisper**)
- Visual VAD (VVAD) becomes an alternative





Related Work

Supervised:

- Navarathna et al. [2011] extracted discrete cosine transform coefficients around mouth and augment them by their derivative.
- Aubery et al. [2007] used active appearance model and retinal filter to detect speech activity based on HMM
- Takeuchi et al. [2009] extracted the variance of optical flow as visual features and proposed audiovisual VAD system

Related Work (Cont.)

Unsupervised:

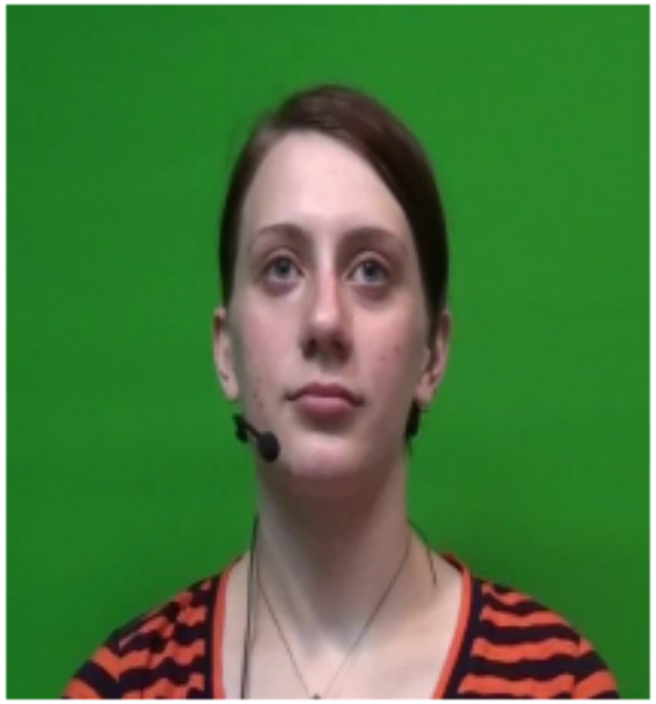
- Sodoyer et al. [2006] proposed an unsupervised method to detect lip activity by adopting a threshold.
- Sadjadi and Hansen[2013] proposed a state-of-the-art unsupervised approach for AVAD

Benefit:

- No training data
- Adapt to testing conditions
- Unsupervised approach offers more flexibility

