Aligning Audiovisual Features for Audiovisual Speech Recognition

Fei Tao and Carlos Busso

Multimodal Signal Processing (MSP) Laboratory
Department of Electrical Engineering,
The University of Texas at Dallas,
Richardson TX-75080, USA

- **Audiovisual approach for robust ASR**



- **DNN emerges for AV-ASR**

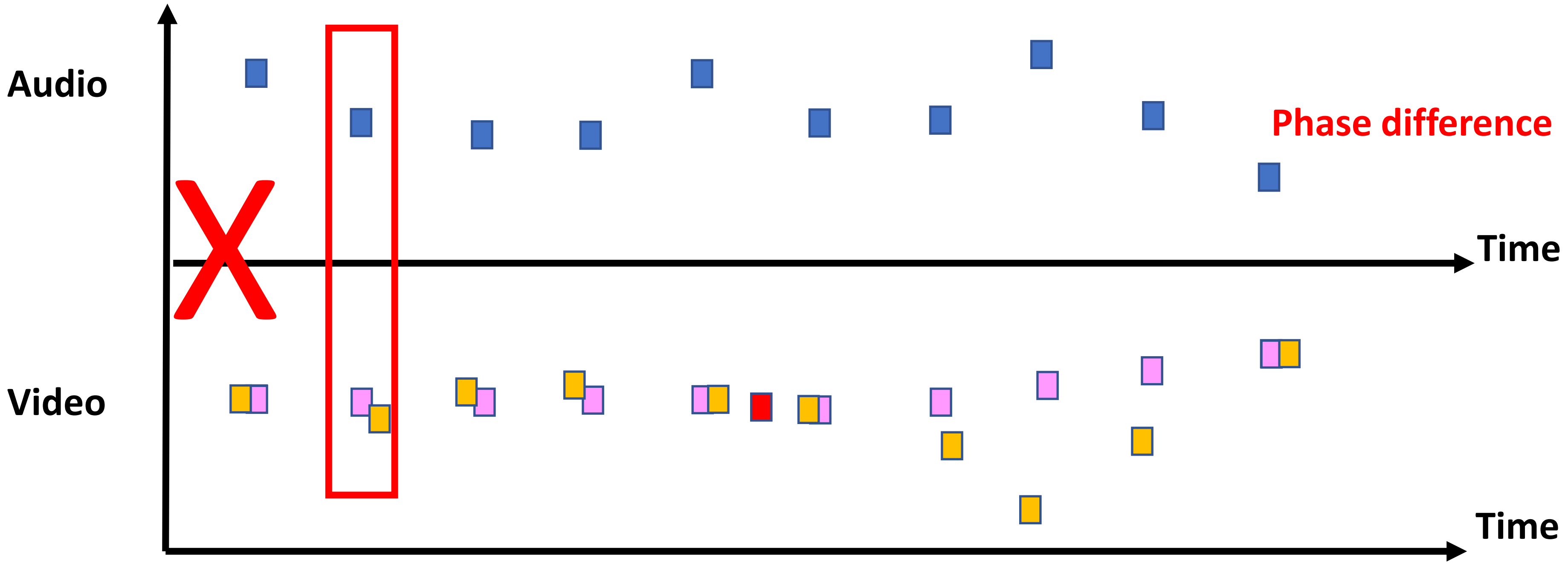
Neti et al. [2000]
GMM-HMM

Ngiam et al. [2011]
Multimodal deep learning

Petridis et al. [2018]
End-to-end AV-ASR

Introduction

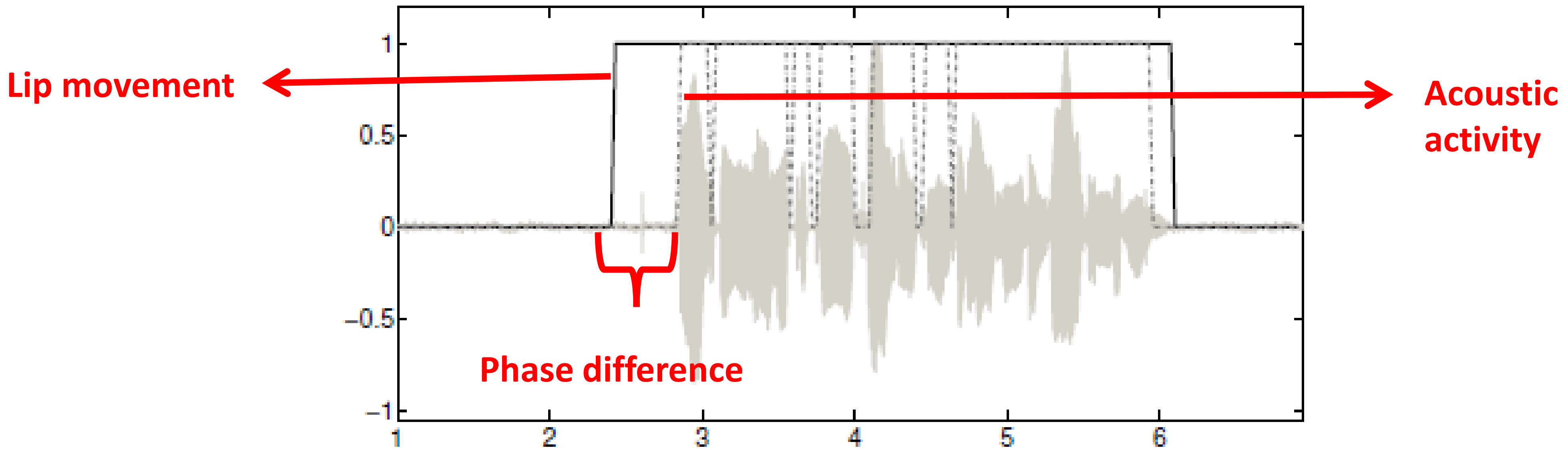
- **Fusing audiovisual features followed static fashion**
  - Linear interpolation (extrapolation) to align
- **Audiovisual modalities fusion on decision, model or feature levels.**



How to align audiovisual modalities?

