END-TO-END MULTIMODAL SPEECH RECOGNITION

Shruti Palaskar*, Ramon Sanabria* and Florian Metze
spalaska@cs.cmu.edu
Carnegie Mellon University

Carnegie Mellon University Language Technologies Institute

Objectives

Subtitling open-domain videos is still a challenge for Automatic Speech Recognition (ASR). In this work, we propose new models for audio-visual speech recognition to improve ASR performance.

- Visual adaptation for two end-to-end models: Connectionist Temporal Classification (CTC) and Sequence-to-Sequence (S2S)
- Using the HowTo dataset: an open-domain dataset of instructional YouTube videos
- Different adaptation strategies for both CTC and S2S models
- Comparison of model behavior on clean, prepared WSJ corpus and the noisy, spontaneous HowTo corpus

Introduction

- **Problem.** Subtitling open-domain videos despite huge acoustic variability, spontaneous speech, unrestricted domain of data
- **Solution.** Adapt ASR models to *visual semantic concepts* extracted from *correlated* visual scenes accompanying speech, different from lip-reading
- **HowTo corpus.** 480h of instructional videos downloaded from YouTube, that are fully transcribed and can be shared

Figure 1: Example of HowTo dataset with visual semantics

Example of error improvement:
Audio: Make sure you have a player
Audio+Visual: Let's show you how we plate it

*equal contribution.

CTC Model

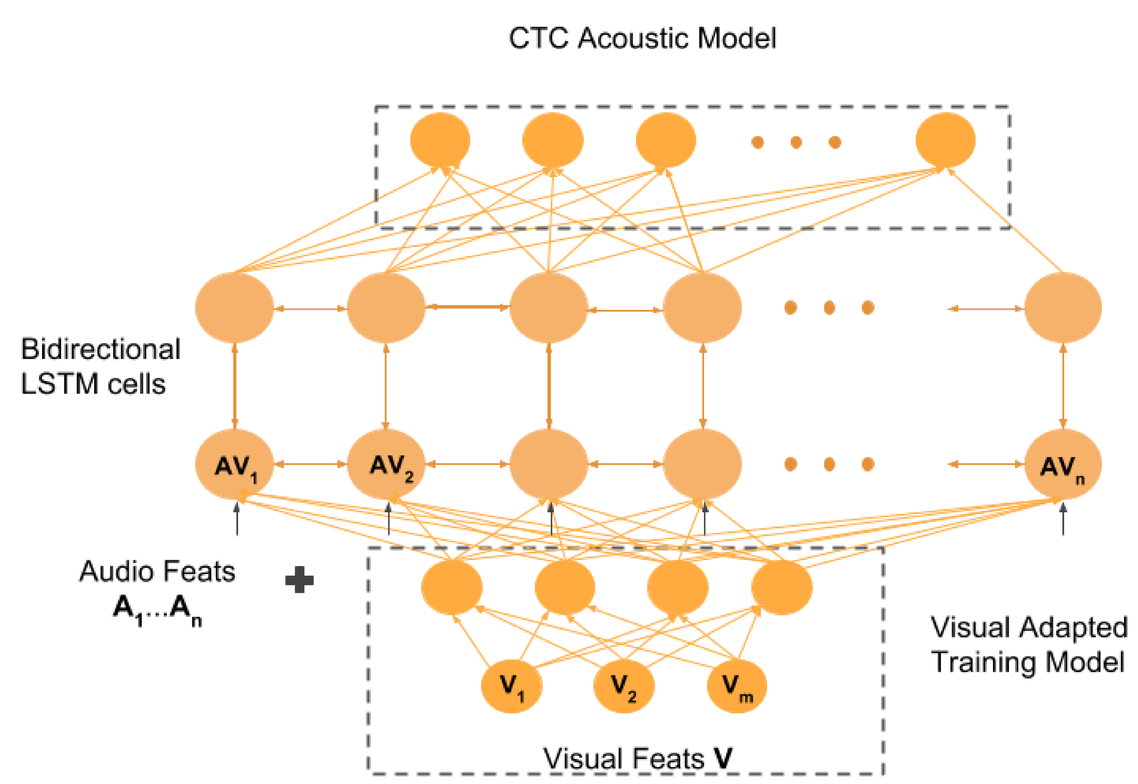

Figure 2: CTC Model Architecture with Adaptation

1. CTC with **Visual Adaptive Training** (VAT).
2. End-to-end training of VAT Multilayer Perceptron and CTC Acoustic Model
3. Separate Language Model Adaptation

