Using Video Information to Improve Automatic Speech Recognition in the Distant Microphone Scenario

Jack Deadman, Jon Barker

{jdeadman1, j.p.barker}@sheffield.ac.uk

Department of Computer Science, University of Sheffield, UK

Introduction

One of the most difficult challenges in Automatic Speech Recognition (ASR) is recognising speech from people who are far away from the microphone. The difficulty comes from the reverberation and competing sound sources in the containing room, corrupting the desired signal. The work in this Ph.D. will explore how the additional modality of video information can be used to improve ASR systems.

Objectives

- Investigate how video information can be used to improve signal enhancement in the ASR frontend.
- Explore how video can be used to improve the acoustic model in the ASR backend.

Dataset

The CHiME-5 [1] dataset is used as the main data source in this work. The dataset consists of 20 dinner parties with 3 distinct stages (cooking, dining and after-dinner socialising), each party is around 2.5 hours long. The parties are recorded using multiple Microsoft Kinect devices which have a 1080p camera and a 4-channel linear microphone array.

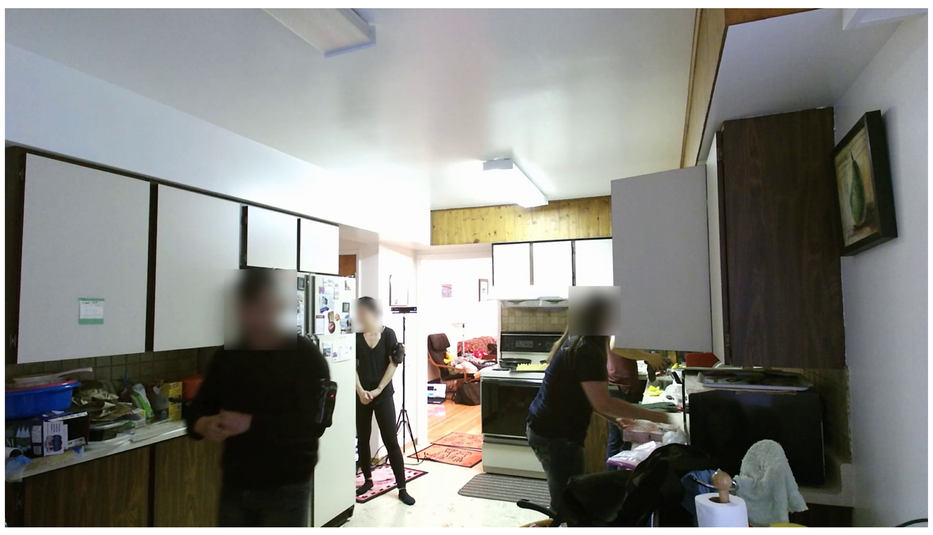

Figure: Example screenshot taken from CHiME-5 dataset during the "Cooking" phase.

Party setup and challenges

- 4 people take part in each party.
- Two Kinect devices capture each stage of the party.
- People are typically not facing the camera when talking.
- People are constantly moving around the rooms and travel between rooms.
- The conversations are unscripted, which involves lots of false starts and interruptions.
- 50 hours of transcribed audio.

The current state of art [2] audio-only system achieved a WER of 41.6%.

References

[1] Barker, Watanabe, Vincent and Trmal, Interspeech 2018
[2] Kanda et al, arXiv,1905.12230, 2019
[3] Cheng et al, Picture Coding Symposium 2018
[4] Hershey, Chen, Le Roux and Watanabe ICASSP 2016

Video Features

The features extracted from the videos are motivated by the speech enhancement algorithms that will be deployed i.e., the location of the speaker and their distance from the microphone is very useful.

(a) Optical Flow (b) Depth Map

(c) Background subtraction

Figure: Examples of generated features from a video in the CHiME 5 dataset.

Due to the high-dimentionality of the video features (compared to speech) an autoencoder [3] is used to reduce the features into a low-dimensional space. These features can be used in acoustic modelling and signal enhancement.

Single Channel Source Separation

- Deep Clustering [4]: Learn embeddings for time-frequency (T-F) bins, which can be used to cluster the T-F bins in an unseen mixture. Video information could be used for learning better embeddings, and to inform the clustering process itself (typically k-means is used).
- Guided Source Separation [2]: Use the timing information in the transcripts to fit a distribution for each of the known speakers in a segment. Video information can be used to improve the initialisation in the process.

Beamforming

When multiple synced microphones are used to capture sound, beamforming algorithms can be used. Beamformers enhance the sound in a direction and suppress sounds in competing directions. Using person tracking techniques the directions of interest can be found, these can complement the audio domain approaches to finding the desired direction.

Acoustic Modelling

The acoustic model can be improved through concatenating the video features with standard acoustic features (e.g., MFCC, i-vectors). A model such as a TDNN can learn to capture the dependencies between location and acoustics.

Evaluation

- Evaluate speech enhancement techniques on simulated mixtures by using standard metrics such as SNR and SDR.
- Evaluate the ASR performance on a simpler task first such as WSJ-2mix.