Motivation

- **Phase between lip motion and speech [Tao et al., 2016]**



- **Bregler and Konig [1994] show the best alignment was with a shift of 120 milliseconds**
  - However, phase is time variant so this may not be the optimum approach

Motivation

- **Audiovisual features concatenated frame-by-frame:**
  - For some phonemes, lip movements precede speech production
  - For other phonemes, speech production precede lip movements

- **In some cases, audiovisual modalities are well aligned [Hazen, 2006]**
  - Pronounce the burst release of /b/
- **Co-articulation effects and articulator inertia may cause phase difference**
  - Lip movement can precedes audio for phoneme /m/ in transition /g/ to /m/ (e.g., word *segment*)

Deep Learning for Audiovisual

- **Deep learning for audiovisual ASR:**
  - Ninomiya et al. (2015) extracted bottleneck feature for audiovisual fusion
  - Ngiam et al. (2011) proposed bimodal DNN for fusing audiovisual modalities
  - Tao et al. (2017) extended to bimodal RNN on AV-SAD problem for modeling audiovisual temporal information
- **Rely on linear interpolation to align audiovisual features**

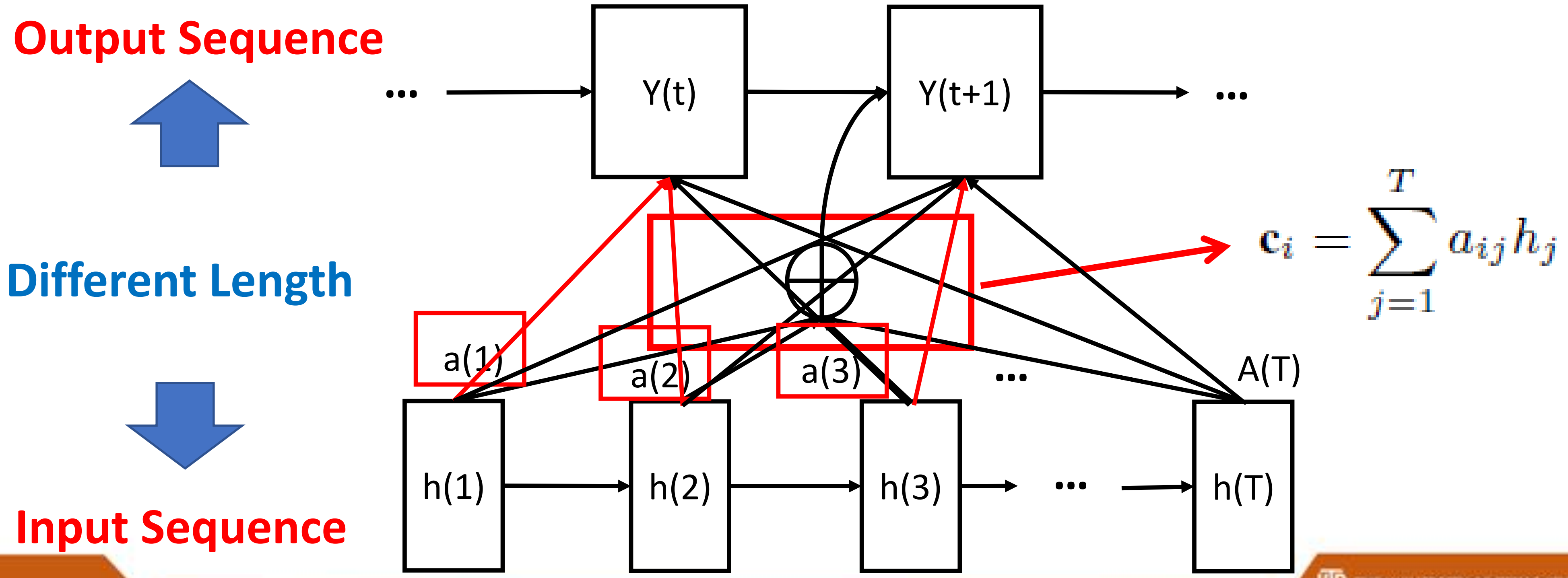
Proposed Approach:
Learn alignment automatically from data using attention model