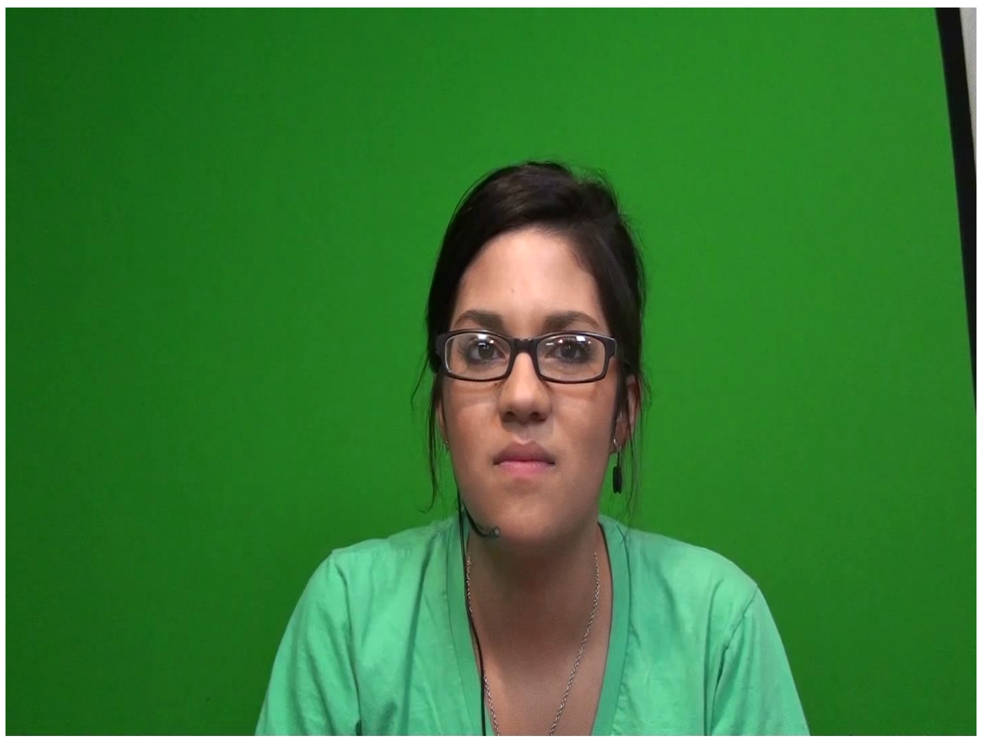
Corpus Description

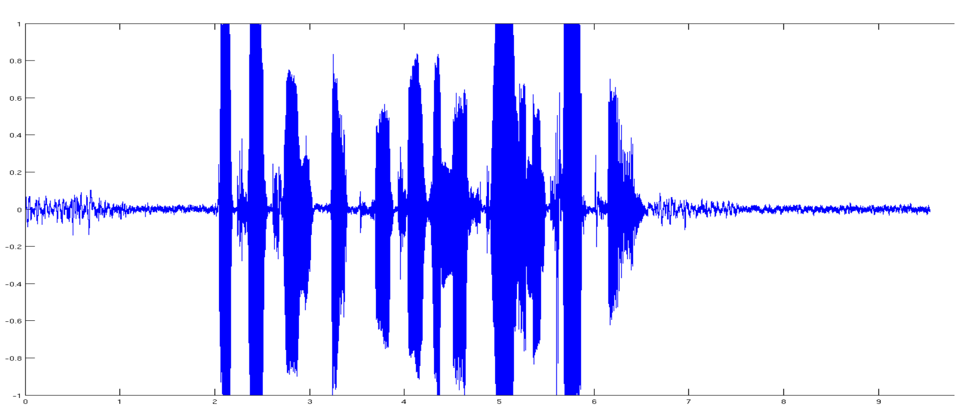
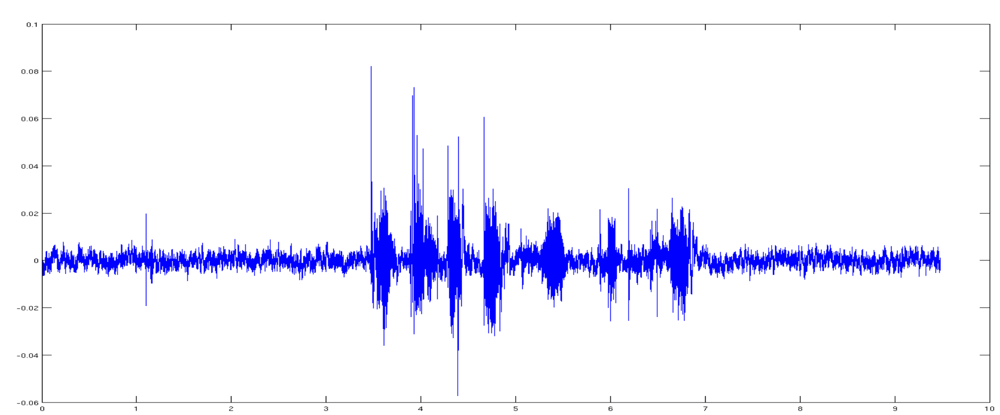
- Audio-visual Whisper (AVW) corpus
- 20 males and 20 females
- Corpus consists of
  - Digits
  - Read sentence (120 TIMIT sentences: 60 in neutral and 60 in whisper)
  - Spontaneous talk
- Audio collected with a SHURE 48 KHz close-talk microphone



Corpus Description

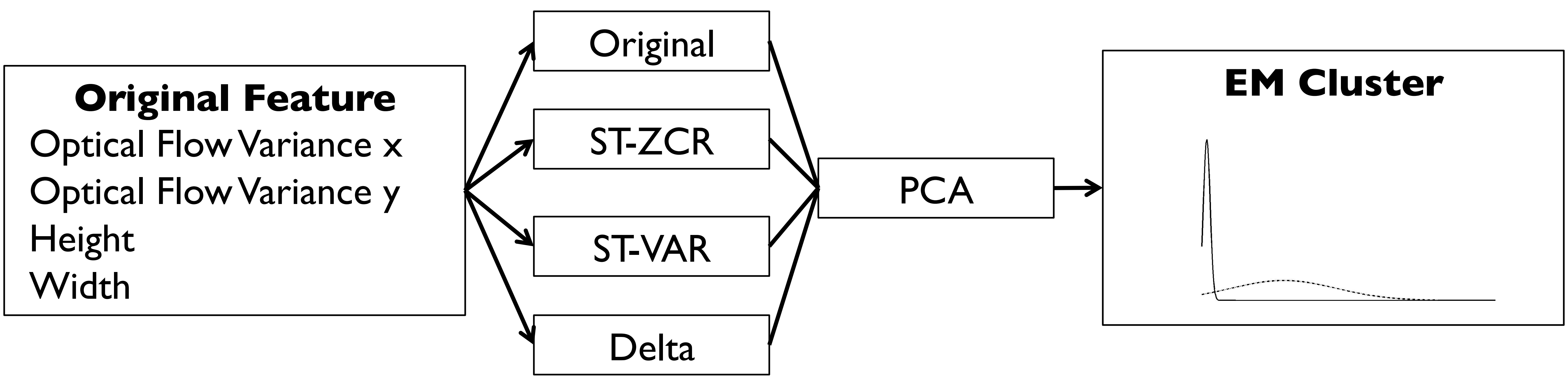
- Video collected with high definition SONY cameras (1440 × 1080) at 29.97 fps (label based on audio)



