CHiME-5 multi-mic `dinner party' corpus

Jon Barker, University of Sheffield, UK; Emmanuel Vincent, Inria, France; Shinji Watanabe, MERL, US

1 Overview

The CHiME-5 corpus is being designed to drive future public distant microphone speech recognition challenges extending the successful CHiME series of ASR evaluations (i.e., CHiME-1 (2011), CHiME-2 (2013), CHiME-3 (2015) and CHiME-4 (2016) challenges[1]). Specifically, we are planning to use the data as the focus for CHiME-5 (2018) and CHiME-6 (2019/20). For these challenges we will be defining an evaluation framework and releasing a Kaldi baseline system.

The CHiME-5 corpus will target the problem of distant microphone conversational speech recognition in everyday home environments. The dataset will be designed for ASR challenges that target multi-microphone solutions with separate tracks for fixed microphone arrays and head-mounted microphones, e.g. binaural ear-bud microphones. Speech material will be elicited using a `dinner party' scenario with efforts taken to capture data that is representative of natural conversational speech.

2 Data Collection

The scenario

- The dataset will be composed of recordings of a number of 4-person dinner parties.
- Each party will take place in the real home of a host person (or host couple). Guests will be invited so that the total number of participants is 4.
- All participants will be native English speakers with US accents.
- To facilitate naturalness all members of the party will be well known to each other.
- Each party will consist of three phases: pre-dinner (meal preparation); dinner; post-dinner (relaxed post dinner conversation).
- Each phase will last a minimum of 30 minutes.
- Each phase will either take place in a different room (kitchen, dining room, living room). Alternatively, in open plan houses/flats, the phases may be in different areas of the same open living space.
- Participants will be free to converse on any topic of their choosing, i.e. no artificial scenario-ization.
- Subjects will be briefed on the need to avoid topics that would be self-identifying or contain private information.
- Subjects will be asked to give consent that the data be made publicly available and that it can be used for commercial purposes.

[1] http://spandh.dcs.shef.ac.uk/projects/chime/

Data collection

Recordings will be made.

- Recordings will be made of a total of 20 parties in 20 different homes.
- Each party will provide between 2 and 3 hours of data, ie. the full set of 20 parties providing between 40 and 60 hours of recording.
- The parties will be split 16/2/2 between training, dev and test data sets.
- The 16 party training set will contain recordings of 8 separate groups of 4 people. Each group of 4 people will participate in two parties in two different homes, i.e. making up 16 parties in 16 homes in total.
- Dev and test sets will each consist of 2 groups of 4 people both recorded in different homes. I.e. 2 homes for dev and 2 homes for test.
- In summary,

Sessions	Homes	Spkrs	Hours (minimum)	Hours (max)
20 (16/2/2)	20 (16/2/2)	48 (32/8/8)	40 (32/4/4)	60 (48/6/6)

Hardware

The hardware will include **fixed mic arrays** and **head worn mics** and **video.** We will use an easily reproducible, easily installable recording set-up, using ‘off-the-shelf’ consumer hardware. A proposed specification would be

- **Fixed arrays**. A total of six separate Microsoft Kinect will be set up with two per each of the three areas of the home.
- The fixed arrays will be positioned to record two contrasting `views’ of the same room with different average speaker distances (medium, ~2m and long, 3-4m).
- **Worn mics**. Each participant will wear a pair of binaural in-ear mics recording audio onto a body-worn mobile device.
- The worn-mics need to be unobtrusive, comfortable, quick to set-up and robust. We plan to use OKM II Classic Studio mics with an A3 adapter (http://www.soundman.de/en/products/) and record onto Tascam DR-05 data recorders.
- **Video** will be captured by each of the 6 Kinect sensors.
- The Kinect will all be set to record for the entire duration of the party so that participants can move between areas (e.g. bringing food from the kitchen to the table), and so that there are no artificial breaks in the party between the kitchen, dinner and post-dinner phases.
- Channels *within* arrays will be sample-synchronized; we will ensure approximate synchronization across arrays and between arrays and the head worn microphones.

Noise background

- There will be no strict control of the noise background - we wish to record whatever occurs naturally, e.g. cooking, eating, music, kids, sound from adjoining rooms.
- To avoid capturing Copyrighted content in the background we will disallow TV. A selection of CD of royalty free music will be provided in case participants wish to play music..
- For each home, the rooms will be sketched and dimensions measured; the locations of the main noise sources will be noted.

Postprocessing

- All speech spoken by the party participants will be end-pointed and transcribed.
- Party participants will be independently transcribed using the audio from the head-worn microphones.
- Speech segment boundaries will be defined.
- Transcription will include laughter, and a unique label for all other non-speech events including filled and unfilled pauses, coughing, breaths, lipsmacks etc.
- Identification of overlapping versus non-overlapping speech segments.
- Transcripts will be used to redact segments containing non-publicly releasable material.
- Transcriptions will be delivered in a suitable format.
- Separate challenge tracks will be designed, e.g. ASR from fixed mic, head-worn mic, overlapping versus non-overlapping

3 Summary of data to be provided

- Recordings of 20 separate parties, each of duration 2-3 hours (i.e. 40-60 hours of recording) and each consisting of 4 speakers.
- For each party there will be:
  - 6 continuous video recordings from 6 continuously recording Kinect sensor
  - 24 (6 x 4) channels of distant mic audio from Kinect sensors
  - 8 (4 x 2) channels of binaural mic audio from microphones worn by participants.
  - A complete transcription of the spoken conversation with speaker identity and timestamps for the start and end of each utterance segment.
  - All audio and video streams will be approximately aligned.

4. Novelty with respect to existing distant mic challenges/datasets

	COSINE	CHiME-1 CHiME-2	CHiME-3 CHiME-4	ICSI AMI(DA) CHIL	REVERB	DIRHA Real	Sheffield War Games	Santa Barbara Corpus	CHiME-5
Natural conversation	✔			✔			✔	✔	✔
Many rooms / places	✔		✔					✔	✔
Many unique speakers	✔			✔				✔	✔
High reverb		✔		✔	✔	✔	✔	✔	✔
Many noises	✔	✔	✔			✔		✔	✔
Mic array	✔		✔	✔	✔	✔	✔		✔
Binaural mics		✔							✔
Lightweight setup			✔		✔				✔
Close-talk channel present	✔		✔	✔			✔		✔