MC-WSJ-AV corpus

Abstract

The MC-WSJ-AV (multi-channel Wall Street Journal Audio-Visual) corpus is a corpus of read speech (WSJ) recorded with close talking and distant microphone (arrays) enabling research in speaker localisation, (blind) speech separation and speech recognition

Authors

Author	Affiliation
Mike Lincoln	The University of Edinburgh
Erich Zwyssig	The University of Edinburgh

The 2nd author only carried out minimal data preparation in the final stages of releasing the corpus.

Data Type

Speech from single, overlapping and moving speakers.

Data Source

The MC-WSJ-AV corpus consists of audio recordings generated at premises of the

The Centre for Speech Technology Research
University of Edinburgh
Informatics Forum
10 Crichton Street
Edinburgh
EH8 9AB
United Kingdom

Recordings were carried out using a headset and lapel microphone and an eight-channel microphone array. Despite the title (AV) no video data is provided. Speaker locations are fixed and can be derived from [1].

Languages and dialects

(Mostly) UK English.

Narrative Description

The MC-WSJ-AV corpus offers researchers an intermediate task between simple digit recognition and large vocabulary conversational speech recognition. It consists of sentences read from the Wall Street Journal (WSJ) taken from the test set of the WSJCAM0 database.

A total of about 45 speakers, male and female, are recorded in three different scenarios, these are:

- single (stationary) speaker
- two (stationary) overlapping speakers
- single moving speaker

The speakers are recorded using a headset and lapel microphone and an eight-channel microphone array, reading WSJ from prompts. In the single speaker scenario participants are asked to read from 6 fixed positions, in the overlapping scenario speaker get assigned a fixed position for the entire recording, and for the moving scenario speakers move from one position to the next while reading.

15 participants were recorded for the single scenario, 9 pairs for the overlapping scenario and 9 for the moving scenario. Each read about 90 sentences which are available for speech separation and recognition experiments.

Task	**Applications**	**Comments**
Single speaker	Distant (automatic) speech recognition	
Overlapping speakers	Speaker localisation, speech separation and distant (automatic) speech recognition	
Moving speaker	Speaker localisation, speech separation and distant (automatic) speech recognition	

Each speaker (pair) read WSJ sentences (WSJCAM0) from script, i.e.

Data set (name)	**Number of sentences**	**Description**
adapt	Approx. 17	TIMIT style, for adaptation
5k	Approx. 40	5,000 word (closed vocabulary) sub corpus of WSJCAM0
20k	Approx. 40	20,000 word (open vocabulary) sub corpus of WSJCAM0

Each sentence is individually split from the recording for recognition and stored in folders following the structure

- ./MC_WSJ_AV/audio/<task>/T<#1>[_T<#2>]/<mic_type>/{adap|5k\20k}/*.wav

... with

  - <tasks > defining the task, i.e. stat (single static), move (single moving) or olap (two overlapping)
  - T<#> defining the participant and his/her number #
  - <mic_type> defining the microphone type, e.g. array1/2, headset1/2 or lapel1/2

.. and *.wav is defined as

- AMI_WSJ<#1>[_<#2>]_<mic_type>_T<#1><ref#1>_T<#2><ref#2>.wav

... where <ref#> determines the correct answer from the mlf file stored in ./MC_WSJ_AV/mlf/MC_WSJ_AV.mlf

The speaker locations are stored in ./MC_WSJ_AV/etc/sentencelocation/T<>.txt

References

[1] The multi-channel Wall Street Journal audio visual corpus (MC-WSJ-AV): Specification and initial experiments,
M. Lincoln, I. McCowan, J. Vepa and H.K. Maganti,
IEEE Workshop on Automatic Speech Recognition and Understanding,
2005