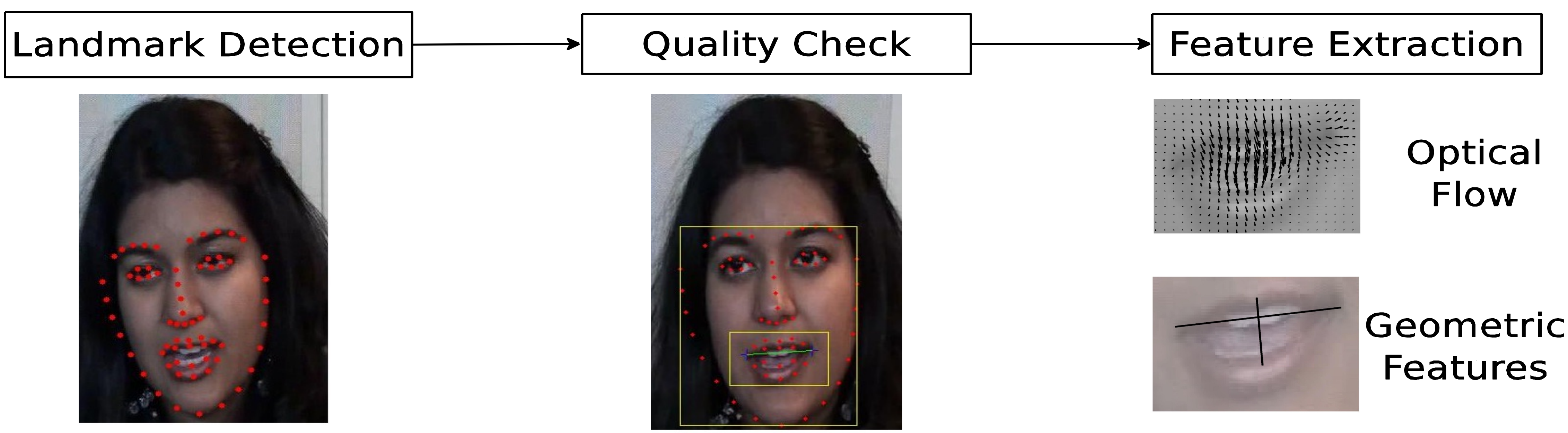
Proposed Approach

- Video processing and facial feature extraction
- Estimation of dynamic and temporal features
- Principle component analysis (PCA)
- Expectation maximum (EM) algorithm for clustering