Outline

1. Introduction
2. Proposed Approach
3. Corpus Description
4. Experiment and Result
5. Conclusion

Proposed Framework

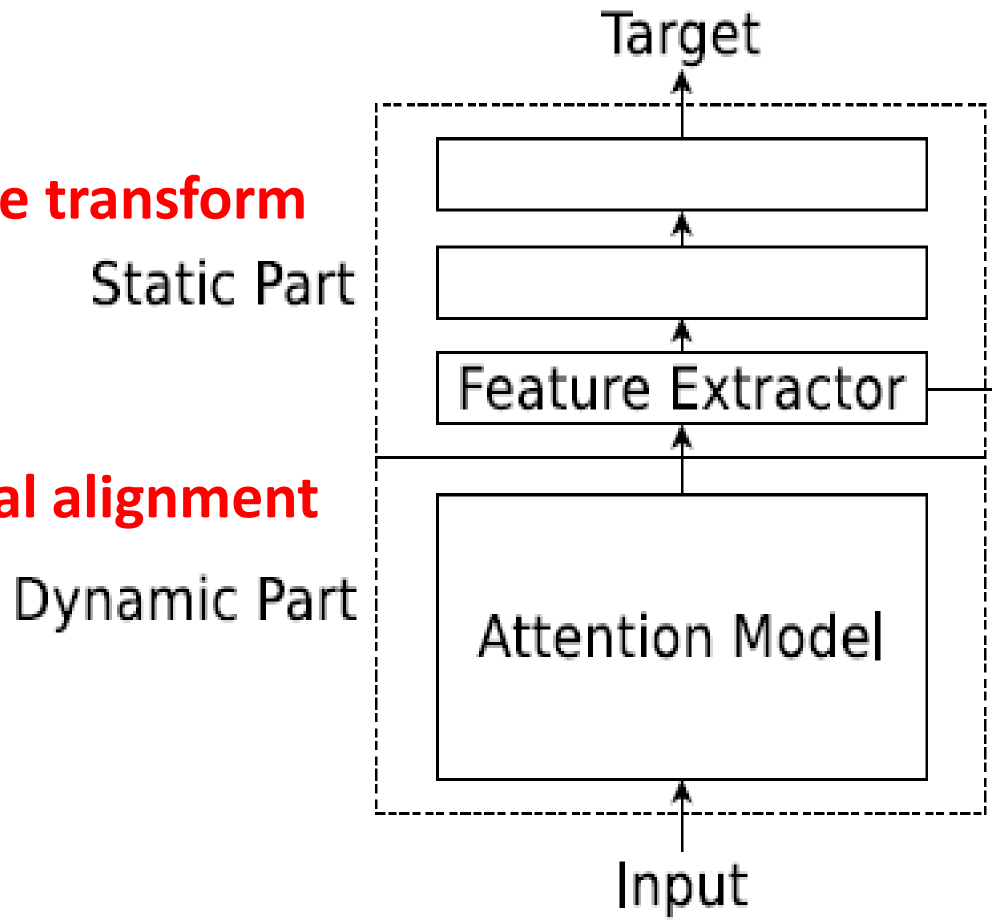
- **Proposed approach relies on attention model**
- **Attention model learns alignment in sequence-to-sequence learning**
  - Output is represented as linear combination of input at all time points
  - Learn the weights in linear combination following a data-driven framework



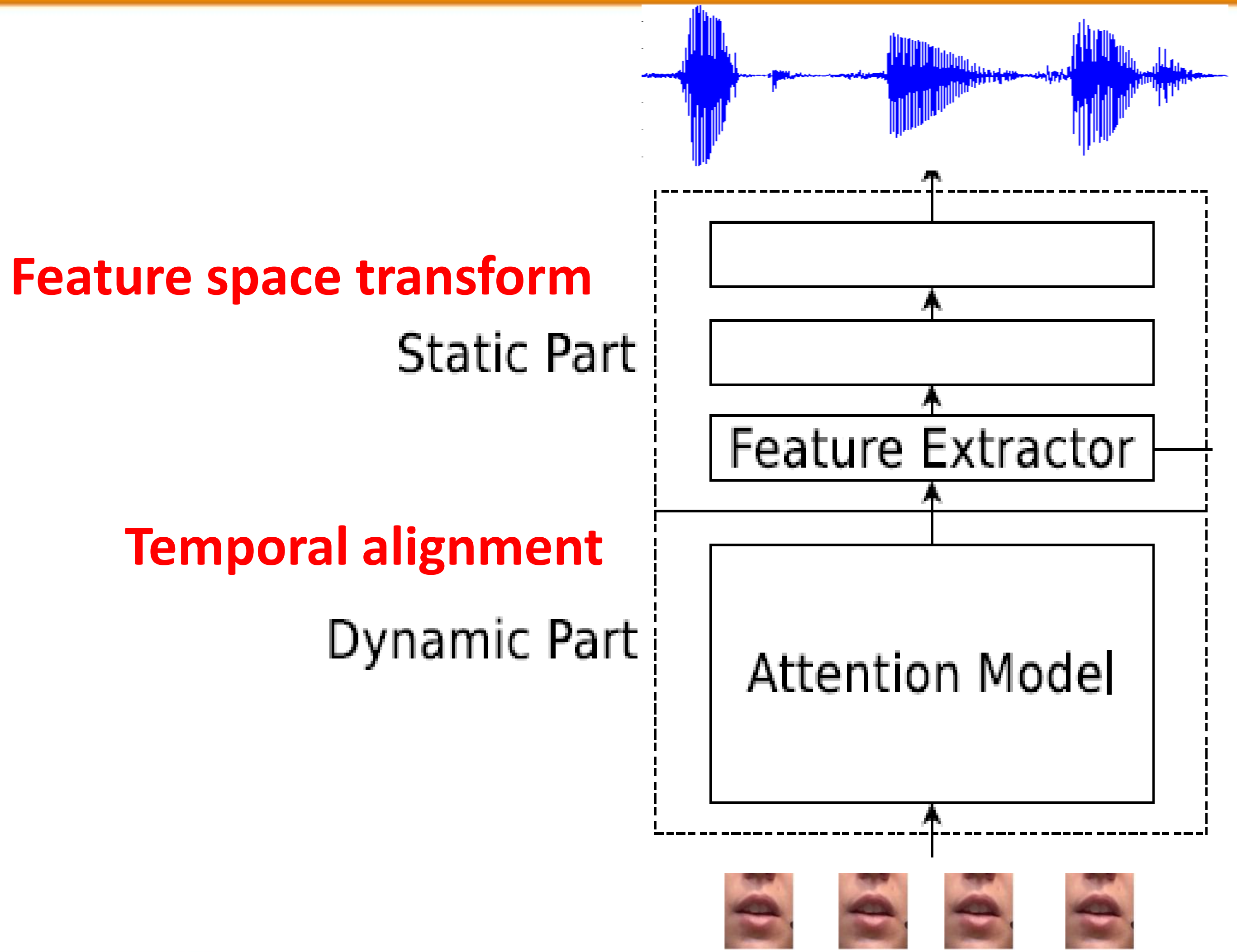
Alignment Neural Network (AliNN)