S2S Model

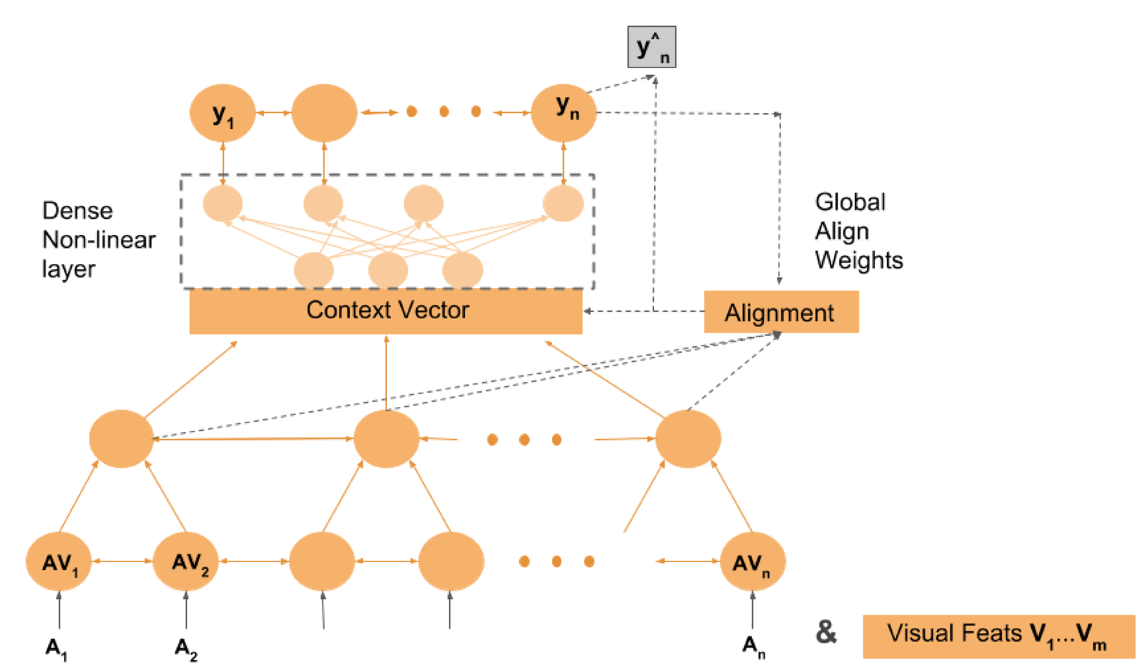

Figure 3: S2S Model Architecture with Adaptation

1. Audio-frame level input concatenation of visual features, Early Fusion
2. Pyramidal encoder with Global Attention mechanism
3. Acoustic Model and Language Model adapted together

Important Results

1. We achieve state-of-the-art performance **and adaptation** of S2S model
2. Image adaptation not only helps in the acoustic and linguistic models separately, but also in a joint architecture such as S2S.
3. End-To-End ASR architectures can be adapted without frame synchronization.

CTC vs. S2S

1. Compare CTC and S2S on standard WSJ dataset
2. Observe huge disparity in the Token Error Rates (TER) of clean and noisy speech corpus
3. Evaluated on 90 hours of HowTo corpus and ~90 hours of WSJ corpus

	CTC	S2S
WSJ	6.9	7.9
How-To	18.5	15.3

Table 1: TER on WSJ (eval92), HowTo(test set)

Audio-Visual Adaptation Results

1. Visual feature adaptation shows steady improvements in the CTC AM (TER)
2. Shows even higher improvement in S2S
3. Large improvements in CTC LM (PPL) establishes strong correlation between speech and visual features.

	A CTC	**A+V** CTC	**A** S2S	**A+V** S2S
TER	15.2	14.1	18.4	16.8
PPL*	113.6	80.6	1.38	1.37

Table 2: Audio(**A**) and Audio-Visual(**A+V**) adaptation. *CTC LM - word-level, S2S LM - character-level.

WSJ vs. HowTo

Variance in minimum and maximum length of transcript affects the S2S model behavior.