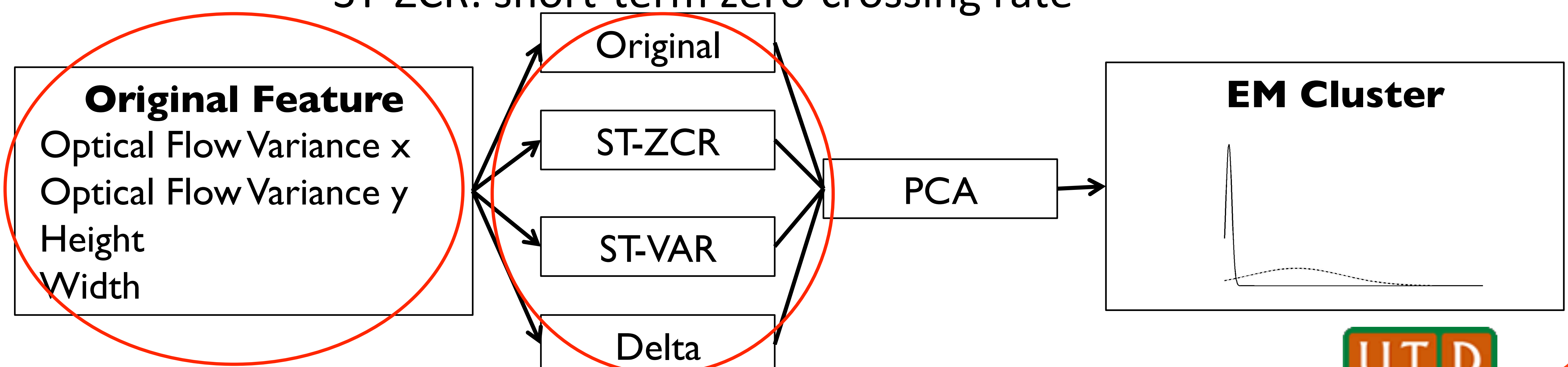
Feature Extraction

- 66 landmarks detected by CSIRO [Cox et al., 2013]
- Quality check with the outputs from another system
- Orofacial feature extraction:
  - height(H) and width(W)
  - variance of optical flow in x direction(OFx) and y direction (OFy)



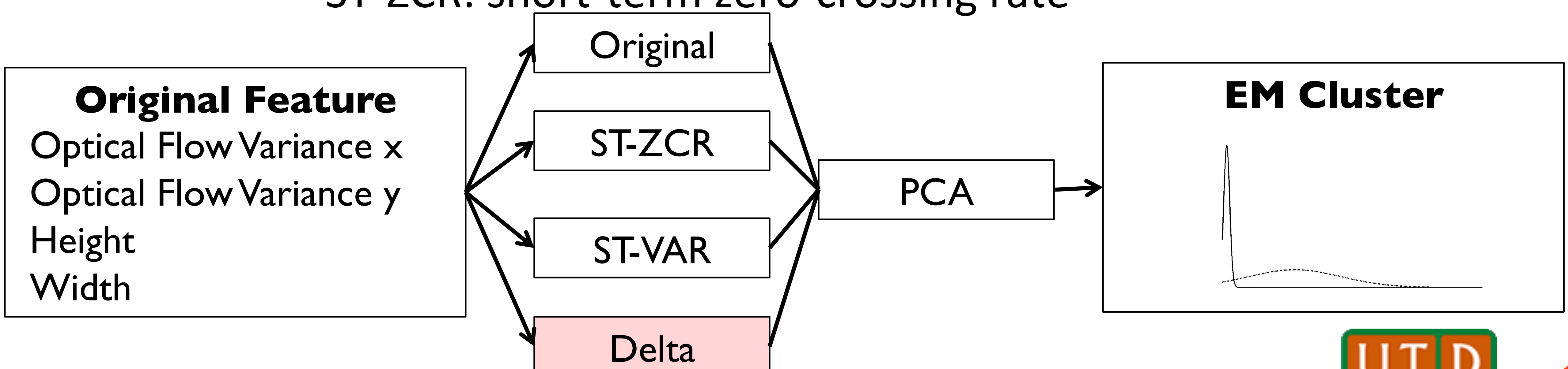
Dynamic and Temporal Features

- Facial feature vector (7D):
  - Overall optical flow variance (OF_{xy}): $OF_x + OF_y$
  - Overall distance (H + W) & approximate area (H × W)
- Statistics over facial feature vector
  - Dynamic features
    - Delta: first order difference
  - Temporal features over 7D vector:
    - ST-VAR: short-term (0.3s) variance
    - ST-ZCR: short-term zero-crossing rate