Feature space transform

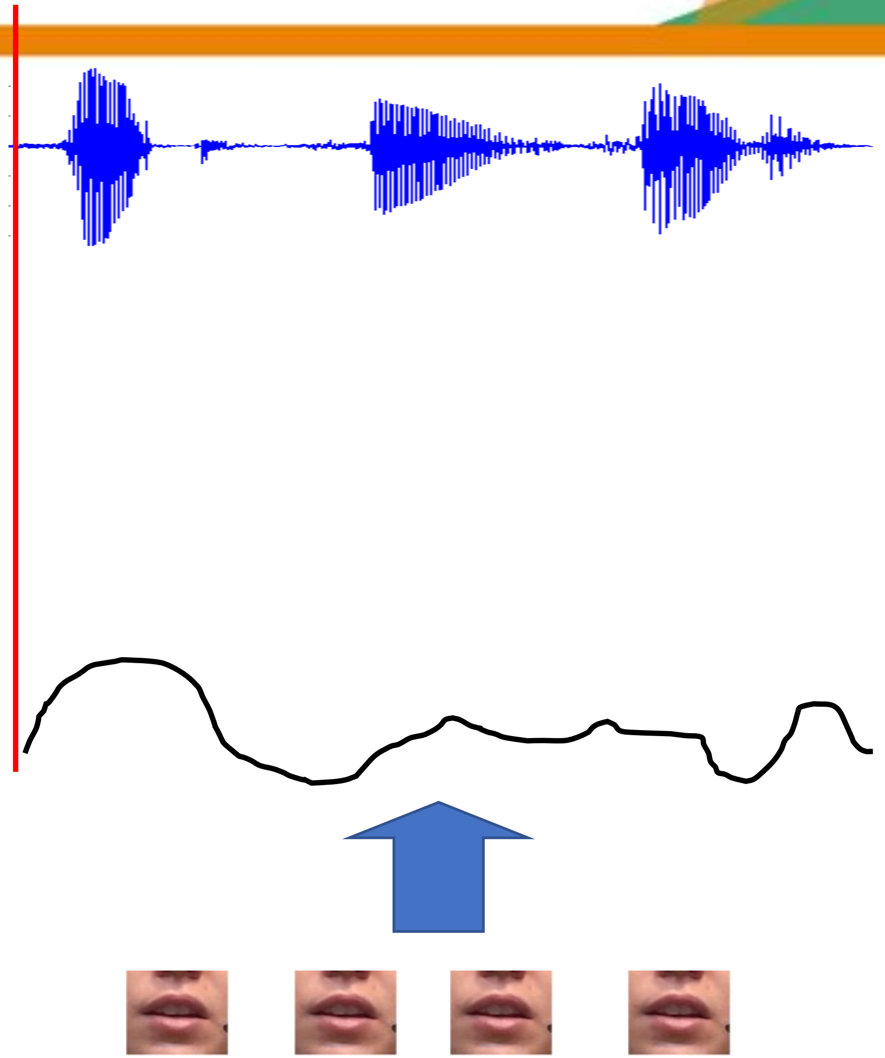
Temporal alignment

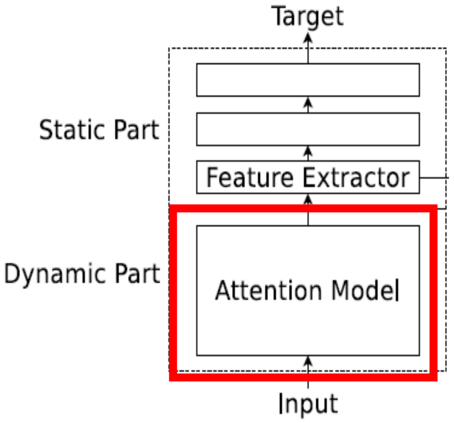


Alignment Neural Network (AliNN)



Alignment Neural Network (AliNN)