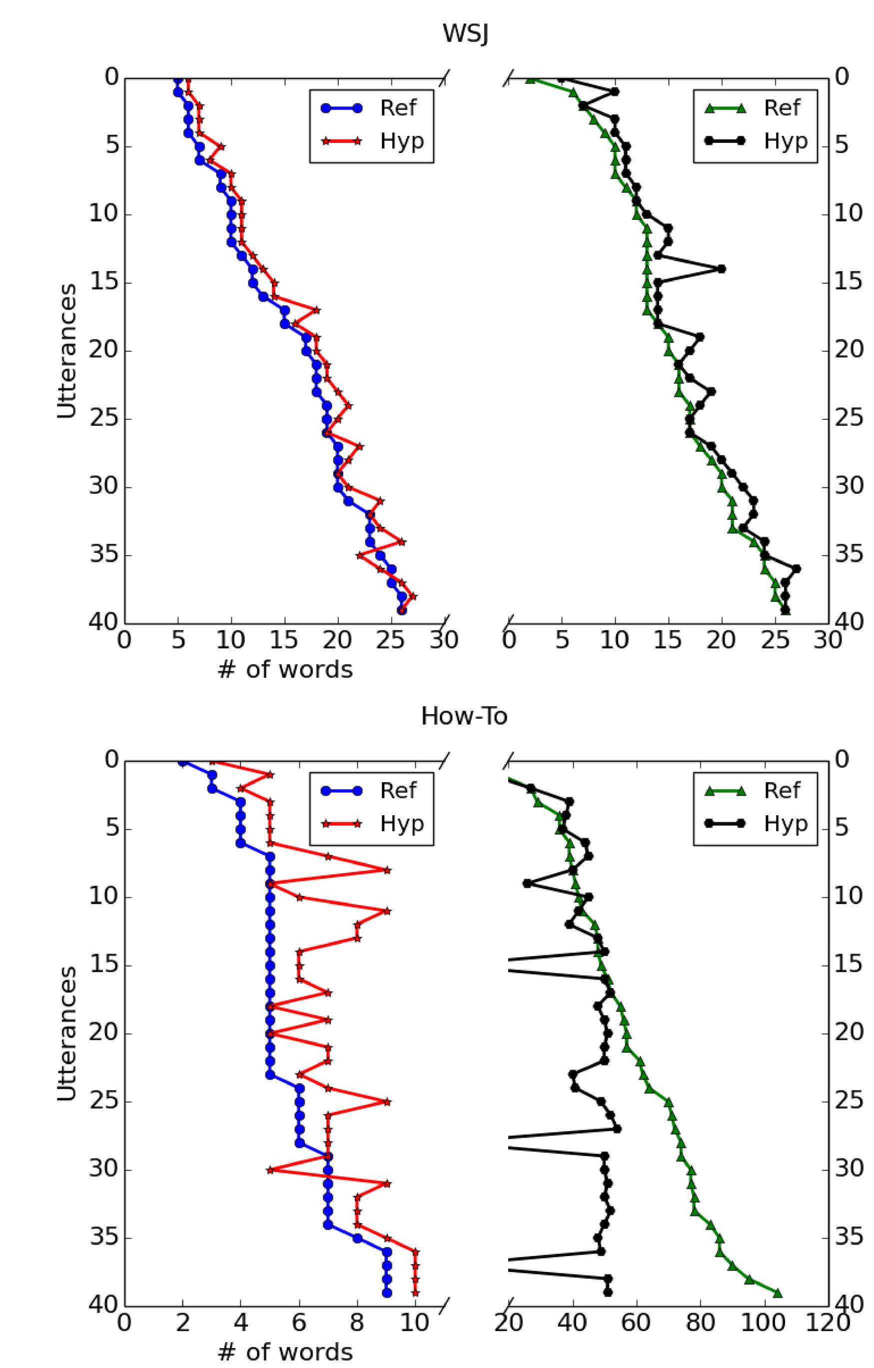

Figure 4: Length normalization by S2S for WSJ and How-To

Conclusion

1. Visual semantic concepts help improve ASR
2. CTC output tends to be very close to the acoustics of the utterance
3. S2S output appears to be closer to the style of the transcriptions

Ongoing & Future Work

1. Many different adaptation strategies for S2S
2. Preparing **public release** of the HowTo dataset, that is ~2000 hours of data
3. Our work will be part of **JSALT** 2018 Workshop at JHU in the team **Grounded Sequence to Sequence Transduction**