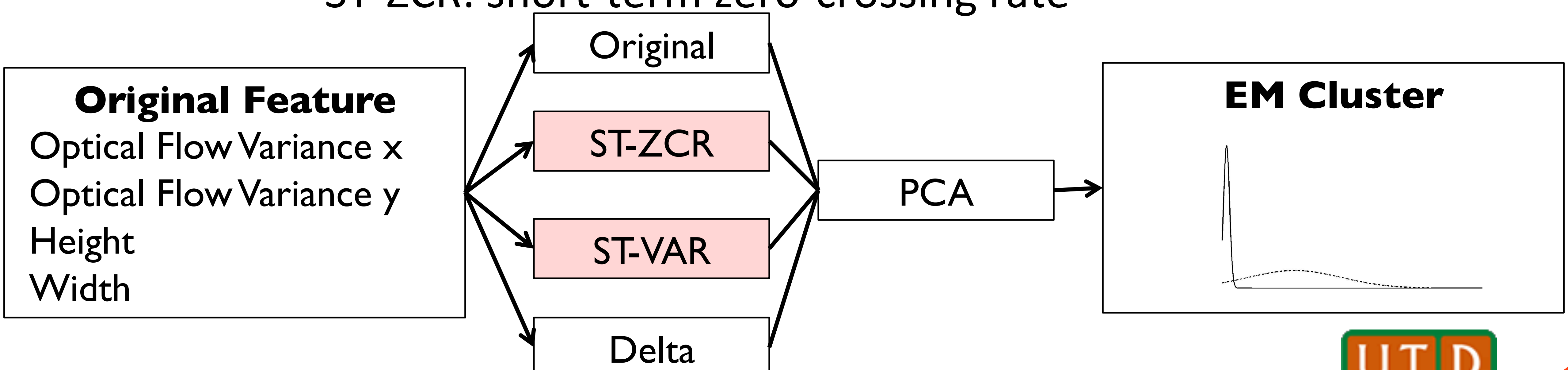
Dynamic and Temporal Features

- Facial feature vector (7D):
  - Overall optical flow variance (OF_{xy}): $OF_x + OF_y$
  - Overall distance (H + W) & approximate area (H × W)
- Statistics over facial feature vector
  - Dynamic features
    - Delta: first order difference
  - Temporal features over 7D vector:
    - ST-VAR: short-term (0.3s) variance
    - ST-ZCR: short-term zero-crossing rate



Dynamic and Temporal Features

- Facial feature vector (7D):
  - Overall optical flow variance (OF_{xy}): $OF_x + OF_y$
  - Overall distance (H + W) & approximate area (H × W)
- Statistics over facial feature vector
  - Dynamic features
    - Delta: first order difference
  - Temporal features over 7D vector:
    - ST-VAR: short-term (0.3s) variance
    - ST-ZCR: short-term zero-crossing rate



Feature Set

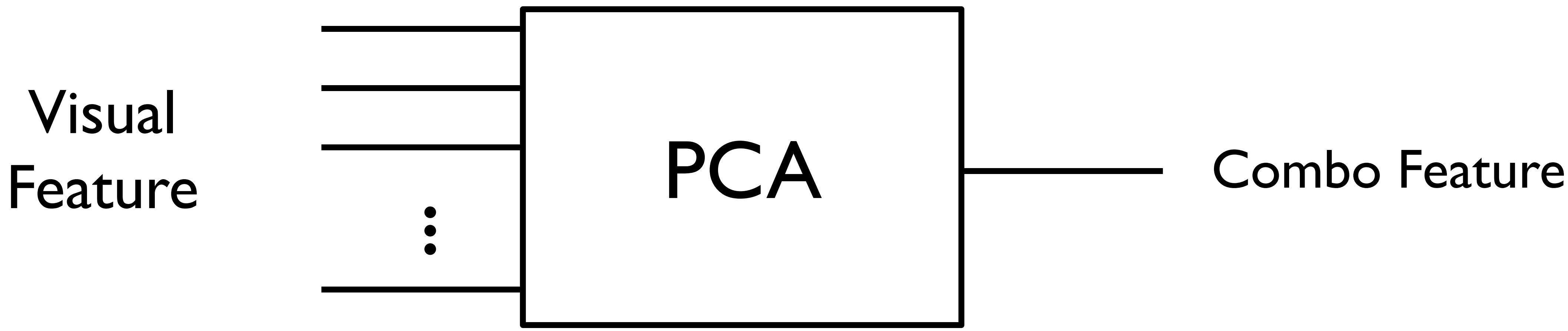
- Final feature vector consists of 19 features

Set	OFx	OFy	OFxy	H	W	H+W	H×W
Original			X				
Delta*				X	X	X	X
ST-VAR*	X	X	X	X	X	X	X
ST-ZCR*	X	X	X	X	X	X	X

•ST-ZCR: short term zero crossing rate;
•ST-VAR: short term variance;
•Delta: first order difference