Temporal align

Dynamic Part

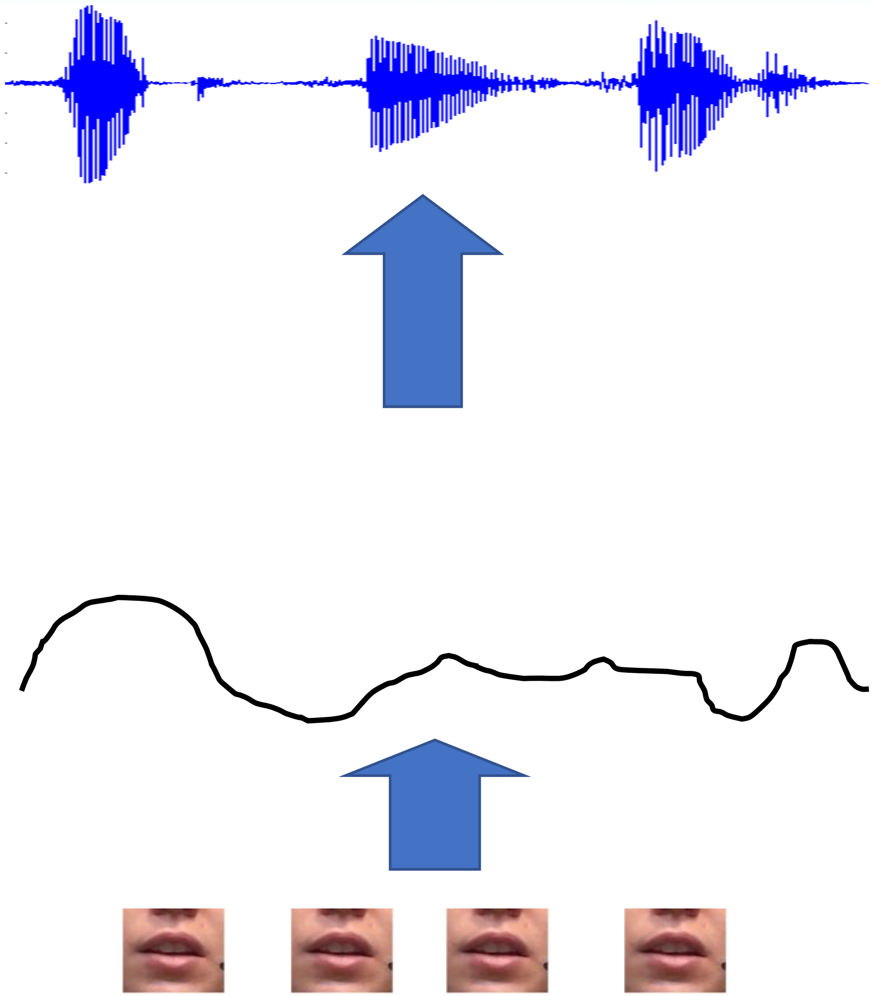
Alignment Neural Network (AliNN)