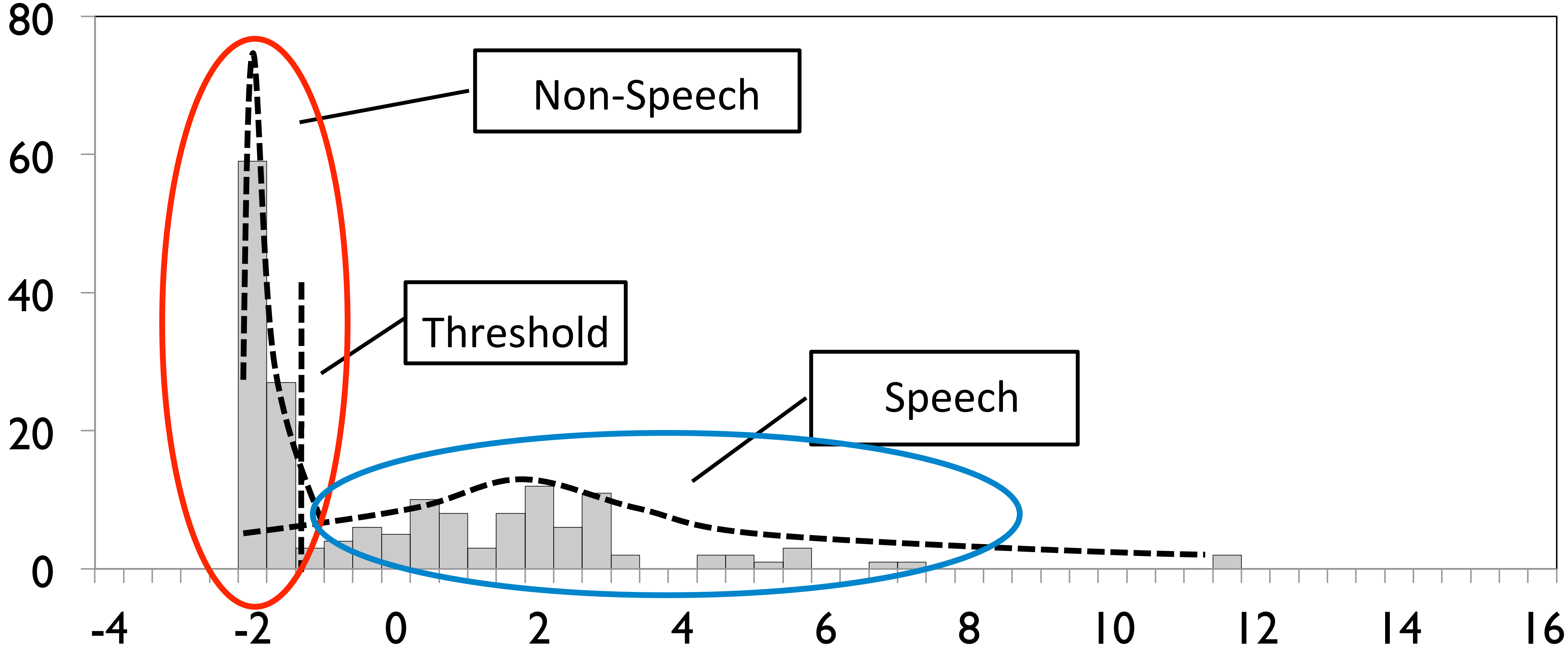
Unsupervised Classification

- Principle component analysis (PCA) applied on final feature to form a 1-D combo feature
  - Inspired by Sadjadi and Hansen [2013]



Unsupervised Classification

- Principle component analysis (PCA) applied on final feature to form a 1-D combo feature
- Expectation maximum (EM) algorithm is run for clustering



Baseline AVAD

- Audio only VAD (proposed by Sadjadi and Hansen [2013]):
  - 5D feature: Harmonicity, Clarity, Prediction Gain, Preodicity, Perceptual Spectral Flux
  - Changing speech mode impair the system performance (20% drop)

Set	Precision[%]	Recall[%]	F-score[%]	Accuracy[%]
Neutral	91.3	98.0	**94.5**	93.9
Whisper	78.7	72.3	**75.3**	74.8

$$\mathbf{F - Score = 2 \times \frac{Precesion \ \times Recall}{Precesion + Recall}}$$

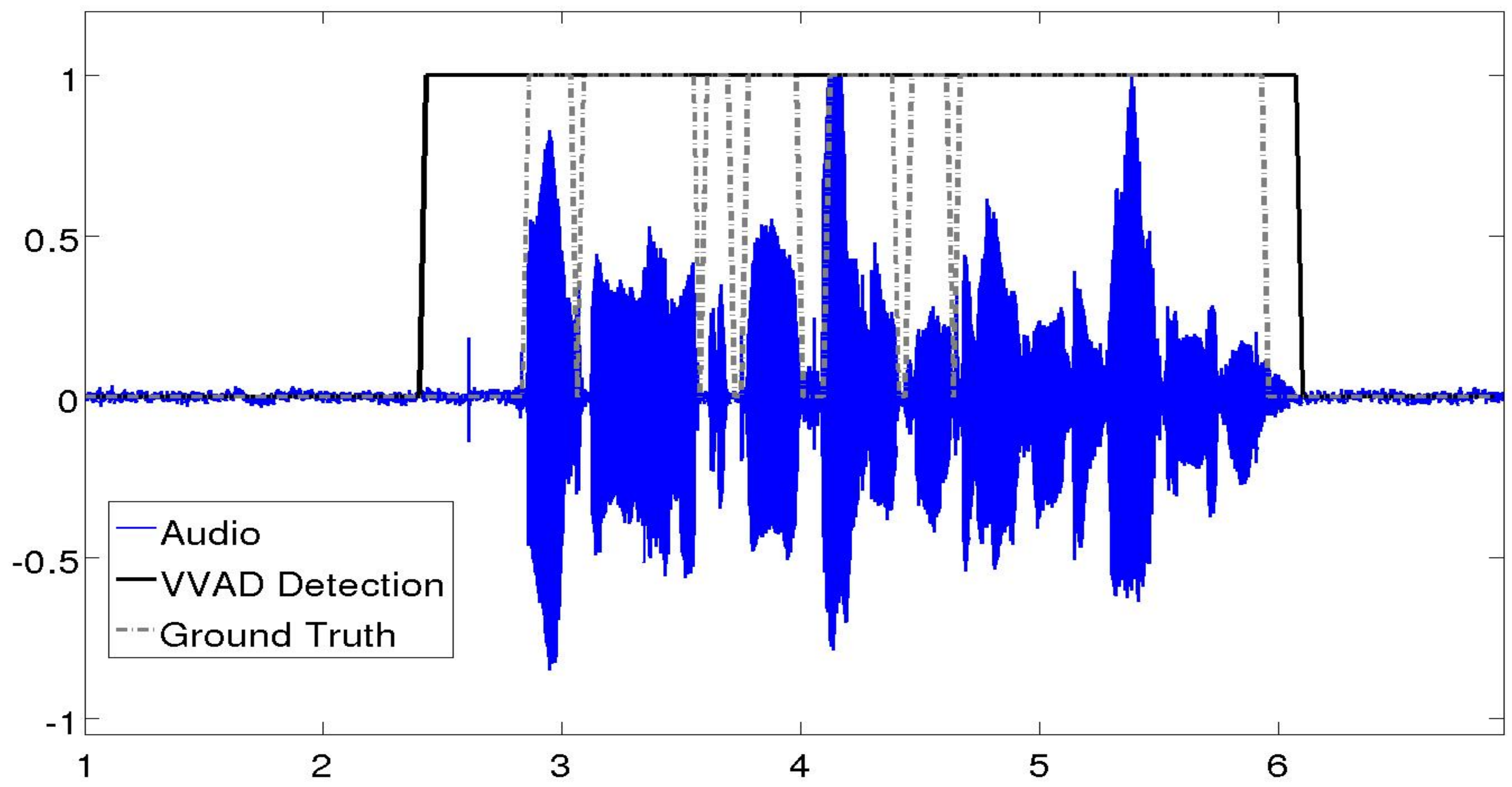
Experiment and Results

- Video only VAD (proposed approach):
  - Visual cues are robust to different speech modes
  - For neutral sentence, the performance is about 13% lower than AVAD system
  - For whispered sentence, the performance is about 6% higher than AVAD system

Set	Precision[%]	Recall[%]	F-score[%]	Accuracy[%]
Neutral	90.7	73.8	**81.4**	80.0
Whisper	90.3	73.5	**81.1**	79.4