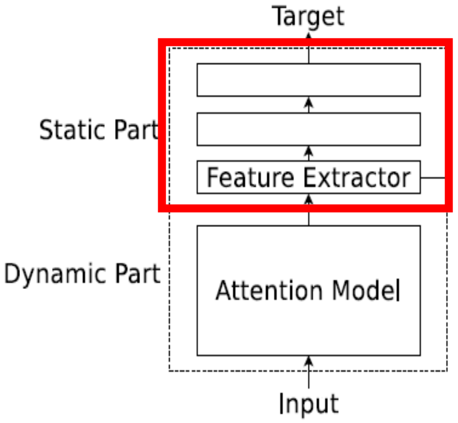
Feature space transform

Static Part

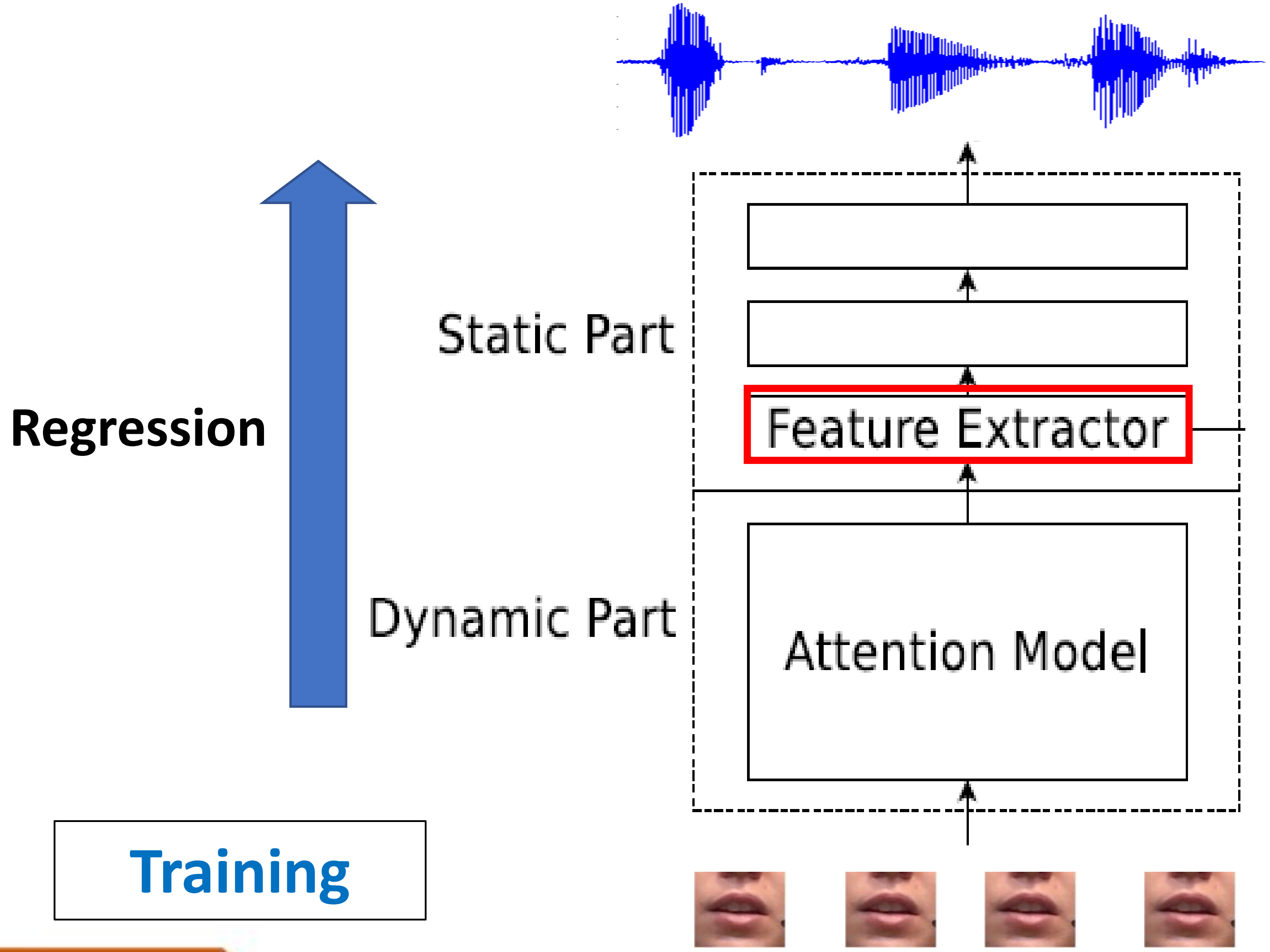
Temporal align

Dynamic Part



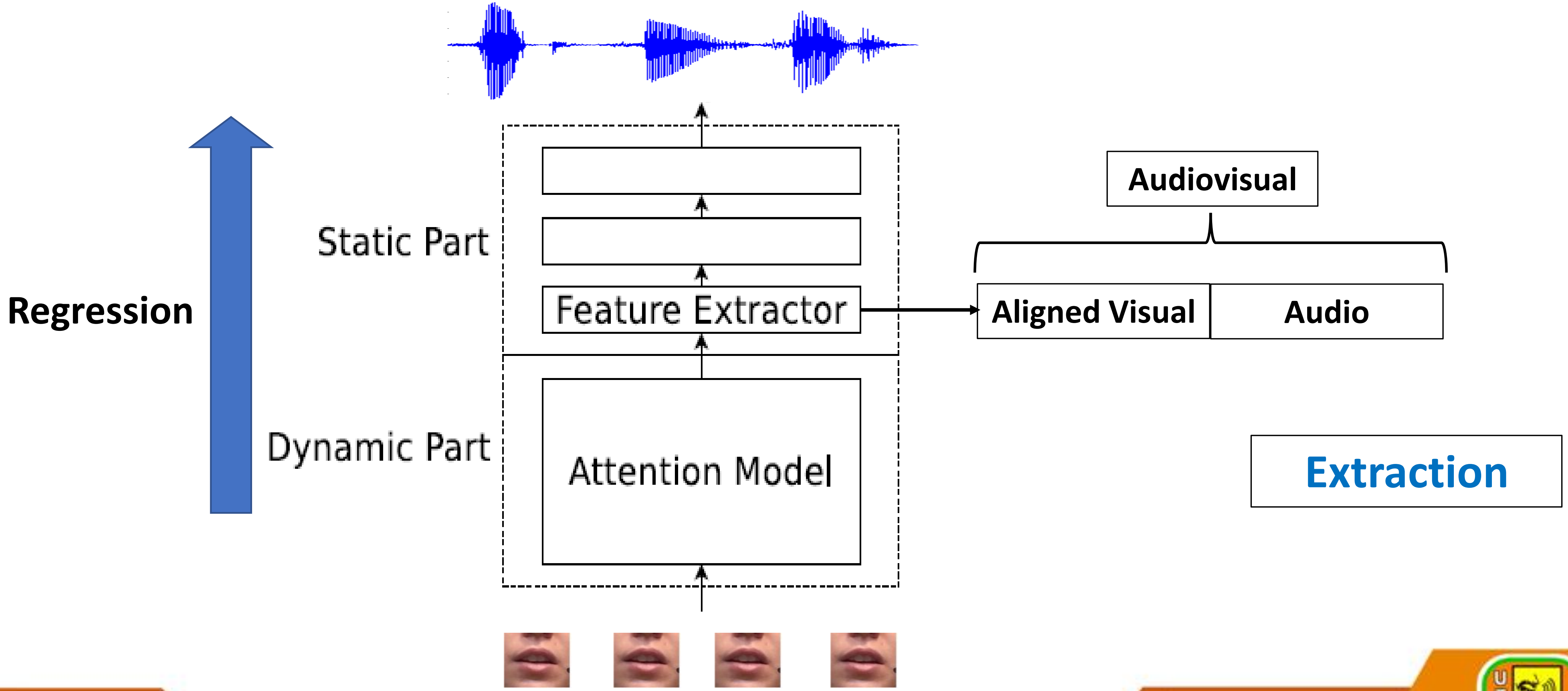


Alignment Neural Network (AliNN)



Training

Alignment Neural Network (AliNN)

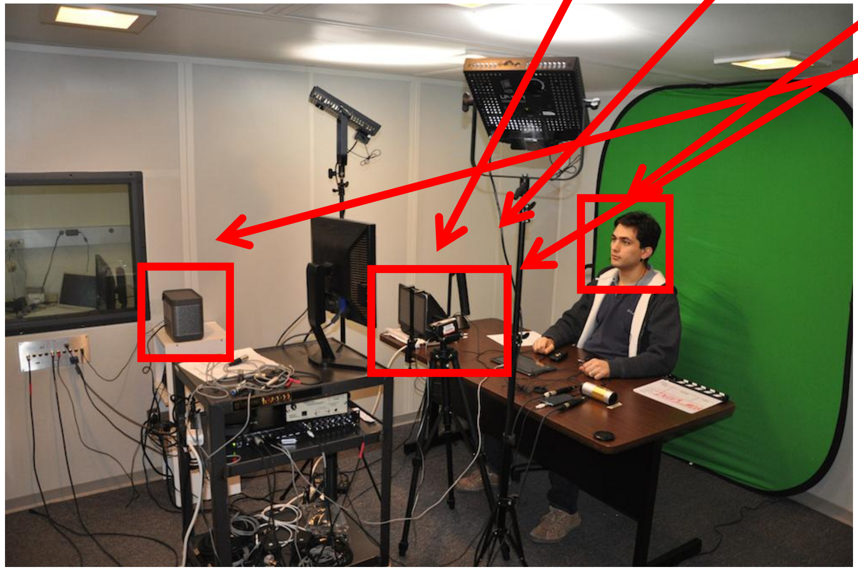


Training AliNN

- **Training AliNN on the whole utterance is computationally expensive**
- **We segment the utterance into small sections**
  - Length of each segment is 1 sec, shifted by 0.5 sec
  - Sequence is padded with zeros if needed

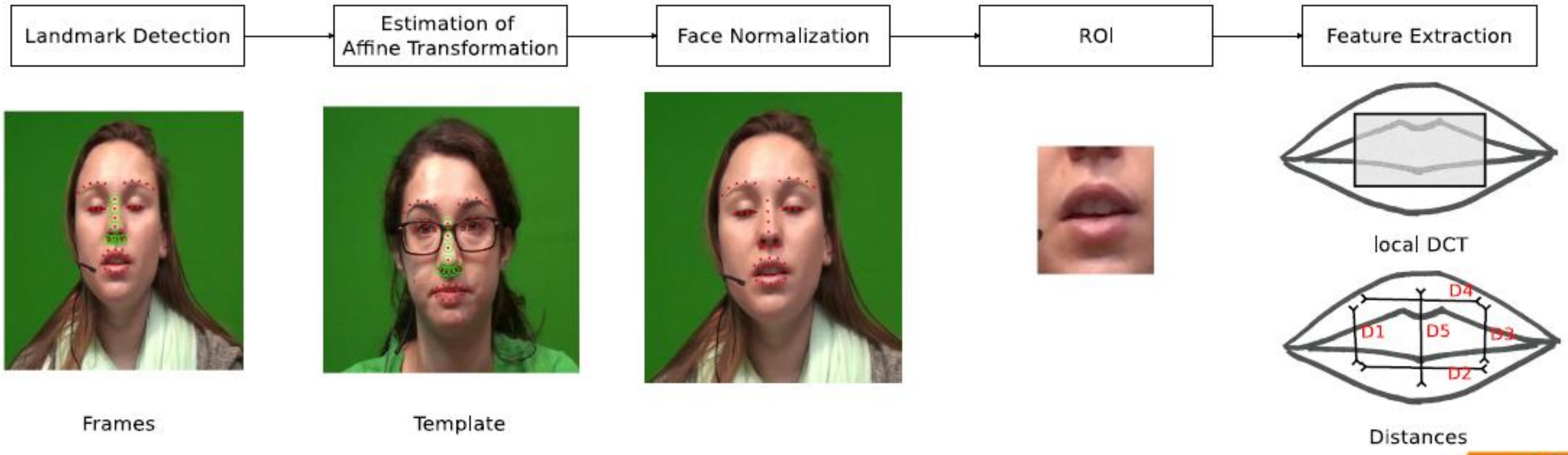
Corpus Description

- **CRSS-4ENGLISH-14 corpus:**
  - 55 females and 50 males (60 hrs and 48 mins)
  - Ideal condition: high definition camera and close-talk microphone
  - Challenge condition: tablet camera and tablet microphone
  - Clean section (read and spontaneous speech) and noisy section (subset of read speech)



Audiovisual Features

- **Audio feature: 13D MFCCs feature (100 fps)**
- **Visual feature: 25D DCT + 5D geometric distance**
  - 30 fps for high definition camera
  - 24 fps for tablet camera



Experiment Setting

- **70 speaker for training, 10 for validation, 25 for testing**
  - Gender balanced
  - Train with ideal condition under clean environment
  - Test with different conditions under different environments
- **Two backend:**
  - GMM-HMM: augmented with delta and delta-delta information
  - DNN-HMM: 15 context frames
- **Data of tablet (24 fps) is linearly interpolated to 30 fps**
- **Linear interpolation for pre-processing as baseline**
- **Focus on word error rate (WER)**