Compare AVAD and VVAD

- Anticipatory movement of lips
- Lower resolution for visual modality



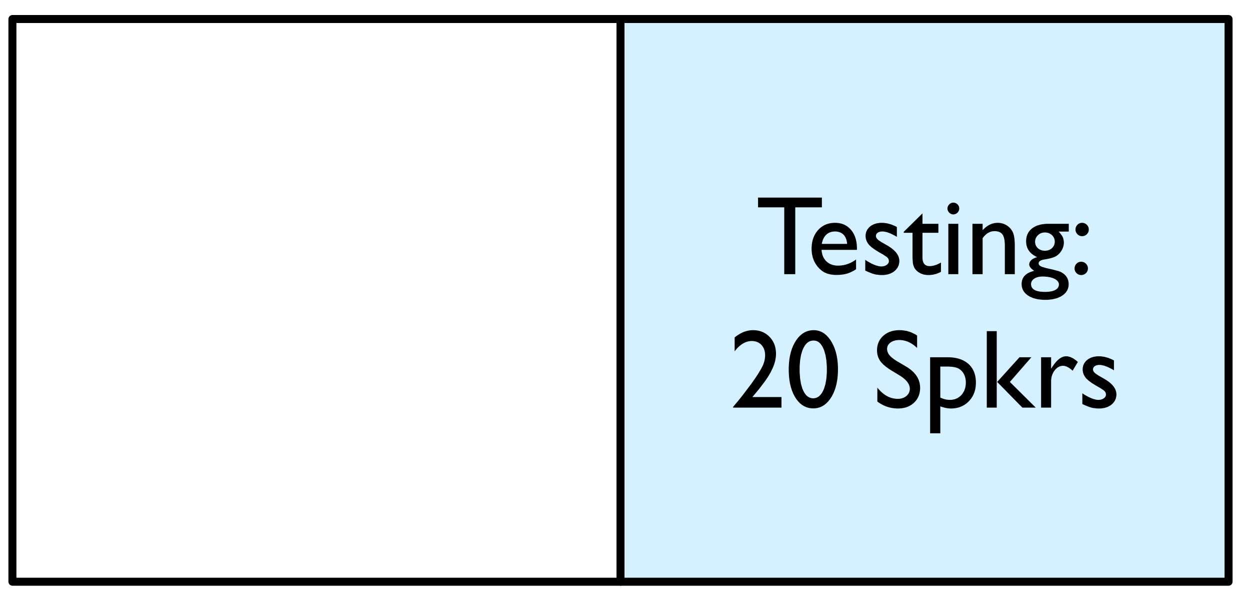
Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set

Training: 20 Spkrs	Testing: 20 Spkrs

Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set



Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set

Training: 20 Spkrs	Testing: 20 Spkrs

Compare Supervised and Unsupervised

- Training set: 20 speakers; testing set: 20 speakers
- Unsupervised setting:
  - Proposed approach is applied on the testing data
- Supervised setting:
  - Linear kernel SVM built with training set

Set	Supervised VVAD				Unsupervised VVAD			
	P[%]	R[%]	F[%]	A[%]	P[%]	R[%]	F[%]	A[%]
Neutral	89.1	84.3	**86.6**	86.9	90.5	73.1	**80.8**	79.1
Whisper	88.7	84.2	**86.4**	86.7	90.1	73.7	**81.1**	79.2

Benefits of Supervised Approach

- Supervised approach is 5% higher than unsupervised approach
  - Trade-off
- Unsupervised approach is 5% higher when tested on a different corpus
- Benefits of supervised approach is gone

Conclusions and Future Work

- A new unsupervised VVAD approach is proposed
- The proposed approach is robust to speech mode changing
- Audiovisual VAD will be explored in future to improve the performance under the neutral mode

Acknowledge

- Thanks to National Science Foundation (NSF)

- Reference:

  - R. Navarathna, D. Dean, S. Sridharan, C. Fookes, and P. Lucey,"Visual voice activity detection using frontal versus profile views," DICTA 2011, Noosa,Queensland, Australia, December 2011, pp. 134–139.
  - A. Aubrey, B. Rivet, Y. Hicks, L. Girin, J. Chambers, and C. Jutten, "Two novel visual voice activity detectors based on appearance models and retinal fillterin," EUSIPCO 2007, Pozna´n, Poland, September 2007.
  - S. Takeuchi, T. Hashiba, S. Tamura, and S. Hayamizu, "Voice activity detection based on fusion of audio and visual information," AVSP 2009, Norwich, United Kingdom, September 2009, pp. 151–154.
  - D. Sodoyer, B. Rivet, L. Girin, J.-L. Schwartz, and C. Jutten, "An analysis of visual speech information applied to voice activity detection," ICASSP 2006, vol. 1, Toulouse, France, May 2006, pp. 601–604.
  - S. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," IEEE Signal Processing Letters, vol. 20, no. 3, pp. 197–200, M
  - M. Cox, J. Nuevo, J.Saragih and S. Lucey, "*CSIRO Face Analysis SDK*", AFGR 2013. March 2013.