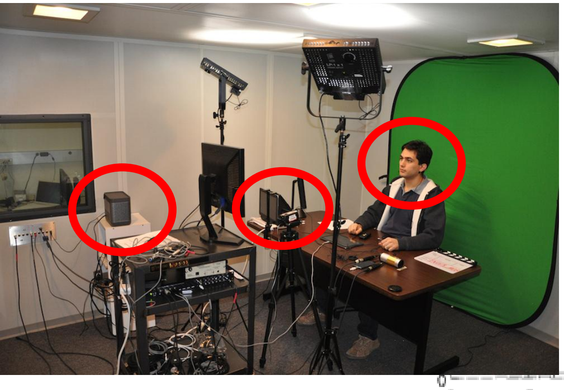
Experiment Results

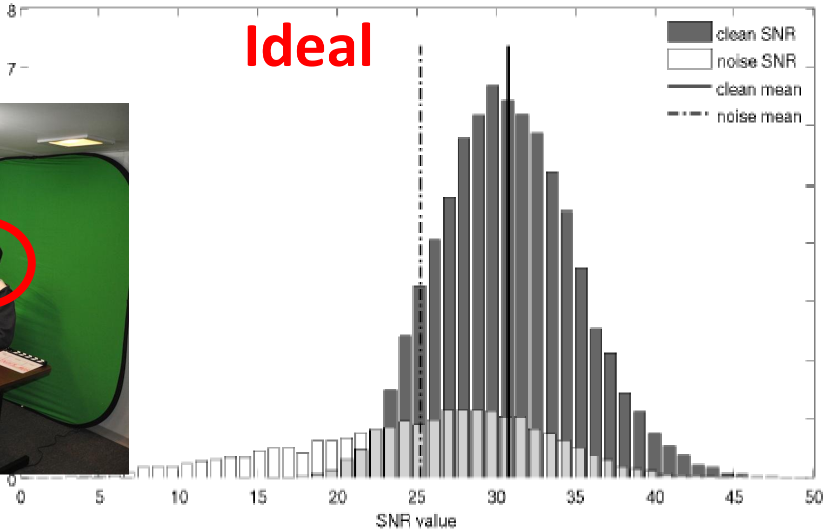
- **Under ideal condition, the proposed front-end always achieves the best performance**
- **Under tablet condition, the proposed front-end achieve the best performance except GMM-HMM backend**
  - ➢ Linear interpolate tablet data to 30 fps may impair the advantage of AliNN

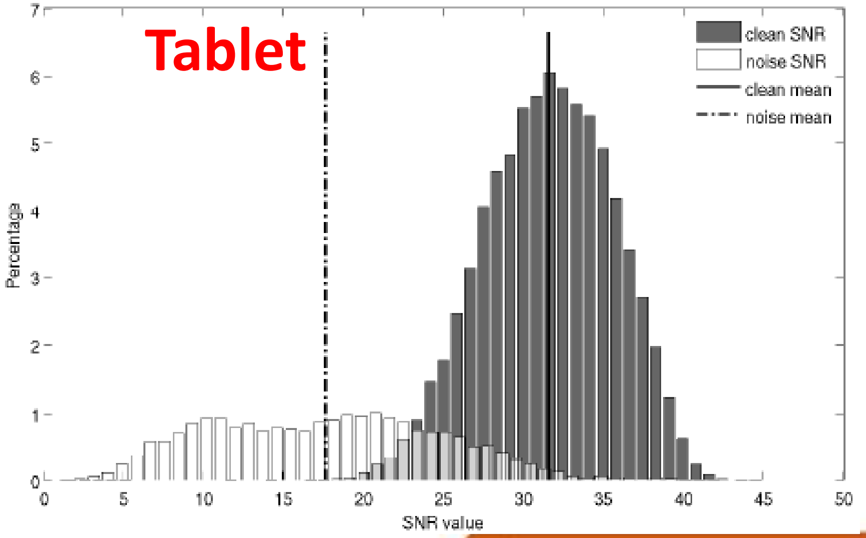
Front-end	MODEL	Ideal Conditions		Tablet Conditions	
		Clean [WER]	Noise [WER]	Clean [WER]	Noise [WER]
LInterp	GMM-HMM	23.3	24.2	24.7	**30.7**
AliNN	GMM-HMM	**17.5**	**19.2**	**22.7**	35.6
LInterp	DNN-HMM	4.2	4.9	15.5	15.9
AliNN	DNN-HMM	**4.1**	**4.5**	**4.6**	**10.0**

Results Analysis

Front-end	MODEL	Ideal Conditions		Tablet Conditions	
		Clean	Noise	Clean	Noise
LInterp	GMM-HMM	23.3	24.2	24.7	**30.7**
AliNN	GMM-HMM	**17.5**	**19.2**	**22.7**	35.6
LInterp	DNN-HMM	4.2	4.9	15.5	15.9
AliNN	DNN-HMM	**4.1**	**4.5**	**4.6**	**10.0**







Conclusions

- **This study proposed the alignment neural network (AliNN)**
  - Learns the alignment between audio and visual modalities from data
  - Does not need alignment or task label
- **The proposed front-end is evaluated on CRSS-4ENGLISH-14 corpus**
  - Large corpus for AV-LVASR (over 60h)
  - The proposed front-end outperforms simple linear interpolation under various conditions
- **Future work will extend approach to end-to-end framework**

Thank you !

References:

- C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, D. Vergyri, J. Sison, A. Mashari, and J. Zhou, "Audio-visual speech recognition," Workshop 2000 Final Report, Technical Report 764, October 2000.
- J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in International conference on machine learning (ICML2011), Bellevue, WA, USA, June-July 2011, pp. 689–696.
- S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic (2018). End-to-end audiovisual speech recognition. arXiv preprint arXiv:1802.06424
- T.J. Hazen, "Visual model structures and synchrony constraints for audio-visual speech recognition," IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 3, pp. 1082–1089, May 2006.
- C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1994)*, Adelaide, Aus- tralia, April 1994, vol. 2, pp. 669–672.
- F. Tao, J.H. L. Hansen, and C. Busso, "Improving bound- ary estimation in audiovisual speech activity detection using Bayesian information criterion," in *Interspeech 2016*, San Francisco, CA, USA, September 2016, pp. 2130–2134.